UNIVERSITAT DE BARCELONA

Network-driven strategies to integrate and exploit biomedical data

Adrià Fernández Torras

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento 4.0. España de Creative Commons***.

This doctoral thesis is licensed under the ***Creative Commons Attribution 4.0. Spain License***.

Network-driven strategies to integrate and exploit biomedical data

Doctoral Thesis | 2022

PhD Program in Biomedicine

Network-driven strategies to integrate and exploit biomedical data

Adrià Fernández Torras

Network-driven strategies to integrate and exploit biomedical data

Memòria presentada per Adrià Fernández Torras
per optar al grau de doctor per la
Universitat de Barcelona.

Patrick Aloy	Miquel Duran-Frigola	Josep Lluís Gelpí	Adrià Fernández-Torras
Director de tesi	Codirector de tesi	Tutor de tesi	Doctorand

Structural Bioinformatics and Network Biology Lab
Institute for Research in Biomedicine (IRB Barcelona)

Programa de Doctorat en Biomedicina (2018-2022)
Universitat de Barcelona (UB)

September 2022

Universitat de Barcelona

IRB Barcelona
Institute for Research in Biomedicine

A tu, lector, per dedicar-me el teu temps.

Agraïments

Al grup que m'ha vist créixer. A tu Patrick, per animar-me a fer el màster i el doctorat i permetre'm viure una aventura que no oblidaré mai. A tu Miquel, per encomanar-me la il·lusió d'aprendre i perseverà, per ser el meu mentor, el meu referent, el meu amic; aquesta tesi es tant meva com teva. A tu Martino, per sempre haver estat incondicionalment al meu costat, em fas feliç. A tu Lídia, per ser la primera que em va rebre quan no sabia on em ficava i per no parar mai de creure en mi. A tu Oriol, per fer-me la vida més fàcil i divertida. A tu Carles, per demostrar-me que l'amistat no entén d'edats. A tu Martina, per donar vida aquesta tesi fora de les meves mans. A tu Nico, per donar-me la oportunitat de complir el meu somni de ser professor. A tu Paula, per compartir el teu viatge amb mi. A tu Angel, per omplir d'optimisme la meva jornada. A tu Aleix, per la teva estima i per ser un veí tant agradable. A tu Nils, per fer-me creure en mi mateix. A tu Angelo, per fer equip. A tu Arnau, per saltar al buit amb mi en cada una de les parides que ens em inventat, ets com un germà. A tu Gema, per mostrar-me la teva dedicació, sensibilitat i compromís, t'admiro. A tu Elena, per tants moments, histories i anècdotes que em viscut, em fas treure la millor versió de mi mateix. I a tants altres que en algun moment heu format part d'aquesta historia: David, Teresa, Csaba, Edu, Sergi, Richa, Francesco, Roberto, Pau, Roger, Mode, Marta.

Als meus amics i companys del IRB, amb els que he compartit moments inoblidables: Hania, Diego, Clara, Paula, Pep, Niko, Elena, Marina S., Claudia, i tants més. Gràcies per recórrer aquest camí al meu costat.

Als meus avis, pares i germans, per educar-me i estimar-me incondicionalment. Segurament no arribareu a entendre més enllà d'aquesta pàgina, però quedeu-vos tranquils, és la més important.

Per últim a tu, Marina, per aguantar-me i estimar-me tal i com soc. Per haver estat la meva companya de vida en aquesta travessia i mai haver-te negat a compartir la càrrega quan pesava. Es un honor que aquesta tesi comenci i acabi amb tu, amb la portada que vas dissenyar. T'estimo.

Estratègies basades en xarxes per integrar i explotar dades biomèdiques

Abstracte

En la cerca d'una millor comprensió dels sistemes biològics complexos, la comunitat científica ha estat aprofundint en la biologia de les proteïnes, fàrmacs i malalties, poblant les bases de dades biomèdiques amb un gran volum de dades i coneixement. En l'actualitat, el camp de la biomedicina es troba en una era de "dades massives" (Big Data), on la investigació duta a terme per ordinadors se'n pot beneficiar per entendre i caracteritzar millor les entitats químiques i biològiques. No obstant, la heterogeneïtat i complexitat de les dades biomèdiques requereix que aquestes s'integrin i es representin d'una manera idònia, permetent així explotar aquesta informació d'una manera efectiva i eficient.

L'objectiu d'aquesta tesis doctoral és desenvolupar noves estratègies que permetin explotar el coneixement biomèdic actual i així extreure informació rellevant per aplicacions biomèdiques futures. Per aquesta finalitat, em fet servir algoritmes de xarxes per tal d'integrar i explotar el coneixement biomèdic en diferents tasques, proporcionant un millor enteniment dels experiments farmaco-omics per tal d'ajudar accelerar el procés de descobriment de nous fàrmacs. Com a resultat, en aquesta tesi hem (i) dissenyat una estratègia per identificar grups funcionals de gens associats a la resposta de línies cel·lulars als fàrmacs, (ii) creat una col·lecció de descriptors biomèdics capaços, entre altres coses, d'anticipar com les cèl·lules responen als fàrmacs o trobar nous usos per fàrmacs existents, (iii) desenvolupat una eina per descobrir quins contextos biològics corresponen a una associació biològica observada experimentalment i, finalment, (iv) hem explorat diferents descriptors químics i biològics rellevants pel procés de descobriment de nous fàrmacs, mostrant com aquests poden ser utilitzats per trobar solucions a reptes actuals dins el camp de la biomedicina.

Network-driven strategies to integrate and exploit biomedical data

Abstract

In the quest for understanding complex biological systems, the scientific community has been delving into protein, chemical and disease biology, populating biomedical databases with a wealth of data and knowledge. Currently, the field of biomedicine has entered a Big Data era, in which computational-driven research can largely benefit from existing knowledge to better understand and characterize biological and chemical entities. And yet, the heterogeneity and complexity of biomedical data trigger the need for a proper integration and representation of this knowledge, so that it can be effectively and efficiently exploited.

In this thesis, we aim at developing new strategies to leverage the current biomedical knowledge, so that meaningful information can be extracted and fused into downstream applications. To this goal, we have capitalized on network analysis algorithms to integrate and exploit biomedical data in a wide variety of scenarios, providing a better understanding of pharmaco-omics experiments while helping accelerate the drug discovery process. More specifically, we have (i) devised an approach to identify functional gene sets associated with drug response mechanisms of action, (ii) created a resource of biomedical descriptors able to anticipate cellular drug response and identify new drug repurposing opportunities, (iii) designed a tool to annotate biomedical support for a given set of experimental observations, and (iv) reviewed different chemical and biological descriptors relevant for drug discovery, illustrating how they can be used to provide solutions to current challenges in biomedicine.

Contents

1 Introduction **1**

1.1 The complexity of biological systems 3
1.2 In the quest for modelling biological systems 6
1.3 Biology meets Big Data 12
1.4 Thesis into context 19

2 Objectives **21**

3 Chapters **25**

Chapter 3.1
Using protein annotations to extract systems-level knowledge from pharmacogenomic screenings **29**

3.1.1 Abstract 30
3.1.2 Introduction 30
3.1.3 Results 32
3.1.4 Concluding remarks 45
3.1.5 Methods 47

Chapter 3.2
The Bioteque, a comprehensive repository of biomedical knowledge descriptors **53**

3.2.1 Abstract 54
3.2.2 Introduction 54
3.2.3 Results 56
3.2.4 Concluding remarks 75
3.2.5 Methods 77

Chapter 3.3
A tool to efficiently annotate biomedical support behind experimental associations **91**
3.3.1 Abstract 92
3.3.2 Introduction 92
3.3.3 Results 93
3.3.4 Concluding remarks 96
3.3.5 Methods 97

Chapter 3.4
Connecting chemistry and biology with descriptors: a future perspective **103**
3.4.1 Abstract 104
3.4.2 Introduction 104
3.4.3 Review's chapters 106
3.4.4 Concluding remarks 118

4 Discussion **121**

5 Conclusions **131**

References **134**

Appendix A Supplementary Figures **193**
A.1 Chapter 3.1 195
A.2 Chapter 3.2 208

Appendix B Publications **215**
B.1 List of publications 217
B.2 Attachment of publications 219

1
Introduction

1.1 The complexity of biological systems

From gene units to functional systems

Genes are fundamental biological units. They encode the instructions to synthesise the biomolecular arsenal of living beings, information that will eventually be transferred to the next generations after being challenged by the natural selection process. Given their primordial role, central fields in biology, such as Genetics, Evolutionary biology, and Molecular biology, have been entirely devoted to their study and comprehension.

Following the central dogma of molecular biology, gene products, namely RNA molecules and proteins, can be considered functional machinery units. Concretely, the functional action is attributed to evolutionary fine-tuned structural domains, able to catalyse chemical reactions. Often, these functional domains originate from high-order structures that assemble after the physical interaction of multiple proteins. This close relation between proteins and biological functions has fostered initiatives such as the Universal Protein Resource (UniProt [1]), committed to thoroughly gathering and annotating protein sequences, structures, and functional evidence from across species.

Nevertheless, the complexity of living systems cannot be reduced to the molecular functions of individual proteins. Rather, proteins usually communicate between them, creating interconnected biological processes or pathways that put the vital processes of living beings in motion. What is more, modifications in the gene identity (e.g., mutations) or in its environment (e.g., epigenetic changes) may affect the functional capabilities of the product protein and, therefore, its interactions. Consequently, the biological system on top is forced to adapt accordingly, accommodating the regulation of other genes and proteins. This action-reaction behaviour creates two-way modulations between gene and protein layers. Moreover, this intimate communication between layers also leads to the emergence of deep dependencies between genes, spawning co-expression patterns (i.e., genes matching their expression levels) and genetic interactions (i.e., genes having an impact on the fitness of an individual when perturbed simultaneously) along the genome. Indeed, all these sophisticated interactions endow genomes with flexible and meticulous control of their regulation, which may partially explain why genome size is not a good indicator of the biological complexity of an organism [2].

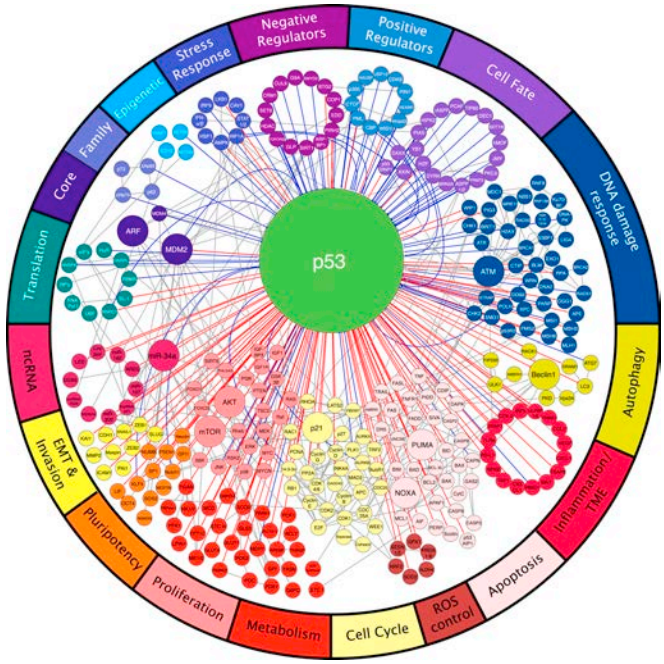


Figure 1.1: The p53 Network. As a master regulator, p53 interacts with an assortment of proteins involved in distinct biological functions, creating an intrinsic communication between many different biological processes. At the same time, positive and negative regulators (on top of the figure) modulate the activity of p53, orchestrating a complex yet powerful biological system. Credit to R. Kastenhuber and Scott W. Lowe's review[3].

In conclusion, from gene units to functional proteins, biology complexity lies in the interplay between different interconnected layers, from which their active communication creates convoluted yet powerful and dynamic systems.

Chemical, disease, and environmental perturbations

Far from being isolated, biological systems are exposed to a changing environment in which other players can interact with them, originating even more convoluted scenarios. Some small molecules, such as metabolites, are essential for the correct functioning of these systems and thus are involved in main processes such as protein signalling, cell structure, or metabolism. On the other hand, other chemicals such as drugs or environmental compounds can directly interact with one or multiple proteins, disrupting their function or forcing the recalibration or rewiring of their interactions.

Perturbed systems, whether due to external insults (e.g., drugs) or internal alterations (e.g., genomic rearrangements), may lead to abnormal conditions or diseases. Cancer, for example, is an alarming pathology originating from multiple system perturbations that, after re-adaptation, lead to 'malfunctioning' cells. Concretely, this malfunction makes these cells inconveniently efficient in their growth and expansion, thus disrupting the healthy environment equilibrium of their niches[4]. Unfortunately, the catalysts of cancer are vast and heterogeneous. To date, causes of cancer include aberrant genetic variations, environmental exposures, chemical perturbations, personal habits, and even dysregulations caused by pathogen infections or other diseases[5]. Usually, it is the combination of multiple factors that eventually triggers the disease, a fact that makes cancer, and other pathologies with heterogeneous origins, be termed 'complex diseases'[6].

Even if a malfunctioning system is, essentially, a divergence from a 'healthy' state, the truth is that the genetic and environmental variability inherent in living organisms adds a variety of shades. Indeed, genetic and environmental differences between individuals originate in specific biological contexts that directly affect how the system responds to a given perturbation. Gut microbiota, for instance, has been broadly associated with the metabolization of drugs, modulating their activity and even compromising their treatment action[7,8]. Since gut microbiota heterogeneity varies between individuals, and dozens of environmental factors are known to rebalance their composition, accounting for (or even tweaking) the microbiome population can potentially lead to better therapy strategies[9]. In the same vein, both the Food and Drug Administration (FDA) and the European Medicines Agency (MDA) are constantly reporting new genetic variations found to affect the therapeutic action of drugs (aka biomarkers)[10,11]. Likewise, different genetic backgrounds can lead to higher risks of diseases[12]. It is thus increasingly common to model disease aetiology and drug treatment based on individual traits so that tailor-made therapies can be designed[13].

Therefore, genetic and environmental backgrounds play a major role in how biological systems behave and respond to perturbation. This fact calls for collecting biological data from individuals so their particular biological context can be described and used to guide the prescription of more effective, personalised therapies.

1.2 In the quest for modelling biological systems

Cellular screenings as surrogates of biological systems

Cells are the main constituents of living organisms. They come as the result of the proper orchestration of the biological machinery, organised in subcellular compartments. As such, their observable traits (aka phenotype) inherently portray the status of their underlying molecular systems. Thus, it is unsurprising that cells and other high-order assemblies, such as organoids, tissue cultures, or animals, have been broadly used as working models to understand biology. Accordingly, an increasing number of omics disciplines have been blooming during the last decades, gathering an extensive and orthogonal variety of cell biology descriptors [14].

Genomics, for example, specialises in the study of genomes. In practice, from the analysis and comparison of DNA sequences, this field has set the basis for the association of gene variations (e.g., mutations) and phenotypes, playing a fundamental role in the diagnosis and prognosis of diseases with a strong genetic origin [15,16,17]. When scaled up to thousand sequences, Genome-Wide Association Studies (GWAS) can mine and contrast genetic variants at single-nucleotide resolution, uncovering those statistically associated with phenotypic traits or abnormal conditions [18,19]. Initiatives such as DisGeNET [20], COSMIC [21], or OMIM [22], collect and supply millions of these gene-disease relationships, paving the way for the scientific community to explore them.

Transcriptomics and Proteomics disciplines, on the other hand, capitalise on the use of gene expression and protein abundance measurements, respectively. These cellular readouts quickly proved to be accurate surrogates of cell biological states, allowing, among other successes, the identification of coordinated actions between pathways [23], the stratification of cancer into subtypes [24], or the materialisation of phenotype-specific gene signatures [25]. More recently, the rise of single-cell technologies led to individual profiling of millions of cells, which provided new insights into cell heterogeneity and spatial organisation [26,27]. Eventually, the accumulation of expression array repositories enabled statistical methods to unearth transcriptional patterns between genes, leading to the systematic collection of gene co-expression networks across different tissues and organisms [28].

Given the rich information provided by these readouts, cellular descriptors started to be implemented to study the effect of external perturbations. Towards this goal, cellular panels such as the Cancer Cell Line

Encyclopedia (CCLE[29]), the Genomics of Drug Sensitivity in Cancer (GDSC[30]) or the Cancer Therapeutics Response Portal (CTRP[31]) collected a wide variety of genomic features for thousands of cell lines together with their phenotypic response to hundreds of small molecules (aka drug sensitivity profiles). These initiatives allowed the direct connection between cellular drug response and cell biological features, thereby boosting the identification of drug biomarkers[32,33] and fostering the implementation of predictive models to anticipate drug sensitivity[34].

Similarly, cell transcriptomics differences before and after chemical treatment have also been exploited to identify new therapies and delve into drug mechanisms of action[35]. In brief, the underlying idea is to connect diseases to drugs based on the similarity (or divergence) of their corresponding gene expression profiles in cell lines. That is, identify chemical compounds that mimic or reverse the transcriptomic phenotype of the disease[36]. This 'connectivity principle' was popularised by the Connectivity Map (CMap[37]) and extended by the LINCS Program[38], currently providing over 1,5 million profiles from tens of cell lines and thousands of molecules and other bio-perturbagens (e.g., shRNAs). Despite the technical challenges and limitations of the approach[39,40], many fruitful studies have benefited from these screenings[41,42,43], identifying potential disease targets[44,45] and opportunities for drug repurposing[46,47].

More recently, the success of the CRISPR technology prompted the implementation of large-scale platforms that started to systematically perturb every single gene in the quest for cell fitness alterations. Consolidated in the Dependency Map (Depmap[48]) resource, cell fitness profiles have enabled the exploration of cell-gene specific vulnerabilities[49,50], the systematic mining of genetic interactions[51,52], and the suggestion of combined therapies[53,54].

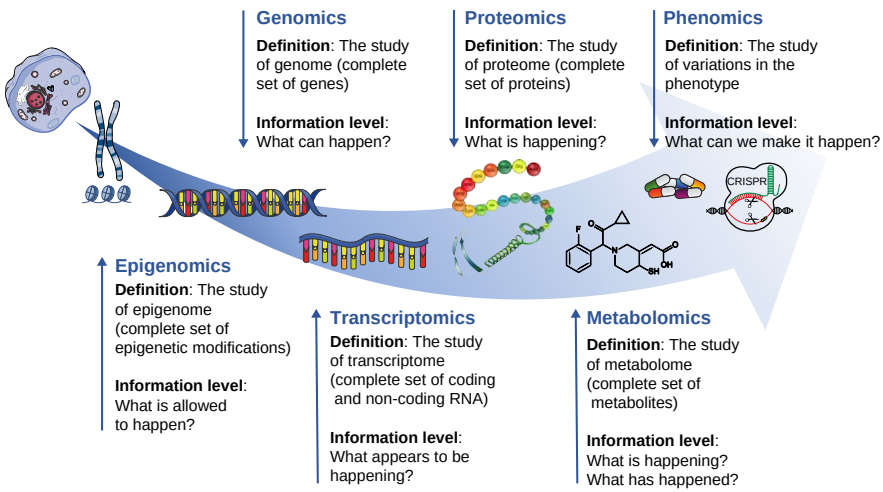


Figure 1.2: The omics revolution. Different omics disciplines screen cell lines at different levels of resolution, gathering complementary descriptors of cell biological state. Inspiration from [55].

Beyond cell-centered descriptors, other high-throughput screenings have also significantly contributed to understanding biological systems. The Human Reference Interactome (HuRI [56]) consortium, for example, has been systematically screening pairwise combinations of human protein-coding genes since 2005. Considering all the releases [56,57,58], this consortium has interrogated over 17,500 proteins and identified more than 60,000 physical protein-protein interactions (PPIs). Other initiatives, such as BioPlex [59], have joined the effort by suggesting alternative screening implementations able to identify other types of physical interactions. Soon, commendable efforts were undertaken to accommodate all these interactions into unified repositories, providing protein interactome landscapes of different types and natures. For example, the IntAct database [60] focused on collecting physical interactions, giving evidence for binary contacts and complex assemblies. On the other hand, other resources such as STRING [61] or BioGrid [62] also incorporate functional associations, including biological pathway membership and co-expression interactions. These resources fostered network biology studies to explore these interactomes in all their forms, revealing valuable insights from these associations while producing all sorts of biomedical applications [63,64].

Network biology as means to rationalise biological data

Network structures have been fundamental in understanding complex molecular interactions and modelling functional cellular systems. In fact, from gene-gene interactions to chemical-protein dysregulations, most of the mechanisms orchestrating cellular biology can be naturally represented as networks. Therefore, network biology opens the door for a more holistic understanding of biological processes, moving from decades of successful but limited reductionist biology (e.g., one gene, one protein, one function) to a scenario where genes, proteins, and other biological entities are not treated in isolation but as a part of a functional system [65,66]. Moreover, the adoption of algorithms from Network Science smoothened the implementation of network-based techniques able to gain relevant insights from biological networks, eventually giving birth to a new field of research [67,68].

Network Biology bore its first fruits from gene and protein networks obtained from model organisms. For instance, Schwikowski et al. [69] constructed a *Saccharomyces cerevisiae* (yeast) protein-protein interaction (PPI) network, covering 2k proteins and 2.7k interactions. After annotating functional information, the authors found that proteins with similar functions and cellular locations tend to cluster together. This made it possible to infer molecular functions for uncharacterised proteins based on the annotations of their interacting partners. In another work, Costanzo et al. [70] gathered nearly 1M of genetic interactions (GI) in yeast, spanning almost every known gene. From this network, they not only detected core essential genes, usually taking shape as network hubs [71], but also were able to map core bioprocesses, identify pleiotropic genes, and discover functionally related proteins. Moving to human networks, in an outstanding joint effort finalized only a few months after the first Covid-19 pandemic outbreak, host-coronavirus protein interactions were characterised with the hope of shedding light on the SARS-Cov-2 infection mechanism [72]. Once the host factors hijacked by the virus were identified, the authors implemented systematic genetic screenings to perturb functionally relevant host proteins, finding protein targets interfering with the virus replication.

Besides revealing functional insights, protein networks have also been mined to uncover new interactions. In a study carried out by Kovács et al., they noticed the following network pattern: proteins tend to interact not if they are similar to each other (i.e., sharing the same interactions) but rather if one of them is similar to the other's partners [73]. In other words,

two proteins (A and B) are more likely to interact if multiple network paths connect them with two intermediary proteins (A-X-Y-B). Willing to exploit this pattern, the authors systematically traversed the whole interactome following this motif, predicting new PPIs with accuracies that outperformed the state of the art.

The interconnectivity of biological systems makes it apparent that activity deviations in one gene will also affect those genes and proteins connected to it. Following this rationale, biological networks have been extensively used as functional layouts to map omics experiments, offering means to trace and measure the impact of gene perturbations based on their propagation through the network[74]. In this vein, F. Vandin et al. conceived HotNet[75], a network algorithm able to find significantly altered subnetworks in genome-scale interactomes. More specifically, following the mapping of patient cancer mutations into genome networks, the authors used HotNet to recover biological pathways known to be altered in specific cancer types. Noteworthy, the algorithm identified cancer-associated pathways that were overlooked (not annotated) in the patient samples, thereby providing a deeper characterisation of the disease for each individual.

On the other hand, the annotation of disease-associated genes on top of gene interactomes has revealed interesting network properties underlying disease mechanisms. Concretely, many studies[76,77,78] have shown that disease-relevant genes are not highly connected (hubs) in the network but rather tend to co-localize in specific functional regions of the interactome, clustering into network modules. Indeed, by exploiting these network motifs, it has been possible to connect diseases between them and to therapeutic-aimed perturbations. On that subject, Menche et al.[79] proposed a distance metric based on the average path separation between two network modules. Guided by this distance, the authors found that diseases with overlapping modules were prone to exert similar pathobiological outcomes, leading to shared symptoms and more likely comorbidities[80]. One year later, Gunney et al.[81] used several network distance measures to connect hundreds of drug targets to different disease modules, finding that shorter drug-disease distances were mostly corresponding to known drug-disease indications. Leveraging this property, they were able to suggest new drug repurposing candidates and anticipate adverse effects.

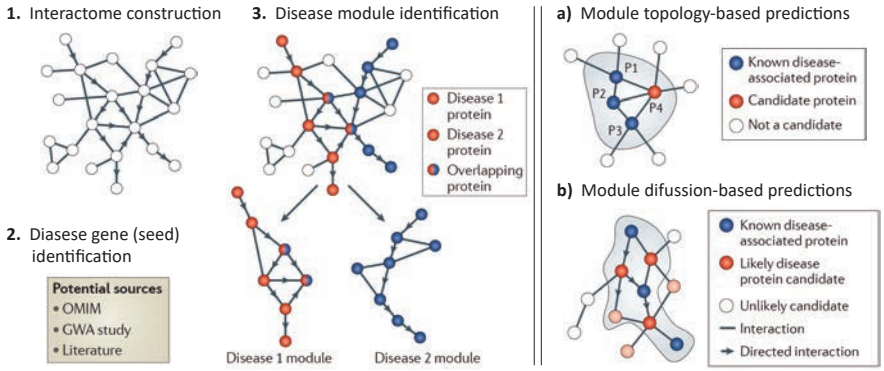


Figure 1.3: Disease module identification. By mapping disease genes into a protein interactome, one can identify those proteins clustering in a sub-network (module) region. Once disease modules are identified, one can predict new disease protein candidates. Some methods directly identify statistically significant associations between protein candidates and disease modules (a). In contrast, others try to expand the disease modules by propagating their protein members through the network (b). Figure adapted from Barabási et al.[68].

Lastly, the inspection of these network patterns led to the design of network-driven strategies for drug combinations. In this area, Cheng et al.[82] quantified the network-based relationship between drug targets and disease protein modules and identified up to six different drug-combination network motifs. Interestingly, this analysis revealed that drug combinations targeting complementary regions of the disease module were indeed more likely to manifest higher synergistic effects. A few years later, the same authors exploited this 'complementary exposure' pattern to identify potential drug combinations for treating the SARS-CoV-2 coronavirus[83].

1.3 Biology meets Big Data

The deluge of data floods biomedical databases

In addition to high-throughput screenings, the ever-growing studies carried out by the scientific community, who often supplement their publications with large tables of experimental results, are flooding every corner of biology with data and knowledge. Accordingly, the bulk of supplied information is rather unwieldy, being the growth of biological databases steeper than ever before[84]. Indeed, data storage in the EMBL-EBI increased sixfold in the last years (from 40 petabytes in 2014 to 250 in 2021)[85], while in 2021 the NAR online Molecular Biology Database Collection was already holding 1,641 different databases[86]. Even more overwhelming, as data are scattered among a myriad of resources, researchers have to deal with a variety of nomenclatures, identifiers, levels of resolution (e.g., protein isoforms or gene splicing), and experiment-tailored conditions, making data integration across platforms a cumbersome process[87].

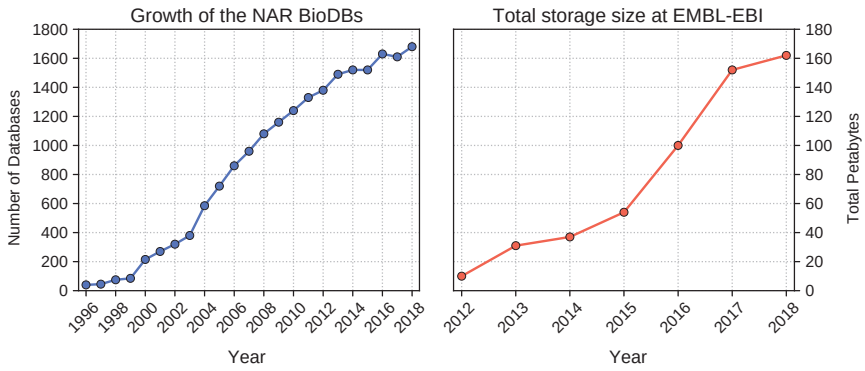


Figure 1.4: Growth of biodata in the 21st century. On the left, is the number of biological databases published in Nucleic Acid Research (NAR) since 1996. On the right, is the total number of petabytes used by the European Bioinformatics Institute (EBI) to store biological data since 2012. Adapted from Cook et al.[88] and Digital World Biology[89]

Aware of this situation, some initiatives have tried to unify vocabularies and formats to fuse biological data. For example, the Harmonizome resource[90] abstracts experiment-specific aspects from gene-centred datasets to represent them as simple gene-attribute pairs (e.g., gene-cell, gene-pathway, or gene-drug). This strategy allowed to harmonise over 125 heterogeneous databases into a standard network-like format. In fact, networks, and especially Knowledge Graphs (KGs), have proven to be conve-

nient structured infrastructures to integrate data and knowledge. In a KG, each node represents an entity of interest (e.g., a protein, drug, or disease), whereas associations (edges) describe semantic relationships between entities (e.g., drug – treats– disease). Thus, KGs enable the representation of heterogeneous entities and their relationships in a unified and intuitive format[91].

Indeed, the flexibility and expressive power of KGs have driven the development of many applications in the biomedical field[92,93,94], and are often incorporated in pharma companies to drive R&D[95]. Concerning the public domain, Himmelstein and colleagues assembled Hetionet[93], a KG built from 29 publicly available resources that cover 11 biological entities and 24 biomedically relevant relationships (e.g., drug-treats-disease, disease-upregulates-gene or gene-participates-pathway). Hetionet not only served as an accessible exploratory tool to query biomedical associations but also proved valuable in making predictions. Concretely, the authors enumerated a list of sequential associations in the KG (aka metapaths) to connect compounds and diseases through various biomedical contexts (e.g., 'drug–resembles–drug–treats–disease', or 'drug–downregulates–gene–upregulates–disease'). After proper statistical modelling, they identified a combination of metapaths able to distinguish treatments from non-treatments and, therefore, suitable for predictive tasks.

Even if network representations smoothen the accommodation of heterogeneous information, traditional graph analytics are not optimised to deal with large volumes of data[96]. Modern computing relies on distributed computing systems to handle large-scale data, where data is split into different shards or servers to be processed simultaneously and boost computational efficiency[97]. Unfortunately, given that network nodes are not independent but coupled to each other by design, data parallelisation in networks becomes extremely problematic. Alternatively, networks are represented by their adjacency matrix, in which both rows and columns represent the nodes of the graph, and edges are annotated in the matrix's cells (i.e., where both nodes intersect). However, although this vectorial representation decouples nodes from each other, these vectors still suffer from a high dimensionality as each node vector must annotate the status of every possible interaction. Moreover, given the sparsity nature of biological networks[98], these annotations are usually highly scattered along the vectors. This is significantly exacerbated in big graphs as network sparsity scales proportionally to the number of nodes[98]. Taken together, these vector properties pose serious technical challenges for their downstream

processing and analysis[99]. Importantly, these network representations directly collide with Machine Learning (ML) implementations, which rely on data parallelisation and a moderate number of non-sparse features to process the data effectively. Considering the impact that ML is having in almost every scientific field, including chemistry[100,101], biology[102,103], and biomedicine[104,105], it soon became evident the need to represent networks in a format amenable to ML methods.

Formatting biological networks for modern machine learning modelling

The urge for more concise graph representations fostered the development of network embedding techniques that drastically reduce the data's dimensionality while preserving the network's structural information. This is achieved by learning a low-dimensional continuous space in which each node in the network is represented by a numerical vector. Accordingly, relationships (edges) between nodes are captured by distances between their corresponding node vectors. As such, nodes related in the original network will have similar vectors and, therefore, encoded close together in the corresponding embedding space.

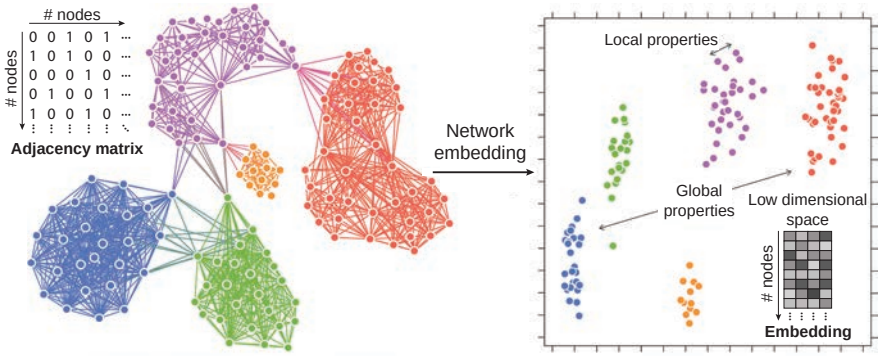


Figure 1.5: The network embedding representation. While traditional graph layouts are based on the network's adjacency matrix, modern representations implement embedding techniques to transform the network into a low-dimensional space. In this space, local (e.g., network edges) and global (e.g., network module separation) properties of the network are captured by the distances between the corresponding node numerical vectors. Adapted from Duran-Frigola et al.[106].

The first sort of network embedding dates back to the early 2000s. Concretely, classical approaches tackle the problem as a structure-preserving dimensionality reduction exercise, which can be solved by Matrix Factorization (MF) algorithms[107]. Intuitively, the goal is to learn a latent representation for each node such that the inner product between them approximates some deterministic measure of node similarity. To this end, the input network is usually first transformed into a similarity matrix, which is afterwards factorised. For instance, in Laplacian eigenmaps[108], one of the most popular methods[109], this similarity matrix is obtained by creating a neighbourhood graph where nearby nodes are connected (e.g., using k-nearest neighbours) and weighted according to their Euclidean distance. Furthermore, some strategies can adjust the learning of the manifold to account for auxiliary information when available[110]. However, despite being a solid field with a long tradition, MF embeddings are mostly restricted to small and medium-sized networks due to the high computational costs demanded by these methods. Besides, these algorithms mainly account for node similarities within the first or second-order neighbours, making the embedding space fail to capture high-order relationships between the nodes and global properties of the network. Moreover, even if some works address the previous limitations[111,112], the deterministic measure of node similarity assumed by MF methods makes them lag behind other strategies that, on the contrary, rely on stochastic measures to learn more continuous and versatile embedding spaces[113].

As an alternative to MF-based methods, embedding techniques based on shallow and deep learning architectures have been gaining popularity during the last few years, quickly becoming state-of-the-art[114]. Within this group, the first remarkable implementations were based on network statistics extracted from Random Walks (RW). Indeed, RWs are a helpful tool for extracting knowledge from biological networks[115,116,117]. The algorithm simulates the behaviour of a walker that moves from node to node stochastically (sometimes with a certain probability of restart). The intuition is that the paths travelled by the RW will sample the close vicinity of every node, thus providing a flexible and stochastic measure of node similarity to other nodes. In practice, the output trajectories (sequences of nodes) of RWs can be seen as a 'text corpus' where each node corresponds to a 'word' and each path to a 'phrase'. This convenient format enabled the adoption of Natural Language Processing (NLP) architectures, which have been optimised for encoding text into latent spaces, giving birth to the first generation of RW-based embedding methods[118]. Since

then, other variants have flourished to 'encourage' the RW to explore local or global regions of the graph (e.g., node2vec[119]), focus on the structural identity of the nodes (e.g., struct2vec[120]), or extend the strategy to heterogeneous networks (e.g., metapath2vec[121]). Unfortunately, most of these methods usually ignore attributes or features associated with network nodes (e.g., molecular weight in a drug-drug interaction network), potentially missing rich information for the embedding process. Worse still, they cannot generate embeddings for nodes not initially found in the network[122].

To overcome this limitation, Graph Neural Network (GNN) encoders leverage node and edge features to learn an embedding function able to generalise to unseen nodes[122]. As a result, these methods can simultaneously learn the topological network structure of each node together with the distribution of node and edge features in the neighbourhood. In brief, GNN methods first sample the neighbourhood nodes in the vicinity of a given anchor node and train an 'aggregation function' to combine the neighbourhood features. The neighbourhood aggregated features are then concatenated with the features of the anchor node, assembling a descriptor that captures both the features of the node and those in its vicinity. Eventually, these vicinity-aware node descriptors will be compressed in a low dimensional space by a neural network encoder, obtaining the node embeddings. Notice that, since GNN models have been trained on node features, they can map new (unseen) network nodes to the learned space *a posteriori*. These promising architectures were introduced in 2017 by W.L. Hamilton et al. with GraphSage[123]. From there, different models have evolved to improve the expressive power of the aggregation functions[124], give different importance (attention) to each neighbour node[125], scale GNN to massive networks[126], or extend them to heterogeneous graph settings[127]. And yet, even if GNN-based embeddings have shown excellent performances in some tasks, their success vastly depends on the availability of meaningful node features and many samples (nodes) to adequately assimilate the network information. Otherwise, these methods are easily surpassed by RW-based or MF-based strategies[128,129].

Regardless of the embedding technique, the resulting latent spaces have several advantages compared to traditional graph representations. First, the dense compression of the dimensional space limits the room for redundant information, reducing the complexity of the space and, thereby, making it more robust to the noise that is inheritably present in networks[130]. On the other hand, as node information is captured in the node vectors,

iterative or combinatorial problems in networks can be directly addressed by either distance metrics, mapping functions, or arithmetic operations on the embedding space, which facilitates node similarity searches and appealing 2D-visualisations[99]. In addition, algorithms based on embeddings are usually faster than their network counterparts, thanks to the low dimensionality of their spaces. Indeed, as network nodes are no longer coupling, parallel computing solutions can further enhance the efficiency of these algorithms. But even more impactful, their condensed numerical representation makes them a natural fit for machine and deep learning models, which can now incorporate network structure information in predictive tasks[122].

Hence, as it could not be otherwise, network embeddings have been extensively implemented in a wide variety of scenarios in biology[131]. Many of these applications capitalise on the so-called 'similarity principle', adopted from fields with a long tradition of exploiting this principle. In cheminformatics, for instance, chemical fingerprints are represented as numerical vectors so that millions of compounds can be efficiently compared, searched, and classified with the underlying idea that similar compounds will have similar descriptors. In great resemblance to chemical embeddings, biological network embeddings can also be used to efficiently screen and cluster biology. For example, Fan et al.[132] encoded protein networks from different species in the same latent space and used the similarity (inner product) between protein embeddings to identify those that were functionally analogous between species, therefore providing means to align their interactomes. Elsewhere in a recent review[133], the authors compared two community detection methods, an embedding-based and a traditional graph-based, on the problem of clustering single-cell RNA-seq profiles, showing how the former outperformed the latter on three of the four datasets tested. Alternatively, Hamilton et al.[134] envisioned a question-answering framework able to represent queries as logical operations in the embedding space, enabling faster screenings of the network (e.g.,'what drugs are likely to target proteins involved with both diseases X and Y?').

When naïve similarity searches are insufficient to produce good predictions, network embeddings are usually plugged into off-the-shelf machine learning methods, which can learn more complex patterns. In this regard, a comprehensive benchmarking of methods to predict new links in biomedical graphs revealed that neural network models fed with precomputed node embeddings outperformed most of the tested baselines[135]. Concretely in the protein domain, Cho et al. devised Mashup[136], a frame-

work that first learns protein embeddings from different networks and then inputs them into a machine learning model to derive functional insights about proteins. Likewise, in the drug discovery domain, Wan et al. developed NeoDTI[137], an end-to-end model that learns how to embed a heterogeneous network populated with drug, protein, and disease associations to predict new drug-target interactions. Overall, network embedding techniques provide a succinct numerical representation that better suits modern computational approaches and machine learning implementations. As a result, embedding spaces allow the generalisation of network algorithms to large-scale scenarios, alleviating some existing limitations inherent in traditional graph representations.

1.4 Thesis into context

Committed to gaining a better understanding of biological systems, the worldwide scientific community has consistently gained knowledge about protein, chemical, and disease biology. Concurrently, with the flourishing of omics technologies, large-scale screening platforms have been implemented to extract quantitative measurements from cellular systems, populating biological repositories with tons of data. Soon, the need to integrate the gathered data into the context of biological systems became apparent. As a natural fit, network architectures proved to be conveniently structured data repositories, able to incorporate a variety of heterogeneous associations in a unified and logical format. Not only that, but network layouts also provide a means to represent complex molecular interactions, offering an analytical framework to understand biological systems and characterise chemical and disease perturbations. Eventually, the gigantic size of these networks fostered the adoption of network embedding techniques that effectively reduce the dimensionality of the data while preserving information in a dense and concise format optimised for computational tasks.

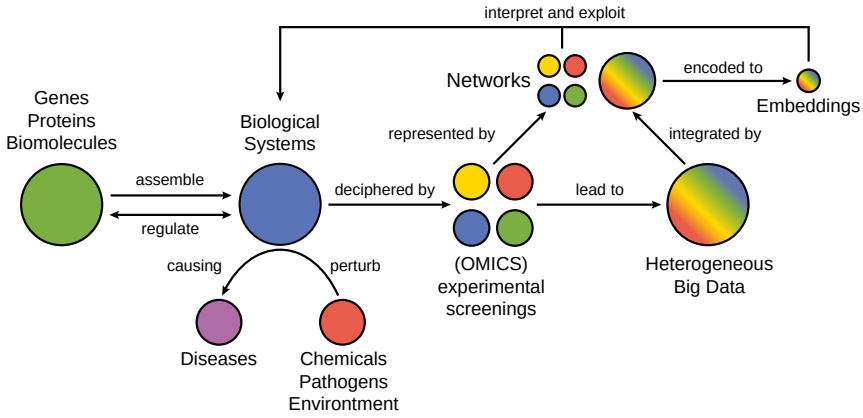


Figure 1.6: Thesis context scheme. Schematic diagram summarising the context of the thesis.

In this scenario of Big Data and technological breakthroughs, computational-driven biomedical research can primarily benefit from the existing biological and chemical information. Particularly in the quest for personalised medicine, the exploitation of existing biological knowledge can help elucidate connections between the vast chemical space and interconnected biological systems, accelerating drug discovery. Unfortunately, biomedical data is still fragmented across different repositories, isolated in individual datasets, and represented in a medley of formats that cannot be naturally assembled. Even when harmonised in format, data in biology is intrinsically complex, vast, and heterogeneous, posing severe challenges for an effective computational exploitation. It is the goal of this thesis to provide strategies, tools, and resources that properly represent and leverage the current biological and chemical data so that meaningful information can be extracted and fused into downstream biomedical applications. Importantly, rather than focusing on a specific gene, disease, or methodology, we seek to identify current challenges and opportunities in biomedicine where incorporating existing knowledge can potentially contribute new insights. And yet, throughout this manuscript, there will be a natural impulse towards using network representations as frameworks for our work and undeniable devotion to the pharmaco-omics field as the focus of research driving our motivation.

2
Objectives

This thesis aims to explore strategies that efficiently and effectively leverage the existing biological knowledge to (i) retrospectively shed light on publicly available experiments and (ii) prospectively exploit the information in different downstream biomedical tasks and applications.

More specifically, the objectives in each chapter are:

1. To map pharmacogenomic screenings on functional interactions to identify gene signatures associated with drug mechanism of action and sensitivity (Chapter 3.1).

2. To systematically format biomedical data into numerical vectors so that the information can be integrated and exploited in downstream computational pipelines (Chapter 3.2).

3. To provide a means of integrating large-scale (omics) experiments to (i) provide biological support to each observation, (ii) quantify the novelty of the dataset as a whole, and (iii) suggest dataset-specific features with potential for predictive analysis (Chapter 3.3).

4. To review descriptors capturing relevant information for drug discovery complementary to the ones produced in Chapter 2, revealing future directions for our work (Chapter 3.4).

3
Chapters

This thesis is organised into four individual chapters. Each chapter is presented as a scientific article, most of them published during the thesis. Accordingly, apart from the results and methodology sections, each chapter includes an abstract, an introduction, and a concluding remarks section which provide greater detail and depth about the proper context of the presented results. In brief, the first three chapters focus on the integration of existing biological knowledge to (i) extract new insights from pharmacogenomic outcomes, (ii) generate and exploit biomedical embeddings in descriptive and predictive downstream tasks, and (iii) devise a tool to annotate biomedical support in experimental screenings. Finally, in the last chapter, we review those protein and chemical representations that provide relevant yet complementary information to the descriptors we produced in previous chapters. Of note, we also discuss our participation in a research community challenge launched during the course of this thesis, in which we formatted and leveraged drug bioactivity information to provide a successful solution.

Chapter 3.1
Using protein annotations to extract systems-level knowledge from pharmacogenomic screenings

Authors	Adrià Fernández-Torras, Miquel Duran-Frigola, Patrick Aloy
Type	Research Article
Stage	Published
Title	Encircling the regions of the pharmacogenomic landscape that determine drug response.
Journal	Genome Medicine
DOI	https://doi.org/10.1186/s13073-019-0626-x
Context	Cell sensitivity profiles can be matched with cell biological traits, such as cell gene expression, to provide links between cellular state and drug sensitivity. However, most of these strategies assume independence between genes, which makes subsequent multiple-correction analysis heavily stringent. Consequently, functional gene systems behind drug response get fragmented into a few statistically significant yet isolated genes, diluting the underlying biological connections between them. In this chapter, we argue that gene knowledge can indeed be used to find functionally coordinated gene expression modules associated with drug response. In this way, drugs can be associated with statistically significant functional gene sets rather than single genes, providing a more comprehensive view of the cellular biological state behind drug response and, therefore, potentially revealing molecular determinants of drug response.
Note	Supplementary data can be accessed at the original publication.

3.1.1 Abstract

The integration of large-scale drug sensitivity screens and genome-wide experiments is changing the field of pharmacogenomics, revealing molecular determinants of drug response without the need for previous knowledge about drug action. In particular, transcriptional signatures of drug sensitivity may guide drug repositioning, prioritise drug combinations, and point to new therapeutic biomarkers. However, the inherent complexity of transcriptional signatures, with thousands of differentially expressed genes, makes them hard to interpret, thus giving poor mechanistic insights and hampering translation to clinics. To simplify drug signatures, we have developed a network-based methodology to identify functionally coherent gene modules. Our strategy starts with the calculation of drug-gene correlations and is followed by a pathway-oriented filtering and a network-diffusion analysis across the interactome. We apply our approach to 189 drugs tested in 671 cancer cell lines and observe a connection between gene expression levels of the modules and mechanisms of action of the drugs. Further, we characterise multiple aspects of the modules, including their functional categories, tissue-specificity, and prevalence in clinics. Finally, we prove the predictive capability of the modules and demonstrate how they can be used as gene sets in conventional enrichment.

3.1.2 Introduction

Gene expression profiling has become a mainstay approach to characterise cell properties and status, unveiling links between gene activities and disease phenotypes. Early efforts were channelled into discovering transcriptional signatures that are specific to a disease state. This work involved the comparison of a relatively small number of diseased and healthy samples[25]. Although providing a rich account of disease biology, these studies have failed to yield better drug therapies, as causality and response to drug perturbations cannot be inferred directly from two-state (diseased vs healthy) differential gene expression analysis[138,139]. To address this issue, initiatives have flourished to profile the basal gene expression levels of hundreds of cell lines, together with their response to treatment over an array of drug molecules using a simple readout such as growth rate[140,29,30,141]. Provided that the panel of cell lines is large enough, this approach allows for a new type of gene expression analysis where basal expression levels are

correlated to drug response phenotypes. A series of recent studies demonstrate the value of this strategy for target identification [33,142], biomarker discovery [143,144], and elucidation of mechanisms of action (MoA) and resistance [145,146].

The largest cell panels available today are derived from cancerous tissues, since a crucial step towards personalised cancer medicine is the identification of transcriptional signatures that can guide drug prescription. However, current signatures are composed of several hundred genes, thereby making them difficult to interpret, harmonise across platforms, and translate to clinical practice [147,148,149]. Recent assessment of sensitivity signatures for over 200 drugs [142] revealed that key genes include those involved in drug metabolism and transport. Intended therapeutic targets, though important, are detected in only a fraction of signatures, and cell line tissue of origin has been identified as a confounding factor throughout the signature detection procedure. In practice, the length of the signatures largely exceeds the number of sensitive cell lines available for each drug, which often yields inconsistent results between cell panels from different laboratories [147]. The current challenge is to filter and characterise transcriptional signatures so that they become robust, informative, and more homogeneous, while still retaining the complexity (hence the predictive power) of the original profiles [32].

Network biology offers means to integrate a large amount of omics data [74]. Most network biology capitalises on the observation that genes whose function is altered in a particular phenotype tend to be co-expressed in common pathways and, therefore, co-localized in specific network regions [150]. Following this principle, it has been possible to convert genome-wide signatures to network signatures, or modules, that are less noisy and easier to interpret [151]. Raphael and co-workers, for instance, developed an algorithm to map cancer mutations on biological networks and identify 'hot' regions that distinguish functional (driver) mutations from sporadic (passenger) ones [75]. Califano's group combined gene expression data with regulatory cellular networks to infer protein activity [152]. Overall, network-based methods come in many flavours and offer an effective framework to organise the results of omics experiments [68].

While many genes and proteins have enjoyed such a network-based annotation (being circumscribed within well-defined modules such as pathways and biological processes), drug molecules remain mostly uncharacterised in this regard. For a number of drugs, the mechanism of action is unclear [139] and off-targets are often discovered [153]. Recent publications of

drug screens against cancer cell line panels, and the transcriptional signatures that can be derived from there, provide a broader view of drug activity and enable the full implementation of network biology techniques. Here we undertake the task of obtaining and annotating transcriptional modules related to 189 drugs. We show how these modules are able to capture meaningful aspects of drug biology, being robust to inherent biases caused by, for example, the cell's tissue of origin, and having a tight relationship to mechanisms of action and transportation events occurring at the membrane. Finally, we perform a series of functional enrichment analyses, which contribute to a better understanding of the molecular determinants of drug activity.

3.1.3 Results

The Genomics of Drug Sensitivity in Cancer (GDSC) is the largest cancer cell line (CCL) panel available to date[33]. This dataset contains drug sensitivity data (growth-inhibition, GI) for 265 drugs screened against 1,001 cell lines derived from 29 tissues, together with basal transcriptional profiles of the cells (among other omics data). Aware of the work by Rees et al.[142], we first looked for the dominant effect of certain tissues in determining associations between drug response and gene expression. We found that CCLs derived from neuroblastoma, hematopoietic, bone, and small cell lung cancers may confound global studies of drug-gene correlations due to their unspecific sensitivity to drugs (Fig. A.1.2a). These tissues were excluded from further analyses. We also excluded genes whose expression levels were low or constant across the CCL panel and drugs tested against fewer than 400 CCLs (see the *Methods* section for details). As a result, we obtained a pharmacogenomic dataset composed of 217 drugs, 15,944 genes, and 671 CCLs.

Following the conventional strategy to analyse pharmacogenomic datasets, we calculated independent drug-gene associations simply by correlating the expression level of each gene to the potency of each drug (area over the growth-inhibition curve; 1-AUC) across the CCL panel. We used a Z-transformed version of Pearson's correlation, as recommended elsewhere[154]. Figure 3.1.1a shows the pairwise distribution of the Z-correlation (zcor) measures between the 15,944 genes and the 217 drugs. We validated the correlations identified in the GDSC panel on an independent set by applying the same protocol to the Cancer Therapeutic Response Portal (CTRP) panel[142] (Fig. A.1.3b). To identify the strongest drug-gene as-

sociations, we set a cutoff of ±3.2 zcor, based on an empirical null distribution obtained from randomised data (see Fig. A.1.1c and the *Methods* section). Please note that this is a widely adopted procedure that is not designed to detect single drug-gene associations (which would require multiple testing correction)[155]. Instead, and similar to signature identification in differential gene expression analysis, the goal is to identify sets of genes that are (mildly) correlated with drug response. For each drug, we obtained a median (Med) of 249 positively correlated genes [first quartile (Q1): 120, third quartile (Q3): 584], and Med of 173 negatively correlated genes [Q1: 59, Q3: 484] (Fig. 3.1.1b). Some drugs, like the BRAF inhibitor dabrafenib, or the EGFR inhibitor Afatinib, had over 1500 positively and negatively correlated genes, while others, like the antiandrogen Bicalutamide or the p38 MAPK inhibitor Doramapimod, had hardly a dozen. We observed that the number of genes that correlate with drug response strongly depends on the drug class (Fig. 3.1.1c), EGFR and ERK-MAPK signalling inhibitors being the classes with the largest number of associated genes, and JNK/p38 signalling and chromatin histone acetylation inhibitors being those with the fewest correlations. This variation may be partially explained by the range of drug potency across the CCL panel, as it is 'easier' to detect drug-gene correlations when the drug has a wide sensitivity spectrum (Fig. A.1.4).

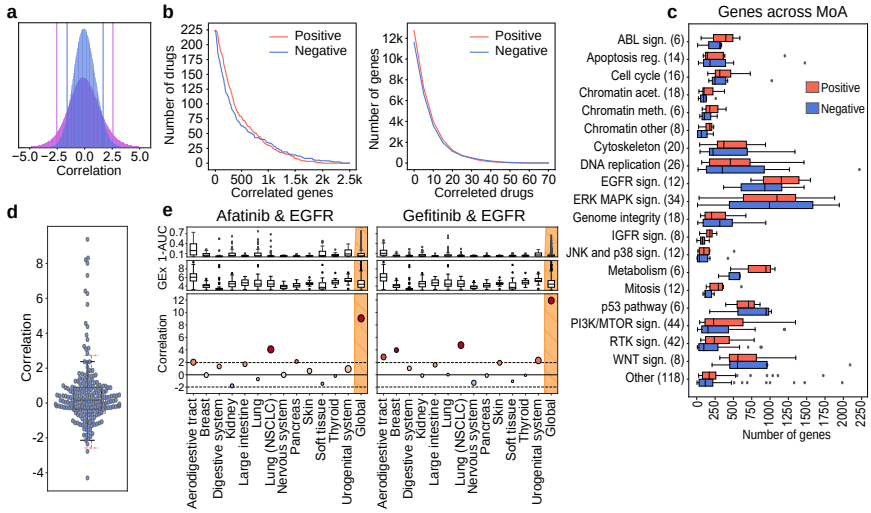


Figure 3.1.1: Analysis of drug-gene correlations. **a** Observed drug-gene correlation distribution (purple) and randomised drug-gene correlation distribution (blue) (random permutation of expression values). Vertical lines denote the percentiles 5th and 95th of each distribution. **b** The left panel shows the 'number of correlated genes per drug', while the right panel shows the 'number of correlated drugs per gene'. In the left panel, one can read, for example, that there are about 25 drugs (y-axis) with at least 1,250 correlated genes (x-axis). Likewise, in the right panel, one can read that about 4,000 genes (y-axis) are correlated to at least 10 drugs (x-axis). **c** Number of positively (red) and negatively (blue) correlated genes across drug classes. **d** Positively correlated targets (see the *Methods* section for details on the *z* score normalisation procedure of this correlation measure). Each dot represents one drug-target correlation. A full account of drug-target annotations is provided in Supplementary Data 8. The red box plot shows the background (random) distribution. **e** Drug-gene correlations (zcor) between Afatinib/Gefitinib and the epidermal growth factor receptor (EGFR) across tissues. In the upper plots, we show the drug sensitivity (1-AUC) across tissues. In the middle plots, we show basal gene expression of EGFR across tissues. The bottom plots show the Afatinib/Gefitinib-EGFR correlation. The rightmost values refer to the correlation when all tissues are considered (Global). The size of the bubbles is proportional to the number of CCLs in each tissue.

Similarly, analysis of independent drug-gene correlations suggests that some genes are positively correlated to many drugs. For instance, we found 5% of the genes to be associated with more than 10% of the drugs (Fig. 3.1.1b and A.1.3a). The transcripts of these 'frequent positively correlated genes' are enriched in membrane processes, specifically focal adhesion (p value $< 5.2\times10^{-12}$) and extracellular matrix (ECM) organisation (p value $< 5\times10^{-16}$), including subunits of integrin, caveolin, and platelet-derived growth factors (PDGFs). These genes determine, among others, the activation of Src kinases [156,157,158,159]. Overall, ECM proteins are known to play an important role in tumour proliferation, invasion, and angiogen-

esis [160,161] and are often involved in the upstream regulation of cancer pathways [162] such as PI3K/mTOR [156,157,158], MAPK [157], and Wnt signalling [163], and in cell cycle and cytoskeleton regulation [164]. It is thus not surprising that ECM genes determine drug response in a rather nonspecific manner.

On the other hand, 'frequent negatively correlated genes' are associated with small molecule metabolism (xenobiotic metabolic processes, p value $< 3.2 \times 10^{-3}$). In this group, we found, among others, the cytochrome CYP2J2 and the GSTK1 and MGST glutathione transferases, which are highly expressed in cancers and known to confer drug resistance through their conjugating activity [165,166,167,168]. Following other studies that reported similar results [142], we checked for the presence of multidrug transporters (MDTs). Reassuringly, we found the efflux pump transporter ABCC3 and a total of 27 different solute carriers (SLCs) to be negatively correlated to the potency of many drugs. Of note, we also found the ABCA1 transporter and other 8 SLCs to be among the frequent positively correlated genes, thus emphasising the key role of transporters and carriers in determining drug potency.

All of the above suggests that systematic analysis of independent drug-gene correlations is sufficient to highlight unspecific determinants of drug sensitivity and resistance (i.e., frequent positively and negatively correlated genes). However, while these determinants are recognized to play a crucial role, they do not inform targeted therapies, as they are usually unrelated to the mechanism of action of the drug. Thus, we assessed whether measuring drug-gene correlations would also be sufficient to elucidate drug targets (i.e., we tested whether the expression level of the target correlates with the potency of the drug). Since most drugs had more than one annotated target, to measure significance, we randomly sampled 1,000 times an equal number of genes and derived an empirical z score (see the *Methods* section). Figure 3.1.1d shows that the expression level of most drug targets did not correlate with drug response. In fact, only ∼10% of the drugs had 'positively correlated targets' (z score > 1.9, p value ∼0.05). Remarkably, the 6 EGF pathway inhibitors in our dataset were among these drugs, as were 3 of the 4 IGF pathway and 3 of the 21 RTK pathway inhibitors. We noticed that the molecular targets for these pathways were usually cell surface receptors (e.g., EGFR, IGFR, ALK, ERBB2, MET, and PDGFRA). Overall, of the 20 drugs with positively correlated targets, 13 bind to cell surface receptors, showing a propensity of drug-gene correlations to capture membrane targets (odds ratio = 15.13, p value = 1.9×10^{-7}). In Fig. A.1.5, we show how this trend is driven mostly by the over-expression of

the target on the cell surface.

The relatively small number of positively correlated targets illustrates how the analysis of expression levels alone is insufficient to reveal MoAs, especially when the drug target is located downstream of the cell surface receptors in a signalling pathway. Some authors have suggested that the tissue of origin of the cells might play a confounding role in defining drug response signatures. To address this notion, we repeated the calculation of Pearson's zcor correlations separately for each of the 13 tissues in our dataset. In general, the trends observed at the tissue level were consistent with the global trends, although tissue-specific correlations were milder due to low statistical power (i.e., few cell lines per tissue) (Fig. A.1.3c, right panel). Accordingly, we confirmed that none of the tissues had a globally dominant effect on the measures of drug-gene correlations (Fig. A.1.2b) and verified that certain tissue-specific associations were still captured by the analysis. For instance, going back to the targeting of EGFR (which was positively correlated with Afatinib and Gefitinib), we show in Fig. 3.1.1e that the 'global' correlation can be partly attributed to non-small cell lung cancer (NSCLC) cells (zcor > 1.96, p value < 0.05). Indeed, Afatinib and Gefitinib have an approved indication for NSCLC. Both drugs correlate with EGFR also in the aerodigestive tract, an observation reported in an independent study dedicated to the discovery of drug-tissue/mutation associations (ACME) [141]. Moreover, and consistent with recent findings [169,170,171,172], Gefitinib has a significant correlation to EGFR in breast cancers, whereas Afatinib correlates with this target in pancreatic CCLs. Afatinib, in turn, is associated with ERBB2 in breast CCLs, as also confirmed by ACME analysis (Fig. A.1.3e).

From drug-gene correlations to drug module

The previous analysis demonstrates that conventional drug-gene correlations do not directly identify drug targets and suggests that standard transcriptional drug signatures contain unspecific and indirect correlations that may mislead mechanistic interpretation. Recent advances in network biology precisely tackle these problems, as they can (i) filter signatures to make them more functionally homogeneous and (ii) allow for the measurement of network distances so that genes proximal to the target can be captured and connected to it, even if the expression of the target itself is not statistically associated with the drug.

Hence, we set to mapping drug-gene correlations onto a large protein-

protein interaction (PPI) network, retaining only genes that can be grouped in network modules (i.e., strongly interconnected regions of the network). In the *Methods* section, we explain in detail the module detection procedure. In brief, starting from drug-gene correlations (Fig. 3.1.2a), we first filtered out those genes whose expression was frequently (and unspecifically) correlated to the potency of many drugs (Fig. A.1.3a). This reduced the number of associations to 182 [median; Q1: 84, Q3: 372] positively and 122 [median; Q1: 41, Q3: 337] negatively correlated genes per drug, respectively. Next, in order to identify genes acting in coordination (i.e., participating in enriched Reactome pathways [173,174]), we adapted the gene set enrichment analysis (GSEA) algorithm [175] to handle drug-gene correlations (instead of gene expression fold-changes) (Fig. 3.1.2b). The resulting GSEA-filtered list of genes kept 100 [median; Q1: 49, Q3: 277] positive and 77 [median; Q1: 30, Q3: 221] negative correlations per drug. After this filtering, we submitted this list to HotNet2 [176], a module detection algorithm that was originally developed for the identification of recurrently mutated subnetworks in cancer patients (Fig. 3.1.2c and A.1.6 show the importance of the Reactome-based filtering previous to HotNet2). As a reference network (interactome) for HotNet2, we chose a high-confidence version of STRING [177], composed of 14,725 proteins and 300,686 interactions. HotNet2 further filtered the list of genes correlated to each drug, keeping only those that were part of the same network neighborhood. Finally, we used the DIAMOnD module expansion algorithm [173] to recover strong drug-gene correlations that had been discarded along the process. Although this step made a relatively minor contribution to the composition of the modules (less than 5% of the genes; Fig. A.1.7), we did not want to lose any strong association caused by the limited coverage of the Reactome database (Fig. 3.1.2d).

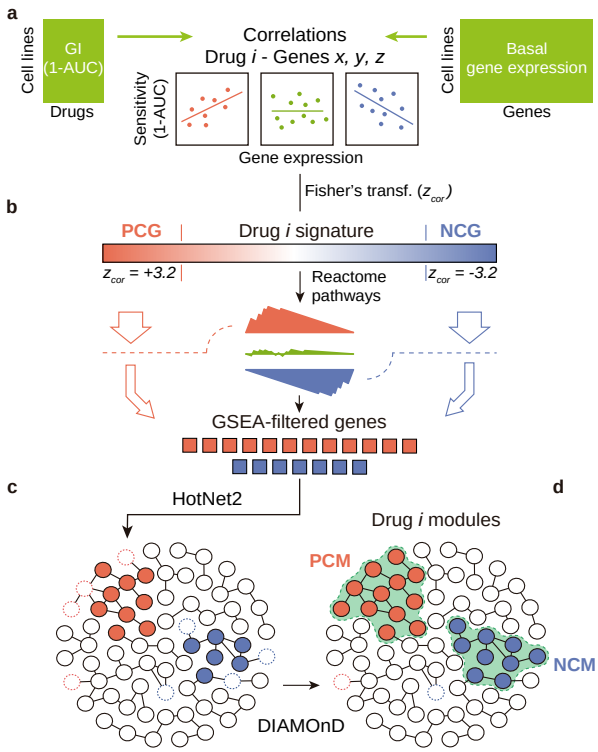


Figure 3.1.2: Methodological pipeline to identify drug modules. **a** The process of obtaining modules starts with the calculation of Z-transformed Pearson's correlation (zcor) between gene expression and drug sensitivity data for each drug-gene pair. Correlations (zcor) beyond ± 3.2 are considered to be significant. **b** We then run a gene-set enrichment analysis (GSEA) for each drug to identify genes that participate in enriched Reactome pathways. **c** GSEA-filtered genes are submitted to HotNet2 on the STRING interactome to identify drug modules. **d** Finally, modules are expanded (when possible) using the DIAMOnD algorithm to recall the few correlated genes that might have been excluded in step *(c)* as a result of the limited coverage of the Reactome database. This final step has a minor impact on the composition of the module (see Fig. A.1.7).

Our pipeline yielded at least one 'positively correlated module' (PCM) for 175 of the 217 drugs (48 genes [median; Q1: 23, Q3: 83]). Similarly, we obtained 'negatively correlated modules' (NCMs) for 154 of the drugs (40 genes [median; Q1:21, Q3:78]). Thus, compared to the original signatures, drug modules are considerably smaller (80% reduction) (Fig. 3.1.3a) and are commensurate with manually annotated pathways in popular databases (Fig. A.1.8). For roughly two thirds of the drugs, we obtained only one PCM and one NCM. For the remaining drugs, a second (usually smaller) module was also identified (Fig. A.1.9a). The complete list of drug modules can be found in Supplementary Data 2. Pairwise drug-gene correlations of the modules are listed as Supplementary Data 3.

Drug modules are tightly related to mechanisms of action

To assess the mechanistic relevance of drug modules, we measured their distance to the corresponding drug targets (i.e., we formulated the hypothesis that drug targets should be 'proximal' to dysregulated network regions). To this end, we used the DIAMOnD algorithm again[173], this time to retrieve, for each drug, a list of genes ranked by their proximity to the corresponding drug module(s) (see the *Methods*). Figure 3.1.3b shows that drug targets are remarkably up-ranked in these lists, making them closer to the drug modules than any other set of random proteins, including druggable genes and pharmacological receptors[178], which usually have prominent positions in the PPI network due to the abundant knowledge available for them. In 82% of the PCMs, the corresponding targets were among proximal proteins (top decile), which means a dramatic increase in mechanistic interpretability compared to the 12.25% of drugs that could be linked to their targets via conventional analysis of drug-gene correlations.

A unique feature of drug modules is that network-based distances can be natively measured between them[79]. We computed the distance between drug modules pairwise (Supplementary Data 4) and grouped them by drug class (Fig. 3.1.3c) (see *Methods* and alternative statistical treatments in Fig. A.1.10). The diagonal of Fig. 3.1.3c clearly indicates that drugs belonging to the same category tend to have 'proximal' modules (some of them in a highly specific manner, like in the case of ERK-MAPK signalling cascade inhibitors). Most interestingly, we could observe proximities between modules belonging to different drug classes. For instance, modules of drugs targeting RTK signalling were 'located' near those of drugs affecting genome integrity, in good agreement with recently reported cross-talk between these two processes[179,180]. Likewise, and as proposed by some studies[181,182,183], IGFR-related drugs were 'proximal' to drugs affecting cell replication events such as mitosis, cell cycle, and DNA replication.

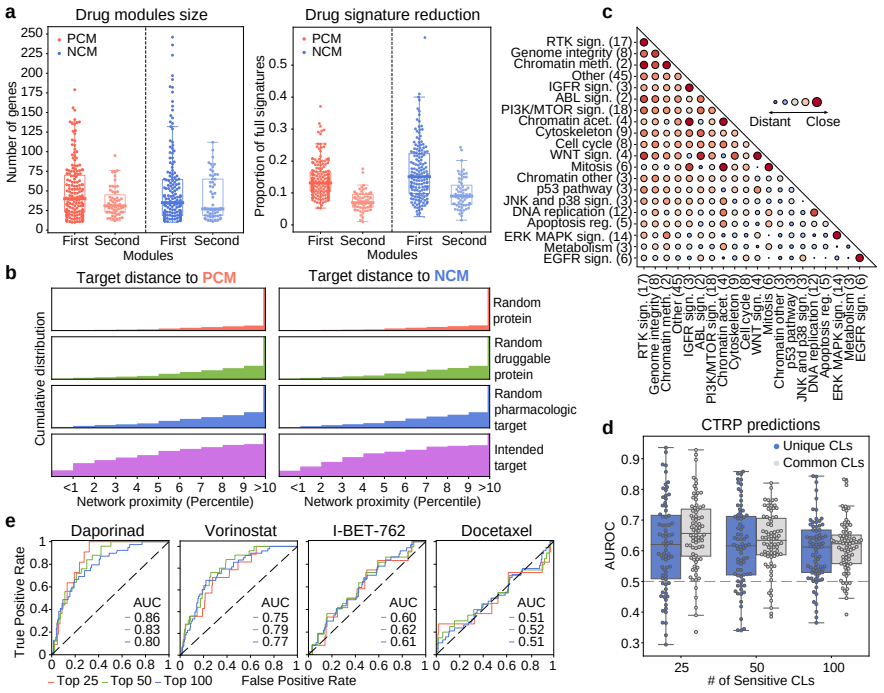


Figure 3.1.3: Global drug module analysis. **a** Number of genes in positively and negatively correlated modules (PCMs and NCMs) (left). Proportion of genes in the modules with respect to PCGs and NCGs (i.e., full signature). **b** Distance between drug targets and PCMs/NCMs (purple cumulative distribution). Results are compared to random proteins from the STRING[177] interactome (red), proteins sampled from the 'druggable proteome' (Target Central Resource Database[184]) (green), and proteins sampled from the pharmacological targets in DrugBank[185] (blue). **c** Network-based distances between drug classes. The bigger the bubble, the closer the distance between drug classes. Drug classes are sorted by specificity in their proximity measures. Please note 'distant' values in the diagonal are possible due to differences in drug modules belonging to the same class. The network-based distance calculation is detailed in the *Methods* section. **d** Predictive performances (AUROC) of the drug modules evaluated in the CTRP panel (top 25, 50, and 100 sensitive CCLs). Blue distributions correspond to results using unique CCLs (i.e., not shared with the GDSC panel). **e** Illustrative ROC curves for Daporinad (FMK866), Vorinostat, I-BET-762, and Docetaxel.

Drug modules retain the ability to predict drug response

We have shown that drug modules are related to the MoA of the drug, but the question remains as to the extent to which they retain the predictive capabilities of the full transcriptional profiles/signatures. In the CCL setting, gene expression profiles are valuable predictors of drug response[29,144,186] and crucially contribute to state-of-the-art pharmacogenomic models. To test whether our (much smaller) drug modules retained

predictive power, we devised a simple drug sensitivity predictor based on the GSEA score (see the *Methods* section). In brief, given a drug, we tested whether cell lines sensitive to a certain drug were enriched in the corresponding drug modules. We expect genes in PCMs to be over-expressed in sensitive cell lines and those in NCMs to be under-expressed. Analogously, we took the positively and negatively correlated genes from the full drug-gene associations (signatures) and also performed a GSEA-based prediction. To nominate a cell 'sensitive' to a certain drug, we ranked CCLs by their sensitivity and kept the top *n* CCLs, *n* being 25, 50, or 100, based on the distribution of sensitive cell lines provided by the authors of the GDSC (Fig. A.1.11a). This simple binarization is, in practice, proportional to more sophisticated 'sensitive/resistant' categorizations such as the waterfall analysis[147], and it yields prediction performance metrics comparable between drugs.

Figure A.1.12 suggests that, when applied to the GDSC, drug module enrichment analysis can classify sensitive cell lines with high accuracy, especially for the top 25 sensitive cell lines (AUROC = 0.77), which is a notable achievement considering that drug modules are 80% smaller than the original signatures. To assess the applicability of our modules outside the GDSC dataset, we performed an external validation with the CTRP panel of cell lines. About 37% of our drugs were also tested in this panel. In CTRP, drug sensitivity is measured independently of GDSC, which poses an additional challenge for prediction as a result of experimental inconsistencies[147]. Of the CCLs, 397 are shared between GDSC and CTRP, and gene expression data are also measured independently. We performed the GSEA-based sensitivity prediction for all CTRP CCLs. Figure 3.1.3d and e show the distribution of prediction performances for the 70 drugs, and illustrative ROC curves corresponding to four drugs (namely Daporinad, Vorinostat, I-BET-762 and Docetaxel), respectively. We found that, when focusing on the top 25 sensitive CCLs, over a quarter of the drugs had AUROC > 0.7, including Daporinad. Acceptable (AUROC > 0.6) predictions were achieved for half the cases (e.g., Vorinostat and I-BET-762), which is a comparable result to recent attempts to translate sensitivity predictors between different CCL panels[187]. For the remaining drugs, predictive performance did not differ to random expectation (AUROC < 0.6) (e.g., Docetaxel). Notably, performance declined only slightly when considering CCLs that were exclusive to the CTRP panel (i.e., not part of the GDSC dataset) (Fig. 3.1.3d, blue boxes). The figure was comparable, if not better, to that obtained using full signatures (PCGs

and NCGs) (Fig. A.1.12, grey boxes). These observations support previous recommendations to pre-filter pharmacogenomic data based on prior knowledge [188] (Fig. A.1.13).

Module-based characterization of drugs

Since drug modules are highly connected in biological networks, they are expected to be (at least to some extent) functionally coherent and easier to interpret. Accordingly, we tested the enrichment of drug modules in a collection of high-order biological processes (the Hallmark gene sets) available from the Molecular Signatures Database (MSigDB) [189]. Figure A.1.14a shows that the number of enriched Hallmark gene sets depends upon the MoA of the drug. The results of the enrichment analysis are given in Supplementary Data 5 and as an interactive exploration tool based on the CLEAN methodology (Supplementary Data 9; `https://figshare.com/s/932dd94520d4a60f076d`) [190]. We chose three drug classes to illustrate how to read these results, namely drugs targeting mitosis, RTK signalling inhibitors, and ERK-MAPK signalling inhibitors (Fig. 3.1.4a).

Drugs targeting mitosis have modules enriched in cell cycle and replication processes (Fig. 3.1.4a, top). Specifically, genes related to the Myc transcription factor are over-represented in three of the drug modules (NPK76-II-72-1, GSK1070916, and MPS-1-IN-1). The modules of these drugs have a rather distinct composition, NPK76-II-72-1 having the largest coverage of Myc-related genes and being, together with MPS-1-IN-1, related to both Myc1 and Myc2 processes. In Fig. A.1.14b-d, we show how, for these two drugs, cell line sensitivity is dependent on Myc expression levels.

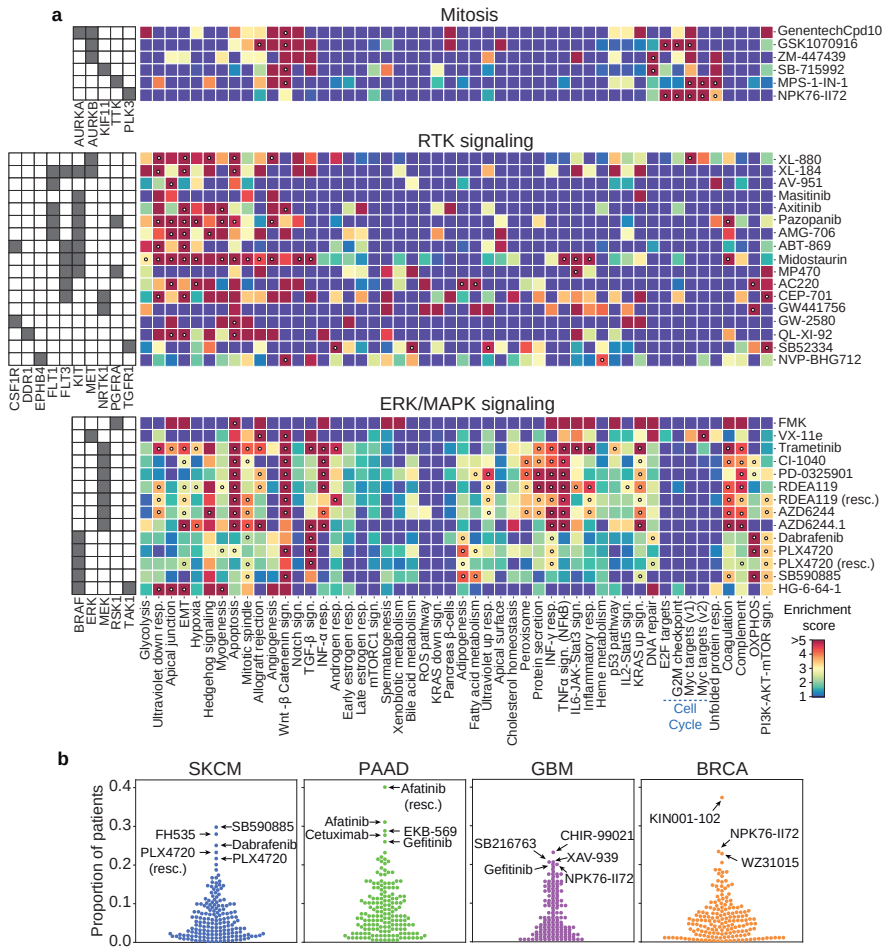


Figure 3.1.4: Drug module characterization. **a** Drug module enrichment analysis based on the Hallmark gene set (odds ratios in colour, p values < 0.05 are marked with a white dot). For simplicity, three drug classes are shown: drugs affecting mitosis, RTK signalling, and ERK/MAPK signalling. **b** TCGA enrichment analysis of PCMs in four types of cancer: SKCM (Cutaneous Melanoma), PAAD (Pancreatic Adenocarcinoma), GBM (Glioblastoma Multiforme), and BRCA (Breast Carcinoma).

In contrast to mitosis inhibitors, drugs targeting the RTK pathway are enriched in biological processes outside the nucleus (Fig. 3.1.4a, middle), among these hypoxia and the epithelial-mesenchymal transition (EMT). Both mechanisms are known to be associated with tyrosine kinases [191,192]. Interestingly, a subgroup of RTK inhibitors (namely ACC220, CEP-701, NVP-BHG712, and MP470) is characteristically associated with the PI3K-AKT-mTOR signalling cascade. With the exception of NVP-BHG712, these inhibitors have the tyrosine kinase FLT3 as a common target [193,194].

Deeper inspection of FLT3 inhibitors reveals module proximities to certain PI3K inhibitors (e.g., GDC0941), and the PI3K-AKT-mTOR pathway is enriched in ERBB2 inhibitors as well (Supplementary Data 4 and 5).

As for ERK-MAPK pathway inhibitors, we observed a total of 17 enriched Hallmarks, making this class of drugs the one with most variability in terms of enrichment signal of the modules (Fig. 3.1.4a and Fig. A.1.14a). However, while some processes like apoptosis are detected in most of the drugs in this category, others are target-specific. Oxidative phosphorylation (OXPHOS), for example, is represented in 3 of the 4 BRAF inhibitors. It is known that, while BRAF inhibitors boost OXPHOS (leading to oncogene-induced senescence), activation of glycolytic metabolism followed by OXPHOS inactivation yields drug resistance [195,196]. Similarly, VX11e (the only drug in our dataset targeting ERK2) shows a distinctive enrichment in Myc-regulated proteins, while FMK (the only drug targeting the Ribosomal S6 kinase) is enriched in p53 signalling pathway and inflammatory response processes. All these observations are consistent with previous studies [197,198,199,200], and Fig. A.1.15b demonstrates that the variability observed between drugs in this class is driven mostly by differences in the sensitivity profiles of the drugs.

Overall, the enrichment signal (i.e., the functional coherence) of drug modules is substantially higher than that of full signatures (PCGs and NCGs) (Fig. A.1.16a). This facilitates, in principle, the mechanistic interpretation of drug-gene correlation results (Fig. A.1.15a). We show an illustrative module (CEP-701) in Fig. A.1.16c.

We next examined whether our results could be extended beyond CCL panels. We found that drug modules are indeed identified (GSEA p value < 0.001) in the majority of patients in the TCGA clinical cohort (Fig. A.1.15c; see the *Methods* section for details). Closer inspection by TCGA tumour type further supports the clinical relevance of our results (Supplementary Data 6). For example, drugs affecting MAPK signalling (specifically, BRAF inhibitors, e.g., Dabrafenib) have a tendency to 'occur' in skin cutaneous melanomas (SKCM), as expected (Fig. 3.1.4e, blue). Of note, one PPAR inhibitor (FH535) was also found enriched in a high number of SKCM patients, in good agreement with work by others proposing the use PPAR inhibitors to treat skin cancer [201,202]. Similarly, we observed an abundance of EGFR inhibitor modules among pancreatic cancers (PAAD) (Fig. 3.1.4e, green), in line with the known crucial role of EGFR in pancreatic tumorigenesis [203,204]. As for glioblastomas (GBMs)

(Fig. 3.1.4e, purple), we found two GSK3 inhibitors (CHIR-99021 and SB216763) and one TNKS inhibitor (XAV939), all of them targeting Wnt signalling, which is a potential mechanism against this tumour type[205]. We also found one EGFR inhibitor (Gefitinib) and the PLK inhibitor NPK76-II-72-1 mentioned above in the context of Myc enrichment analysis. Both mechanisms have shown promise in EGFR- and Myc-activated gliomas, respectively[206,207]. Finally, we encountered a more heterogeneous pattern in breast cancer patients (BRCA) (Fig. 3.1.4e, orange), including mechanisms supported by the literature, such as AKT, IRAK1, and PLK3 inhibition[208,209,210].

Beyond the tumour-type level, we looked for modules that were significantly enriched (odds ratio > 2, p value < 0.001) in patients harbouring specific driver mutations (see the *Methods* section). A full account of this enrichment analysis is given in Supplementary Data 7. We found, for instance, that modules of drugs targeting ERK/MAPK signalling are related to patients with mutations in HRAS and BRAF[211,212] and that, in turn, BRAF is (together with KRAS) frequently mutated in patients 'expressing' modules of EGFR signalling inhibitors[213]. Taken together, and although TCGA treatment response data is too scarce to allow for prediction assessment[214], these results indicate that the drug modules identified in CCLs hold promise for translation to clinical practice.

3.1.4 Concluding remarks

Two limitations of large-scale pharmacogenomic studies are the difficulty to reproduce results across screening platforms and the eventual translation to clinics, as it remains unclear whether immortalised cells are able to model patient samples[215]. Another important limitation is the overwhelming number of drug-gene correlations that can be derived from these experiments, yielding signatures of drug sensitivity that are almost impossible to interpret. We have shown, for example, that (i) the number of correlated genes is highly variable across drugs, (ii) some genes are unspecifically correlated to many drugs, and (iii) not all drug-gene pairs are equally correlated in every tissue. We propose that converting transcriptional signatures to network modules may simplify the analysis, since network modules are smaller, more robust, and functionally coherent. We have validated this strategy by proving that drug response modules, which are enriched in biological processes of pharmacological relevance and exhibit comparable predictive power to the full signatures, are tightly related

to the MoA. Further, we have characterised the modules extensively (Supplementary Data 8 and e.g., Fig. A.1.17) and confirmed their occurrence in the TCGA clinical cohort (Supplementary Data 6 and Supplementary Data 10).

However, our approach does have some of the limitations of ordinary transcriptomic analyses. Expression levels of mRNA do not perfectly match protein abundance, nor are they able to capture post-translational modifications such as phosphorylation events, which are key to some of the pathways studied here. Moreover, wide dynamic ranges in gene expression and drug sensitivity data are necessary for drug-gene correlations to be captured, thus requiring, in practice, considerably large panels of CCLs, which limits the throughput of the technique to a few hundred drugs. In particular, one cannot precisely measure correlations within poorly represented tissues, which in turn makes it difficult to disentangle tissue-specific transcriptional traits that may be irrelevant to drug response. Our module-based approach partially corrects for this confounding factor, although the integration of other CCL omics data (such as mutations, copy number variants and chromatin modifications) could further ameliorate these issues and also provide new mechanistic insights. In this context, systems biology tools that learn the relationships between different layers of biology are needed. Along this line, the release of CCL screens with readouts other than growth inhibition or proliferation rate[38,216] will help unveil the connections between the genetic background of the cells and the phenotypic outcome of drug treatment.

All in all, transcriptomics is likely to remain the dominant genome-wide data type for drug discovery, as recent technical and statistical developments have drastically reduced its cost[217]. The L1000 Next-generation Connectivity Map, for instance, contains about one million post-treatment gene expression signatures for 20,000 molecules[38]. These signatures await to be interpreted and annotated, and more importantly, they have to be associated with pre-treatment signatures in order to identify therapeutic opportunities. We believe that network biology strategies like the one presented here will enable this connection, encircling relevant 'regions' of the signatures and measuring the distances between them.

3.1.5 Methods

Data download and pre-processing

We collected gene expression and drug response data from the GDSC resource (https://www.cancerrxgene.org). We first discarded those genes whose expression levels were low or stable across cell lines (Fig. A.1.1a). To this end, we analysed the distribution of basal expression of each gene in every CCL and filtered out those with an expression level below 4.4 (log2 units) across the panel (see Fig. A.1.1b for a robustness analysis). Regarding drug response data, GDSC provides measurements of cell survival at a range of drug concentrations (area under the dose-response curve (AUC)). Since this measure is inversely proportional to drug sensitivity (i.e., the more sensitive the cell, the shorter its survival), we used the 1-AUC as a measure of potency. Thus, positive correlations denote drug sensitivity caused by gene overexpression while negative correlations indicate that sensitivity is associated with gene underexpression.

Recent studies report a confounding effect of certain tissues in the global analysis of drug-gene correlations [142]. In order to identify these potential biases in our dataset, we performed a principal component analysis (PCA) on the matrix of raw drug-gene correlations (Pearson's correlation between 1-AUC and gene expression units). Then, we correlated the loadings of the first PC with gene expression values for each CCL. Finally, we filtered out CCLs belonging to tissues that were strongly correlated to the drug-gene correlation profiles (Fig. A.1.2). We removed leukaemia, myeloma, lymphoma, neuroblastoma, small cell lung cancer (SCLC), and bone CCLs. In addition, we considered only drugs with sensitivity measurements available for at least 400 CCLs, as recommended by Rees et al. [142].

Drug gene correlations

After this filtering process, we recalculated, for each drug-gene pair, the Pearson's correlation between basal gene expression and 1-AUC drug potencies across CCLs. We applied Fisher's Z-transformation to the correlation coefficients in order to account for variation in the number of CCLs available for each drug [154]. Overall, we obtained positive and negative drug-gene correlations for 217 drugs and 15,944 genes across a total of 671 CCLs. Drug-gene correlations (zcor) beyond ± 3.2 were considered to be significant (Fig. A.1.1c-d show that this cutoff is a robust choice).

For each gene, we counted the number of correlated drugs (zcor beyond ±3.2) and inspected the resulting cumulative distribution (Fig. A.1.3a). Genes at the 5% end of the distribution were considered to be 'frequently correlated genes' (FCGs). We found 869 positive and 799 negative FCGs, which were removed from further analyses. Finally, we performed enrichment analyses (hypergeometric tests) on those genes using the Gene Ontology database[218] and the DAVID toolbox (https://david.ncifcrf.gov/summary.jsp).

To obtain tissue-specific correlations we first split the CCL panel into sets of CCLs belonging to the same tissue. We then calculated drug-gene correlations (zcor) separately for each of the 13 tissues represented in our dataset. In order to verify that measures of positively correlated genes (PCGs) and negatively correlated genes (NCGs) were consistent across tissues, we calculated the median zcor across tissues for each drug-PCG/NCG pair. In general, tissue-specific correlations had the same 'direction' (i.e., same sign of zcor) as the global correlation used throughout the study (Fig. A.1.3c).

To obtain drug-target correlations we first obtained drug targets from the GDSC resource (disambiguating them with DrugBank[219], when necessary). We assigned at least one target to 202 of the 217 drugs. We focused on the zcor correlation of the targets to check whether target expression (positively) correlates with drug sensitivity. When more than one target was annotated per drug, we kept the maximum correlation. To validate the statistical significance of this measure, we randomly sampled genes (corresponding to the number of known targets per drug; here again, we kept the maximum correlation). This process was repeated 1000 times for each drug. The mean and the standard deviation of this null distribution were used to derive a z score, making results comparable between drugs.

Drug module detection

After removing frequently correlated genes from the list of drug-gene correlations, we kept 182 [median; Q1: 84, Q3: 372] positively and 122 [median; Q1: 41, Q3: 337] negatively correlated genes (PCGs, NCGs) per drug. Further, we used correlation values (zcor) to run a gene-set enrichment analysis (GSEA)[175] for each drug and identify the genes that participate in enriched Reactome pathways[173,174]. We only considered Reactome pathways composed of at least 5 genes. Then, for each drug, we kept the significantly correlated genes found in any of the enriched path-

ways (p value <0.01). The resulting GSEA-filtered list of genes retained 100 [median; Q1: 49, Q3: 277] positive and 77 [median; Q1: 30, Q3: 221] negative correlations per drug. Then, taking the zcor values as input scores, we submitted the GSEA-filtered list of genes to HotNet2 [176], using a high-confidence version of STRING [177] (confidence score >700). We ran HotNet2 iteratively, keeping the largest module and removing its genes for the next iteration, until the modules had fewer than 5 genes or were not statistically significant (p value >0.05). To recall strong drug-gene correlations 'proximal' to the drug modules (missed, most likely, by the incomplete coverage of Reactome), we used the DIAMOnD module-expansion algorithm [173]. We considered only genes that (i) were correlated to the drug response, (ii) were not present in any of the Reactome pathways, and (iii) were in the top 200 closest genes to the module, according to DIAMOnD (this cutoff was proposed by the authors of DIAMOnD based on orthogonal functional analyses). Hence, we obtained at least one positively correlated module for 175 of the drugs (48 genes [median; Q1: 23, Q3: 83]) and one negatively correlated module for 154 of the drugs (40 genes [median; Q1: 21, Q3: 78]). Robustness analysis of this procedure is found in Fig. A.1.1d. A GMT formatted list of the drug modules can be found in Supplementary data 2. The correlation values of the genes in the drug modules are available in Supplementary Data 3.

Distances between drug modules and targets

DIAMOnD [173] provides a list of genes sorted by their network-based proximity to the module. Accordingly, we retrieved from the STRING interactome the top closest 1450 genes (~10% of the largest connected component of the network) for every drug module. We then checked the ranking of drug targets in the resulting DIAMOnD lists, (conservatively) taking the median value when more than one target was available. To assess the proximity of drug targets to the modules, we measured distances to three different sets of random proteins. The first random set corresponded to the STRING proteome. For the second, we collected all genes defined as Tclin or Tchem in the Target Central Resource Database [178] (i.e., 'druggable proteins'). Finally, the third random set included all pharmacologically active drug targets reported in DrugBank (`https://www.drugbank.ca/`).

Distances between modules

We calculated distances between positively and negatively correlated modules separately using the network distance proposed by Menche et al.[79]. This distance measure is sensitive to the number of genes (size) included in the modules. To normalise this measure, we devised the following procedure. First, we grouped drug modules on the basis of their size. Then, for each module, we calculated the distribution of shortest distances from each gene to the most central one[220]. We used this distribution to sample random modules from the network. When the distribution constraint could not be fully met, we used the DIAMOnD algorithm[173] to retrieve the remaining genes (50% of the genes at maximum). We repeated this process to obtain 10 random modules of each size. Next, we distributed the random modules into ranges (intervals) of 5 (i.e., from 10 to 14 genes, from 15 to 19, etc.; 50 random modules per interval). Then, for each pair size, we randomly retrieved 100 pairs of modules and calculated the network-based distance between them. The mean and standard deviation of the distances at each pair size were used to normalise the observed distances, correspondingly (z score normalisation) (we checked that 100 random pairs were sufficient to approximate the mean and standard deviation of the population). The more negative the network distance (dnet), the more proximal the modules are. We provide the network distances as Supplementary Data 4.

Drug response prediction

We performed drug response predictions in the GDSC dataset by using drug modules (only first PCMs and NCMs, to make results comparable between drugs). We devised a simple GSEA-like predictor in which CCLs were evaluated for their up-/downregulation of the modules, correspondingly. To this end, we first normalised the expression of each gene across the CCL panel (z score). Then, for each drug, we ranked CCLs based on the GSEA enrichment scores (ES), taking drug modules as gene sets. To evaluate the ranking, we chose the top 25, 50, and 100 CCLs based on the known drug sensitivity profile. Performance was evaluated using the AUROC metric. Results were compared to those obtained with positively and negatively correlated genes (PCG, NCG) from the full signatures (zcor beyond $\pm$ 3.2).

To check whether modules derived from GDSC generalise to other CCL panels, we applied the same procedure to the Cancer Therapeutics

Response Portal (CTRP). As done with the GDSC panel, we removed all CCLs derived from neuroblastomas, hematopoietic, bone, and small cell lung cancer tissues, leaving a total of 636 CCLs, 397 in common with our GDSC panel (67 drugs in common). Drug response predictions for CTRP were performed as detailed above. We used the best ES among all modules associated with the drug. In addition, we did the analysis using CCLs exclusive to CTRP (i.e., not shared with the GDSC panel).

Module enrichment in Hallmark gene sets

We downloaded the Hallmark gene set collection from the Molecular Signature Database (MSigDB) of the Broad Institute (`http://software.broadinstitute.org/gsea/index.jsp`). We evaluated each gene set independently using a hypergeometric (Fisher's exact) test (first and second modules were merged, when applicable; the gene universe was that of GDSC). Enrichments can be found in the Supplementary Data 5.

Module enrichment in TCGA cohort

We downloaded gene expression data (median z scores) for 9,788 patients and 31 cancer tissues from the PanCancer Atlas available in the cBioPortal resource (`http://www.cbioportal.org`). 'Presence' or 'expression' of the module in each patient was evaluated using GSEA (p value <0.001), ensuring that the direction (up/down) of the enrichment score corresponded to the 'direction' of the module (PCM/NCM). For a complete list of enrichment results, please see Supplementary Data 6 (results are organized by tumour type). Further, to identify associations between drug modules and cancer driver genes, we checked whether patients 'expressing module of drug X' (p value <0.001) were 'harbouring a mutation in driver gene Y' (Fisher's exact test). We considered 113 driver genes (obtained as described in [221], using the 'known' flag) (Supplementary Data 7).

Annotating biological and network features

To characterise drug modules, we designed 21 features belonging to the following categories: (i) general features derived directly from the pharmacogenomics panel, (ii) network features related to network measures such as topological properties, and (iii) biological features encompassing a series of orthogonal analyses related to drug biology. For more information, please see Supplementary Data 8 and its corresponding legend.

Chapter 3.2
The Bioteque, a comprehensive repository of biomedical knowledge descriptors

Authors	Adrià Fernández-Torras, Miquel Duran-Frigola, Martino Bertoni, Martina Locatelli, Patrick Aloy
Type	Research Article
Stage	Published
Title	Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque.
Journal	Nature Communications
DOI	https://doi.org/10.1038/s41467-022-33026-0
Context	In the previous work we showed how biological knowledge, encapsulated as a form of a network, can be used to extract meaningful information from omics experiments. Nevertheless, although networks offer the means to mine insights from biological data, the systematic implementation of downstream analysis is unfeasible, hampering the comprehensive integration and exploitation of biomedical resources. In this chapter, we have developed a strategy to systematically encode network biomedical information into low-dimensional numerical vectors (embeddings). As a result, we have created a resource of pre-calculated biomedical descriptors that offers a thoroughly processed, tractable, and highly optimised assembly of the biomedical knowledge available in the public domain.
Note	Supplementary data can be accessed at the original publication.

3.2.1 Abstract

Biomedical data is accumulating at a fast pace and integrating it into a unified framework is a major challenge, so that multiple views of a given biological event can be considered simultaneously. Here we present the Bioteque, a resource of unprecedented size and scope that contains precalculated biomedical descriptors derived from a gigantic knowledge graph, displaying more than 450 thousand biological entities and 30 million relationships between them. The Bioteque integrates, harmonises, and formats data collected from over 150 data sources, including 12 biological entities (e.g., genes, diseases, drugs) linked by 67 types of associations (e.g., 'drug treats disease', 'gene interacts with gene'). We show how Bioteque descriptors facilitate the assessment of high-throughput protein-protein interactome data, the prediction of drug response and new repurposing opportunities, and demonstrate that they can be used off-the-shelf in downstream machine learning tasks without loss of performance with respect to using original data. The Bioteque thus offers a thoroughly processed, tractable, and highly optimised assembly of the biomedical knowledge available in the public domain.

3.2.2 Introduction

Systematic measurements of biological samples through omics technologies, together with efforts to distil the scientific literature into structured databases, are providing an ever-growing corpus of biomedical and biomolecular information[222]. Indeed, the data stored in the EMBL-EBI has increased sixfold in the last few years, from 40 petabytes in 2014 to over 250 in 2021[85]. Associated with this phenomenon, a variety of nomenclatures have been proposed, along with identifiers, levels of resolution (e.g., protein isoforms or gene splice variants) and experimental conditions, making data integration and harmonisation across platforms a challenging step[87]. As a result, even though as many as 1,641 resources were listed in the 2021 Online Molecular Biology Database Collection[86], only a small portion are broadly used, and hundreds remain isolated with their own particular formats[223,224]. Aware of the situation, several initiatives have emerged to standardise biological data by establishing common vocabularies and formats. For instance, the pioneering Harmonizome[90] was able to integrate knowledge from several gene-centric databases by representing data (e.g., gene expression, disease genetics, etc.) in a simple discretized format

that was applicable to each type of data.

Nowadays, in an attempt to capture the complexity of biological systems, multiple omics profiles are often measured simultaneously (i.e., trans-omics analyses)[225,226] so that complementary views of a given phenotype or event can be considered in parallel and as a whole[227]. However, current methods mainly adapt and combine existing strategies developed to analyse individual omics data, and often the net result is that most conclusions are drawn from the most informative single data type, while the rest are used as support. It is thus fundamental to devise strategies able to capture the coordinated interplay of the many regulatory layers present in biological systems. Himmelstein et al. suggested the use of knowledge graphs (KG) as a tool to integrate heterogeneous biomolecular data[228,93]. In a biomedical KG, nodes represent biological or chemical entities (e.g., genes, cell lines, diseases, drugs, etc.), and edges capture the interactions or relationships between them (e.g., 'drug treats disease' or 'cell upregulates gene'). This concept has recently been expanded to include clinical entities[92].

However, large biomedical networks are intractable by conventional graph analytics techniques[109], thus prompting the development of dimensionality reduction techniques that learn numerical feature representations of nodes and links in a low dimensional space (aka network embeddings). As a result, network embeddings reduce the dimensionality of the data while preserving the topological information and the connectivity of the original network[131]. Moreover, the vectorial format of the nodes resulting from network embedding approaches is better suited as an input for machine learning algorithms. For instance, Zitnik and Leskovek presented a set of protein embeddings that consider the protein interactions within each human tissue, as well as inter-tissue relationships, and showed their potential to predict tissue-specific protein functions[229]. Later on, the same authors embedded several networks (i.e., protein-protein, drug-target and disease-gene interactions) to explore the mechanisms of drug action[230]. Recently, Cantini et al. evaluated the capacity of several dimensionality reduction methods to integrate continuous multi-omics data (e.g., gene expression, copy number variation, miRNAs and methylation)[231], assessing their ability to preserve the structure of the original data and their prediction performance in different tasks. Overall, embedding-based descriptors provide a scalable and standard means to capture complex relationships between biological entities and they integrate the myriad of omics experiments associated with them[106,232].

To make biomedical knowledge embeddings available to the broad scientific community, we have developed the Bioteque, a resource of unprecedented size and scope that contains pre-calculated embeddings derived from a gigantic heterogeneous network (more than 450k nodes and 30M edges). The Bioteque harmonises data extracted from over 150 data sources, including 12 distinct biological entities (e.g., genes, diseases, compounds) linked through 67 types of relationships (e.g., 'compound treats disease', 'gene interacts with gene'). We demonstrate that Bioteque embeddings retain the information contained in the large biological network and illustrate with examples how this concise representation of the data can be used to evaluate, characterise and predict a wide set of experimental observations. Finally, we offer an online resource to facilitate access and exploration of the pre-calculated embeddings (https://bioteque.irbbarcelona.org).

3.2.3 Results

A comprehensive biomedical knowledge graph (KG)

To build a KG that integrates biological and biomedical knowledge available in the public domain, we first defined the basic entities (nodes) of the network and the relationships between them (edges). As shown in Fig. 3.2.1a, the resource is gene-centric. Thus, genes and gene products (GEN) are represented in the centre of the KG scheme and are involved in most associations. To better characterise genes and proteins, we collected their molecular function (MFN), cellular component localization (CMP), functional structure or domains (DOM), and biological processes or pathways (PWY). Additionally, we included information on cell lines (CLL), one of the most studied entities in biology, as well as their anatomical ensembles, namely the tissues (TIS). Analogously, chemical compounds (CPD) are depicted together with pharmacological classes (PHC) and chemical entities (CHE), two common vocabularies for medicinal compounds. Diseases (DIS) are abnormal conditions that have been widely studied in various fields, giving rise to a wide diversity of interactions between different nodes. Furthermore, although CPD and DIS are two of the major perturbational agents found in repositories like GEO[233] and LINCS[38], we also considered other biological entities such as miRNA, shRNA and overexpression vectors that can also act as perturbagens (PGN). To connect the entities in the Bioteque, we defined 67 types of associations re-

flecting biological relationships between them. An example of such an association would be a gene that is associated with a given pathway (GEN-ass-PWY) and might be downregulated in a certain cell (GEN-dwr-CLL) or tissue type (GEN-dwr-TIS), or a drug compound that is used to treat a disease (CPD-trt-DIS). A comprehensive list of all the biological and chemical entities included in the Bioteque, as well as the different associations, are summarised in Fig. 3.2.1a and Table 3.2.1 and provided in Supplementary Data 1 and 2.

Table 3.2.1: Biological and Chemical entities in the Knowledge Graph (KG). We show the number of nodes, metaedges and edges contained in the KG for each metanode.

Metanode	Abbreviation	Nodes	Metaedges	Edges
Cell	CLL	40,681	15	7,512,366
Cellular component	CMP	3,992	2	3,461,731
Chemical entity	CHE	115,002	2	435,011
Compound	CPD	153,279	12	5,713,785
Disease	DIS	10,144	10	5,037,293
Domain	DOM	16,913	2	85,747
Gene	GEN	20,229	42	25,788,255
Molecular function	MFN	11,006	2	164,447
Pathway	PWY	1,585	4	133,851
Perturbagen	PGN	66,988	7	2,889,047
Pharmacological class	PHC	6,072	2	31,004
Tissue	TIS	2,157	8	4,928,112

Having defined the biological entities and their interactions, we populated the Bioteque with data collected from representative datasets and resources. We first incorporated data from the Harmonizome[90], the most complete compendium of biological datasets to date, and added data from another 100 reference datasets. Each dataset was mapped to the KG scheme (or metagraph) depicted in Fig. 3.2.1a. Inspired by the Harmonizome strategy, we processed each dataset separately following author guidelines, when possible (*Methods*). In brief, we binarized continuous data so that it could be represented in a network format, and we standardised identifiers from multiple sources.

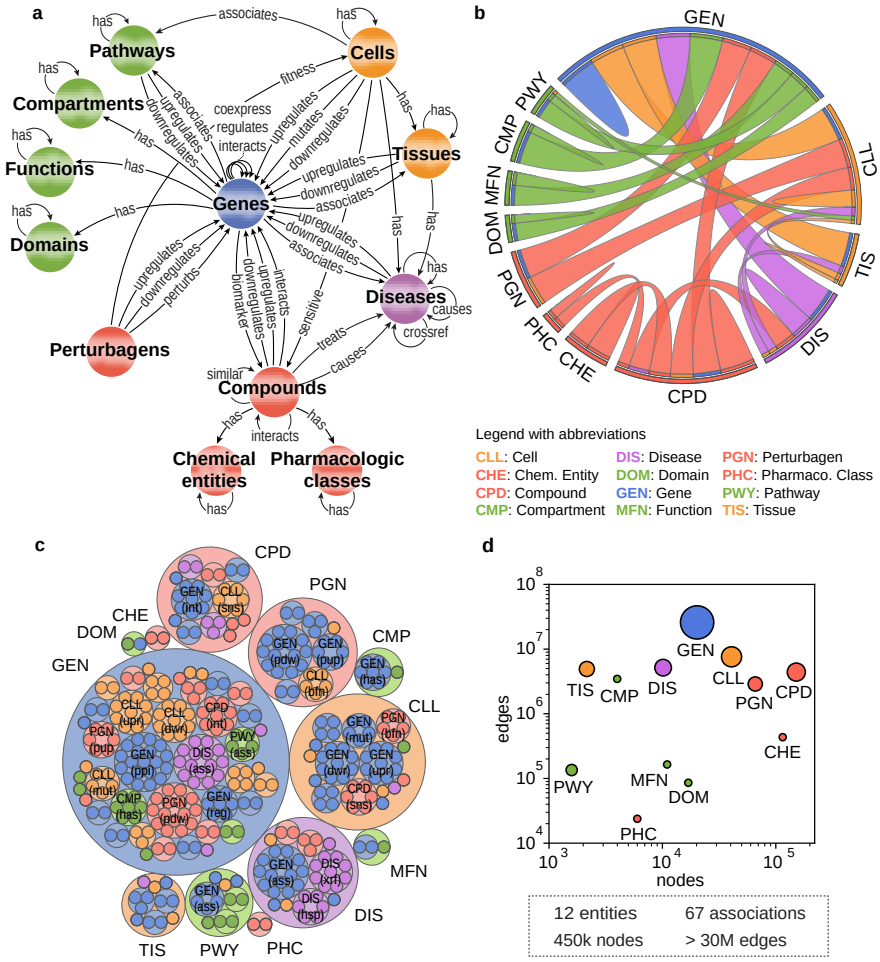


Figure 3.2.1: Building the Bioteque Knowledge Graph (KG). a Metagraph of the Bioteque, showing all the entities and the most representative associations (metaedges) between them. **b** Circos plot representation of the KG, showing the relationships between nodes. **c** Treeplot showing the number of datasets used to construct each metaedge. **d** Total number of nodes (x-axis) and edges (y-axis) available for each entity type. The size of the circles is proportional to the number of metaedges in which the entities participate.

The current version of the KG contains over 450k nodes, belonging to 12 types of biological entities (metanodes), and over 30M edges, representing 67 types of relationships (metaedges) (Fig. 3.2.1b). In general, the size of our KG is comparable to other recently published biomedical KGs[92,234,235,236]. In fact, taking as a reference the comparison made by Bonnet et al[237], our KG is the most comprehensive in the number of processed datasets, the second most comprehensive with respect to enti-

ties, edges, and relation types, and the third regarding entity types (Table A.2.1). Not surprisingly, genes and proteins account for most of the edges (25M) and metaedges (42) in the graph (Fig. 3.2.1c-d). In terms of the number of reference datasets, protein interactions (GEN-ppi-GEN) and gene-disease associations (GEN-ass-DIS) are the most represented metaedges, supported by 17 and 15 datasets, respectively (Fig. 3.2.1c). A comparison of data extracted from each dataset revealed that, although there is some overlap, most sets cover distinct associations, probably due to differences in the focus of the underlying experiments (i.e., physical[60] vs. functional[61] PPIs or drug-driven[238] vs. genomics-driven[20] gene associations) (Fig. A.2.1a).

Calculation of network embeddings across the KG

To integrate the biological knowledge gathered, we devised an approach to obtain, for a given node in the KG, a set of embeddings capturing different contexts defined by one or more types of relationships between this node and other entities (Fig. 3.2.2a). For example, the pharmacological context of a certain compound can be captured by 'compound interacts with protein' associations, while its clinical context may be captured by 'compound treats disease' links. The embedding procedure is as follows. We first define the types of biological entities (metanodes) to be connected and the sequence of relationships (metaedges) between them that we wish to explore. This sequence of relationships is called metapath. We then systematically examined all possible paths from the source and target nodes of the metapath, down-weighting highly connected nodes to ensure exhaustive exploration of the network[228]. This step yields a simplified homogeneous (when source and target metanodes belong to the same type) or bipartite (when source and target metanodes belong to different types) graph that can be explored with conventional network embedding techniques. We chose to use a random walk method, where the trajectories of an agent that explores the network are retained and eventually fed into a text-embedding algorithm[119]. As a result, for each node in the network, a 128-dimensional vector (i.e., an embedding) is obtained, so that similar vectors are given to nodes that are proximal in the network. During this process, we mostly keep different datasets separately (i.e., without merging equivalent networks in different sources) to preserve the original information captured in them[239]. A more detailed description of the protocol is provided in the *Methods* section.

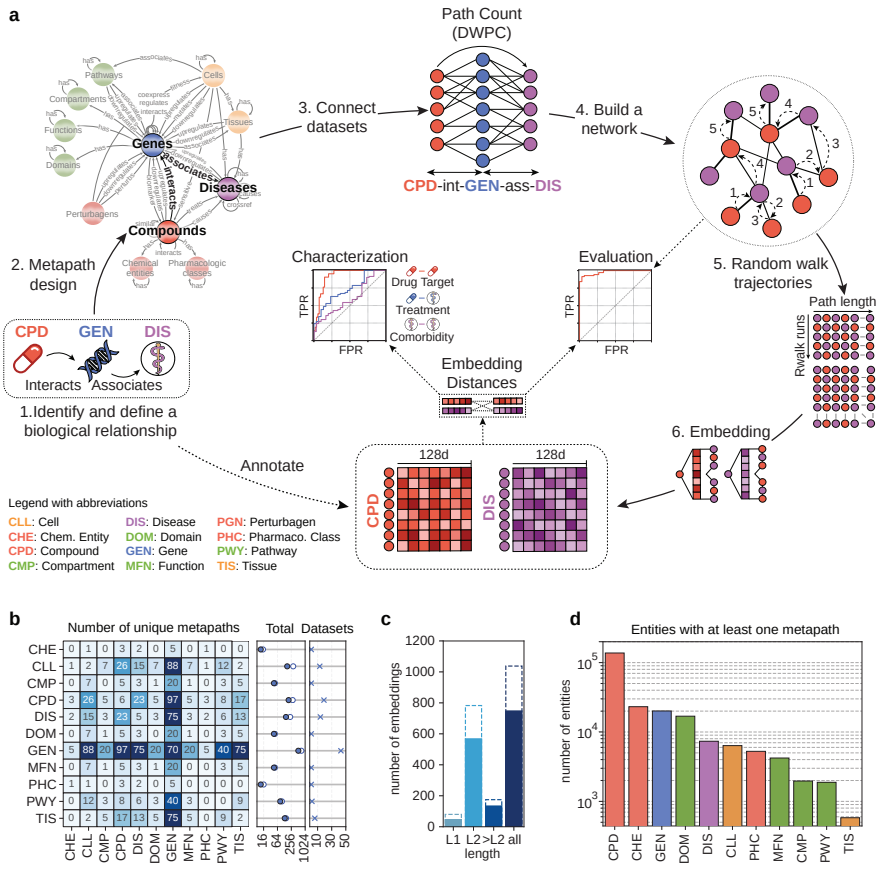


Figure 3.2.2: Generating the Bioteque embeddings. **a** Scheme of the methodology. First, we define the biological entities to be connected and the specific context to be explored. Then a source-target network is derived by traversing all the paths available from the source to the target nodes of a given metapath. The vicinity of each node in the network is then explored by a random walker and embedded in a 128-dimensional space. Finally, embeddings are evaluated and characterized. **b** Number of unique metapath embeddings linking each entity. In the middle plot, the filled dots indicate the total number of unique metapaths while the empty dots show the total number of metapath-dataset combinations. In the rightmost plot, we show the number of entity-specific datasets used in the metapaths. **c** Number of metapath-dataset embedding combinations obtained at each metapath length. Solid bars highlight the number of unique metpaths. **d** Number of nodes within each entity with at least one embedding in the Bioteque resource. Note that during metapath construction, perturbagen (PGN) entities are always mapped to the corresponding perturbed genes. Thus, although used to construct several metapaths, PGN nodes are not explicitly embedded (i.e., they are not the first or last nodes in the metapaths).

We have created a resource of pre-calculated biomedical embeddings, the Bioteque, where we have exhaustively considered most metapaths of length 1 and 2 extracted from the KG (i.e., direct connections between

source and target nodes, or with one intermediate node between them). In addition, we have curated a collection of 135 metapaths of length ≥3. Overall, the Bioteque currently holds a total of 81, 785, and 175 embeddings of length 1, 2, and ≥3, respectively (Fig. 3.2.2c and Supplementary Data 3). Length 1 (L1) metapaths correspond to direct associations in the knowledge graph and provide the simplest domain knowledge representations of the entities. Larger metapaths (>L1), on the other hand, are either dedicated to connecting different entities through a third one (i.e., CPD-int-GEN-ass-DIS) or extend L1 associations to similar entities (i.e., CPD-int-GEN-ppi-GEN or CPD-trt-DIS-ass-GEN-ass-DIS), allowing the identification of more complex relationships between biological entities (i.e., two compounds may target different proteins yet affect the same pathway, or CPD-int-GEN-ass-PWY).

Given that the constructed KG is gene-centric, genes (GEN) are the most frequently embedded biological entity in the resource (515 unique metapaths from 43 different datasets) followed by compounds (CPD), cell lines (CLL), and diseases (DIS) (198, 168 and 150 unique metapaths, respectively) (Fig. 3.2.2b). Furthermore, most of the metapaths used gene entities, such as those derived from omics experiments or literature curated annotations, as bridges to connect distinct entities (Fig. A.2.2). Compounds also play an important role, connecting pharmacological classes and chemical entities to the rest of the graph and being a major source of metapaths embedding cell lines, diseases and tissues.

Overall, the Bioteque provides a collection of 1,041 embeddings obtained from 746 unique metapaths, covering all entities defined in the biological KG (Fig. 3.2.2d).

Embeddings retain the interactions in the original KG and reveal new insights

Having obtained embeddings for all nodes in the KG, we performed a set of analyses to, on the one hand, validate that the embeddings retained the connectivity observed in the KG and, on the other, to characterise each embedding space in the light of other (orthogonal) datasets in the Bioteque. As an illustrative example, Fig. 3.2.3 shows the analysis of the metapath CPD-int-GEN-ass-DIS, corresponding to compounds that interact with genes, which are, in turn, associated with a disease.

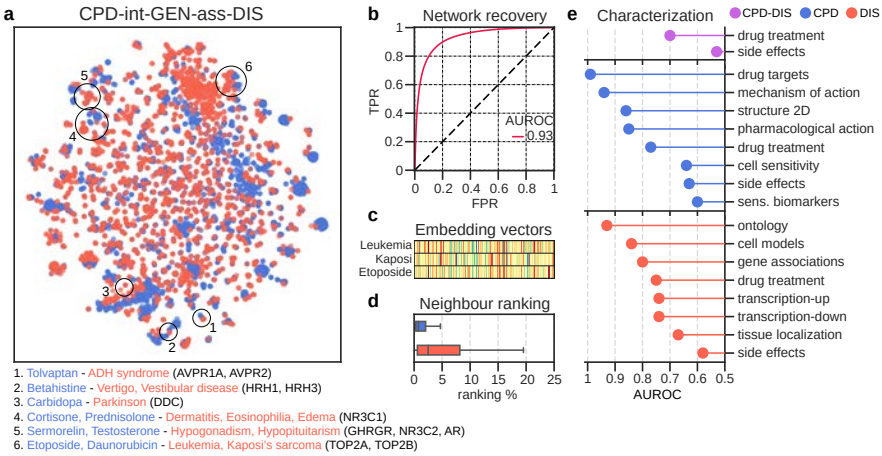


Figure 3.2.3: A Bioteque embedding summary card. **a** 2D projection (opt-SNE) of the compound (CPD, blue) and disease (DIS, red) embeddings from the metapath 'compound interacts protein associates with disease' (CPD-int-GEN-ass-DIS). We highlight clusters of compounds and diseases sharing treatment evidence. We highlight some representative compounds and diseases found in these clusters, together with the drug targets associated with the diseases. **b** ROC curve validation when reconstructing the original network with the corresponding embeddings. **c** Visual representation of the embedding vectors of Leukaemia (top) and Kaposi's sarcoma (middle), together with the drug Etoposide (bottom). **d** Ranking proportion in which the putative CPD (n = 131,648) and DIS (n = 134,997) neighbours are found. Box plots indicate median (middle line), 25th, 75th percentile (box) and max value within the 1.5*75th percentile (whiskers). **e** Recapitulation of orthogonal associations by using embedding distances. The AUROC (x-axis) summarizes the performance obtained when ranking the orthogonal associations. Drug targets are collected from Drugbank[185], the Drug Repurposing Hub[240] and PharmacoDB[241], and gene-disease associations are obtained from Open Targets[238].

To validate the embeddings, we calculated their cosine distances pairwise, and checked that proximal embeddings corresponded to edges in the KG (Fig. 3.2.3b), measured with the Area Under the Receiver Operating Characteristic (AUROC) metric. Similarly, when we used the embedding distances to rank entity pairs, we found their known neighbours in the closest 10% of possible nodes (Fig. 3.2.3d). Note that the goal of this study is not to benchmark the embedding method (which is already a well-accepted implementation in the field[119]), but to provide an assessment of the approach across a comprehensive set of cases.

Analogously, distances between embeddings can be used to measure whether the dimensional space preserves similarities among entities that share biological traits (i.e., cell lines sharing tissue of origin or genes sharing molecular functions). Following this rationale, we can characterise the type of biological signal captured by a given metapath by comparing its

embeddings to a battery of reference biological traits, an approach already used to benchmark drug-drug similarities on the basis of shared chemical features[242]. The use of embeddings allows for straightforward comparison of entities of the same type (for example, similarity of cell lines according to their up-regulated genes can be measured by computing distances of CLL entities in the CLL-upr-GEN embedding). Likewise, it is easy to compare and uncover correlations between different types of associations. For instance, the correlation between copy number amplification and upregulation can be assessed by considering similarities in the CLL-cnu-GEN and CLL-upr-GEN embedding spaces. In the CPD-int-GEN-ass-DIS example, drug targets and gene-disease associations are among the biomedical traits that are better recapitulated by the compound and disease embeddings (Fig. 3.2.3e). Accordingly, we see how compounds and diseases associated with similar treatments are close in the embedding space. We also observe that compound-disease treatment similarity is achieved at the edge level (AUROC = 0.7), suggesting that not only compounds and diseases with similar treatments are close in the embedding space, but also that compound-disease treatment pairs are often found in the same vicinity. Indeed, compound and disease-associated genes have proven useful in drug treatment prediction exercises[93,243].

A projection of the 128-dimensional embeddings onto a 2D space reveals clusters of drugs and treatments which, by the definition of the metapath, have identifiable targets (Fig. 3.2.3a). We find, for instance, drug-disease groups associated with the treatment of leukaemia (e.g., Etoposide and Daunorubicin), hormonal disorders (e.g., Somatostatin and Sermorelin), nervous system disorders (e.g., Carbidopa, Betahistine, and Protriptyline), and inflammatory conditions (e.g., Cortisone and Prednisolone). We observe that most of these drugs target a small subset of proteins or protein families directly related to the diseases, such as the growth hormone-releasing hormone receptor (GHRHR) for hypogonadism treatment, the somatostatin receptor (SSTR) for acromegaly treatment, and the DOPA decarboxylase to prevent dopamine formation in the treatment of Parkinson's disease. Additionally, the analysis reveals that drugs approved to treat either leukaemia or Kaposi's sarcoma cluster, sharing the Topoisomerase II alpha (TOP2A) enzyme as target (Fig. 3.2.3c). Indeed, comorbidity between these two diseases has been reported in several studies[244,245,246], although, to the best of our knowledge, the role of TOP2A in this comorbidity has not been yet described.

The repertoire of embeddings encoded in the Bioteque enables explo-

ration of a given biomedical entity from multiple perspectives, often corresponding to different biological contexts, such as genes with the same biological role yet expressed in different tissues, or cell lines with similar transcriptional profiles but dissimilar at the proteome and drug response levels (Fig. 3.2.4a). When performed systematically, this analysis quantifies the relationship of a certain metapath with the other metapaths in our collection, which in turn helps assess the types of biological traits that it captures. Figure 3.2.4b shows ten of the top metapaths recapitulating gene molecular function and compound pharmacological class. We see that genes targeted by the same compounds or having similar domains tend to share molecular function while, as expected, sets of interacting compounds, or those with similar binding profiles, tend to belong to the same pharmacological class.

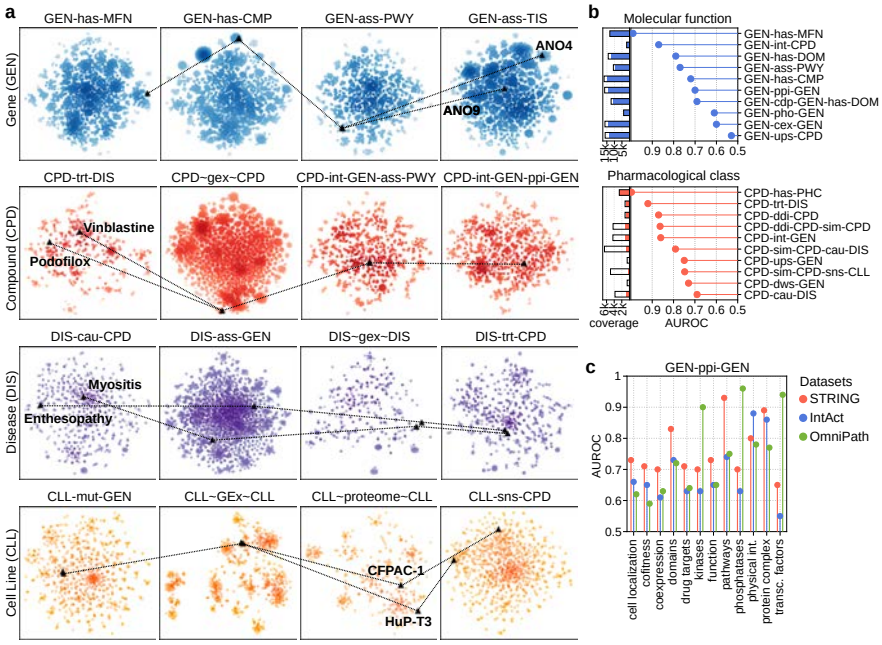


Figure 3.2.4: Comparison of embeddings built from different metapaths and datasets. a Four illustrative examples showing pairs of genes (GEN), compounds (CPD), diseases (DIS) and cell lines (CLL) with similarities or differences depending on the metapaths. The extended nomenclature of each metapath can be found in Supplementary Data 2. **b** Top metapaths (y-axis) recapitulating (AUROC, x-axis) gene molecular function (MFN, blue) and compound pharmacological class (PHC, red). The coloured bars indicate the proportion of nodes in the metapath that could be assessed (i.e., with annotated molecular function or pharmacological classes). **c** Gene embedding characterization of three reference PPI datasets, namely STRING, IntAct and OmniPath. We limited the analysis to the common gene universe (9,395 genes) between the three sources.

Additionally, one can explore differences among datasets within a single metapath. In Fig. 3.2.4c, we embedded three well-known protein-protein interaction (PPI) networks, representing functional interactions (STRING[61]), physical interactions (IntAct[60]), and protein-signalling interactions (OmniPath[247]), and quantified the capacity of these networks to capture a variety of biological features, from cellular localization to protein complexes. The diversity of functional interactions contained in STRING favours recapitulation of most of the features explored, especially those involving similar biological pathways (AUROC = 0.93), protein complexes (AUROC = 0.89) and protein domains (AUROC = 0.83). Not surprisingly, IntAct better preserves physical interactions (AUROC = 0.88) and shows good performance with protein complexes (AUROC = 0.86). Finally, OmniPath shows an enrichment in signalling processes such as kinase-substrate (AUROC = 0.9), phosphatase-substrate (AUROC = 0.96) and transcription factor interactions (AUROC = 0.94), in good agreement with the type of resources used to build this network.

In general, the different considerations followed to populate these networks may favour some domains of knowledge, hence suiting different tasks, which can be efficiently and systematically revealed by transforming them into embeddings. In the next sections, we present three illustrative examples on how these biological embeddings can be used off-the-shelf in a variety of tasks.

Gene expression embeddings as biological descriptors of cell lines

Gene Expression (GEx) experiments have been widely used to characterise cellular identity and state, as they broadly recapitulate tissues of origin[248] and they are notable genomic biomarkers for anticipating drug response[33]. However, these experiments typically measure the expression of 15-20k genes, yielding numerical profiles that are computationally demanding and prone to overfitting problems when used as input in machine learning approaches with limited data[249,250].

We thus explored whether our more succinct 128-dimensional vectors were able to retain the information contained within the full GEx profile. Taking the Genomics of Drug Sensitivity in Cancer (GDSC)[33] panel as a reference, we collected, for each cell line, the basal (raw) GEx (17.7K genes) and the corresponding Bioteque metapath embedding CLL-dwr+upr-GEN-dwr+upr-CLL (hereafter CLL-gex-CLL), aimed at capturing gene

expression similarities between cell lines.

We first examined the similarity landscape of the cell lines by performing a 2D projection of the raw and embedded GEx. By colouring the cell lines according to their tissue of origin, we visually verified the capacity of the CLL-gex-CLL embedding to resemble the raw GEx data (Fig. 3.2.5a). Indeed, cosine similarities between CLL-gex-CLL vectors up-ranked CLLs sharing tissue of origin with a similar rate as when using correlations between raw GEx vectors (AUROC = 0.75 and 0.76, respectively) (Fig. 3.2.5b).

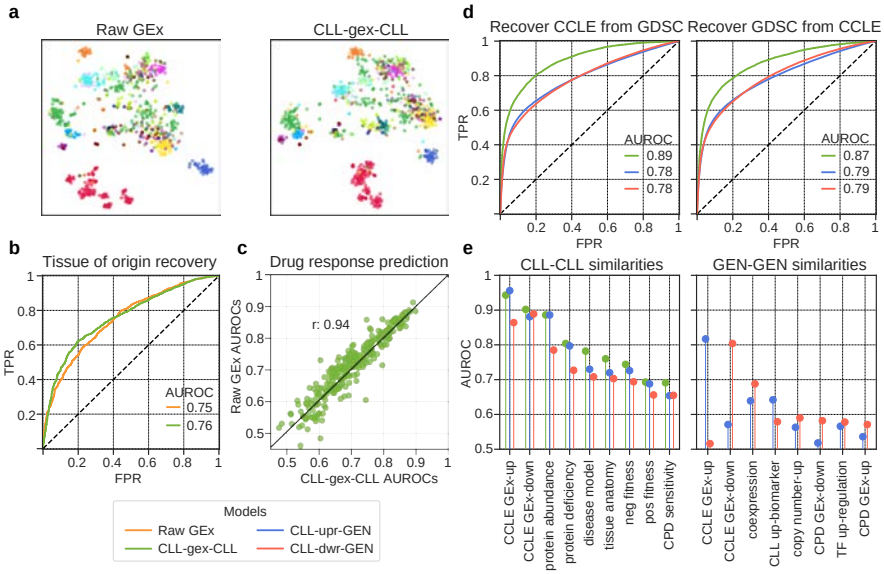


Figure 3.2.5: Analysis of gene expression (GEx) embeddings. **a** 2D projection of the raw GEx (left) and the corresponding Bioteque 'cell has similar gex cell' (CLL-gex-CLL) embedding (right). Each dot corresponds to one cell line and is coloured by tissue of origin. **b** Tissue recovery by the raw GEx and the CLL-gex-CLL embedding. **c** Drug response prediction performance (AUROC) for each drug in the GDSC resource from models trained with either the raw GEx (y-axis) or the CLL-gex-CLL embeddings (x-axis). **d** Recovering CCLE (left) and GDSC (right) cell-cell (CLL-CLL) similarities (green), cell-gene (CLL-GEN) upregulation (upr) similarities (blue) and CLL-GEN downregulation (dwr) similarities (red) using embedding distances from the GDSC and the CCLE embedding spaces, respectively. **e** Characterization of the CLL-CLL (left) and GEN-GEN (right) embedding similarities for three metapaths: CLL-gex-CLL (green), CLL-upr-GEN (blue) and CLL-dwr-GEN (red).

Next, we assessed the capacity of our embeddings to predict the drug response of each cell line. To this end, we trained a standard machine learning model (a random forest classifier) for each of the 262 drugs in the panel and predicted sensitive/resistant responses using the raw GEx and

our embeddings independently (*Methods*). Indeed, we found that the capacity of the CLL-gex-CLL embedding to recapitulate drug response is equivalent to that observed when the raw GEx data is used (average AUROC = 0.70 and 0.71, respectively). Moreover, the models based on embeddings had strong concordance with the raw GEx model (0.94 Pearson's correlation) (Fig. 3.2.5c). This level of agreement is remarkable and represents a clear advantage for the embeddings since they are smaller, easier to handle and do not require expert knowledge to pre-process the raw data. A disadvantage of the embedding approach is the less obvious interpretability of predictions.

After verifying that the Bioteque GEx embeddings retain the basal transcriptional information from the cell lines, we used them to compare profiles obtained from different cell line panels. Specifically, we compared the GDSC with the Cancer Cell Line Encyclopedia (CCLE)[251]. In agreement with previous reports, we observed a strong correspondence between the two panels, measured as CLL-gex-CLL similarities in the embedding space (AUROC = 0.89) (Fig. 3.2.5d). To assess whether these similarities were driven by the up- or down-regulation of the same genes, we repeated the analysis focusing on the CLL-upr-GEN and CLL-dwr-GEN embeddings and checked whether the CLL-GEN similarities in the GDSC panel were also preserved in the CCLE. In general, the recovery score of cell line-specific up-/down-regulated genes (i.e. CLL-GEN pairs) was lower (AUROC = 0.78) (Fig. 3.2.5d). We obtained similar results when we reversed the exercise and used CCLE embeddings to recapitulate GDSC similarities (Fig. A.2.3). This finding suggests that, while cell line similarities between panels are robust (i.e., cell lines sharing similar transcriptional signatures in one panel also share similar ones in the other), the specific transcriptional changes of a given cell line may differ. The characterization of the CLL-CLL and GEN-GEN distances further confirmed the better recapitulation of cell line similarity in comparison to gene similarity between panels (AUROC = 0.9 and 0.8 for the CLL-CLL and GEN-GEN similarities, respectively) (Fig. 3.2.5e). Furthermore, the CLL-CLL similarity characterization revealed a strong concordance between protein and transcript levels (AUROCs = 0.9 and 0.8 for protein abundance and deficiency, respectively), which was partially driven by the same CLL-GEN pairs (AUROCs = 0.72 and 0.63 for the protein abundance and protein deficiency CLL-GEN pairs, respectively) (Fig. A.2.3c). In addition to tissue of origin, we also observed resemblances between cell lines used to model a given disease (AUROC = 0.78), sharing fitness profiles (AUROC

= 0.72 for negative and 0.69 for positive fitness profiles) and similar drug responses (AUROC = 0.7). Finally, the GEN-GEN similarities also revealed a mild recapitulation of known co-expressed gene pairs (AUROC = 0.64 and 0.69, for the up- and down-regulated gene similarities, respectively), thereby suggesting that some of the genes commonly up- or down-regulated in the same cell lines from different panels may share the same transcriptional regulatory programs.

On the whole, our approach retains meaningful information from the original data into a reduced number of dimensions (128 vs ∼20k), even when the data comes from a much noisier source such as transcriptomic technologies. We believe that the standardised and dense format of our embeddings provides a by-default way to integrate and compare omics datasets.

Assessing the uniqueness of new omics datasets

Since the consolidation of high-throughput omics technologies, several long-term initiatives have been established to comprehensively characterise certain levels of biological systems (i.e., genetic interactions in yeast[252] or the transcriptomes of cell line panels and human tissues[251,253]). After several years running, all these efforts have had to balance a potential decrease in novelty and an increase in costs as the screens approach saturation. The Bioteque provides a corpus of biological data that is cast to a single format and, as such, it offers a means to quantify the degree of novelty of new data releases of omics experiments. As an illustrative example, we analyse the systematic charting of the Human Reference Interactome (HuRI) with the yeast two-hybrid methodology, which has already identified over 50,000 protein-protein interactions (PPIs) of high quality over the last 15 years[56,57,58].

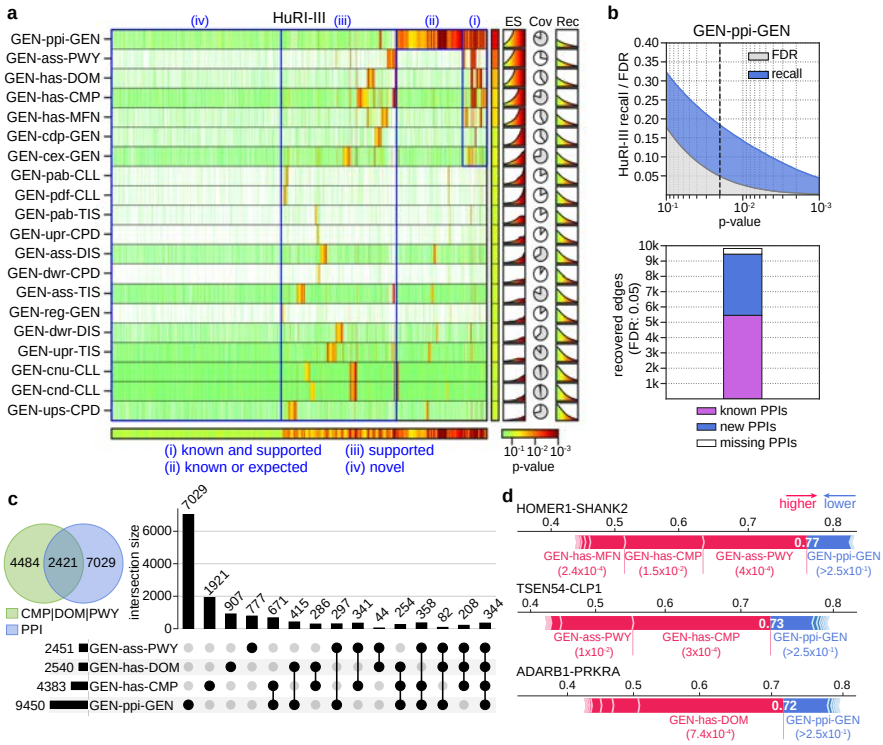


Figure 3.2.6: Assessing the novelty of the HuRI-III interactome. **a** Embedding distance *p* values are calculated for each PPI in HuRI-III (x-axis) using the corresponding gene-gene (GEN-GEN) embeddings from a subset of metapaths (y-axis). Please, note that these *p* values do not reflect the significance of any statistical test, but indicate the normalized quantile rank position of a given observation in a background distance distribution (*Methods*). Red tones (lower *p* values) indicate similarity according to a given embedding space. The column and row next to the heatmap show the 10th percentile of the *p* value distribution for each metapath and the lowest *p* value for each edge, respectively. In blue, we grouped edges according to four levels of support. On the right, it is shown the enrichment scores (ES) (capped between 1 to 5 on the y-axis) across *p* values, the coverage (Cov), and the cumulative recall (Rec) across *p* values. **b** (Top) Recovery of HuRI-III edges (recall) and randomly permuted edges (FDR) by 'protein interacts protein' (GEN-ppi-GEN) embeddings across the *p* values (x-axis). The dashed line is placed at the 0.05 FDR (corresponding to a *p* value of 0.02). (Bottom) Number of HuRI-III interactions recovered by the GEN-ppi-GEN embedding at 0.05 FDR stratified by those covered in the original network (known PPIs), those not available in the network, hence, predicted by the embeddings (new PPIs), and those present in the original network but not covered at the given *p* value (missing PPIs). **c** Number of unique HuRI-III edges recovered at 0.05 FDR by the GEN-ppi-GEN and/or the three most supportive metapaths, including 'gene has cellular components' (GEN-has-CMP), 'protein has domain' (GEN-has-DOM), and 'gene associates with pathway' (GEN-ass-PWY). **d** Shapley force plots corresponding to the prediction of three PPIs with no direct evidence of physical interaction before HuRI-III was released. Red segments are metapath-specific *p* values that pushed predictions toward a high probability of interactions, while blue segments pulled predictions towards a low probability. The length of the segments is proportional to their impact on the prediction. The final output probability given by the model is found where both forces equalize (shown in white).

To estimate the level of support from different experiments and assess the novelty of the latest HuRI release (HuRI-III[58], we used the embedding space of relevant metapaths to determine the biological context of each pair of interacting proteins. In brief, for each gene-gene pair, we calculated an empirical p value corresponding to the measured similarity in the embedding space, which allowed for commensurate comparison of distance/similarity measures performed in different embedding spaces (see *Methods*). Note that, to have a fair representation of the known physical interactions, we embedded an older version of the protein-interaction network, without including any of the entries from HuRI-III. We then categorised each interaction in HuRI-III into four groups, depending on the level of support contained in the Bioteque embeddings. In this regard, we labelled them as (i) known and supported interactions (covered by GEN-ppi-GEN and at least another metapath), (ii) known interactions (only covered by GEN-ppi-GEN), (iii) supported interactions (covered by other metapaths but not GEN-ppi-GEN) and (iv) potentially novel interactions (with no apparent support in any of the metapaths screened) (Figure 3.2.6a). Remarkably, after three updated versions of HuRI, almost half of the interactions can be classified as potentially novel according to the selected metapaths. Moreover, although only 5,825 (11%) of the interactions were supported by GEN-ppi-GEN embeddings, mostly coming from previous versions of HuRI[56,57], our analysis suggests that a higher proportion can be recovered. In fact, at 0.05 FDR (*Methods*), the GEN-ppi-GEN embedding recovered 18% of HuRI-III, retrieving 5,456 (94%) of previously known interactions while finding 3,994 new pairs (Figure 3.2.6b). On the other hand, we observed a substantial number of physical interactions presumably involved in similar pathways (GEN-ass-PWY), cellular components (GEN-has-CMP), or protein domains (GEN-has-DOM). At 0.05 FDR, these metapaths alone recovered 6,905 unique interactions of which 4,484 (65%) were not obvious from the physical interaction space (Figure 3.2.6c).

To delve into the correlation and relative importance of the metapath for explaining PPIs, we used the p values as features for a tree-based machine learning model trained to identify HuRI-III edges. We then assessed the importance of each metapath for the prediction using Shapley values[254]. As visually anticipated from the heatmap, the model achieved a reasonable performance (AUROC = 0.69), mostly relying on previously known physical interactions, cellular components, protein domains, and pathways, all of them showing a certain degree of agreement (Fig. A.2.4).

Interestingly, we also identified successfully predicted cases with little to no evidence from physical PPIs. For instance, our metapath distance-based model predicted the interaction between the neuronal proteins HOMER1 and SHANK2, the tRNA-splicing endonuclease TSEN54 and the polyribonucleotide CLP1, and the adenosine deaminase ADARB1 and the protein kinase PRKRA, none of which had any reported evidence in protein interaction databases but showed strong positive support in the GEN-ass-PWY, GEN-has-CMP, and GEN-has-DOM metapaths, respectively (Fig. 3.2.6d). Indeed, some of these associations have been related in other contexts[255,256,257], but with no indication of physical interactions before HuRI-III.

We have shown how the continuous and interpretable dimensional space of the Bioteque embeddings provides a powerful framework for characterising individual observations, which can, in turn, be exploited to guide the interpretation of the entire dataset and, to some extent, assess the novelty of the data.

Discovery of drug repurposing opportunities

Drug repurposing is often regarded as an attractive opportunity to quickly develop new therapies[258]. However, perhaps with the exception of cancer, where abundant models and molecular data are available, it is difficult to generate data-driven predictors to suggest new uses for approved or investigational drugs, mainly due to the lack of disease descriptors and the small number of known drug-disease indications. Indeed, according to the last update of repoDB, half of the drugs (1,097) have only one approved indication, and a third of the diseases (458) are treated with only one drug (Fig. A.2.5). Thus, training models with all the known drug-disease associations and later transfer of the insights gained to underexplored treatment areas would be highly desirable[259,260].

To explore whether the Bioteque could be useful in this scenario, we set out to predict new compound-disease indication pairs introduced in repoDB in 2020 (v2) training a model on the previous version (v1), launched in 2017 (*Methods*). We mapped all disease terms to the Disease Ontology, removed redundant indications (according to the ontology), and trained a conventional random forest classifier to predict whether a given CPD-DIS corresponds to a true therapeutic indication. We used two sets of metapath embeddings: one in which we used L1 metapaths (*Short*) based on the drug targets (CPD-int-GEN) and gene associations (DIS-ass-GEN), and

another in which we used L3 metapath (*Long*) linking the pharmacological class and the treatment of known CPD and DIS to those sharing drug target (CPD-int-GEN-int-CPD-has-PHC) or gene associations (DIS-ass-GEN-ass-DIS-trt-CPD), respectively. We chose to use drug targets and gene associations because we observed that their embeddings broadly recapitulate the pharmacological class and the disease treatment for a sufficient number of nodes (Fig. A.2.5). Moreover, to assess the capacity of the gene-based similarities to correctly infer the treatment, we also tested a metapath (*Long-b*) in which we prevented the CPDs and DISs from being linked, thus making the association with PHC or treatment purely based on the gene-driven similarity to other CPD or DIS. To avoid trivial predictions, we removed associations with PHCs or treatments for drugs and disease unique to the repoDB v2 in all *Long* metapaths. As a basal model, we used chemical fingerprints (ECFP4, 2048 bits) for the CPDs and either one-hot identity vectors (*Basal1*) or binary gene annotations (*Basal2*) for the DISs.

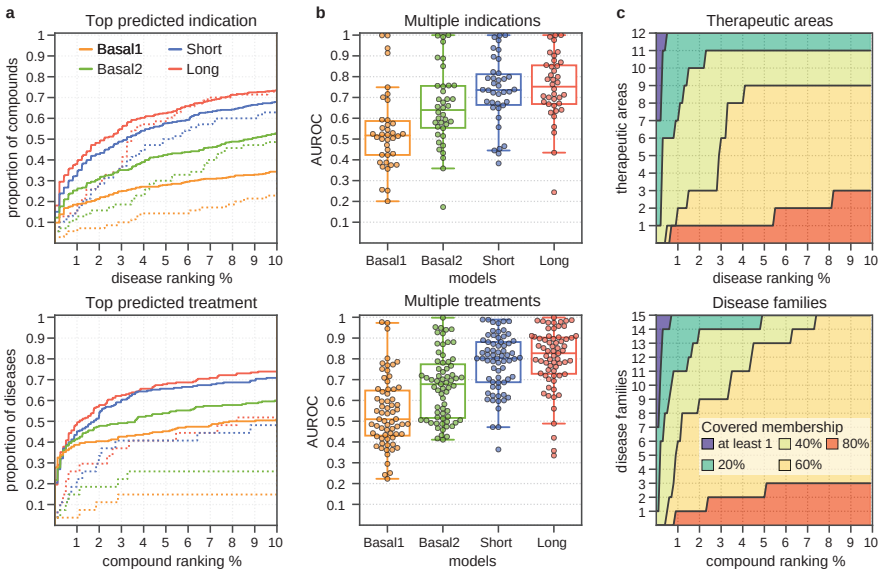


Figure 3.2.7: Prediction of drug indications and disease treatments from repoDB. a Cumulative distribution (y-axis) of compounds (top) and diseases (bottom) according to the ranked position (x-axis) of the top predicted disease indication (top) or compound treatment (bottom) for the four tested models. The rankings are shown in percentages and only for the first 10% of compound/disease predictions (corresponding to the top 50 and 80 diseases and compounds, respectively). Dotted lines show the distribution for those compounds or diseases with only one positive indication in repoDB v1. **b** Classification performance obtained for each compound (n = 38, top plot) and disease (n = 67, bottom plot) with multiple (≥5) new indications reported in repoDB v2. Box plots indicate median (middle line), 25th, 75th percentile (box), and max and min value within the 1.5*25th and 1.5*75th percentile range (whiskers). **c** Number of different therapeutic areas (top) and disease families (bottom) covered by the predictions of the *Long* model. We considered a given therapeutic area or disease family to be covered when the model predicted one true indication or treatment (as in panel *(a)*) for at least 1%, 20%, 40%, 60%, or 80% of its instances.

We considered two use cases: a drug repurposing exercise, in which we ranked all the diseases predicted to be potentially treated with a given compound, and a prescription exercise, in which we ranked all compounds that might be useful to treat a given disease. In both scenarios, the three metapath embeddings showed remarkable predictive power compared to the basal models, with the model built from *Long* embeddings being the one with superior performance (Fig. 3.2.7a). Specifically, for half the tested compounds, the *Long* embeddings model found a new validated therapeutic purpose within the top 2% of disease predictions (corresponding to the top 10 ranked diseases). Analogously, for roughly 50% of the diseases, the model found a correct treatment within the top 1% of com-

pound predictions (corresponding to the top 8 ranked compounds). Furthermore, although with poorer performance, our biological embeddings were able to yield correct predictions for compounds and diseases with minimal evidence available (i.e., with only one known indication or treatment in repoDB v1) (Fig. 3.2.7a, dotted lines). In contrast, the best performing basal model (*Basal2*) found correct predictions for 32% of the compounds and 41% of the diseases within the same ranking range. Moreover, the Bioteque-based models were better at consistently up-ranking indications (or treatments) of compounds (or diseases) with multiple new annotations in repoDB v2 (Fig. 3.2.7b). In fact, among our top predictions, we found repurposing cases that reached clinical trials (Fig. A.2.6a). For instance, while both Verapamil and Ranolazine drugs have been approved for the treatment of angina pectoris, our model correctly predicted the repurposing effect of Verapamil in the treatment of ischemic stroke (clinical trial: NCT02823106) and Ranolazine in the treatment of atrial fibrillation (clinical trial: NCT03162120) in the top 1 and 2 positions, respectively (Fig. A.2.6b). Interestingly, our model highlights hyperinsulinemia as the top repurposing for Ranolazine. While this link is not included in repoDB, we have found diverse studies supporting the correlation of Ranolazine with insulin levels[261,262,263].

Finally, we verified that these predictions covered a broad range of therapeutic areas and disease families. Indeed, we found that within the top 1% of predictions, the *Long* model successfully predicted one indication or treatment for 20% of all the compounds and diseases in each therapeutic area or disease family (Fig. 3.2.7c and A.2.6e). These results were reproduced with the *Long-b* model, showing that, as expected, the genes associated with drugs or diseases of known treatment can indeed be used to better infer the activity of drugs and diseases with unknown indication (Fig. A.2.6c-d).

Overall, we showed how Bioteque embeddings can be directly plugged into machine learning models, and how, by combining different context associations into larger metapaths, they can increase the performance of drug-disease prediction models.

The bioteque resource

We built an online resource to facilitate access to all the pre-calculated Bioteque embeddings (https://bioteque.irbbarcelona.org). The Bioteque web offers a visual way to explore over one thousand metapaths by selecting the nodes to connect, as well as the type of relationship between them. For a selected metapath, we provide an analytical card displaying a 2D representation of the embedding, a ROC curve assessing the preservation of the original network, distance distributions of the embedding space, and biological associations that are best recapitulated by the metapath of interest.

Furthermore, the web page also offers a section where metapath embeddings and other metadata can be downloaded. The generated file contains the embeddings for each node, the nearest neighbours of each node in the space, and the analytical card displayed on the web. Additionally, we make available executable notebooks showing how to download our embedding resource programmatically as well as how to perform most of the downstream analyses presented throughout this manuscript. More specifically, we illustrate how to (i) generate 2D (interactive) visualisations that can be coloured and annotated according to side information (e.g., colour cell lines by tissue of origin), (ii) identify similar nodes (close neighbours) for a given entity of interest, (iii) cluster the embedding space, and (iv) build a predictor model trained on our embeddings.

The Bioteque web also provides information on the specific sources used to construct each metapath, and some general statistics on the contents of the current version of this web resource. We also provide a link to our GitLab repository, which contains the full code necessary to preprocess the data to generate and analyse biological embeddings (http://gitlabsbnb.irbbarcelona.org/bioteque). The entire resource, including the underlying data and biological embeddings, will be updated once per year, or as soon as a major dataset is released.

3.2.4 Concluding remarks

With the accumulation of large-scale molecular and cell biology datasets, coming from ever-growing literature, omics experiments and high-throughput screenings, new frameworks for integrative data analysis are necessary. For a given biological entity (e.g., a gene), we are now able to stack multiple layers of its biological complexity (e.g., its structure, function, regulation,

or interactions), which offers an opportunity for a more complete, systemic view of biological phenomena, but brings along several challenges, including the handling of different data structures, nomenclatures, signal strengths, and variable dimensionalities.

To tackle these challenges, we have developed the Bioteque, a resource of pre-calculated, fixed-format vector embeddings built from a comprehensive biomedical knowledge graph (KG). The KG contains physical entities like genes, cell lines, and compounds, as well as concepts like pathways, molecular functions, and pharmacological classes. Embeddings capture the connections between nodes in the KG according to a certain metapath (i.e., a sequence of semantic and/or mechanistic relationships between entities). We have shown how this approach is useful to (i) produce compact descriptors that broadly preserve the original data, (ii) systematically characterise biological datasets such as cancer cell line transcriptional signatures, (iii) assess the novelty of a given omics experiment, and (iv) mine for drug repurposing opportunities based on multiple associations between drugs and diseases.

In the Bioteque, we have incorporated datasets from over 150 distinct sources, keeping the integrity of the original data to a feasible extent and applying standard transformations when required. Note that the accuracy of the Bioteque is determined by the quality of the source data. As experimental technologies continue to evolve, new information will populate these databases and novel standards will emerge, opening the door for more comprehensive and higher quality embeddings. In addition, as a first attempt, we used a network embedding technique that purely relies on the graph topology built from the biomedical data, in contrast to other techniques that also leverage node and edge attributes (e.g., Graph Neural Networks, GNN). While these methods may contribute to improving the embedding space, their quality depends on the availability of enough data and meaningful node features, while requiring a thorough fine-tuning of the hyperparameters [129,128]. Taken together, the proper implementation of these methods becomes unfeasible for the systematic embedding of thousands of networks. Additionally, the incorporation of external node features in the network could compromise the controlled identity of the metapaths. Nevertheless, Bioteque descriptors can be easily recycled as node features for new task-specific networks, thus transferring the learning encoded from orthogonal biomedical datasets to more complex, attribute-aware models.

Finally, we would like to point out that there are parts of the current

biomedical knowledge that have not yet been included in the resource, such as antibody-target interactions and metabolomics. As a molecular/cell-centric resource, the Bioteque also lacks patient-derived data[92], including interactions with the microbiome[264]. Updated versions of the Bioteque will have to be complemented with the incorporation of other fields of biological knowledge, the re-accommodation of the datasets in the resource (based on updated standards), and the improvement of embedding strategies to account for side-features of the nodes or incorporate unseen (external) nodes in the embedding space. Moreover, future developments will explore the adoption of biological descriptors as features for a variety of downstream-specific tasks, including a systematic screening of the biological support of wet lab experiments or the modelling of complex diseases to guide the generation of new chemical entities to tackle them[232].

3.2.5 Methods

Building the metagraph: Nodes (entites)

The nodes in the graph can belong to one of 12 types (aka metanodes). For each entity type, we predefined a universe of nodes and chose a reference vocabulary based on standard terminologies. These 12 entity types are (in alphabetical order):

Chemical entities (CHE): Chemistry terminologies extracted from the Chemical Entities of Biological Interest (ChEBI) ontology[265].

Cells (CLL): Cell lines used in biomedical research and extracted from the Cellosaurus resource[266].

Cellular Components (CMP): Biomolecular structures and complexes as defined by the Gene Ontology[267] (extracted from the basic filtered ontology).

Compounds (CPD): Small molecules codified with the standard InChIKey. As we do not use any predefined library of compounds, the universe will be determined by the union of compounds included in other datasets (e.g., drug-target interactions).

Diseases (DIS): Abnormal conditions, drug side effects and symptoms. We used the Disease Ontology[268] as a reference vocabulary.

Domains (DOM): Functional and structural protein domains extracted from InterPro[269].

Genes and Proteins (GEN): Genes and proteins were unified and stored by Uniprot[270] accession code (UniProtAC). We worked on the reviewed Human proteome.

Molecular Functions (MFN): Biological function of the proteins defined by the basic Gene Ontology[267].

Perturbagens (PGN): CRISPR, overexpression, and shRNA perturbations. Note that PGNs are always mapped to the corresponding perturbed gene when constructing the metapath. Therefore, instead of providing PGN labels, we provide the UniProtAC of the perturbed genes.

Pharmacologic Classes (PHC): Pharmacologic classes defined by the Anatomical Therapeutic Chemical (ATC) code (`http://www.whocc.no`).

Pathways (PWY): Biological pathways and processes. We used Reactome[271] as a reference vocabulary.

Tissues (TIS): Anatomical tissues and cell types defined by the BRENDA Tissue Ontology[272]. Please note that in the datasets containing ontological terms (CMP, DIS, MFN and PWY), we removed the least informative terms (i.e., those that are higher up in the ontology). These terms were identified by calculating the information content[273]. The node universe for each entity and the list of removed terms are available in Supplementary Data 1.

Building the metagraph: Vocabulary mapping

To integrate terminologies, we extracted curated cross-references from the official terminology sources and associated ontologies. As the nomenclatures used to identify diseases and pathways were particularly diverse and rarely cross-referenced, we further increased the mapping of these terms by inferring similarities within concepts as detailed below.

Diseases were mapped by calculating disease term similarities through shared cross-references to the Unified Medical Language System (UMLS), obtained from the DisGeNET mapping resources (https://www.disgenet.org/downloads). Specifically, we encoded each disease term into a binary vector spanning the universe of UMLS terms of all nomenclatures. We then transformed the binary vectors with the corresponding term frequency-inverse document frequency (TF-IDF) values and computed pairwise cosine distances between the Disease Ontology and the rest of the vocabularies. Using the similarities obtained from curated cross-references as reference, we found a cosine similarity cutoff of 0.5 to correspond to an empirical p value of $5x10^{-4}$.

Pathway cross-references were extracted from the ComPath resource[274] and extended following the PathCards[275] approach. This approach first clusters the pathways into SuperPaths based on overlapping genes and then uses Jaccard similarities between the SuperPaths genes to define pathway similarity. We used the same parameters described in the PathCards paper (0.9 for the overlap cutoff, 20 minimum genes in the pathways, and a Jaccard similarity of at least 0.7).

Building the metagraph: Edges (associations)

Edges in the graph are used to link biological and/or chemical entities. Since two entities may be connected by multiple edge types (i.e., 'compound treats disease' or 'compound causes disease'), we define the associations as triplets (metapaths) of entity-relationship-entity (CPD-trt-DIS, CPD-cau-DIS).

Homogeneous associations are those concerning entities (metanodes) of the same type (e.g., 'gene is coexpressed with gene', GEN-cex-GEN), while heterogeneous associations are related to entities of different types (e.g., 'tissue has cell', TIS-has-CLL). Note that we annotated only one direction of the heterogeneous associations (in fact, we kept CLL-has-TIS instead of TIS-has-CLL), although both directions are valid when defining metapaths. On the other hand, edges were treated as directional whenever a homogeneous association had only one valid directionality, like in the case of kinase-substrate interactions ('gene phosphorylates gene', GEN-pho-GEN) or transcription factor regulations ('gene regulates gene', GEN-reg-GEN). Finally, edges corresponding to similarity measures required a predefined set of nodes for pairwise comparison, and they were computed only after the rest of the graph was populated.

Populating the knowledge graph with data

For each type of association or metaedge, we can have one or more datasets (Supplementary Data 2). Datasets are not merged but kept as individual sources so that they can be embedded individually or in combination within a given metapath. The dataset processing pipeline consisted of two steps. In the first step, nomenclatures were standardised and cutoffs were applied. In the second, applied only to ontological data, terminologies were mapped and the network was pruned.

Dataset standardisation

We processed each dataset individually in order to handle the diversity of formats and data types. The guiding principles of data processing were those defined by the Harmonizome[90].

Datasets that already provided binary data were integrated naturally by converting them into the network format of the KG. If the database provided a measure of confidence (e.g., edge weights or p values), we applied default cutoffs (if given) and/or followed author recommendations in order to remove spurious interactions. To build the network, we did not use any edge weight coming from the original source during the embedding process. This was motivated by the observation that most of these weights are based on a measure of support or confidence, which does not necessarily reflect biological significance/strength. Instead, these scores usually capture biases on the knowledge annotation (e.g., associations for under-studied diseases will be less covered among the different sources and, therefore, are prone to have lower confidence scores) or detectability limitations of the experimental screening (e.g., the abundance level of some proteins are more difficult to detect than others). While weighted edges could provide valuable information for the embedding, we could not find a general way to treat them across the diverse and heterogeneous associations in our resource.

Occasionally, the same dataset can be further divided into different subsets on the basis of a given categorical variable (e.g., curated/inferred). We kept these subsets as independent datasets when applicable. For instance, there is a curated version of DisGeNET and an inferred version of it.

Continuous data requires the application of a cutoff before its integration in the KG. Below, we detail how these cutoffs were chosen depending on the nature of the data.

Transcriptomics and proteomics data: We adapted the strategy followed by Harmonizome, which is based on traditional statistical treatment of gene expression profiles. More specifically, we first mapped the samples and genes to our reference vocabulary and collapsed the duplicates by their mean value. A log2 transformation was then applied followed by a quantile normalisation of the genes (unless the dataset was already transformed by the data providers). Next, we subtracted the median and scaled the data according to the quantile range of each gene. Finally, the top 250 most positive and negative genes were selected for each sample and kept in the corresponding metaedges (e.g., CLL-upr-GEN and CLL-dwr-GEN).

Drug sensitivity: To binarize drug sensitivity data, we used the waterfall method first described by Barretina et al.[29], and used since then in different subsequent works (e.g., [147,148,276]). This method ranks cell lines on the basis of a drug response measure, for instance, the area under the growth inhibition curve (AUC), and uses the shape of the plot to define a sensitivity threshold. The waterfall method was applied for each compound in the dataset, keeping at least 1% but no more than 20% of sensitive cell lines and requiring an AUC sensitivity value lower than 0.9.

Perturbation experiments: Gene perturbation data required a preliminary step to differentiate the type of perturbation (e.g., 'CRISPR modification silences gene A') from its outcome (e.g., 'silencing gene A results in overexpression of gene B'). First, for each perturbation in the dataset, we created a perturbagen (PGN) node with a unique identifier. We then simplified the two-step relationship (e.g., 'perturbagen that silences gene A upregulates gene B') into a 'perturbagen upregulates gene B' association (PGN-upr-GEN).

Other datasets: For some datasets containing continuous data, we had to apply customised approaches to convert them into a network format. Details about the pre-processing of each particular dataset are provided in Supplementary Data 2, while the corresponding Python scripts can be found on `https://bioteque.irbbarcelona.org/sources`.

Terminologies and pruning

Six terminologies (namely, CMP, DOM, MFN, PHC, and PWY) had semantic relationships between them. In these cases, we propagated all the reported relationships with other terms (e.g., GEN) through the parents of their corresponding ontologies. To maximise coverage, propagation was done before cross-referencing.

Selection of metapaths

We chose a controlled set of metapaths for which we precomputed embeddings. These are the embeddings that are deposited in the Bioteque resource. The metapaths were selected as follows.

Length 1 (L1): All possible metapaths of length 1 are embedded except for those capturing cross-references (DIS-xrf-DIS), ontologies (PWY-hsp-PWY), compound-compound similarities (CPD-sim-CPD), and PGN associations. Note that PGN nodes are mapped to the corresponding perturbed genes through the PGN-pdw-GEN or PGN-pup-GEN metapaths (thus, >L1 metapaths).

Length 2 (L2): Only the mimicking (e.g., CLL-dwr+upr-GEN-dwr+upr-CLL) or reversion (CLL-upr+dwr-GEN-dwr+upr-CLL) of both directions (up/down) are used for metapaths connecting entities through transcriptomic, proteomic or transcription factor signatures. CLL and TIS are always connected through the CLL-has-TIS association. Finally, only the following associations are allowed when linking cells and genes within a metapath: CLL-upr-GEN, CLL-dwr-GEN, CLL-mut-GEN.

Length 3 (L3): L3 metapaths are constructed by linking L1 metapaths with any of the following L2 metapaths: CLL-dwr+upr-GEN-dwr+upr-CLL; CLL-has-TIS-has-CLL; CMP-has-GEN-has-CMP; CPD-has-PHC-has-CPD; CPD-int-GEN-int-CPD; DIS-ass-GEN-ass-DIS; DOM-has-GEN-has-DOM; MFN-has-GEN-has-MFN; TIS-dwr+upr-GEN-dwr+upr-TIS; or PWY-ass-GEN-ass-PWY. GENs from the PGN-pup-GEN or PGN-pdw-GEN are linked through heterogeneous or directed homogeneous associations but not through undirected homogeneous associations.

Length >3 (>L3): Generated when mapping the source or target PGN to the perturbed genes in L3 metapaths. In the case of directed homoge-

neous associations, we used the '_' mark next to the entity that acted as the source of the association. For instance, GEN-_pho-GEN-ass-PWY links the kinases to the pathways associated with their substrates while GEN-pho_-GEN-ass-PWY links the substrates with the pathways associated with their kinases.

Finally, metapaths whose embedding did not preserve the original network or that failed to keep most of the nodes in a single connected component were removed as described in the following section. The entire list of the embedded metapaths is provided in Supplementary Data 3.

Obtaining Bioteque embeddings

To obtain the embeddings we used the node2vec algorithm [119], a well-accepted approach based on random walk trajectories [277], in which metapaths are used as single networks and fed to the node2vec algorithm. We acknowledge that there are embedding methods that allow a direct embedding of the network from metapath walks (e.g., metapath2vec [121]). However, we decided to first pre-compute the source-target networks using the DWPC method, since the resulting network already weighs those source-target associations that are more strongly connected according to the metapath, thus requiring fewer random walker steps to learn the relationship between the source and target nodes. Moreover, this pre-computed network encourages the embedding model to only focus on source-target relations, giving us more control about what information we are encoding in the embedding space while allowing an easier generalisation of the model's hyperparameters across different metapaths lengths (i.e., the source and target nodes are always one-hop apart regardless of the metapath length). Notice that, since all our metapath networks are either homogeneous or bipartite, the default skip-gram implementation of metapath2vec is equivalent to node2vec.

Embedding homogeneous and bipartite networks

L1 metapaths already correspond to homogenous or bipartite networks. For >L1 metapaths, the source and target nodes were connected by computing degree-weighted path counts (DWPC) [228] through the corresponding datasets and associations in the metapath. To this end, we sorted the datasets according to the associations of the metapath, represented them as adjacency matrices and kept the same source (rows) and target (columns)

node universe as the target and source nodes of the previous and following datasets, respectively. Following the DWPC method, we first down-weighted the degree of the nodes in each of the datasets by raising the degrees to the *-0.5* power. We then calculated the DWPC values by concatenating the matrix multiplication from the source to the target dataset. As a result, we obtained a new *n* x *m* matrix where *n* are the source nodes of the first dataset and *m* are the target nodes of the last dataset. The values of the matrix are the DWPC between the source and target nodes, which are used as weights during the random walker exploration. Finally, we limited the number of edges for each node to 5% of the total possible neighbours (with a minimum of 3 and maximum of 250 edges per node).

Occasionally, we used more than one dataset within the same association or we combined two metapaths into one. This is a common case for >L1 metapaths with transcriptomic signatures where the two directions (CLL-upr-GEN and CLL-dwr-GEN) are often combined (CLL-dwr+upr-GEN-dwr+upr-CPD). To handle these cases, we first obtained an individual network for each metapath or dataset following the approach detailed above. We then merged all the networks by taking the union of the edges (L1 metapaths) or adding the DWPC values (>L1 metapaths).

At the end of the process, we removed network components that cover less than 5% of the entities from the network. And we also removed from the source metapaths that fail to retain 50% of the total nodes within their network components.

Node2vec parameters

The node2vec algorithm consists of a random walk-driven exploration of the network followed by a feature vector learning through a skip-gram neural network architecture. We implemented a custom random walker (with the node2vec parameters *p* and *q* set to 1) and ran 100 walks of length 100 for each node of the network. For >L1 metapaths, we scaled the DWPC values for each node to sum 1 and used them as probabilities to bias the random walker. We used the C++ skip-gram implementation provided by Dong et al.[121] with default parameters to obtain a 128-dimensional vector for each node.

Accounting for node degree biases

The uneven distribution of information across the different knowledge domains and data sources incorporated in our KG inevitably leads to an uneven number of associations across entities, introducing a bias towards nodes with higher degrees. We implemented several measures to mitigate these biases, not only during the generation of the embeddings, but also in the way distances are calculated.

To control the degree of the metapath networks, we implemented the DWPC method (as described in the previous section), which was specifically developed to account for degree biases. Furthermore, we also limited the number of connections a given node can have at the end of the metapath to 5% of the total possible neighbours (with a minimum of 3 and maximum of 250 edges per node). This was implemented since we observed that nodes in longer metapaths often find at least one spurious path to connect to every other node in the network. Although most of them end up having very low weights, the resulting network is very dense, requiring a much larger number of random-walks for the skip-gram model to learn the weight distribution of the network. All these cutoffs were chosen based on the thought exploration made by Himmelstein et al. and after optimising for different metapaths in our resource. Importantly, the effect of controlling the degree of the network was fundamental for having embedding spaces of good quality, especially for longer metapaths where these biases get exacerbated due to the combination of high-degree nodes from different datasets (Fig. A.2.7).

Additionally, we removed from the KG those nodes whose meaning was too general according to the information content provided in the ontology. This prevented those nodes to attract many connections in the network at the cost of providing very little information (e.g., disease terms such as 'cancer', 'syndrome' or 'genetic disease'; or cell compartments terms such as 'cell', 'membrane' or 'cell periphery'). All the pruned terms are provided in Supplementary Data 1.

Most downstream analyses rely on distances between the embeddings. However, even if we have implemented measures to control the degree of the network when producing the embedding, it is expected that nodes having more general implications will be generally closer to the rest than others that are more specific (e.g. 'Brain disease' (`https://disease-ontology.org/term/DOID:936`) will be closer to a much broad set of genes than 'Migraine' (`https://disease-ontology.org/term/DOID:6364`) which is

a specific condition comprised within the family of Brain diseases). Therefore, some terms may be biassed to have a closer distance distribution than others just because their edges define broader associations. Although encoding this can be useful in some downstream analysis (e.g., identifying drugs that target proteins specifically associated with particular brain diseases) it also may introduce biases when comparing distance distributions between terms (Fig. A.2.7).

To address these biases, we first assessed how different distances differentiate between these terms, finding that cosine distances provided more comparable distributions between terms while still preserving the (expected) enrichment of small distance associations of broader terms. Moreover, in order to add a measure of specificity in the distance, we also opted to compute co-ranks quantiles, which requires both nodes to be close to each other in order to consider they are sharing a close relationship (this was used in the HuRI-III exercise and the procedure is detailed in the corresponding section). By doing that, we can normalise the distance values of all entities, making them comparable (e.g., having a 0.1 co-rank quantile means the same regardless of the disease node).

Additionally, network permutations can be used in downstream analysis to control spurious observations made in networks that are being analysed with our embeddings. In fact, in the HuRI-III analysis, we randomly permuted the HuRI-III network (as detailed in the corresponding section) and used the permuted network as a reference to derive statistical significance cutoffs for the embedding distances we calculated.

Embedding evaluation and characterization

We used opt-SNE[278] to generate the 2D representation of the embeddings. To assess the quality of the embeddings, we reassembled the network obtained from the metapath using the embedding vectors. To this end, we first computed the cosine distance of each edge in the network using the embedding vectors of the nodes. Next, we generated 100 random permutations for each edge in the network and calculated the cosine distances between them. Finally, we sorted all the distances and computed the Area Under the Receiver Operating Characteristic (AUROC) curve using the network edges and the random permutations as the positive and negative sets, respectively. When assessing >L1 metapaths, we repeated the same exercise using 3 extra network subsets obtained by keeping, for each node, the top 1%, 25% and 50% closest neighbours according to the

DWPC weights of their edges. Embeddings with an AUROC below 0.8 were removed from the resource.

To characterise the embeddings, we first preselected a collection of reference networks representing commonly used biological associations. Then, given a set of embeddings corresponding to a certain metapath, we tested their capacity to recapitulate edges from other (orthogonal) datasets (i.e., the reference networks). Two measures were kept, the coverage (i.e., the number of overlapping nodes) and the AUROC, following the approach described above.

Aiming to extend this characterization, for each metapath we sought to characterise nodes separately, based on their entity type. We first calculated the term frequency-inverse document frequency (TF-IDF) values of the nodes from each reference network in our collection. Next, within the same entity type and network, we used the TF-IDF-transformed vectors to compute pairwise cosine similarities between nodes. Finally, we built the entity similarity network by keeping the top 5 closest neighbours for each node. Note that from one heterogeneous (bipartite) network this process yields two homogeneous networks, one for each entity type.

Some of the networks in our collection required customised pre-processing. To represent perturbation associations, we directly linked the perturbed genes (PGN-pup-GEN or PGN-pdw-GEN) and the outcome of such perturbation (e.g., PGN-bfn-CLL or PGN-upr-GEN) through the corresponding associations and datasets. We computed the CHE-has-CPD similarity networks by directly linking each node with the top 3 partners that shared more neighbours. Additionally, some entity similarity networks were gathered from other sources, like the CPD-CPD mechanism of action similarity obtained from our Chemical Checker resource[279].

Embedding-based gene expression analysis of cell lines

We downloaded the RMA-normalised gene expression (GEx) and the drug sensitivity data from the GDSC1000[33] web resource (https://www.cancerrxgene.org). We mapped the cell lines and genes to our reference vocabularies and took the mean value whenever duplicates occurred. We used the tissue of origin annotations from the CLUE cell app (https://clue.io/cell-app), which were already part of our graph (CLL-has-TIS, cl_tissue_clueio). Regarding CCLE data, we used the next-generation data[251] from the Broad Institute Portal (https://portals.broadinstitute.org/ccle/about). We processed the RNAseq data and produced

three embeddings (CLL-upr-GEN, CLL-dwr-GEN and CLL-dwr+upr-GEN-dwr+upr-CLL) following the pipeline detailed in the *Dataset standardisation* and *Obtaining the embeddings* sections.

In the drug sensitivity prediction exercise, we trained a random forest (RF) classifier for each drug and each GEx input data (i.e., the raw GEx or any of the GEx-derived embeddings). After removing drugs with less than 10 sensitive or resistant cell lines, we modelled 262 drugs. We used the SciKit Learn implementation of the RF algorithm, with a 10-fold stratified cross-validation scheme, and optimised RF hyperparameters over 20 iterations of Hyperopt[280].

Analysis of the HuRI-III interaction network

We downloaded HuRI-III from the Human interactome atlas (`http://www.interactome-atlas.org/`). Next, we considered all L1 metapaths containing a GEN metanode, keeping the dataset with higher coverage for each metapath and discarding those covering less than 10% of the HuRI-III network. As a representative of PPI interactions (GEN-ppi-GEN), we used a version of IntAct dated December 2019 (before publication of the HuRI-III network) from which we removed all entries belonging to the HuRI-III screening (IMEX: IM-25472). Next, we calculated the cosine distance between each PPI in each of the metapath embedding spaces and ranked the distances according to the distance distribution of each of the proteins. Distances and rankings were obtained with FAISS[281]. To derive empirical p values, we transformed the rankings into percentiles by normalising them by the total number of covered genes in each metapath and kept the geometric mean of the normalised co-ranked pairs.

In parallel, we generated 1000 random permutations of HuRI-III by randomly swapping each of the HuRI-III edges 10 times using the BiRewire bioconductor package (`https://doi.org/doi:10.18129/B9.bioc.BiRewire`) and, likewise, calculated p values for each metapath. For each permuted network, we calculated the recovery of the edges with a sliding p value cutoff (between 1 and 0.001) and averaged the counts at each cutoff. After repeating this process with the HuRI-III network, we were able to derive, for each metapath, the expected fold change (FC) across different p value cutoffs (i.e., the number of covered HuRI-III edges at a given p value cutoff divided by the average number of covered edges in the permuted networks). Moreover, the permuted networks were also used to estimate an empirical FDR for a given p value. For instance, for each metapath,

we found the p value cutoff associated with a 0.05 FDR by calculating the minimum p value needed to cover no more than 5% of the permuted network edges. Finally, to build the matrix shown in Fig. 3.2.6a, we selected the top 20 metapaths with the highest FC (i.e., FC average in the p value range between 0.1 and 0.001), and used their p values to cluster the PPIs with the fastcluster package[282] and the ward distance update formula.

To obtain the Shapley values, we trained a XGBoost model to classify GEN-GEN edges as positive (i.e., present in HuRI-III) or negative (i.e., not present in HuRI-III) using the p values across metapaths as features. To sample negative pairs, we used the instance of the permuted networks hitting fewer HuRI-III edges (~3%) in order to avoid having the same edge as positive and negative at the same time. Furthermore, since the objective of this exercise was to study the interplay between the metapaths, we removed edges that were covered by less than 10 (50%) metapaths, resulting in a dataset of 60k positive and negative pairs. A simple mean imputation was applied to the missing p values. At training time, we implemented a 20-fold stratified cross-validation split scheme and fine-tuned the hyperparameters using 20 iterations of Hyperopt[280]. Finally, we obtained the Shapley values from the test splits by implementing the TreeExplainer method[254]. All subsequent analyses and figures were obtained using the SHAP package (`https://github.com/slundberg/shap`).

Drug repurposing based on drug and disease embeddings

The first release of the repoDB (v1) data was downloaded from `http://apps.chiragjpgroup.org/repoDB` while the updated release (v2) was obtained from `https://unmtid-shinyapps.net/shiny/repodb`. Compounds were mapped to InChIKeys and diseases to the Disease Ontology (DO) forcing a 1:1 mapping. As features, we used the following metapaths (datasets) from the Bioteque resource: CPD-int-GEN (curated_targets); DIS-ass-GEN (disgenet_curated+disgenet_inferred); CPD-int-GEN-int-CPD-has-PHC (curated_targets-curated_targets-atc_drugs); and DIS-ass-GEN-ass-DIS-trt-CPD (disgenet_curated+disgenet_inferred-disgenet_curated+disgenet_inferred-repodb).

Additionally, we obtained the 2,048-bit Morgan fingerprints (ECDF4) of the compounds using RDKIT (`http://rdkit.org`) and used the adjacency matrix of the disease-gene network from DisGeNET as binary descriptors of diseases. Having defined the features of the model, we filtered out those drugs and diseases from repoDB that fell outside the embedding

universe and removed redundant pairs by de-propagating the associations to the most specific drug-disease terms according to the Disease Ontology. As a result, the train (repoDB v1) and test (repoDB v2) splits consisted of 2,522 and 1,187 unique drug-disease associations, respectively (Fig. A.2.5). Additionally, to prevent the model from focusing on the most frequently annotated drug and disease entities, we further processed the train data to balance the number of associations (degree of the nodes). More specifically, we capped the number of drug or disease associations to 5% of all possible associations (44 diseases and 26 drugs, respectively). Therefore, the associations of those drugs or diseases exceeding this limit were subsampled by performing a K-means clustering (where K was set to the capping limit) using the CPD-int-GEN or DIS-ass-GEN embeddings as features, and by randomly selecting a representative association from each of the clusters (Fig. A.2.5). This step slightly decreased the number of training data to 2,326 drug-disease associations.

Next, we produced train negative pairs by aggregating 20 negative networks obtained by randomly swapping the edges of the training data (thus, forcing a ratio of 1:20 between the positive and negative instances), while preventing inconsistencies in the Disease Ontology (i.e., having a negative association that would be obtained by propagating a positive drug-disease association through the ontology). Note that, to comply with the time-split scenario, we did not remove any negative drug-disease pair reported to be positive in the repoDB v2 release.

We ran a RF classifier for each feature set using 20 iterations of hyperopt[280] to fine-tune the hyperparameters. At prediction time, drug-disease associations in repoDB v2 were treated as positive pairs, while all the remaining drug-disease pairwise combinations were treated as negative pairs. To avoid inconsistencies, we removed those negative pairs that were semantically related to positive pairs according to the Disease Ontology. As a result, we obtained between [460-500] diseases and [750-800] drug predictions for each drug and disease, respectively. As most of the drugs and diseases only had one or two positive instances, we assessed the performance of the models by ranking all the predictions individually for each entity. Additionally, we calculated ROC curves for those drugs and diseases that had at least 5 positive instances. Finally, we obtained the pharmacological action of the drugs by mapping them to the uppermost level of the Anatomical Therapeutic Chemical (ATC) classification, when available. Likewise, disease families were derived by propagating the disease terms to the first and second levels of the Disease Ontology.

Chapter 3.3
A tool to efficiently annotate biomedical support behind experimental associations

Authors	Adrià Fernández-Torras, Martina Locatelli, Martino Bertoni, Miquel Duran-Frigola, Patrick Aloy
Type	Application Note
Stage	In preparation
Title	BQsupports: Systematic annotation of biomedical support for binary data.
Journal	N/A
DOI	N/A
Context	In the last chapter, the common standardisation of the embeddings enabled us to suggest a systematic approach to quantify the 'novelty' of the recently published Human Reference Interactome. Following this proof of concept, we developed an automatic pipeline to assess the biomedical support of experimental binary (network) datasets. Overall, this tool enables the screening of experimental data across a diverse set of biological scenarios, identifying biomedical contexts that potentially recapitulate the given observations.
Note	This chapter does not have supplementary material.

3.3.1 Abstract

Within a Big Data era in Biomedicine, there is an unmet need to systematically assess experimental observations in the context of available information. This assessment not only would offer a means for an unbiased validation of the results but also may provide an initial estimate of the potential novelty of the findings. Here we present BiotequeSupports (BQsupports), a web-based tool built upon a resource of biomedical descriptors that systematically annotate the biomedical support behind a given set of observations. The tool relies on biomedical descriptor spaces from which we know the biological traits encoded, covering over 11 different biological and chemical entities, including genes, cell lines, and small molecules. By assessing hundreds of descriptors, BQsupports provide support scores for each observation across a wide variety of biomedical contexts. These scores are then aggregated to summarise the biomedical support of the dataset as a whole. Finally, the tool also identifies predictive features of the given dataset, which can be exploited in downstream machine learning applications.

3.3.2 Introduction

Since the popularisation of high-throughput experiments to obtain an unbiased and more comprehensive description of biology, many initiatives have massively gathered data from biological systems[283]. Initial efforts relied on model organism screenings to uncover protein-protein interactions[56] and gene co-expression patterns[284]. Concurrently, drug repositories began to annotate bioactivity data for hundreds of drugs[285,286,287]. With the consolidation of large cell line panels, traditional omics started to gather all sorts of biological descriptors, from mutations in the genome to protein abundances[29,30]. The next generations incorporated global biological responses to small molecules and genetic perturbations[141,37,48]. Eventually, the accumulation of genomic data enabled the statistical exploration of gene-phenotype associations, leading to the extensive identification of disease-associated genes[288,289].

All these initiatives are populating biomedical repositories with hundreds of datasets, many of which are part of monumental efforts that still keep providing new releases to date (e.g., [185,251,20,58]). However, while first-in-class datasets may offer a wealth of new biological findings, the previously unknown insights extracted from subsequent releases or analogous

studies will inevitably saturate. Thus, it is paramount to have the means to contextualise new data in light of current biomedical knowledge. Indeed, it is a common practice to contextualise experimental results based on existing data, as it helps to validate some of the results, thereby, gaining confidence in the provided insights. Unfortunately, no standard exists for this analysis, which inevitably hampers the unbiased comparison with existing resources. Besides, these assessments usually use previous releases or analogous datasets as a reference, missing whether similar relationships have been found in orthogonal studies.

Here we present BiotequeSupports (BQsupports), a web tool to systematically quantify the support of novel biological associations based on the current biomedical knowledge pre-encoded in the Bioteque[290].
BQSupports measures the similarity of each pair of biomedical entities given by the user within a collection of diverse biomedical descriptors, providing support scores across various biomedical contexts. Additionally, by identifying the biomedical descriptors that better explain each novel type of association, BQsupports suggests biomedical features that can be used to predict relationships between entities that the screens might have missed. Overall, BQsupports takes a list of biomedical entity pairs as input and returns a detailed analysis of the biomedical context shared between them.

3.3.3 Results

BQsupports description

BQsupports annotates biomedical support scores between pairs of entities provided by the user. This support derives from biomedical knowledge descriptors (aka embeddings) gathered in the Bioteque resource[290], which can assess links between 11 different types of biomedical entities, namely: genes/proteins (GEN), cell lines (CLL), tissues (TIS), small molecule compounds (CPD), diseases (DIS), pharmacological classes (PHC), chemical entities (CHE), pathways (PWY), cellular components (CMP), protein domains (DOM) and molecular functions (MFN).

Given a set of node pairs covered by the resource, BQsupports automatically identifies contextual biomedical descriptors (termed 'metapaths') potentially related to the input data. Then, it measures the distance of each node pair within these descriptors to provide a support score for each pair. These individual scores are further aggregated to obtain a single support estimate for the entire dataset. Moreover, the tool uses net-

work permutations to (i) derive the expected support of the dataset and (ii) detect entity pairs that are significantly close (supported) in a particular biomedical context (quantified by an enrichment score). Additionally, BQsupports identifies metapaths able to distinguish the user's data from random permutations, thus providing means to prospectively predict new associations. Eventually, all these results are provided in different tables and summarised in a canvas picture. The entire pipeline is detailed in the *Methods* section.

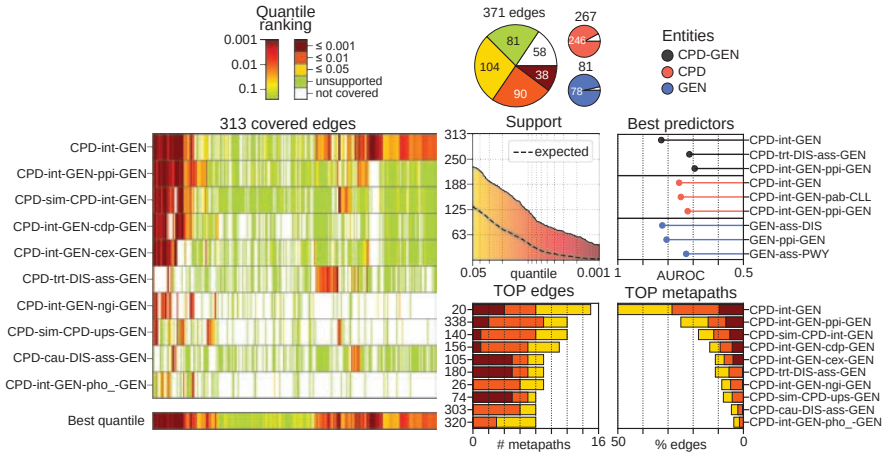


Figure 3.3.1: BQsupports analytical canvas for compound-gene biomarker pairs reported by FDA. On the left, we show the quantile ranking (support score) for all the input relationships (y-axis) covered by the top 10 most supportive metapaths (x-axis). The lower (redder) the quantile rank the higher the support score. On the right, we stratify the support scores according to different cutoffs and summarise the results in a pie chart, along with the coverage of the input dataset (smaller pie chart). Below, we quantify the support of the input dataset across different quantiles, showing the 'expected support' achieved by permuted networks (i.e., random expectation). Next to it, we rank the most predictive metapaths (quantified by the Area Under the Receiver Operating Characteristic (AUROC) curve). We perform this analysis for the specific pairs provided in the input dataset (black) and for entity types pairs sharing a similar association profile (see *Methods*). Finally, we show the top most supported input edges and the metapaths that most support the dataset at the bottom right corner.

As an illustrative example, Fig. 3.3.1 shows the BQsupports analysis for the pharmacogenomic biomarkers pairs provided by the FDA[10]. The heatmap on the left shows the level of support each metapath (rows) gives to each compound-gene (CPD-GEN) pair in the dataset (columns). When considering all the metapaths together (last row), over 74% of the covered interactions are supported by a sort of biomedical context (quantile rank ≤ 0.05). The expected support suggests that the support levels

for this dataset are higher than expected by chance. According to the TOP metapath ranking, compounds often interact with their biomarkers (CPD-int-GEN) or proteins associated with them (e.g., CPD-int-GEN-ppi-GEN). Interestingly, biomarkers are directly associated with the disease treated by the compound in 10% of the cases (CPD-trt-DIS-ass-GEN). Indeed, according to the best predictors of the dataset, compounds targeting the same proteins (CPD-int-GEN) tend to have similar biomarkers. Likewise, genes associated with the same diseases (GEN-ass-DIS) are prone to be biomarkers of the same compounds.

Coverage and Salability

BQsupports fetches biomedical descriptors from the Bioteque resource[290]. At the date of publication, it covers 11 different entities, allowing for the screening of 11 homogeneous and 110 bipartite entity-entity combinations. Each entity type has its identifier vocabulary, enabling a harmonised data integration. The available entities, vocabularies, and metapath descriptors are specified in the web resource (`https://bioteque.irbbarcelona.org`).

We tested the tool with tens of different datasets and confirmed that it scales well with hundreds of thousand edges. However, we capped the maximum number of edges to 1M to control the computational resources. With the default number of permutations (20), the entire pipeline takes between 1h to 2h to run, depending on the entity types and number of edges (Fig, 3.3.2, left). However, we observed that the number of permuted networks significantly impacts the computational time. For example, in the Bioplex-III network (~70k edges), moving from 20 to 1000 random permutations added 20 hours of extra computation (Fig. 3.3.2, right).

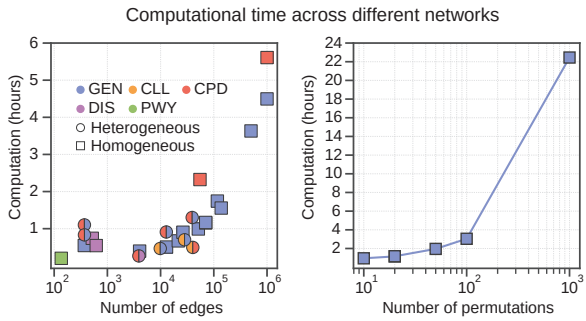


Figure 3.3.2: The computational time of the BQsupports pipeline. On the left, we show the computational time in hours (y-axis) that is needed to run the whole pipeline for different networks of varying sizes (x-axis) and types (shape and colour). On the right, we show the computational time in hours (y-axis) taken to run the BQsupports pipeline on the Bioplex-III [59] network (∼70k edges) using 10, 20, 50, 100, and 1000 network permutations (x-axis).

3.3.4 Concluding remarks

BQsupports is a tool intended to help scientists to better understand the biomedical relationships existing between their experimental observations. It annotates support scores for each association in a given dataset, providing insights for hypothesis generation and offering a means to systematically assess the novelty of the data. Additionally, it also identifies biomedical descriptors with predictive capabilities, which can then be downloaded from the Bioteque resource and used downstream to prospectively predict associations that might have been missed in the dataset. BQsupports is available as a web-based tool, where the user is only asked to (i) provide the dataset associations with proper identifiers, (ii) specify the entity types and (iii) optionally tune some parameters (e.g., the number of permuted networks). At the end of the process, BQSupports returns a canvas figure summarising the results (Fig. 3.3.1) and three (.tsv) tables files providing the quantile ranking score for each association-metapath combination, a digested summary counts for each descriptor, and the estimated performance of these metapaths in downstream predictive tasks. Moreover, as support scores are supplied for each association-metapath pair, the user can easily recompute most of the presented analyses according to custom needs (e.g., limiting the support score to a particular set of biomedical contexts or requiring a minimum enrichment score). The tool is accessible from the main Bioteque page (`https://bioteque.irbbarcelona.org/bqsupports`).

3.3.5 Methods

Data input and user interface

BQsupports takes pairs of associations (networks) as input data. Users can provide the data either explicitly to the web or by uploading an edge file. Next, the user has to specify the type of entities provided. When providing homogeneous networks, it is possible to specify whether the associations are undirected (e.g., protein-protein interactions) or directed (kinase-substrate interactions). This only affects the network permutation process (e.g., in a directed kinase-substrate network, random permutations will always produce kinase-substrate pairs). Users can also vary the number of permuted networks from 10 to 1000. By default, the tool uses 20 network permutations, allowing a p value resolution of 0.05. Notice that the number of permuted networks directly impacts the enrichment score, where statistical power increases proportionally to the number of permutations (i.e., enrichment scores tend to be more significant and accurate). However, increasing the number of permuted networks will also affect the computational time of the pipeline (Fig. 3.3.2, right).

Support score calculation

The pipeline starts by listing all the metapaths connecting the entities specified by the user. It considers metapaths of any length available in the Bioteque resource, except for GEN-GEN associations, which are limited to L1 metapaths. BQsupports omits metapaths covering less than 10% of the data.

Once the metapath universe is defined, it computes cosine distances between each provided association for each metapath space and ranks them according to the metapath distance distribution. To obtain these rankings efficiently, only the top 25% closest neighbours (first quartile) for each node are retrieved using FAISS[291]. Accordingly, the node found in the first quartile sets the maximum ranking distance in the metapath. Next, BQsupports transforms rankings into quantiles by dividing them by the number of nodes in the embedding space. Finally, as this process generates two quantiles (i.e., we obtain one ranking for each node), it derives an edge-level quantile by keeping the geometric mean of the pair (i.e., the normalised co-rank). This process is repeated independently for each metapath-dataset descriptor in the pre-selected universe.

Random permutations and enrichment score calculation

To generate random permutations of the data, we perform n random swaps of the network using the BiRewire Bioconductor package[292], where n is fixed to be ten times the number of edges in the dataset. Then, quantile ranking scores are calculated independently for each network permutation following the pipeline described in the previous section.

Enrichment scores are computed for each metapath and different support scores. More specifically, given a metapath-source descriptor space and a quantile cutoff (tested range between 1 and 0.001), the pipeline first annotates the number of associations in the given dataset that score lower than the given quantile cutoff. Then, it obtains a Fold Change (FC) by dividing this number by the median number of associations obtained from the random permutations. Additionally, it derives an empirical p value by counting the proportion of permuted networks with equal or more associations than the original dataset. Notice that the resolution of this p value will depend on the number of permuted networks (e.g. given 20 random permutations, the lowest computable p value is < 0.05).

Identifying potential predictors

To suggest potential predictors, the tool evaluates the capacity of the metapath descriptor to distinguish the dataset associations from random permutations. To this aim, BQsupports ranks all the associations according to their cosine similarities. Suppose 'real' edges (i.e., given by the user) are up-ranked before random permutations in a given metapath embedding space. In that case, it means that the space preserves the structure of the dataset, thus, descriptors (embeddings) of this space likely hold predictive power. This is quantified by computing the Area Under the Receiver Operating Characteristic (AUROC) curve between the user edges and 10 random permutations. To prevent an association from being counted as a positive and negative instance simultaneously, the pipeline generates new random permuted networks without allowing them to overlap with the user's data. At the end of the process, BQsupports provide the AUROC average across the 10 permutations together with the universe of each metapath and the covered portion of the user's dataset. Notice that the covered data represent the applicability domain of the computed AUROC.

Additionally, the pipeline also seeks metapath descriptors that, while not directly preserving the associations provided by the user, retain the

nodes' neighbourhood similarity. In other words, BQsupports first identifies pairs of nodes with similar interactions and then tries to find metapath descriptors that recapitulate these pairs. To identify these common interactors, the pipeline builds a similarity network by linking the nodes with other nodes of the same entity type that are significantly associated with the same nodes. This similarity network is created by (i) representing each node with a binary vector annotating their interactions (that is to say, the adjacency matrix), (ii) calculating term frequency-inverse document frequency (TF-IDF) values between the vectors and (iii) keeping the top 3 neighbours with highest TF-IDF similarity for each node. As a result, a new homogenous network is obtained, whose edges capture the most similar nodes from the user's network. Next, the pipeline lists a new metapath universe for this network and computes the recapitulation (AUROC) scores of these metapaths (using the same approach described in the previous paragraph). As a result, it identifies metapaths that keep the similarities between the nodes. Of note, metapaths identified in this analysis may differ from the ones in the previous analysis, even for homogeneous networks.

Generating the summary figure

Heatmap

To generate the heatmap matrix, BQsupports first aggregates the scores by keeping the best quantile ranking among the sources belonging to the same metapath, obtaining a unique score per metapath. Then, it ranks metapaths according to the number of interactions they support with a quantile lower than 0.05, selecting the top 10 for the heatmap. However, it provides the best quantile across all the screened metapaths in the last row. Associations not covered by a given metapath are left blank. Notice, though, that quantile scores are capped at 0.25 (1st quartile) as higher quantiles are ignored by the pipeline (see the *Metapath universe selection and distance-based quantile ranking calculation* section).

Pie Charts

Support scores for each dataset association are aggregated by selecting the best score across metapaths. Next, they are stratified into four groups according to their quantile: ≤ 0.001, ≤ 0.01, ≤ 0.05, and *unsupported* (quantile > 0.05). The pie chart reports the counts of each group, together with those not covered by the resource (if any). Additionally, a minor pie chart depicts the fraction of nodes covered for each entity, coloured according to the colour code used in the Bioteque resource.

Dataset support

The total number of supported associations in the dataset is reported across the range of significant quantile rankings (from 0.05 to 0.001). Additionally, BQsupports annotates the mean and standard deviation achieved with permuted networks (dashed line), providing the expected supportiveness according to the dataset's applicability domain (universe).

Edge and metapath ranking

The top 10 most supported edges and supportive metapaths are ranked according to the number of metapaths (or edges) with a quantile lower than 0.05. BQsupports uses the index provided in the original network (starting the count from 1) to label the edges in the plot (shown on the y-axis).

Best predictors

The canvas shows the top 3 metapath for each tested network (i.e., the one provided by the user and each entity-entity similarity network generated by BQsupports). In this case, the reported scores are not aggregated by metapath and correspond to a specific metapath-source combination. Furthermore, the tool only shows significant and relevant metapath-source combinations, that is, those whose average AUROC value (after subtracting their standard deviation) is higher than 0.6 and cover at least 20% of the dataset.

Generating the output of the pipeline

In addition to the summary canvas, the pipeline outputs the results in different (.tsv) files. A first file provides, for each association-metapath-source triplet combination, (i) the quantile ranking score, (ii) cosine distance (with its corresponding z score transformation), and (iii) the inferred enrichment score (with its corresponding p value) as detailed in the *Random permuted networks and enrichment score inference* section. A second one summarises the number of edge counts supported by each metapath-dataset across different support scores. Lastly, an additional file provides the recapitulation score (AUROCs) for each metapath assessed as a potential predictor, together with the coverage of the dataset and other practical information (e.g., metapath universe size).

Chapter 3.4
Connecting chemistry and biology with descriptors: a future perspective

Authors	Adrià Fernández-Torras, Arnau Comajuncosa-Creus, Miquel Duran-Frigola, Patrick Aloy
Type	Review
Stage	Published
Title	Connecting chemistry and biology through molecular descriptors
Journal	Current Opinion in Chemical Biology
DOI	https://doi.org/10.1016/j.cbpa.2021.09.001
Context	While Bioteque embeddings capture relationships between biological entities, very recent advances in representation learning are producing descriptors able to capture other aspects of biological and chemical entities, such as their sequence or structure. More importantly, the homogeneous format of these embeddings enables a natural concatenation with other descriptors, potentially creating more holistic representations. In this last chapter, we review the most promising chemical and biological embeddings that can complement the ones provided in the Bioteque resource. Besides, we illustrate a successful example in which different types of drug bioactivity data were formatted and integrated to provide a winning solution in an open community competition.
Note	This chapter does not have supplementary material.

3.4.1 Abstract

Through the representation of small molecule structures as numerical descriptors and the exploitation of the similarity principle, chemoinformatics has made paramount contributions to drug discovery, from unveiling mechanisms of action and repurposing approved drugs to *de novo* crafting of molecules with desired properties and tailored targets. Yet, the inherent complexity of biological systems has fostered the implementation of large-scale experimental screenings seeking a deeper understanding of the targeted proteins, the disrupted biological processes and the systemic responses of cells to chemical perturbations. After this wealth of data, a new generation of data-driven descriptors has arisen providing a rich portrait of small molecule characteristics that goes beyond chemical properties. Here, we give an overview of biologically relevant descriptors, covering chemical compounds, proteins and other biological entities, such as diseases and cell lines, while aligning them to the major contributions in the field from disciplines, such as natural language processing or computer vision. We now envision a new scenario for chemical and biological entities where they both are translated into a common numerical format. In this computational framework, complex connections between entities can be unveiled by means of simple arithmetic operations, such as distance measures, additions, and subtractions.

3.4.2 Introduction

Small molecules are an excellent tool to probe biological functions and the primary choice of pharmaceutical companies, as they are easy to manufacture, store, and distribute, and synthetic chemists can conceive a broad variety of them [293]. Some commercial and public chemical collections include up to 109 compounds, with the number increasing to 1,020 for proprietary libraries, which means that the chemical space available to researchers is essentially infinite [294]. Moreover, new strategies based solely on the combination of two- or three-step reaction sequences estimate that it would be possible to readily synthesise ∼29 billion compounds [295]. The size of the accessible chemical space easily explodes if fewer constraints are applied, with some plausible estimates exceeding 1,060 compounds for molecules under 500 Da [296]. In addition, and perhaps more importantly, in the last years high-throughput screening (HTS) assays have penetrated the public research sector (e.g., the study by Subramanian et al. [38]

and Corsello et al.[297]), providing depth of annotation to the compound collections. This is reflected in the increasing number of bioactive small molecules catalogued in open databases, which already amount to over two million entries[298,299].

Querying compounds in these databases differ greatly from querying proteins or genes. Biological sequences are richly annotated, and even when they are not, evolutionary and structural domains help link them to molecular functions, which, in turn, contributes to our understanding of higher-order biological processes[300]. Compared to biological sequences, small molecules spell a much more complicated code which, for the most part, has not been explored by the rules of natural evolution. In consequence, there is no clear and continuous connection between structure and function, which converts an apparently simple task, such as measuring similarity between two molecules into an open problem driving a whole field of research.

In practice, representing chemical compounds in a meaningful way (for compound similarity measures or other computational chemistry calculations) requires the selection of a small molecule descriptor. Among the classical chemical notations, we find the simplified molecular input line entry system (SMILES) that, although it might be ambiguous (i.e., one molecule can be described with multiple SMILES), it is very intuitive and widely used[301]. Other popular molecular descriptors encode the structural, topological and/or physicochemical properties of the compounds. These descriptors can account for the presence or absence of a specific set of pre-defined chemical groups, like in the case of the molecular access system keys[302], defined dynamically by listing the 2D structural elements encountered in a molecule. For example, in the extended connectivity fingerprints atoms are enumerated, and neighbouring elements and bonds are captured. Other complex descriptors broaden the structural information by capturing the spatial 3D coordinates of the atoms[303] or go beyond molecular geometry and consider environment-dependent properties, such as the active site of the receptor[304] or those derived from molecular simulations[305], within a given radius[306]. These and other similar descriptors have been at the core of chemoinformatics and are still the first choice in most applications (see the study by David et al.[307] for a recent and very comprehensive review). However, the last years have witnessed the expansion of a new generation of molecular descriptors, deemed to be 'data-driven' and based on deep learning approaches, that are engineered on the basis of large-scale chemistry databases and are thus

adaptable to a given task or region of the chemical space[101]. In particular, graph and text-based autoencoders are able to embed the information provided by 2D structures and SMILES strings, respectively, into a dense numerical vector belonging to a 'latent space'[308]. Simple measures such as Euclidean distances within the latent space are able to capture chemical similarity and, when coupled to machine learning algorithms, these descriptors have shown state-of-the-art performance in several biophysics and physiological benchmark datasets[309].

A natural extension of this first generation of data-driven descriptors is to include the wealth of bioactivity information available in the databases, to encapsulate, in the form of 'bioactivity descriptors', the experimental evidence gathered over years of research. Here, we review some recent attempts to provide these biologically relevant molecular descriptors and discuss how a descriptor-based approach may help integrate small molecules with larger biomolecules in a common framework able to capture several layers of biological complexity encompassing protein targets to cellular pathways and disease phenotypes.

3.4.3 Review's chapters

Extending the similarity principle beyond chemical structures

Chemical descriptors, in their different flavours, encode the physicochemical and structural properties of small molecules and provide a computer-friendly format to represent and compare them (Fig. 3.4.1). However, these descriptors do not incorporate bioactivity information explicitly, which handicaps the discovery of links between small molecules and other entities, such as proteins or cells. In pioneering work, instead of focusing on chemical structures, Kauvar et al.[310] characterised a set of compounds according to their ability to bind a panel of 18 receptors and used these affinity profiles to assess similarities between them. The idea of relating small molecules based on their target profiles was further developed over the next years[311,312], enhancing the performance in classical chemoinformatics tasks (e.g., target prediction). In a more complex attempt to capture phenotypic effects induced by drug activity in cells, MacDonald et al.[313] used a protein complementation assay to monitor the status of several cellular pathways after compound perturbation. Then, they derived pathway activity fingerprints for over a hundred compounds and found

that pathway-based similarities strongly correlated with known structure–activity relationships. Similarly, Young et al.[314] combined automated microscopy with image analysis to profile the biological effects of a compound library. They integrated the resulting phenotypic profiles with the chemical structure of the compounds and their predicted targets and found that the combination of the three features had a substantially higher capacity to identify mechanisms of action than either one in isolation.

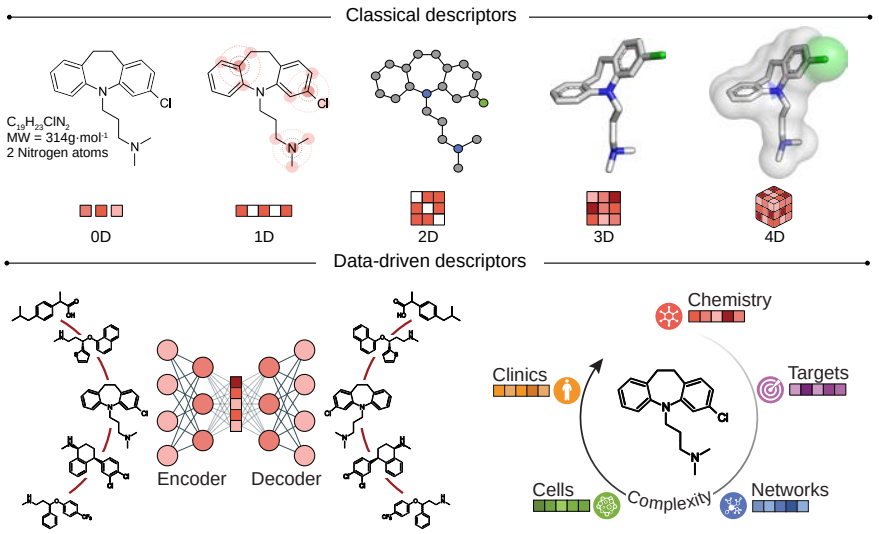


Figure 3.4.1: Encoding chemical molecules through their chemistry and bioactivity. Molecular descriptors allow for the mathematical treatment of chemical and structural features of molecules. There is a wide range of strategies to generate such descriptors. Simple approaches account for global molecular properties (0D, e.g. molecular weight) or the presence of particular structural features (1D, e.g. encoding circular environment of each atom up to a specific radius). The molecular topology (2D, e.g. distance matrices between atoms) or the spatial information of the atoms (3D, e.g. cartesian coordinates) can be encapsulated by conveniently representing molecules as chemical graphs. In addition, there are sophisticated methods that capture environment-dependent properties, such as functional regions or intramolecular interactions (4D, e.g. energetically favourable binding sites or multiple conformational states). Driven by the bloom of high-throughput assays and the following population of compound libraries, a new generation of data-driven descriptors based on deep learning strategies encode molecules into abstract latent spaces, representing molecular similarities as simple distance measures between numerical vectors. Furthermore, molecular descriptors have expanded beyond chemistry, integrating relevant biological data from heterogeneous bioactivity assays and providing a complementary framework to assess molecular similarity.

Indeed, the popularity of HTS assays has revealed that it is possible to establish relationships between compounds based on their functional activity rather than their chemical structure. For instance, it was suggested

that molecules triggering similar transcriptional responses in cell lines might share mechanisms of action, an observation that inspired the implementation of the connectivity map[37] and the following library of integrated network-based cellular signatures (LINCS L1000)[38] initiatives. These libraries provide a catalogue of transcriptional signatures in different cell lines, measured as a result of a systematic screening of genetic (CRISPR or shRNA) and pharmacological perturbations, which has been exploited, for instance, to suggest potential targets for a given compound[315]. Likewise, molecules that inhibit the growth of a similar subset of cell lines (i.e. that have similar sensitivity profiles)[316] or drugs that elicit similar side effects, also tend to share mechanisms of action[317], even if their 2D or 3D structures appear to be unrelated.

Building upon these seminal works, the Chemical Checker (CC) integrates the major chemogenomics and drug activity repositories and represents the largest collection of small molecule bioactivity signatures available to date[279]. The CC gathers experimentally determined bioactivity data for about 1M small molecules in the medicinal chemistry space and provides bioactivity descriptors in five levels of increasing biological complexity. The first level of descriptors characterises the chemical properties of the compounds, including their 2D and 3D structures, scaffolds, functional groups, and physicochemical properties. The second level captures information on the protein receptors of the molecules, including known mechanisms of action, metabolising enzymes and HTS binding assays. Descriptors in the third level of complexity address the propagation of the target perturbations triggered by the small molecules, including protein–protein interactions and pathways provided by several types of biological networks. The fourth level of signatures captures the bioactivity of the compounds measured at the cellular level, with assays including differential gene expression and sensitivity profiles in cancer cell line panels. Finally, for the few compounds that reached clinical stages, the fifth level of CC signatures encodes details on their therapeutic areas, adverse side effects and drug–drug interactions. A known limitation of the CC was that the number of molecules with reported bioactivities diminished at each level of complexity, and thus, it could only derive a limited set of bioactivity descriptors corresponding to a minority of well-characterised compounds. To extend the coverage of bioactivity descriptors to uncharacterised molecules, the authors trained a collection of deep neural networks (i.e., 'signaturizers') that are able to infer bioactivity signatures for any compound of interest, even when only its chemical structure is avail-

able. Noteworthy, they were able to assign a confidence score to the predictions of the signaturizers and systematically apply them to sets of compounds beyond drug molecules, including plant metabolites and food ingredients[318].

Overall, bioactivity signatures provide a complementary means to describe small molecules, focusing on the integration of multiple types of experimental data[319]. Indeed, these descriptors have proven useful to navigate the chemical space in a biologically relevant manner and boost the performance in many drug discovery tasks that typically rely on chemical descriptors, for example, target identification or toxicity prediction[318].

Target descriptors to complement small molecule signatures

In the quest to predict small-molecule bioactivities, often through machine learning approaches, the chemical compounds represent only one part of the equation. To match the rich chemical representations described previously, researchers are also developing methods to encapsulate information available for the biomolecular targets (Fig. 3.4.2). Protein sequence descriptors, for example, annotate the identity and the physicochemical properties of each amino-acid (e.g., the study by Hellberg et al.[320]) or measure general features of the full-length sequence, such as global residue composition and distribution (e.g., the study by Xiao et al.[321]). In any case, these relatively simple representations have been used in a battery of bioinformatics tasks, including protein engineering[322] or function prediction[323].

Like in the case of 'data-driven' descriptors for small molecules, deep learning is providing new ways to describe biological sequences. For instance, in a recent study, Alley et al.[324] applied deep neural networks to a vast set of unlabeled sequences, yielding semantics-rich descriptors that capture structural, evolutionary and biophysical properties of proteins. These descriptors have proven their value to predict the stability of *de novo* designed proteins, but their agnostic nature and versatile format make them a suitable input for almost any machine learning task involving proteins. In general, protein sequences are treated as text data, which allows for borrowing techniques from natural language processing, a discipline that has made extraordinary progress for knowledge representation[325,326]. In a first attempt to systematically benchmark language models (LMs) for protein modelling, Rao et al.[327] designed a set of tasks assessing protein embeddings and reported promising results for a variety of models

involving evolutionary understanding and protein engineering. Earlier this year, Elnaggar et al.[328] explored the limits of up-scaling LMs trained on protein sequences achieving, for the first time, performances competitive with evolutionary models, but requiring much less time to compute. Just recently, Bepler and Berger[329] extended their previous work and pretrained a protein LM conditioned to structure prediction tasks (e.g., the model was forced to predict residue contacts and structural similarity during training)[330]. By including evolutionary and structural information, they not only showed improvements in downstream tasks (e.g., protein function prediction) but also evidenced that hybrid approaches leveraging both data-driven sequences and physics-based domains can help LM to better embrace the sequence-structure–function paradigm. In another fresh work, Rao et al.[331] trained an LM taking multiple sequence alignments as input, conversely to the single sequence approach. Their model showed a better recapitulation of evolutionary variation and set a new state-of-the-art on unsupervised protein structure prediction[332]. It is worth noting that learning from both the multiple sequence alignments and the interplay between protein sequence and structure has been paramount to AlphaFold2 success in achieving outstanding accurate 3D protein structure predictions[333].

Most of these successful models are based on transformers, such as the Bidirectional Encoder Representations from Transformers (BERT), a widely used architecture in text recognition[334]. However, as with almost any method involving deep learning, the interpretability of these protein LMs is very limited. In a remarkable attempt to shed light on the biological and biophysical information captured by bidirectional encoder representations from transformers-based descriptors, Vig et al.[335] analysed the inner layers of the deep neural network thoroughly and found that they uncovered relevant associations in the 3D space, such as residues that were far apart in the sequence but spatially close in the structure or those constituting the protein binding sites. We refer the reader to the study by Bepler and Berger[330] for an insightful review of LMs in protein biology.

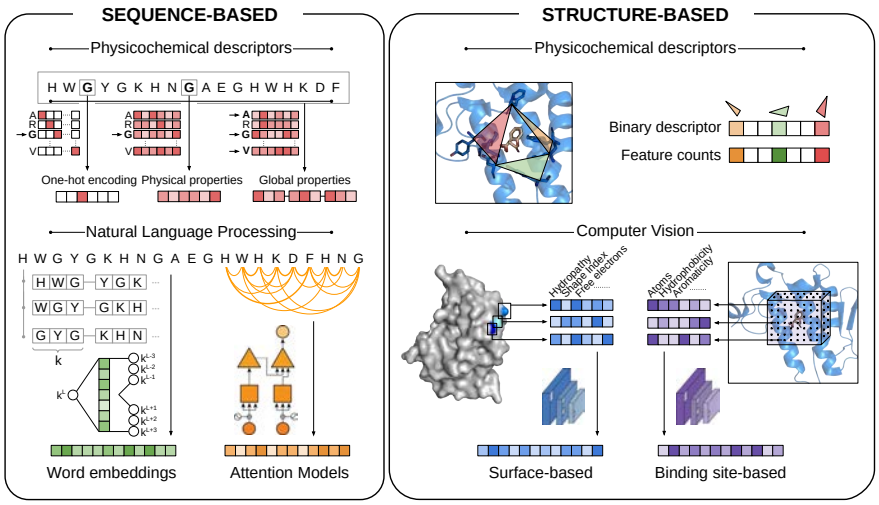


Figure 3.4.2: Target and binding pocket descriptors. The simplest way to represent a target protein sequence is by encoding the identity or the physicochemical properties of its amino-acids, either individually (i.e., one-hot encoding) or using sliding windows to capture their short-range environment. To account for more distant amino-acid relationships, proteins can be encoded using techniques borrowed from natural language processing (i.e., word embeddings or attention models), where sequences are often treated as a set of constant-length overlapping fragments or k-mers. Whenever high-resolution models of target proteins are available, these can be used to derive structure-based descriptors. The classical ones consider the geometry and physicochemical properties of the binding pockets by calculating distances between pharmacophoric points and transforming them into high–dimensional profiles, accounting for the presence or absence of a given pharmacophoric geometry. More recently, computer vision and deep learning techniques have been adapted to embed structural properties of protein surfaces and specific binding pocket features.

Binding between targets and ligands is determined by the biophysical properties of protein 3D structures and, in particular, the surface residues where potentially druggable pockets are found. Indeed, while a study exploring the binding promiscuity of over 160 drugs could not identify correlations between drug promiscuity and their chemical features (e.g., hydrophobicity), it did reveal structural similarities amongst their protein targets, highlighting the need to study binding site similarity across the proteome[336]. Thus, whenever high-resolution structures of the target proteins are available, more specific descriptors can be developed. Classic pocket descriptors measure the geometrical and electrostatic features of small molecule binding sites and translate them into binary fingerprints that just account for the presence or absence of a given structural motif (e.g., the study by Weill and Rognan[337], Siragusa et al.[338]). This is similar to what the extended connectivity fingerprint or molecular access system

descriptors do for chemical compounds. Cavity similarities based on these binding pocket fingerprints have unveiled interesting cases of remote homology between proteins[339] and are the basis for several polypharmacology strategies[340,341]. The popularity of methods to compare druggable pockets prompted the creation of thorough benchmark datasets, such as TOUGH-M1[342] and the protein site pairs for the evaluation of cavity comparison tools[343], which pointed out the strengths and weaknesses of a variety of descriptor types and approaches, and provided a gold standard to validate pocket comparison strategies to come. Systematic evaluation has revealed that some descriptors are better suited than active sites of related proteins, while others perform better to describe macromolecular binding interfaces, being the latter more appropriate for drug polypharmacology and repurposing studies[344]. If progress in natural language processing has enabled sequence-based descriptors, progress in image analysis and computer vision has prompted the development of 3D structure-based descriptors. For instance, Gainza et al.[344,345] devised a novel strategy to segment high-resolution protein surfaces into overlapping radial patches, mapping chemical, and geometrical features onto them. These data are then transferred into a convolutional neural network (CNN) to generate the descriptors, which can be fine-tuned for specific tasks, such as ligand-binding pocket similarity or protein–protein interaction interface comparisons. DeeplyTough is another recent method that also uses CNNs to encode 3D characteristics of protein binding pockets[346]. The peculiarity of DeeplyTough is that it has been trained to ensure that similar pockets are encoded into similar descriptors, while retaining the ability to account for small structural variations and differentiate closely related binding sites. In a recent protein site pairs for the evaluation of cavity comparison tools benchmark, pocket comparisons based on these descriptors scored among the best[343].

The significant improvement of both chemical and protein descriptors has prompted the development of proteochemometric strategies, where machine learning models are trained on a combination of ligand and target representations[347]. Indeed, these kinds of approaches have already shown superior performances in multi-target bioactivity prediction compared with classical methods[348], although some results may be over-optimistic due to bias in the training datasets as pointed out in the study by Chen et al.[349]. Moreover, Bongers et al.[347] showed that structure-based descriptors are often superior when a detailed definition of the target is needed (i.e., to distinguish drug selectivity among members of the same protein fam-

ily), while sequence-based ones are better suited for more generic models, especially when key structural details are lacking.

Capturing biological complexity in biomolecular descriptors

From a drug discovery perspective, genomic initiatives are providing new target opportunities[49,48], but many of these correspond to gene products thought to be undruggable, and the avalanche of data has not spurred the development of truly personalised, or even precision, therapies based on the exquisite interaction between a drug and an optimal target[350]. In fact, whole-cell phenotypic screenings continue to be the approach that contributes the most to the discovery of first-in-class medicines, while target-centric approaches appear more useful only for the development of follow-on products[139,351]. Thus, to tackle complex phenotypes, we need to move away from the 'one disease, one target, one drug' paradigm and consider the complexity of human pathologies from the early stages of the drug development process. Indeed, a growing fraction of recently approved drugs is associated with pharmacological biomarkers at the genomic scale[10], meaning that omics experiments are able to identify links between biomolecular profiles and drug action. This evidence is often complementary to the modulation of the intended therapeutic target and thus offers a more systemic view of drug activity.

In an attempt to capture this systemic complexity, it is increasingly common for HTS experiments to simultaneously characterise multiple omics profiles (i.e., trans-omics analyses)[225,226] so that several views of small molecule action can be analysed in parallel. New methodologies are flourishing to deal with such data (e.g., the study by Argelaguet et al.[227]) and yet, these methods mainly adapt existing strategies developed in the past for single omics experiments, and often draw conclusions from the most informative data type, while the rest are used as support. It is, thus, fundamental to come up with strategies able to capture the coordinated interplay of the many regulatory layers present in biological systems (Fig. 3.4.3).

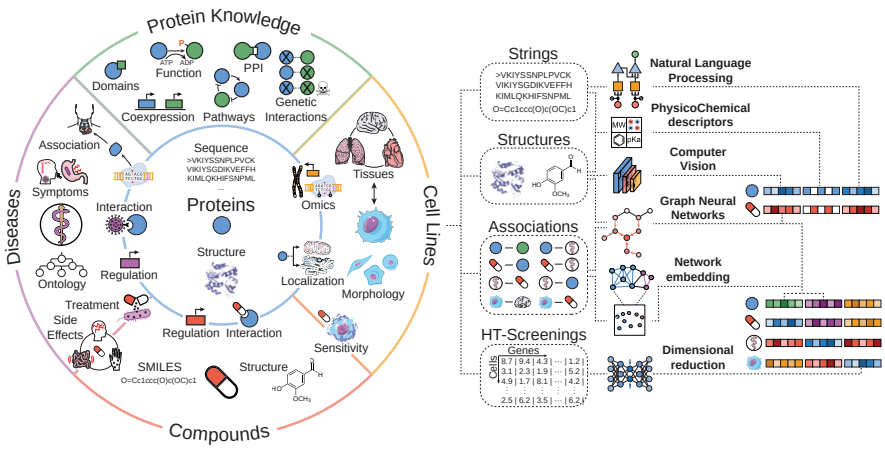


Figure 3.4.3: Capturing biological complexity in the form of descriptors. Bioactive chemical compounds often interact with their molecular targets to exert their function. However biological complexity spans far beyond protein targets, and long-range effects have a clear impact on drug action. At a molecular level, genes and proteins interact, forming complex networks that regulate physiology. Many of these physical or functional connections and their effects can be captured by individual biology experiments, while the integration of multi-omic unmasks the interrelations between different regulation layers. However, there is a resolution gap where we lose causality and all we can measure are somehow vague associations between molecules and higher-order phenotypic observations, such as a disease state. Depending on the nature of each experimental readout, different encoding strategies have been optimised to condense such complex biological data in the form of vector-like descriptors suitable for modern machine learning. String-like data, such as gene sequences or compound SMILES, are often encoded through the use of natural language models. Structural data, like the one representing protein and chemical structures or cellular morphology, is better suited for convolutional or graph neural networks. Alternatively, if the data to be encoded represent relationships between different biological entities, such as protein networks or compound–gene associations, network embedding techniques seem to yield the best results. Finally, as the readout of high-throughput screening experiments, such as drug sensitivity or cell transcriptomics, yields big numerical matrices, they are best condensed through the use of autoencoders.

Integrating many levels of biology into a single resource is a daunting task because one needs to standardise data formats and identifiers, normalise records across different resources and categorise the observations by applying significance cutoffs (e.g. of differential gene expression). Unlike chemical data, where we often have millions of molecules with relatively poor annotations, biological databases annotate a relatively small set of biomolecules with a large number of interactions between them and associations with other biological entities, such as diseases, pathways, molecular functions, cells, and tissues. The first successful attempts to organise multiple databases into a single resource (e.g., Harmonizome[90] and

Hetionet[93]) have structured the information in the form of a network, or knowledge graph, focused on the relationships (edges) between biological entities (nodes). However, the magnitude of biological networks is computationally intractable by traditional graph analysis techniques[109] which, also, in this case, has boosted the development of graph embedding approaches to reduce the dimensionality of the data while preserving the structural information and properties of the network[131]. Thanks to these advances, we have been able to release the Bioteque, a resource of biological network embeddings of unprecedented size and scope[290]. Bioteque descriptors are derived from a gigantic heterogeneous network (more than 450k nodes and 30M edges) that harmonises data extracted from >150 data sources, including 12 different biological entities (e.g., genes, diseases, drugs) linked through 67 types of relationships (e.g., 'drug treats disease', 'gene interacts with gene'). We have shown that this concise representation of the data can be used to evaluate and characterise a wide array of experimental observations (e.g., drug sensitivity assays), and have illustrated how these omics-based descriptors can be plugged into machine learning tasks, similar to what is done with their counterparts centred on proteins and chemical compounds. Also recently, Cantini et al.[231] evaluated the performance of several embedding methodologies to integrate continuous multi-omics data (e.g., gene expression, copy number variation, methylation and miRNA expression). In addition to evaluating the preservation of the original (raw data) structure, the authors also assessed their performance in predicting clinical outcomes in a cancer cohort, as well as classifying multi-omics single-cell data from cancer cell lines. They found that, while the performance of each method significantly changed depending on the task, a concomitant analysis of multiple datasets (i.e., multiple co-inertia analysis)[352] was the most consistent across different benchmarks.

While omics data has provided us with a broad understanding of biological phenomena, there are biological entities that are not easy to describe from a molecular perspective, as they usually involve ontological concepts or high-order functions. Biological pathways, often represented by gene ontology terms, are commonly embedded by grouping genes that participate in similar biological processes or have related functional categories[353]. Recently, Wang et al.[354] introduced an approach in which multiple gene sets are represented together in the embedding space, using a protein–protein interaction network as a measure of proximity between genes. This type of gene set descriptors has shown an improved

capacity to identify new functionally related gene set members and reveal subnetworks with clinical prognostic capacity in sarcoma samples. At a cellular level, Schubert et al.[355] trained a CNN to learn embeddings of neuron images, where each embedding represented a fragment of the cell thus capturing the neuron morphology. They proved the power of these embeddings to identify subcellular compartments, cell types and, more importantly, detect neuron reconstruction errors. Going one step up in the hierarchy of the biological organisation, Zitnik and Leskovec[229] developed OhmNet, a set of protein descriptors that take into consideration the specific protein–protein interactions within each human tissue, as well as the inter-tissue relationships, so that proteins with similar network neighbourhoods in similar tissues are placed proximally in the embedding space. Then, they showed that these tissue-aware protein descriptors provide more accurate predictions of tissue-specific protein functions than alternative approaches, making them a powerful tool to transfer these learned functions to the lesser characterised tissues. In related work, the same authors have embedded different networks (i.e., protein–protein, drug-target and disease-gene interactions) to explore the mechanisms of action of drugs[230]. Here, they modelled how drug effects spread through a hierarchy of biological functions coordinated by the underlying protein–protein interaction network. Thus, for each drug and disease, they learnt a diffusion profile to identify the key proteins and biological functions involved in treatment providing a transparent interpretation of the drug therapy.

Formatting heterogeneous data for effective downstream applications

The undeniable success of multi-omics and multi-modal implementations highlights the benefits of leveraging different layers of information in downstream tasks. And yet, biomedical data is as rich in content as heterogeneous in format (Fig. 3.4.3). From 1D sequences and 3D structures to binary associations and high-throughput screenings, the medley of formats makes data integration far from being a simple task. As already evidenced in this review, different areas in machine learning are crafting specialised architectures tailored to extract meaningful information from particular data formats. Indeed, the latent representation learned and outputted by these methods offers a convenient way to portray an extensive landscape of information in a homogeneous, concise, and amenable format for ma-

chine learning pipelines. Additionally, these vectors can be concatenated to obtain holistic representations of biological and chemical entities that can capture different facets of their nature, function, and interactions in a single numerical vector. Downstream models can then leverage the potential complementary information from various data sources by finding and exploiting connections between the vectors' dimensions. Ultimately, capturing orthogonal information in a harmonised format can conceivably lead to more effective solutions when addressing current challenges in drug discovery.

As a proof of concept, we would like to share our participation in the 'Pancancer Drug Activity DREAM challenge' launched in 2019 and published recently[356]. Briefly, DREAM challenges are conceived as community competitions that, by formatting fundamental questions into time-limited challenges, foster the scientific community to join and compete for the best contribution or solution. In the drug activity challenge case, participants were asked to predict the putative protein targets for a set of drugs solely from cell post-perturbational readouts, namely cell transcriptomics and cell growth inhibition, being blind to the chemical identity. We approached this challenge as a data integration exercise, where we tried to represent the transcriptional and sensitivity profiles of the challenge with a format compatible with cell readouts available from compound libraries with known targets. In a nutshell, by finding a common representation for all the drugs, we could train a model to extract insights from drug annotations and then predict the target profile of the compounds in the challenge. While conceptually simple, the particularities existing in the challenge data raised practical limitations. To begin with, the screening technology and experimental conditions used in the challenge did not agree with other publicly accessible screenings, hampering the direct integration between platforms. This forced us to develop small-scale predictive models to first transform the DREAM challenge data into readout profiles consonant with those in external resources (we refer the reader to the original paper[356] for a detailed description of the methods). Unfortunately, differences (batch effects) were still perceptible between different data sources, compromising the generalisability of the final model. Not only that, the dimensionality between both bioactivity sources was largely imbalanced (e.g., a few hundred sensitive cell lines against tens of thousands of genes' transcriptomics), biassing bioactivity information content[357]. To address these inconveniences, we used the Chemical Checker[279] pipeline to encode these bioactivities into low-dimensional vectors. In these new

spaces, the challenge and annotated drugs were represented in a homogeneous and compact vector format (128-dimensions), integrating the different drug sources in a shared space while balancing the volume of information from different bioactivity screenings. Reassuringly, we confirmed that the representation of both bioactivity types in this numerical format significantly boosted the prediction accuracy of the final model, eventually scoring among the top performers of the challenge.

Overall, embedding-based descriptors provide a scalable and intuitive means to capture complex data for biological and chemical entities, being convenient for integrating heterogeneous levels of information in a format readily optimised for downstream machine learning applications.

3.4.4 Concluding remarks

In this article, we have provided an overview of methods to represent chemical and biological entities in a common framework based on numerical descriptors. Although the approach may strike as too abstract to researchers uninitiated in data science, it has the unique advantage of capturing a number of data points that would otherwise be intractable. On top of that, this type of representation helps uncover links between entities by means of simple arithmetic calculations, such as similarity and distance measures between descriptors or additions to represent higher–order processes. The strategy can be applied at the atomistic level (e.g., compound similarity), as well as the phenotypic level, as first demonstrated by the connectivity map and LINCS L1000[38,37] in the context of gene expression data. Indeed, dissimilarities between chemical and disease perturbation signatures can be leveraged to find small molecules that potentially revert a specific disease gene expression profile, hence providing support for drug-disease indications[44].

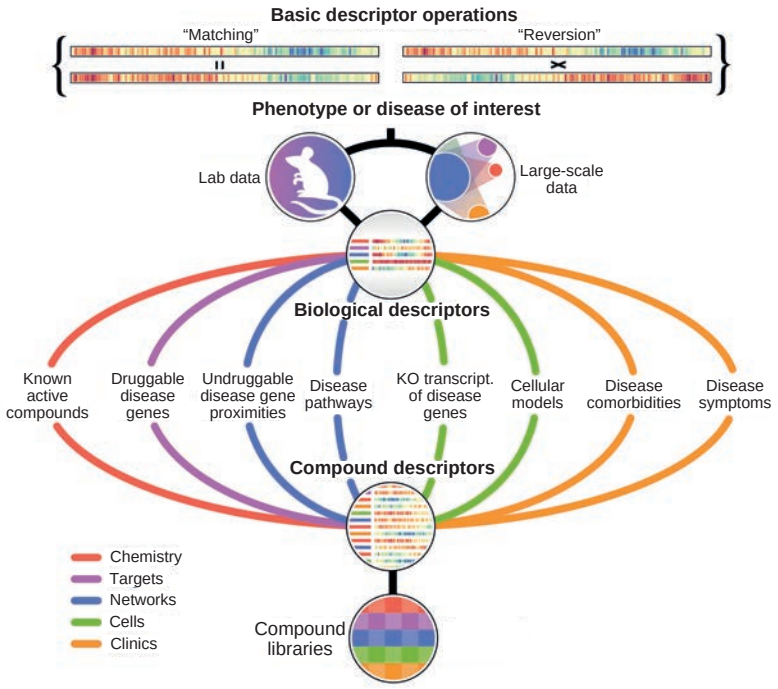


Figure 3.4.4: Connecting biology and chemistry through molecular descriptors. A common framework for small molecule and biological descriptors will enable a direct comparison between compound structures, bioactivity data and biological entities such as protein targets, cell lines or disease symptoms.

We envisage a scenario for computational chemistry and biology where drug candidates and biological entities will be first described with numerical vectors in the light of the available data, coming either from public repositories or in-house experiments (Fig. 3.4.4). These data would include structural features of the molecules and the targets, together with omics profiles, such as gene expression data, as well as large-scale biological networks and ontologies. Data will be linked at different levels with relatively simple operations, allowing for ultra-large, unbiased and systematic identification of the existing connections between the chemical space and the intricate biological space defined by disease biology.

4
Discussion

Biological systems are complex. Fortunately, global efforts driven by the scientific community are providing a wealth of data and knowledge that can be adopted as meaningful descriptors of biology. Moreover, the proper integration and harmonisation of this information allows for depicting a more comprehensive representation of biological systems, thereby providing a means to characterise and face their complexity. On that subject, network architectures have emerged as appealing frameworks due to their natural way of representing biological connections, which are intrinsic in biological systems. Especially useful has been the network representation of protein-protein interactions, as it provides a functional map where biological measurements (e.g., cell omics readouts) can be annotated and interpreted in the context of the underlying protein systems. Overall, networks offer a practical and versatile framework to represent and traverse biological information. In the context of this thesis, we have adopted network representations as the main frameworks to describe and exploit the existing biomedical knowledge, willing to offer strategies and solutions that address current challenges in the biomedicine field.

We started this journey intending to uncover functional insights from the outcome of pharmacogenomic experiments. In fact, with the identification of biologically functional determinants of drug response, one can shed light on the mechanism of action (MoA), favouring the design of more precise therapies. Unfortunately, while biological networks have been extensively exploited to delve into protein and disease biology, their role regarding chemical compounds has been mostly limited to their therapeutic connection to diseases[81,82]. To bridge this gap, we were inspired by network-driven advances in characterising diseases' protein landscapes[75,79,78] to develop a strategy to identify functional gene sets (network modules) associated with drug sensitivity. We applied this to the GDSC dataset[33], one of the most extensive pharmacogenomic panels to date, identifying functional modules for 189 drugs. Subsequent analysis demonstrated that these modules capture relevant genes of the drugs, being tightly related to the drug MoA and capable of anticipating drug response. More importantly, we could use these modules to functionally characterise the drugs, uncovering biological processes associated with their efficacy. Ultimately, we found these drug modules expressed in transcriptomics profiles from patient cohorts. According to scientific literature, the expression of these modules specifically linked the drug's MoA to relevant cancer-type-specific biological processes, highlighting the potential applicability of these modules beyond cellular screenings.

A limitation of our approach is that it is confined to the few tens of drugs accessible in pharmacogenomics panels. The truth is that these numbers are still distant from the thousands of drugs that are currently approved[358], and practically negligible if one considers the billions of compounds that are theoretically accessible from the vast chemical space[295]. Unfortunately, the high cost and involvement of these experimental screenings make the situation unlikely to change in the short term. Still, some initiatives[38,359] screen thousands of small molecules at the expense of covering much fewer cell lines. However, given that cell lines define the biological contexts available in the experiment, a poor representation of their biological diversity can compromise the robustness of the derived results. In fact, the lack of cellular variability has been identified as one of the main culprits behind the weak reproducibility of cell perturbational screenings[39,40]. In addition, some cellular properties, such as their division rate, have been proved to interfere with cell response readouts, raising serious concerns about the agreement between different pharmacogenomic panels[147,149,360].

Nevertheless, even if significant improvements are still to come to the field, our strategy only provides a partial view of the whole picture. Indeed, other omics may provide complementary cellular portraits playing a role in drug response[361,362,363], being recently combined in multi-omics settings to improve prediction performance[364,365]. However, multi-omics approaches often require sophisticated deep learning architectures to properly integrate and extract the biological information needed for a specific task of interest. Data obtained at different layers in biology will inevitably have their particularities, namely distinct numeric representations, distribution shapes, and screening resolutions (e.g., while cell mutational data is discrete and sparse, cell transcriptomics data is continuous and multimodal). Consequently, biological data becomes highly scattered and heterogeneous, posing challenges for their integration. More importantly, this limits our ability to unveil associations between different biological layers and establish connections between these biological systems and other domains of knowledge within life sciences.

To ameliorate this situation, we started collecting, harmonising and integrating data from recognised resources in the biomedical field. This endeavour led to the consolidation of a gigantic Knowledge Graph (KG), representing 12 entities and over 60 biomedical associations while covering about half a million nodes and ~30M edges from tens of data sources. To exploit this database effectively, we devised a pipeline that systematically encodes rational-guided biomedical network paths (metapaths) into

low-dimensional numerical representations (embeddings). As a result, we assembled the Bioteque, a resource of precomputed embeddings that, in its first release, provides over 1,000 descriptors depicting more than 700 distinct biomedical relationships. A significant added value of these representations is their ability to retain meaningful information from the KG in a much-reduced space. By illustration, we showed how our much reduced (128 dimensions) cell transcriptomics embeddings could recapitulate cells' tissue of origin and anticipate drug response with a similar level as raw gene expression profiles ($\sim$15-20k genes). Besides, we also notice that our descriptors often capture orthogonal information in their dimensional space. For instance, the compound-disease embedding space based on gene-driven associations partially recapitulates approved drug-disease treatment pairs, even if the metapath used for this space did not explicitly explore this information.

The fact that these descriptors can retain relevant biological relationships in their succinct space allows for highly efficient screening of biomedical knowledge. For example, by transforming different PPI network datasets into embeddings, we quickly uncovered domains of knowledge that were favoured in each dataset, thereby identifying those downstream tasks in which they are more performant. Likewise, we can itemise this characterisation to each individual association in a network dataset. Indeed, we used a repertoire of gene-based metapath embeddings to annotate and characterise the Human Reference Interactome (HuRI). Interestingly, by revealing biological contexts shared within PPIs, we could highlight those interactions that were particularly surprising considering the biological evidence collected in our resource. In other words, our descriptors proved to be an effective tool for identifying (orthogonal) evidence supporting a given biological connection.

It is a common practice to contextualise experimental data based on existing knowledge, as it helps to validate some of the results, hence, gaining confidence in the provided insights. Perhaps more interesting, this assessment provides an initial estimate of the novelty of the data. This novelty estimate can be particularly crucial for initiatives that periodically invest time and money to enlarge their datasets. Concretely, the new insights offered in further releases may no longer justify the cost of the assay in the long run, according to the Pareto principle[366]. Nevertheless, and despite its practical importance, the juxtaposition of new data with established knowledge is often done unsystematically, which inevitably increases the risk of biases during the assessment of the results.

This observation led us to extend our work and build an automatic web-based tool (named BQsupports) to annotate biomedical support systematically. By capitalising on the Bioteque resource, this tool can deal with more than 100 different entity-entity associations and screen, in a matter of a few hours, up to a million pairs across our collection of metapath descriptors. Besides, this tool not only gathers biomedical evidence behind each observation, helping validate the results, but also reveals the proportion of potentially novel data. Furthermore, metapath descriptors able to significantly recapitulate dataset associations will, by definition, hold great potential for predicting new ones. The intuition behind this is as follows: if entities sharing a trait of interest are close together in a given dimensional space, it should be possible to learn simple arithmetic rules (e.g., distance boundaries) to explore these spaces and, thus, identify other entities meeting those requirements. As shown in chapter 2, given a set of interactions sharing a trait of interest (e.g., compounds treating similar diseases), one can systematically quantify how this trait is preserved in different metapaths spaces and identify those more likely to work in downstream predictive tasks. Notice that this provides an efficient way to identify features with predictive power for a given dataset interest. Thus, by incorporating this functionality into BQsupports, this tool can be used to identify suitable descriptors for machine learning pipelines.

Of note, embeddings are especially optimised for computational approaches, as illustrated in chapter 2 with the task of predicting new drug-disease indications. Concretely, models trained with Bioteque embeddings showed a much superior performance than those trained with traditional descriptors, highlighting the benefits of having concise and homogeneous representations in the joint modelling of different modalities (aka multimodal learning[357]). The joint modelling of drug and disease pairs was crucial to tackling this exercise. Combining all drug-disease indications in a single training set allowed the model to uncover and learn universal relationships between these pairs. These insights could then be 'transferred' to predicting compounds and diseases with few annotations. Known as 'transfer learning'[367], this feature can be of utmost importance in scenarios where the lack of positive or negative instances prevents building robust individual models (i.e., remember that only one-third of the diseases and half of the drugs had at least one approved indication in the drug repurposing exercise). Additionally, as a single model is used to learn the relationship between two entities (e.g., drugs and disease), it can make predictions for any new pair combination, even if none of the entities in

the pair has been seen by the model before. Of course, the model may provide unfounded predictions if the descriptors of the new pairs are very distinct from the ones in the training data (i.e., outside the model's applicability domain). However, there are ways to assess this limitation *a posteriori*[318]. Conversely, this is impossible in unimodal settings, where individual models are built independently for each sample and can only provide predictions for pairs involving the modelled sample.

While it is true that multimodal approaches may offer substantial benefits, it is not always a trivial matter to assemble and train them in practice. Indeed, one of the major challenges comes from the distinct nature of the data that have to be fused. Consider, for example, the multimodal modelling of drug-protein interactions. While classical chemical fingerprints (e.g., Morgan fingerprints) describe drugs as flat binary vectors of a few thousand dimensions, proteins are traditionally described by 2D matrices representing physicochemical properties in their columns and the protein sequence in their rows. To properly integrate and leverage all these features conjointly, one first needs to deal with the heterogeneity and unbalance of their representations (i.e., drug descriptors are flat binary vectors of thousand dimensions, while proteins are 2D continuous matrices of few tens of dimensions). From a modelling perspective, one can address this challenge by designing deep learning architectures that will first intake these representations separately to integrate them later in the model[368]. However, this often comes at the expense of computational complexity, which usually demands technical understanding and computational power. In this regard, a significant added value of our embedding collection is that their standardised and reduced format enables the early integration of heterogeneous biomedical data, smoothing the implementation of multimodal-based strategies. Not only that, but these embeddings can also be concatenated to other numerical descriptors beyond our resource, creating even more comprehensive representations. As a result, predictive models can easily benefit from the complementarity and cooperation of different features to tackle more complex tasks.

Indeed, there is essential biochemical information that, despite being undoubtedly informative, cannot be easily incorporated into our network framework. A clear example is the chemical structure of the compounds, which defines their physicochemical properties, or the amino acid sequence of the proteins, which intrinsically captures their structure, thus their function. This data is more naturally represented in the form of text sequences (e.g., line notation of amino acids or chemical atoms), images (e.g., 3D co-

ordinates of amino acids or chemical atoms), or graph representations at the molecular level (e.g., connecting amino acids or chemical atoms by edges). Accordingly, recent and startling advances in natural language processing, computer vision, and graph representation are pioneering a revolutionary generation of data-driven encoders, able to extract meaningful insights from text, images, and graphs. Eventually, life science disciplines embraced these breakthroughs and started yielding information-rich descriptors based on sequences, structures, and physicochemical properties of biological and chemical entities. More importantly, the numerical properties of these representations are analogous to the ones obtained from network embeddings (i.e., both provide low-dimensional continuous spaces), making the concatenation of different descriptors straightforward. This opens the door for holistic representations, able to characterise, in a numerically condensed manner, multiple layers of biologically relevant information (e.g., from protein sequence and structure to their biological functions and interactions). Lured by these appealing representations, we devoted our last chapter to reviewing the latest advances in the characterisation of chemicals, proteins, and other biomedical entities that can significantly complement the descriptors provided in the Bioteque resource.

Finally, we have shown how the proper formatting and integration of different descriptors can be exploited to provide effective solutions to current challenges in Biomedicine. In particular, we participated in the Pan-cancer Drug Activity DREAM challenge, aimed at inferring the protein target profile of drugs solely from cell perturbational readouts. To address this challenge, we (i) harmonised and encoded bioactivity data, (ii) gathered drug target annotations from public libraries, and (iii) trained a multi-task deep learning model from these data to predict the targets of the DREAM challenge drugs. In the end, we managed to score among the top performers of the challenge, and consequently, our contribution was presented and discussed in a later manuscript[356]. Although not discussed in the manuscript, we would like to highlight that a preliminary version of the Bioteque resource had a significant contribution during the development of the final model. Concretely, the KG database proved to be a convenient entry framework to accommodate the small-scale experimental data provided by the challenge into the rich amalgam of biomedical evidence we collected when building the resource. Once accommodated, we could use our metapath approach to traverse the network and connect the drugs of the challenge to other small molecules, creating a joint embed-

ding space for them. Indeed, during the first validation phase of the challenge, a first prediction based on these embeddings achieved promising results, ranking among the best performers at that moment. Unfortunately, the constraints imposed by the low data coverage of the challenge vastly restricted the information we could exploit from the KG, limiting further advances in this strategy. However, after exploring the Bioteque embedding spaces, we could identify protein families over-represented within the top predictions. Naturally, we used this information to guide important decisions in our strategy, allowing us to boost our final model's performance significantly.

Overall, during the course of this thesis, we have proven how the existing biological knowledge can be used to digest omics readouts, characterise new biological observations and predict new associations of biomedical relevance. And among our contributions, particularly noteworthy has been the consolidation of a gigantic resource of precomputed biomedical descriptors, conceived with the underlying desire to provide biomedical knowledge in an effective format for its downstream exploitation. This was in part motivated by the successful publication of the Chemical Checker[279], a resource that formats bioactivity data from public repositories to describe small molecules. Indeed, the retrospective exploitation of previously published data, often in ways not intended by the original authors, is fundamental in the pursuit of scientific advancement. As it happens, during the time this thesis was carried out, at least two crucial scientific breakthroughs powerfully illustrate this fact. In the first quarter of 2020, an unprecedented pandemic shook the globe, becoming the worst health and economic threat in the 21st century. In a titanic joined response, the scientific community dug into their domain of knowledge and expertise and started to leverage sequences[369], structures[370], interactomes[83], omics readouts[371,372], drug libraries[373], and biomedical literature[374] to provide data, tools, and solutions to face the situation. This commitment was materialised in the first covid vaccine one year after the outbreak, a scientific milestone often referred to as a 'miracle' by the public opinion[375,376,377]. Concurrently at that time, the AlphaFold2 article and code were released to the community[378], making accessible the insights and details that allowed solving a so-called '50-year-old grand challenge in biology'. Since its publication, an 'AlphaFold mania'[379] haunted scientists, who started recycling the model in a plethora of new tasks, including the prediction of protein interactions[380], the study of disordered regions[381], or the 'hallucination' of new proteins[382].

Open-source and open-access trends in science encourage the scientific community to think out of the box, uncovering new findings while bypassing the time and money investment needed for data generation. It also fosters initiatives to assemble computational tools, models, and benchmarks into harmonised ecosystems, serving state-of-the-art solutions in a unified framework for their exploitation [383]. Fundamentally, these computational ecosystems significantly smoothen the implementation of cutting-edge technologies, bringing them closer to low-resourced countries [384]. Taken together, we can state that 'data parasitism' is emerging as a fruitful paradigm of research [385]. Indeed, the nature of this thesis deeply resonates with the data parasite ambition. We have devoted ourselves to elaborating tools and strategies that maximise the information available in biomedical databases to better comprehend and predict the present data. But at the same time, we have collected, harmonised, and reshaped biomedical data into a more convenient representation. And thus, we hope other data scientists will benefit from the data and insights presented in this thesis, contributing, in this way, to future research.

5
Conclusions

As stated in the Objectives section, this thesis aims to explore strategies that efficiently and effectively leverage the existing biological knowledge to (i) retrospectively shed light on publicly available experiments and (ii) prospectively exploit the information in different downstream biomedical tasks and applications.

Accordingly, we have made the following contributions:

1. The implementation of a new strategy to derive functional coordinated gene expression modules, able to characterise drugs' mechanisms of action through sensitivity-driven gene signatures.

2. The design of a pipeline to systematically encode heterogeneous biomedical data, yielding a resource of precomputed embeddings for a dozen of biological and chemical entities that proved to be valuable in multiple tasks and scenarios.

3. An automated web-based tool to mine existing biomedical knowledge and provide biomedical support to new experimental observations.

4. A review of state-of-the-art computational representations of proteins, small molecules, and other biological entities, that provide a complementary view of relevant information not covered in previous chapters.

5. The development of a computational model able to predict drug target profiles solely from cell perturbational readouts, scoring with it among the top performers of a DREAM community challenge.

Overall, we have illustrated how the annotated biological knowledge can be used as a means to effectively mine insights from biological screenings and provided a comprehensive collection of biomedical descriptors along with tools to better exploit the available information in biological and chemical databases.

References

[1] UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, 49(D1):D480–D489, January 2021.

[2] Geoffrey M Cooper. *The Complexity of Eukaryotic Genomes*. Sinauer Associates, 2000.

[3] Edward R Kastenhuber and Scott W Lowe. Putting p53 in context. *Cell*, 170(6):1062–1078, September 2017.

[4] Kit Curtius, Nicholas A Wright, and Trevor A Graham. An evolutionary perspective on field cancerization. *Nat. Rev. Cancer*, 18(1):19–32, January 2018.

[5] Vincent James Cogliano, Robert Baan, Kurt Straif, Yann Grosse, Béatrice Lauby-Secretan, Fatiha El Ghissassi, Véronique Bouvard, Lamia Benbrahim-Tallaa, Neela Guha, Crystal Freeman, Laurent Galichet, and Christopher P Wild. Preventable exposures associated with human cancers. *J. Natl. Cancer Inst.*, 103(24):1827–1839, December 2011.

[6] Allan Balmain. Cancer as a complex genetic trait: tumor susceptibility in humans and mouse models. *Cell*, 108(2):145–152, January 2002.

[7] Henry J Haiser, David B Gootenberg, Kelly Chatman, Gopal Sirasani, Emily P Balskus, and Peter J Turnbaugh. Predicting and manipulating cardiac drug inactivation by the human gut bacterium eggerthella lenta. *Science*, 341(6143):295–298, July 2013.

[8] Timothy A Scott, Leonor M Quintaneiro, Povilas Norvaisas, Prudence P Lui, Matthew P Wilson, Kit-Yi Leung, Lucia Herrera-Dominguez, Sonia Sudiwala, Alberto Pessia, Peter T Clayton,

Kevin Bryson, Vidya Velagapudi, Philippa B Mills, Athanasios Typas, Nicholas D E Greene, and Filipe Cabreiro. Host-Microbe co-metabolism dictates cancer drug efficacy in c. elegans. *Cell*, 169(3):442–456.e18, April 2017.

[9] Anna E Lindell, Maria Zimmermann-Kogadeeva, and Kiran R Patil. Multimodal interactions of drugs, natural compounds and pollutants with the gut microbiota. *Nat. Rev. Microbiol.*, 20(7):431–443, July 2022.

[10] U.S. Food and Drug Administration. Table of pharmacogenomic biomarkers in drug labeling. https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling.

[11] Virginia B Kraus. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat. Rev. Rheumatol.*, 14(6):354–362, June 2018.

[12] John S Witte, Peter M Visscher, and Naomi R Wray. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.*, 15(11):765–776, November 2014.

[13] Euan A Ashley. Towards precision medicine. *Nat. Rev. Genet.*, 17(9):507–522, August 2016.

[14] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nat. Rev. Genet.*, 19(5):299–310, May 2018.

[15] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, 42(1):30–35, January 2010.

[16] Euan A Ashley, Atul J Butte, Matthew T Wheeler, Rong Chen, Teri E Klein, Frederick E Dewey, Joel T Dudley, Kelly E Ormond, Aleksandra Pavlovic, Alexander A Morgan, Dmitry Pushkarev, Norma F Neff, Louanne Hudgins, Li Gong, Laura M Hodges, Dorit S Berlin, Caroline F Thorn, Katrin Sangkuhl, Joan M Hebert, Mark Woon, Hersh Sagreiya, Ryan Whaley, Joshua W Knowles, Michael F Chou, Joseph V Thakuria,

Abraham M Rosenbaum, Alexander Wait Zaranek, George M Church, Henry T Greely, Stephen R Quake, and Russ B Altman. Clinical assessment incorporating a personal genome. *Lancet*, 375(9725):1525–1535, May 2010.

[17] Elizabeth A Worthey, Alan N Mayer, Grant D Syverson, Daniel Helbling, Benedetta B Bonacci, Brennan Decker, Jaime M Serpe, Trivikram Dasu, Michael R Tschannen, Regan L Veith, Monica J Basehore, Ulrich Broeckel, Aoy Tomita-Mitchell, Marjorie J Arca, James T Casper, David A Margolis, David P Bick, Martin J Hessner, John M Routes, James W Verbsky, Howard J Jacob, and David P Dimmock. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.*, 13(3):255–262, March 2011.

[18] Teri A Manolio. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.*, 14(8):549–558, August 2013.

[19] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, and Tuuli Lappalainen. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, August 2021.

[20] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, 48(D1):D845–D855, January 2020.

[21] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 47(D1):D941–D947, January 2019.

[22] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. McKusick's online mendelian inheritance in man

(OMIM). *Nucleic Acids Res.*, 37(Database issue):D793–6, January 2009.

[23] David P Nusinow, John Szpyt, Mahmoud Ghandi, Christopher M Rose, E Robert McDonald, 3rd, Marian Kalocsay, Judit Jané-Valbuena, Ellen Gelfand, Devin K Schweppe, Mark Jedrychowski, Javad Golji, Dale A Porter, Tomas Rejtar, Y Karen Wang, Gregory V Kryukov, Frank Stegmeier, Brian K Erickson, Levi A Garraway, William R Sellers, and Steven P Gygi. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, 180(2):387–402.e16, January 2020.

[24] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X Zhu, Per E Lønning, Anne-Lise Børresen-Dale, Patrick O Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

[25] Joseph R Nevins and Anil Potti. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.*, 8(8):601–609, August 2007.

[26] Sarah Aldridge and Sarah A Teichmann. Single cell transcriptomics comes of age. *Nat. Commun.*, 11(1):4307, August 2020.

[27] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, 12(3):e694, March 2022.

[28] Takeshi Obayashi, Yuki Kagaya, Yuichi Aoki, Shu Tadaka, and Kengo Kinoshita. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.*, 47(D1):D55–D62, January 2019.

[29] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger,

John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Jr, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi A Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012.

[30] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(Database issue):D955–61, January 2013.

[31] Amrita Basu, Nicole E Bodycombe, Jaime H Cheah, Edmund V Price, Ke Liu, Giannina I Schaefer, Richard Y Ebright, Michelle L Stewart, Daisuke Ito, Stephanie Wang, Abigail L Bracha, Ted Liefeld, Mathias Wawer, Joshua C Gilbert, Andrew J Wilson, Nicolas Stransky, Gregory V Kryukov, Vlado Dancik, Jordi Barretina, Levi A Garraway, C Suk-Yee Hon, Benito Munoz, Joshua A Bittker, Brent R Stockwell, Dineo Khabele, Andrew M Stern, Paul A Clemons, Alykhan F Shamji, and Stuart L Schreiber. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, August 2013.

[32] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu,

Francesco Iorio, Didier Surdez, Li Chen, Randy J Milano, Graham R Bignell, Ah T Tam, Helen Davies, Jesse A Stevenson, Syd Barthorpe, Stephen R Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A Engelman, Sreenath V Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S Gray, Jeffrey Settleman, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, March 2012.

[33] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, July 2016.

[34] Hossein Sharifi-Noghabi, Soheil Jahangiri-Tazehkand, Petr Smirnov, Casey Hon, Anthony Mammoliti, Sisira Kadambat Nair, Arvind Singh Mer, Martin Ester, and Benjamin Haibe-Kains. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief. Bioinform.*, 22(6), November 2021.

[35] Zichen Wang, Alexander Lachmann, Alexandra B Keenan, and Avi Ma'ayan. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics*, 34(12):2150–2152, June 2018.

[36] Karel K M Koudijs, Anton G T Terwisscha van Scheltinga, Stefan Böhringer, Kirsten J M Schimmel, and Henk-Jan Guchelaar. Transcriptome signature reversion as a method to reposition drugs against cancer for precision oncology. *Cancer J.*, 25(2):116–120, 2019.

[37] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S Lander, and Todd R Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.

[38] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017.

[39] Mario Niepel, Marc Hafner, Caitlin E Mills, Kartik Subramanian, Elizabeth H Williams, Mirra Chung, Benjamin Gaudio, Anne Marie Barrette, Alan D Stern, Bin Hu, James E Korkola, LINCS Consortium, Joe W Gray, Marc R Birtwistle, Laura M Heiser, and Peter K Sorger. A multi-center study on the reproducibility of Drug-Response assays in mammalian cell lines. *Cell Syst*, 9(1):35–48.e5, July 2019.

[40] Nathaniel Lim and Paul Pavlidis. Evaluation of connectivity map shows limited reproducibility in drug repositioning. *Sci. Rep.*, 11(1):17624, September 2021.

[41] Alexandra B Keenan, Megan L Wojciechowicz, Zichen Wang, Kathleen M Jagodnik, Sherry L Jenkins, Alexander Lachmann, and Avi Ma'ayan. Connectivity mapping: Methods and applications. *Annu. Rev. Biomed. Data Sci.*, 2(1):69–92, July 2019.

[42] Bence Szalai, Vigneshwari Subramanian, Christian H Holland, Róbert Alföldi, László G Puskás, and Julio Saez-Rodriguez. Signatures of cell death and proliferation in perturbation transcriptomics data-from confounding factor to effective prediction. *Nucleic Acids Res.*, 47(19):10010–10026, November 2019.

[43] Andrew Jones, Aviad Tsherniak, and James M McFarland. Post-perturbational transcriptional signatures of cancer cell line vulnerabilities. March 2020.

[44] Bin Chen, Li Ma, Hyojung Paik, Marina Sirota, Wei Wei, Mei-Sze Chua, Samuel So, and Atul J Butte. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat. Commun.*, 8:16022, July 2017.

[45] Song Yi, Shengda Lin, Yongsheng Li, Wei Zhao, Gordon B Mills, and Nidhi Sahni. Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.*, 18(7):395–410, July 2017.

[46] Vasileios Stathias, Anna M Jermakowicz, Marie E Maloof, Michele Forlin, Winston Walters, Robert K Suter, Michael A Durante, Sion L Williams, J William Harbour, Claude-Henry Volmar, Nicholas J Lyons, Claes Wahlestedt, Regina M Graham, Michael E Ivan, Ricardo J Komotar, Jann N Sarkaria, Aravind Subramanian, Todd R Golub, Stephan C Schürer, and Nagi G Ayad. Drug and disease signature integration identifies synergistic combinations in glioblastoma. *Nat. Commun.*, 9(1):5315, December 2018.

[47] Thai-Hoang Pham, Yue Qiu, Jucheng Zeng, Lei Xie, and Ping Zhang. A deep learning framework for high-throughput

mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell*, 3(3):247–257, March 2021.

[48] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, Robin M Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F J Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A Garraway, David E Root, Todd R Golub, Jesse S Boehm, and William C Hahn. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, July 2017.

[49] Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, Rizwan Ansari, Sarah Harper, David Adam Jackson, Rebecca McRae, Rachel Pooley, Piers Wilkinson, Dieudonne van der Meer, David Dow, Carolyn Buser-Doepner, Andrea Bertotti, Livio Trusolino, Euan A Stronach, Julio Saez-Rodriguez, Kosuke Yusa, and Mathew J Garnett. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature*, 568(7753):511–516, April 2019.

[50] Elaine M Oberlick, Matthew G Rees, Brinton Seashore-Ludlow, Francisca Vazquez, Geoffrey M Nelson, Neekesh V Dharia, Barbara A Weir, Aviad Tsherniak, Mahmoud Ghandi, John M Krill-Burger, Robin M Meyers, Xiaofeng Wang, Phil Montgomery, David E Root, Jake M Bieber, Sandi Radko, Jaime H Cheah, C Suk-Yee Hon, Alykhan F Shamji, Paul A Clemons, Peter J Park, Michael A Dyer, Todd R Golub, Kimberly Stegmaier, William C Hahn, Elizabeth A Stewart, Stuart L Schreiber, and Charles W M Roberts. Small-Molecule and CRISPR screening converge to reveal receptor tyrosine kinase dependencies in pediatric rhabdoid tumors. *Cell Rep.*, 28(9):2331–2344.e8, August 2019.

[51] Joo Sang Lee, Avinash Das, Livnat Jerby-Arnon, Rand Arafeh, Noam Auslander, Matthew Davidson, Lynn McGarry, Daniel James, Arnaud Amzallag, Seung Gu Park, Kuoyuan Cheng, Welles Robinson, Dikla Atias, Chani Stossel, Ella Buzhor, Gidi Stein, Joshua J Waterfall, Paul S Meltzer, Talia Golan, Sridhar Hannenhalli, Eyal Gottlieb, Cyril H Benes, Yardena Samuels, Emma

Shanks, and Eytan Ruppin. Harnessing synthetic lethality to predict the response to cancer treatment. *Nat. Commun.*, 9(1):2546, June 2018.

[52] Edmond M Chan, Tsukasa Shibue, James M McFarland, Benjamin Gaeta, Mahmoud Ghandi, Nancy Dumont, Alfredo Gonzalez, Justine S McPartlan, Tianxia Li, Yanxi Zhang, Jie Bin Liu, Jean-Bernard Lazaro, Peili Gu, Cortt G Piett, Annie Apffel, Syed O Ali, Rebecca Deasy, Paula Keskula, Raymond W S Ng, Emma A Roberts, Elizaveta Reznichenko, Lisa Leung, Maria Alimova, Monica Schenone, Mirazul Islam, Yosef E Maruvka, Yang Liu, Jatin Roper, Srivatsan Raghavan, Marios Giannakis, Yuen-Yi Tseng, Zachary D Nagel, Alan D'Andrea, David E Root, Jesse S Boehm, Gad Getz, Sandy Chang, Todd R Golub, Aviad Tsherniak, Francisca Vazquez, and Adam J Bass. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature*, 568(7753):551–556, April 2019.

[53] Kyuho Han, Edwin E Jeng, Gaelen T Hess, David W Morgens, Amy Li, and Michael C Bassik. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.*, 35(5):463–474, May 2017.

[54] Medina Colic, Gang Wang, Michal Zimmermann, Keith Mascall, Megan McLaughlin, Lori Bertolet, W Frank Lenoir, Jason Moffat, Stephane Angers, Daniel Durocher, and Traver Hart. Identifying chemogenetic interactions from CRISPR screens with drugz. *Genome Med.*, 11(1):52, August 2019.

[55] Thao V Nguyen and Andrea C Alfaro. Applications of omics to investigate responses of bivalve haemocytes to pathogen infections and environmental stress. *Aquaculture*, 518:734488, March 2020.

[56] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute,

Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[57] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E Begg, Pascal Braun, Marc Brehme, Martin P Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J Gutierrez, Madeleine F Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R Murray, Alexandre Palagi, Matthew M Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruyssinck, Julie M Sahalie, Annemarie Scholz, Akash A Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O Tejeda, Shelly A Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E Cusick, Yu Xia, Albert-László Barabási, Lilia M Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A Calderwood, David E Hill, Tong Hao, Frederick P Roth, and Marc Vidal. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, November 2014.

[58] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G Coté, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F Hardy, Nishka Kishore, Jennifer J Knapp, István A Kovács, Irma Lemmens, Miles W Mee, Joseph C Mellor, Carl Pollis, Carles Pons, Aaron D Richardson, Sadie Schlabach, Bridget Teeking, Anupama Yadav, Mariana Babor, Dawit Balcha, Omer Basha, Christian Bowman-Colin, Suet-Feung Chin, Soon Gang Choi, Claudia Colabella, Georges Coppin, Cassandra D'Amata, David De Ridder, Steffi De Rouck, Miquel Duran-Frigola, Hanane Ennajdaoui, Florian

Goebels, Liana Goehring, Anjali Gopal, Ghazal Haddad, Elodie Hatchi, Mohamed Helmy, Yves Jacob, Yoseph Kassa, Serena Landini, Roujia Li, Natascha van Lieshout, Andrew MacWilliams, Dylan Markey, Joseph N Paulson, Sudharshan Rangarajan, John Rasla, Ashyad Rayhan, Thomas Rolland, Adriana San-Miguel, Yun Shen, Dayag Sheykhkarimli, Gloria M Sheynkman, Eyal Simonovsky, Murat Taşan, Alexander Tejeda, Vincent Tropepe, Jean-Claude Twizere, Yang Wang, Robert J Weatheritt, Jochen Weile, Yu Xia, Xinping Yang, Esti Yeger-Lotem, Quan Zhong, Patrick Aloy, Gary D Bader, Javier De Las Rivas, Suzanne Gaudet, Tong Hao, Janusz Rak, Jan Tavernier, David E Hill, Marc Vidal, Frederick P Roth, and Michael A Calderwood. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, April 2020.

[59] Edward L Huttlin, Raphael J Bruckner, Jose Navarrete-Perea, Joe R Cannon, Kurt Baltier, Fana Gebreab, Melanie P Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, John Szpyt, Brandon M Gassaway, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Guha Thakurta, Tian Zhang, Ramin Rad, Joshua Pan, David P Nusinow, Joao A Paulo, Devin K Schweppe, Laura Pontano Vaites, J Wade Harper, and Steven P Gygi. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040.e28, May 2021.

[60] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, 42(Database issue):D358–63, January 2014.

[61] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47(D1):D607–D613, January 2019.

[62] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-Aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, 47(D1):D529–D541, January 2019.

[63] Trey Ideker and Ruth Nussinov. Network approaches and applications in biology. *PLoS Comput. Biol.*, 13(10):e1005771, October 2017.

[64] Edwin K Silverman, Harald H H W Schmidt, Eleni Anastasiadou, Lucia Altucci, Marco Angelini, Lina Badimon, Jean-Luc Balligand, Giuditta Benincasa, Giovambattista Capasso, Federica Conte, Antonella Di Costanzo, Lorenzo Farina, Giulia Fiscon, Laurent Gatto, Michele Gentili, Joseph Loscalzo, Cinzia Marchese, Claudio Napoli, Paola Paci, Manuela Petti, John Quackenbush, Paolo Tieri, Davide Viggiano, Gemma Vilahur, Kimberly Glass, and Jan Baumbach. Molecular networks in network medicine: Development and applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 12(6):e1489, November 2020.

[65] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, March 2002.

[66] T Ideker, T Galitski, and L Hood. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2:343–372, 2001.

[67] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, February 2004.

[68] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, 12(1):56–68, January 2011.

[69] B Schwikowski, P Uetz, and S Fields. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18(12):1257–1261, December 2000.

[70] Michael Costanzo, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D Lee, Vicent Pelechano, Erin B Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S Piotrowski, Tharan Srikumar, Sondra Bahr, Yiqun Chen, Raamesh Deshpande, Christoph F Kurat, Sheena C Li, Zhijian Li, Mojca Mattiazzi Usaj, Hiroki Okada, Natasha Pascoe, Bryan-Joseph San Luis, Sara Sharifpoor, Emira Shuteriqi, Scott W Simpkins, Jamie Snider, Harsha Garadi Suresh, Yizhao Tan, Hongwei Zhu, Noel Malod-Dognin, Vuk Janjic, Natasa Przulj, Olga G Troyanskaya, Igor Stagljar, Tian Xia, Yoshikazu Ohya, Anne-Claude Gingras, Brian Raught, Michael Boutros, Lars M Steinmetz, Claire L Moore, Adam P Rosebrock, Amy A Caudy, Chad L Myers, Brenda Andrews, and Charles Boone. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), September 2016.

[71] Patrick Goymer. Network biology: why do we need hubs? *Nat. Rev. Genet.*, 9(9):650, September 2008.

[72] David E Gordon, Joseph Hiatt, Mehdi Bouhaddou, Veronica V Rezelj, Svenja Ulferts, Hannes Braberg, Alexander S Jureka, Kirsten Obernier, Jeffrey Z Guo, Jyoti Batra, Robyn M Kaake, Andrew R Weckstein, Tristan W Owens, Meghna Gupta, Sergei Pourmal, Erron W Titus, Merve Cakir, Margaret Soucheray, Michael McGregor, Zeynep Cakir, Gwendolyn Jang, Matthew J O'Meara, Tia A Tummino, Ziyang Zhang, Helene Foussard, Ajda Rojc, Yuan Zhou, Dmitry Kuchenov, Ruth Hüttenhain, Jiewei Xu, Manon Eckhardt, Danielle L Swaney, Jacqueline M Fabius, Manisha Ummadi, Beril Tutuncuoglu, Ujjwal Rathore, Maya Modak, Paige Haas, Kelsey M Haas, Zun Zar Chi Naing, Ernst H Pulido,

Ying Shi, Inigo Barrio-Hernandez, Danish Memon, Eirini Petsalaki, Alistair Dunham, Miguel Correa Marrero, David Burke, Cassandra Koh, Thomas Vallet, Jesus A Silvas, Caleigh M Azumaya, Christian Billesbølle, Axel F Brilot, Melody G Campbell, Amy Diallo, Miles Sasha Dickinson, Devan Diwanji, Nadia Herrera, Nick Hoppe, Huong T Kratochvil, Yanxin Liu, Gregory E Merz, Michelle Moritz, Henry C Nguyen, Carlos Nowotny, Cristina Puchades, Alexandrea N Rizo, Ursula Schulze-Gahmen, Amber M Smith, Ming Sun, Iris D Young, Jianhua Zhao, Daniel Asarnow, Justin Biel, Alisa Bowen, Julian R Braxton, Jen Chen, Cynthia M Chio, Un Seng Chio, Ishan Deshpande, Loan Doan, Bryan Faust, Sebastian Flores, Mingliang Jin, Kate Kim, Victor L Lam, Fei Li, Junrui Li, Yen-Li Li, Yang Li, Xi Liu, Megan Lo, Kyle E Lopez, Arthur A Melo, Frank R Moss, 3rd, Phuong Nguyen, Joana Paulino, Komal Ishwar Pawar, Jessica K Peters, Thomas H Pospiech, Jr, Maliheh Safari, Smriti Sangwan, Kaitlin Schaefer, Paul V Thomas, Aye C Thwin, Raphael Trenker, Eric Tse, Tsz Kin Martin Tsui, Feng Wang, Natalie Whitis, Zanlin Yu, Kaihua Zhang, Yang Zhang, Fengbo Zhou, Daniel Saltzberg, QCRG Structural Biology Consortium, Anthony J Hodder, Amber S Shun-Shion, Daniel M Williams, Kris M White, Romel Rosales, Thomas Kehrer, Lisa Miorin, Elena Moreno, Arvind H Patel, Suzannah Rihn, Mir M Khalid, Albert Vallejo-Gracia, Parinaz Fozouni, Camille R Simoneau, Theodore L Roth, David Wu, Mohd Anisul Karim, Maya Ghoussaini, Ian Dunham, Francesco Berardi, Sebastian Weigang, Maxime Chazal, Jisoo Park, James Logue, Marisa McGrath, Stuart Weston, Robert Haupt, C James Hastie, Matthew Elliott, Fiona Brown, Kerry A Burness, Elaine Reid, Mark Dorward, Clare Johnson, Stuart G Wilkinson, Anna Geyer, Daniel M Giesel, Carla Baillie, Samantha Raggett, Hannah Leech, Rachel Toth, Nicola Goodman, Kathleen C Keough, Abigail L Lind, Zoonomia Consortium, Reyna J Klesh, Kafi R Hemphill, Jared Carlson-Stevermer, Jennifer Oki, Kevin Holden, Travis Maures, Katherine S Pollard, Andrej Sali, David A Agard, Yifan Cheng, James S Fraser, Adam Frost, Natalia Jura, Tanja Kortemme, Aashish Manglik, Daniel R Southworth, Robert M Stroud, Dario R Alessi, Paul Davies, Matthew B Frieman, Trey Ideker, Carmen Abate, Nolwenn Jouvenet, Georg Kochs, Brian Shoichet, Melanie Ott, Massimo Palmarini, Kevan M Shokat,

Adolfo García-Sastre, Jeremy A Rassen, Robert Grosse, Oren S Rosenberg, Kliment A Verba, Christopher F Basler, Marco Vignuzzi, Andrew A Peden, Pedro Beltrao, and Nevan J Krogan. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*, 370(6521), December 2020.

[73] István A Kovács, Katja Luck, Kerstin Spirohn, Yang Wang, Carl Pollis, Sadie Schlabach, Wenting Bian, Dae-Kyum Kim, Nishka Kishore, Tong Hao, Michael A Calderwood, Marc Vidal, and Albert-László Barabási. Network-based prediction of protein interactions. *Nat. Commun.*, 10(1):1240, March 2019.

[74] Eric E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, September 2009.

[75] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.*, 18(3):507–522, March 2011.

[76] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proc. Natl. Acad. Sci. U. S. A.*, 104(21):8685–8690, May 2007.

[77] Igor Feldman, Andrey Rzhetsky, and Dennis Vitkup. Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U. S. A.*, 105(11):4323–4328, March 2008.

[78] Mika Gustafsson, Colm E Nestor, Huan Zhang, Albert-László Barabási, Sergio Baranzini, Sören Brunak, Kian Fan Chung, Howard J Federoff, Anne-Claude Gavin, Richard R Meehan, Paola Picotti, Miguel Àngel Pujana, Nikolaus Rajewsky, Kenneth Gc Smith, Peter J Sterk, Pablo Villoslada, and Mikael Benson. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med.*, 6(10):82, October 2014.

[79] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Disease networks. uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, February 2015.

[80] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, 17(10):615–629, October 2016.

[81] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nat. Commun.*, 7:10331, February 2016.

[82] Feixiong Cheng, István A Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. *Nat. Commun.*, 10(1):1197, March 2019.

[83] Yadi Zhou, Yuan Hou, Jiayu Shen, Yin Huang, William Martin, and Feixiong Cheng. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov*, 6:14, March 2020.

[84] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, June 2013.

[85] Gaia Cantelli, Alex Bateman, Cath Brooksbank, Anton I Petrov, Rahuman S Malik-Sheriff, Michele Ide-Smith, Henning Hermjakob, Paul Flicek, Rolf Apweiler, Ewan Birney, and Johanna McEntyre. The european bioinformatics institute (EMBL-EBI) in 2021. *Nucleic Acids Res.*, 50(D1):D11–D19, January 2022.

[86] Daniel J Rigden and Xosé M Fernández. The 2021 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.*, 49(D1):D1–D9, January 2021.

[87] Andrew D Rouillard, Zichen Wang, and Avi Ma'ayan. Reprint of "abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction". *Comput. Biol. Chem.*, 59 Pt B:123–138, December 2015.

[88] Charles E Cook, Rodrigo Lopez, Oana Stroe, Guy Cochrane, Cath Brooksbank, Ewan Birney, and Rolf Apweiler. The european bioinformatics institute in 2018: tools, infrastructure and training. *Nucleic Acids Res.*, 47(D1):D15–D22, January 2019.

[89] Bio databases 2018: How do they taste? https://digitalworldbiology.com/blog/bio-databases-2018-how-do-they-taste, January 2018.

[90] Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016, July 2016.

[91] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514, February 2022.

[92] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, and Matthias Mann. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.*, 40(5):692–702, May 2022.

[93] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6, September 2017.

[94] Xi Yang, Wei Wang, Jing-Lun Ma, Yan-Long Qiu, Kai Lu, Dong-Sheng Cao, and Cheng-Kun Wu. BioNet: a large-scale and heterogeneous biological network model for interaction prediction with graph convolution. *Brief. Bioinform.*, 23(1), January 2022.

[95] Finlay MacLean. Knowledge graphs and their applications in drug discovery. *Expert Opin. Drug Discov.*, 16(9):1057–1069, September 2021.

[96] Jaroslav Pokorný. Graph databases: Their power and limitations. In *Computer Information Systems and Industrial Management*, pages 58–69. Springer International Publishing, 2015.

[97] Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C Stern, and Artem Cherkasov. The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, March 2022.

[98] Daniel M Busiello, Samir Suweis, Jorge Hidalgo, and Amos Maritan. Explorability and the origin of network sparsity in living systems. *Sci. Rep.*, 7(1):12323, September 2017.

[99] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Trans. Knowl. Data Eng.*, 31(5):833–852, May 2019.

[100] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[101] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, July 2018.

[102] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nat. Genet.*, 51(1):12–18, January 2019.

[103] Artificial intelligence in structural biology is here to stay. http://dx.doi.org/10.1038/d41586-021-02037-0, July 2021.

[104] Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J Frey. Deep learning in biomedicine. *Nat. Biotechnol.*, 36(9):829–838, October 2018.

[105] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.*, 18(6):463–477, June 2019.

[106] Miquel Duran-Frigola, Adrià Fernández-Torras, Martino Bertoni, and Patrick Aloy. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 9(6):e1408, November 2019.

[107] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, Cambridge, England, January 2020.

[108] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[109] H Cai, V W Zheng, and K Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge & Data Engineering*, 30(09):1616–1637, 2018.

[110] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression: a unified subspace learning framework for content-based image retrieval. In *Proceedings of the 15th ACM international conference on Multimedia*, MM '07, pages 403–412, New York, NY, USA, September 2007. Association for Computing Machinery.

[111] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale information network embedding. March 2015.

[112] Shaosheng Cao, Wei Lu, and Qiongkai Xu. GraRep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 891–900, New York, NY, USA, October 2015. Association for Computing Machinery.

[113] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. May 2017.

[114] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. May 2020.

[115] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, 82(4):949–958, April 2008.

[116] Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Peng Ni, Kaijie Zhao, Fang-Xiang Wu, and Yi Pan. Computational drug repositioning with random walk on a heterogeneous network. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 16(6):1890–1900, November 2019.

[117] Kyle C Chipman and Ambuj K Singh. Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, 10:17, January 2009.

[118] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. March 2014.

[119] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. July 2016.

[120] Leonardo F R Ribeiro, Pedro H P Savarese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. April 2017.

[121] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 135–144, New York, NY, USA, August 2017. Association for Computing Machinery.

[122] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *CoRR*, abs/1709.05584, 2017.

[123] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[124] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? October 2018.

[125] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. October 2017.

[126] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. GraphSAINT: Graph sampling based inductive learning method. July 2019.

[127] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 793–803, New York, NY, USA, July 2019. Association for Computing Machinery.

[128] Megha Khosla, Vinay Setty, and Avishek Anand. A comparative study for unsupervised network representation learning. *IEEE Trans. Knowl. Data Eng.*, 33(5):1807–1818, May 2021.

[129] Tobias Schumacher, Hinrikus Wolf, Martin Ritzert, Florian Lemmerich, Jan Bachmann, Florian Frantzen, Max Klabunde, Martin Grohe, and Markus Strohmaier. The effects of randomness on the stability of node embeddings. May 2020.

[130] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[131] Michelle M Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine. *arXiv*, April 2021.

[132] Jason Fan, Anthony Cannistra, Inbar Fried, Tim Lim, Thomas Schaffner, Mark Crovella, Benjamin Hescott, and Mark D M Leiserson. A Multi-Species functional embedding integrating sequence and network structure. March 2018.

[133] Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg, and Roded Sharan. To embed or not: Network embedding as a paradigm in computational biology. *Front. Genet.*, 10:381, May 2019.

[134] William L Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. June 2018.

[135] Gamal Crichton, Yufan Guo, Sampo Pyysalo, and Anna Korhonen. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, 19(1):176, May 2018.

[136] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact integration of Multi-Network topology for functional analysis of genes. *Cell Syst*, 3(6):540–548.e5, December 2016.

[137] Fangping Wan, Lixiang Hong, An Xiao, Tao Jiang, and Jianyang Zeng. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, 35(1):104–111, July 2018.

[138] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.*, 10(6):428–438, June 2011.

[139] David C Swinney and Jason Anthony. How were new medicines discovered? *Nat. Rev. Drug Discov.*, 10(7):507–519, June 2011.

[140] A Monks, D Scudiero, P Skehan, R Shoemaker, K Paull, D Vistica, C Hose, J Langley, P Cronise, and A Vaigro-Wolff. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J. Natl. Cancer Inst.*, 83(11):757–766, June 1991.

[141] Brinton Seashore-Ludlow, Matthew G Rees, Jaime H Cheah, Murat Cokol, Edmund V Price, Matthew E Coletti, Victor Jones, Nicole E Bodycombe, Christian K Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J Wawer, Nurdan Kuru, Joanne D Kotz, C Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A Bittker, Michelle Palmer, James E Bradner, Alykhan F Shamji, Paul A Clemons, and Stuart L Schreiber. Harnessing connectivity in a Large-Scale Small-Molecule sensitivity dataset. *Cancer Discov.*, 5(11):1210–1223, November 2015.

[142] Matthew G Rees, Brinton Seashore-Ludlow, Jaime H Cheah, Drew J Adams, Edmund V Price, Shubhroz Gill, Sarah Javaid, Matthew E Coletti, Victor L Jones, Nicole E Bodycombe, Christian K Soule, Benjamin Alexander, Ava Li, Philip Montgomery, Joanne D Kotz, C Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Daniel A Haber, Clary B Clish, Joshua A Bittker, Michelle Palmer, Bridget K Wagner, Paul A Clemons, Alykhan F Shamji, and Stuart L Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.*, 12(2):109–116, February 2016.

[143] Paul Geeleher, Zhenyu Zhang, Fan Wang, Robert F Gruener, Aritro Nath, Gladys Morrison, Steven Bhutra, Robert L Grossman, and R Stephanie Huang. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res.*, 27(10):1743–1751, October 2017.

[144] Su-In Lee, Safiye Celik, Benjamin A Logsdon, Scott M Lundberg, Timothy J Martins, Vivian G Oehler, Elihu H Estey, Chris P Miller, Sylvia Chien, Jin Dai, Akanksha Saxena, C Anthony Blau, and Pamela S Becker. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.*, 9(1):42, January 2018.

[145] Xiaoming Liu, Jiasheng Yang, Yi Zhang, Yun Fang, Fayou Wang, Jun Wang, Xiaoqi Zheng, and Jialiang Yang. A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.*, 6:22811, March 2016.

[146] Mario Niepel, Marc Hafner, Qiaonan Duan, Zichen Wang, Evan O Paull, Mirra Chung, Xiaodong Lu, Joshua M Stuart, Todd R Golub, Aravind Subramanian, Avi Ma'ayan, and Peter K Sorger. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat. Commun.*, 8(1):1186, October 2017.

[147] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo J W L Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, December 2013.

[148] Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84–87, December 2015.

[149] Paul Geeleher, Eric R Gamazon, Cathal Seoighe, Nancy J Cox, and R Stephanie Huang. Consistency in large pharmacogenomic studies. *Nature*, 540(7631):E1–E2, November 2016.

[150] Yoo-Ah Kim and Teresa M Przytycka. Bridging the gap between genotype and phenotype via network approaches. *Front. Genet.*, 3:227, 2012.

[151] L H Hartwell, J J Hopfield, S Leibler, and A W Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, December 1999.

[152] Mariano J Alvarez, Yao Shen, Federico M Giorgi, Alexander Lachmann, B Belinda Ding, B Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, 48(8):838–847, August 2016.

[153] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 25(2):197–206, February 2007.

[154] Vlado Dančík, Hyman Carrel, Nicole E Bodycombe, Kathleen Petri Seiler, Dina Fomina-Yadlin, Stefan T Kubicek, Kimberly Hartwell, Alykhan F Shamji, Bridget K Wagner, and Paul A Clemons. Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. *J. Biomol. Screen.*, 19(5):771–781, June 2014.

[155] Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res.*, 46(D1):D994–D1002, October 2017.

[156] Michael Z Gilcrease. Integrin signaling in epithelial cells. *Cancer Lett.*, 247(1):1–25, March 2007.

[157] Vered Givant-Horwitz, Ben Davidson, and Reuven Reich. Laminin-induced signaling in tumor cells. *Cancer Lett.*, 223(1):1–10, June 2005.

[158] P J Keely, J K Westwick, I P Whitehead, C J Der, and L V Parise. Cdc42 and rac1 induce integrin-mediated cell motility and invasiveness through PI(3)K. *Nature*, 390(6660):632–636, December 1997.

[159] J A Gelderloos, S Rosenkranz, C Bazenet, and A Kazlauskas. A role for src in signal relay by the platelet-derived growth factor alpha receptor. *J. Biol. Chem.*, 273(10):5908–5915, March 1998.

[160] Manuel Patarroyo, Karl Tryggvason, and Ismo Virtanen. Laminin isoforms in tumor invasion, angiogenesis and metastasis. *Semin. Cancer Biol.*, 12(3):197–207, June 2002.

[161] P Keely, L Parise, and R Juliano. Integrins and GTPases in tumour cell growth, motility and invasion. *Trends Cell Biol.*, 8(3):101–106, March 1998.

[162] Soo-Hyun Kim, Jeremy Turnbull, and Scott Guimond. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. *J. Endocrinol.*, 209(2):139–151, May 2011.

[163] Jing Du, Yan Zu, Jing Li, Shuyuan Du, Yipu Xu, Lang Zhang, Li Jiang, Zhao Wang, Shu Chien, and Chun Yang. Extracellular matrix stiffness dictates wnt expression through integrin pathway. *Sci. Rep.*, 6:20395, February 2016.

[164] Paulina Moreno-Layseca and Charles H Streuli. Signalling pathways linking integrins with cell cycle progression. *Matrix Biol.*, 34:144–153, February 2014.

[165] Caroline A Lee, David Neul, Andrea Clouser-Roche, Deepak Dalvie, Michael R Wester, Ying Jiang, J P Jones, 3rd, Sascha Freiwald, Michael Zientek, and Rheem A Totah. Identification of novel substrates for human cytochrome P450 2J2. *Drug Metab. Dispos.*, 38(2):347–356, February 2010.

[166] Ulrich M Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.*, 138(1):103–141, April 2013.

[167] Nerino Allocati, Michele Masulli, Carmine Di Ilio, and Luca Federici. Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis*, 7(1):8, January 2018.

[168] John D Hayes, Jack U Flanagan, and Ian R Jowsey. Glutathione transferases. *Annu. Rev. Pharmacol. Toxicol.*, 45:51–88, 2005.

[169] M D Green, P A Francis, V Gebski, V Harvey, C Karapetis, A Chan, R Snyder, A Fong, R Basser, J F Forbes, and Australian New Zealand Breast Cancer Trials Group. Gefitinib treatment in hormone-resistant and hormone receptor-negative advanced breast cancer. *Ann. Oncol.*, 20(11):1813–1817, November 2009.

[170] Xia Zhang, Bin Zhang, Jie Liu, Jiwei Liu, Changzheng Li, Wei Dong, Shu Fang, Minmin Li, Bao Song, Bo Tang, Zhehai Wang, and Yang Zhang. Mechanisms of gefitinib-mediated reversal of tamoxifen resistance in MCF-7 breast cancer cells by inducing ERα re-expression. *Sci. Rep.*, 5:7835, February 2015.

[171] Florence Huguet, Marie Fernet, Nicole Giocanti, Vincent Favaudon, and Annette K Larsen. Afatinib, an irreversible EGFR family inhibitor, shows activity toward pancreatic cancer cells, alone and in combination with radiotherapy, independent of KRAS status. *Target. Oncol.*, 11(3):371–381, June 2016.

[172] N Ioannou, A G Dalgleish, A M Seddon, D Mackintosh, U Guertler, F Solca, and H Modjtahedi. Anti-tumour activity of afatinib, an irreversible ErbB family blocker, in human pancreatic tumour cells. *Br. J. Cancer*, 105(10):1554–1562, November 2011.

[173] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.*, 11(4):e1004120, April 2015.

[174] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica

Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res.*, 46(D1):D649–D655, January 2018.

[175] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.

[176] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, 47(2):106–114, February 2015.

[177] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45(D1):D362–D368, January 2017.

[178] Tudor I Oprea, Cristian G Bologa, Søren Brunak, Allen Campbell, Gregory N Gan, Anna Gaulton, Shawn M Gomez, Rajarshi Guha, Anne Hersey, Jayme Holmes, Ajit Jadhav, Lars Juhl Jensen, Gary L Johnson, Anneli Karlson, Andrew R Leach, Avi Ma'ayan, Anna Malovannaya, Subramani Mani, Steven L Mathias, Michael T McManus, Terrence F Meehan, Christian von Mering, Daniel Muthas, Dac-Trung Nguyen, John P Overington, George Papadatos, Jun Qin, Christian Reich, Bryan L Roth, Stephan C Schürer, Anton Simeonov, Larry A Sklar, Noel Southall, Susumu Tomita, Ilinca Tudose, Oleg Ursu, Dušica Vidovic, Anna Waller, David Westergaard, Jeremy J Yang, and Gergely Zahoránszky-Köhalmi. Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.*, 17(5):377, May 2018.

[179] Kiran Mahajan and Nupam P Mahajan. Cross talk of tyrosine kinases with the DNA damage signaling pathways. *Nucleic Acids Res.*, 43(22):10588–10601, December 2015.

[180] Mei-Kuang Chen and Mien-Chie Hung. Regulation of therapeutic resistance in cancers by receptor tyrosine kinases. *Am. J. Cancer Res.*, 6(4):827–842, March 2016.

[181] D Ish-Shalom, C T Christoffersen, P Vorwerk, N Sacerdoti-Sierra, R M Shymko, D Naor, and P De Meyts. Mitogenic properties of insulin and insulin analogues mediated by the insulin receptor. *Diabetologia*, 40(2):S25–S31, June 1997.

[182] Nikolay Borisov, Edita Aksamitiene, Anatoly Kiyatkin, Stefan Legewie, Jan Berkhout, Thomas Maiwald, Nikolai P Kaimachnikov, Jens Timmer, Jan B Hoek, and Boris N Kholodenko. Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. *Mol. Syst. Biol.*, 5:256, April 2009.

[183] Georges Mairet-Coello, Anna Tury, and Emanuel DiCicco-Bloom. Insulin-like growth factor-1 promotes G(1)/S cell cycle progression through bidirectional regulation of cyclins and cyclin-dependent kinase inhibitors via the phosphatidylinositol 3-kinase/akt pathway in developing rat cerebral cortex. *J. Neurosci.*, 29(3):775–788, January 2009.

[184] Dac-Trung Nguyen, Stephen Mathias, Cristian Bologa, Soren Brunak, Nicolas Fernandez, Anna Gaulton, Anne Hersey, Jayme Holmes, Lars Juhl Jensen, Anneli Karlsson, Guixia Liu, Avi Ma'ayan, Geetha Mandava, Subramani Mani, Saurabh Mehta, John Overington, Juhee Patel, Andrew D Rouillard, Stephan Schürer, Timothy Sheils, Anton Simeonov, Larry A Sklar, Noel Southall, Oleg Ursu, Dusica Vidovic, Anna Waller, Jeremy Yang, Ajit Jadhav, Tudor I Oprea, and Rajarshi Guha. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, 45(D1):D995–D1002, January 2017.

[185] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin,

Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082, January 2018.

[186] Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.*, 15(3):R47, March 2014.

[187] Teresa Juan-Blanco, Miquel Duran-Frigola, and Patrick Aloy. Rationalizing drug response in cancer cell lines. *J. Mol. Biol.*, 430(18 Pt A):3016–3027, September 2018.

[188] Dana Ferranti, David Krane, and David Craft. The value of prior knowledge in machine learning of complex network systems. *Bioinformatics*, 33(22):3610–3618, November 2017.

[189] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6):417–425, December 2015.

[190] Johannes M Freudenberg, Vineet K Joshi, Zhen Hu, and Mario Medvedovic. CLEAN: CLustering enrichment ANalysis. *BMC Bioinformatics*, 10:234, July 2009.

[191] Astrid A Glück, Daniel M Aebersold, Yitzhak Zimmer, and Michaela Medová. Interplay between receptor tyrosine kinases and hypoxia signaling in cancer. *Int. J. Biochem. Cell Biol.*, 62:101–114, May 2015.

[192] Jean Paul Thiery. Epithelial-mesenchymal transitions in development and pathologies. *Curr. Opin. Cell Biol.*, 15(6):740–746, December 2003.

[193] O Lindblad, E Cordero, A Puissant, L Macaulay, A Ramos, N N Kabir, J Sun, J Vallon-Christersson, K Haraldsson, M T Hemann, Å Borg, F Levander, K Stegmaier, K Pietras, L Rönnstrand, and J U Kazi. Aberrant activation of the PI3K/mTOR pathway promotes resistance to sorafenib in AML. *Oncogene*, 35(39):5119–5131, September 2016.

[194] Ayako Nogami, Gaku Oshikawa, Keigo Okada, Shusaku Fukutake, Yoshihiro Umezawa, Toshikage Nagao, Tetsuya Kurosu, and Osamu Miura. FLT3-ITD confers resistance to the PI3K/Akt pathway inhibitors by protecting the mTOR/4EBP1/Mcl-1 pathway through STAT5 activation in acute myeloid leukemia. *Oncotarget*, 6(11):9189–9205, April 2015.

[195] Rizwan Haq, Jonathan Shoag, Pedro Andreu-Perez, Satoru Yokoyama, Hannah Edelman, Glenn C Rowe, Dennie T Frederick, Aeron D Hurley, Abhinav Nellore, Andrew L Kung, Jennifer A Wargo, Jun S Song, David E Fisher, Zolt Arany, and Hans R Widlund. Oncogenic BRAF regulates oxidative metabolism via PGC1α and MITF. *Cancer Cell*, 23(3):302–315, March 2013.

[196] Rizwan Haq, David E Fisher, and Hans R Widlund. Molecular pathways: BRAF induces bioenergetic adaptation by attenuating oxidative phosphorylation. *Clin. Cancer Res.*, 20(9):2257–2263, May 2014.

[197] Michael D Amatangelo, Shaun Goodyear, Devika Varma, and Mark E Stearns. c-myc expression and MEK1-induced erk2 nuclear localization are required for TGF-beta induced epithelial-mesenchymal transition and invasion in prostate cancer. *Carcinogenesis*, 33(10):1965–1975, October 2012.

[198] Francesco Marampon, Carmela Ciccarelli, and Bianca M Zani. Down-regulation of c-myc following MEK/ERK inhibition halts the expression of malignant phenotype in rhabdomyosarcoma and in non muscle-derived human tumors. *Mol. Cancer*, 5:31, August 2006.

[199] Ugo Moens, Sergiy Kostenko, and Baldur Sveinbjørnsson. The role of Mitogen-Activated protein Kinase-Activated protein kinases (MAPKAPKs) in inflammation. *Genes*, 4(2):101–133, March 2013.

[200] Yong-Yeon Cho, Zhiwei He, Yiguo Zhang, Hong Seok Choi, Feng Zhu, Bu Young Choi, Bong Seok Kang, Wei-Ya Ma, Ann M Bode, and Zigang Dong. The p53 protein is a novel substrate of ribosomal S6 kinase 2 and a critical intermediary for ribosomal S6 kinase 2 and histone H3 interaction. *Cancer Res.*, 65(9):3596–3603, May 2005.

[201] Dirk Schadendorf. Peroxisome proliferator-activating receptors: a new way to treat melanoma? *J. Invest. Dermatol.*, 129(5):1061–1063, May 2009.

[202] Michael G Borland, Ellen M Kehres, Christina Lee, Ashley L Wagner, Brooke E Shannon, Prajakta P Albrecht, Bokai Zhu, Frank J Gonzalez, and Jeffrey M Peters. Inhibition of tumorigenesis by peroxisome proliferator-activated receptor (PPAR)-dependent cell cycle blocks in human skin carcinoma cells. *Toxicology*, 404-405:25–32, July 2018.

[203] Christine M Ardito, Barbara M Grüner, Kenneth K Takeuchi, Clara Lubeseder-Martellato, Nicole Teichmann, Pawel K Mazur, Kathleen E Delgiorno, Eileen S Carpenter, Christopher J Halbrook, Jason C Hall, Debjani Pal, Thomas Briel, Alexander Herner, Marija Trajkovic-Arsic, Bence Sipos, Geou-Yarh Liou, Peter Storz, Nicole R Murray, David W Threadgill, Maria Sibilia, M Kay Washington, Carole L Wilson, Roland M Schmid, Elaine W Raines, Howard C Crawford, and Jens T Siveke. EGF receptor is required for KRAS-induced pancreatic tumorigenesis. *Cancer Cell*, 22(3):304–317, September 2012.

[204] Ching-Wei D Tzeng, Andrey Frolov, Natalya Frolova, Nirag C Jhala, J Harrison Howard, Donald J Buchsbaum, Selwyn M Vickers, Martin J Heslin, and J Pablo Arnoletti. Epidermal growth factor receptor (EGFR) is highly conserved in pancreatic cancer. *Surgery*, 141(4):464–469, April 2007.

[205] Yeri Lee, Jin-Ku Lee, Sun Hee Ahn, Jeongwu Lee, and Do-Hyun Nam. WNT signaling in glioblastoma and therapeutic opportunities. *Lab. Invest.*, 96(2):137–150, February 2016.

[206] Shailaja K Rao, Jennifer Edwards, Avadhut D Joshi, I-Mei Siu, and Gregory J Riggins. A survey of glioblastoma genomic amplifications and deletions. *J. Neurooncol.*, 96(2):169–179, January 2010.

[207] Fumi Higuchi, Alexandria L Fink, Juri Kiyokawa, Julie J Miller, Mara V A Koerner, Daniel P Cahill, and Hiroaki Wakimoto. PLK1 inhibition targets Myc-Activated malignant glioma cells irrespective of mismatch repair Deficiency-Mediated acquired resistance to temozolomide. *Mol. Cancer Ther.*, 17(12):2551–2563, December 2018.

[208] Elisavet Paplomata and Ruth O'Regan. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Ther. Adv. Med. Oncol.*, 6(4):154–166, July 2014.

[209] Zhen Ning Wee, Siti Maryam J M Yatim, Vera K Kohlbauer, Min Feng, Jian Yuan Goh, Yi Bao, Puay Leng Lee, Songjing Zhang, Pan Pan Wang, Elgene Lim, Wai Leong Tam, Yu Cai, Henrik J Ditzel, Dave S B Hoon, Ern Yu Tan, and Qiang Yu. IRAK1 is a therapeutic target that drives breast cancer metastasis and resistance to paclitaxel. *Nat. Commun.*, 6:8746, October 2015.

[210] Yassi Fallah, Janetta Brundage, Paul Allegakoen, and Ayesha N Shajahan-Haq. MYC-Driven pathways in breast cancer subtypes. *Biomolecules*, 7(3), July 2017.

[211] L Van Aelst, M Barr, S Marcus, A Polverino, and M Wigler. Complex formation between RAS and RAF and other protein kinases. *Proc. Natl. Acad. Sci. U. S. A.*, 90(13):6213–6217, July 1993.

[212] S A Moodie, B M Willumsen, M J Weber, and A Wolfman. Complexes of Ras.GTP with raf-1 and mitogen-activated protein kinase kinase. *Science*, 260(5114):1658–1661, June 1993.

[213] Timothy L Fitzgerald, Kvin Lertpiriyapong, Lucio Cocco, Alberto M Martelli, Massimo Libra, Saverio Candido, Giuseppe Montalto, Melchiorre Cervello, Linda Steelman, Stephen L Abrams, and James A McCubrey. Roles of EGFR and KRAS and their downstream signaling pathways in pancreatic cancer and pancreatic cancer stem cells. *Adv. Biol. Regul.*, 59:65–81, September 2015.

[214] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, Larsson Omberg, Denise M Wolf, Craig D Shriver, Vesteinn Thorsson, Cancer Genome Atlas Research Network, and Hai Hu. An integrated TCGA Pan-Cancer clinical data resource to drive High-Quality survival outcome analytics. *Cell*, 173(2):400–416.e11, April 2018.

[215] Samira Jaeger, Miquel Duran-Frigola, and Patrick Aloy. Drug sensitivity in cancer cell lines is not tissue-specific. *Mol. Cancer*, 14:40, February 2015.

[216] Mathias J Wawer, Kejie Li, Sigrun M Gustafsdottir, Vebjorn Ljosa, Nicole E Bodycombe, Melissa A Marton, Katherine L Sokolnicki, Mark-Anthony Bray, Melissa M Kemp, Ellen Winchester, Bradley Taylor, George B Grant, C Suk-Yee Hon, Jeremy R Duvall, J Anthony Wilson, Joshua A Bittker, Vlado Dančík, Rajiv Narayan, Aravind Subramanian, Wendy Winckler, Todd R Golub, Anne E Carpenter, Alykhan F Shamji, Stuart L Schreiber, and Paul A Clemons. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U. S. A.*, 111(30):10911–10916, July 2014.

[217] Idan Frumkin, Dvir Schirman, Aviv Rotman, Fangfei Li, Liron Zahavi, Ernest Mordret, Omer Asraf, Song Wu, Sasha F Levy, and Yitzhak Pilpel. Gene architectures that minimize cost of gene expression. *Mol. Cell*, 65(1):142–153, January 2017.

[218] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[219] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–72, January 2006.

[220] Linton C Freeman. Centrality in social networks conceptual clarification, 1978.

[221] Lidia Mateo, Oriol Guitart-Pla, Miquel Duran-Frigola, and Patrick Aloy. Exploring the OncoGenomic landscape of cancer. *Genome Med.*, 10(1):61, August 2018.

[222] Monya Baker. Big biology: The 'omes puzzle. *Nature*, 494(7438):416–419, February 2013.

[223] Avi Ma'ayan, Andrew D Rouillard, Neil R Clark, Zichen Wang, Qiaonan Duan, and Yan Kou. Lean big data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.*, 35(9):450–460, September 2014.

[224] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.*, 16(6):1069–1080, November 2015.

[225] Kentaro Kawata, Atsushi Hatano, Katsuyuki Yugi, Hiroyuki Kubota, Takanori Sano, Masashi Fujii, Yoko Tomizawa, Toshiya Kokaji, Kaori Y Tanaka, Shinsuke Uda, Yutaka Suzuki, Masaki Matsumoto, Keiichi I Nakayama, Kaori Saitoh, Keiko Kato, Ayano Ueno, Maki Ohishi, Akiyoshi Hirayama, Tomoyoshi Soga, and Shinya Kuroda. Trans-omic analysis reveals selective responses to induced and basal insulin across signaling, transcriptional, and metabolic networks. *iScience*, 7:212–229, September 2018.

[226] Burcu Vitrinel, Hiromi W L Koh, Funda Mujgan Kar, Shuvadeep Maity, Justin Rendleman, Hyungwon Choi, and Christine Vogel. Exploiting interdata relationships in next-generation proteomics analysis. *Mol. Cell. Proteomics*, 18(8 suppl 1):S5–S14, August 2019.

[227] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, 14(6):e8124, June 2018.

[228] Daniel S Himmelstein and Sergio E Baranzini. Heterogeneous network edge prediction: A data integration approach to prioritize Disease-Associated genes. *PLoS Comput. Biol.*, 11(7):e1004259, July 2015.

[229] Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, July 2017.

[230] Camilo Ruiz, Marinka Zitnik, and Jure Leskovec. Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.*, 12(1):1796, March 2021.

[231] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.*, 12(1):124, January 2021.

[232] Adrià Fernández-Torras, Arnau Comajuncosa-Creus, Miquel Duran-Frigola, and Patrick Aloy. Connecting chemistry and biology through molecular descriptors. *Curr. Opin. Chem. Biol.*, 66:102090, February 2022.

[233] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(Database issue):D991–5, January 2013.

[234] Saee Paliwal, Alex de Giorgio, Daniel Neil, Jean-Baptiste Michel, and Alix Mb Lacoste. Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Sci. Rep.*, 10(1):18250, October 2020.

[235] David Geleta, Andriy Nikolov, Gavin Edwards, Anna Gogleva, Richard Jackson, Erik Jansson, Andrej Lamov, Sebastian Nilsson, Marina Pettersson, Vladimir Poroshin, Benedek Rozemberczki, Timothy Scrivener, Michael Ughetto, and Eliseo Papa. Biological insights knowledge graph: an integrated knowledge graph to support drug development. November 2021.

[236] Daniel N Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, and Russ B Altman. A Literature-Based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In *Biocomputing 2020*, pages 463–474. WORLD SCIENTIFIC, November 2019.

[237] Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L

Hamilton. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. February 2021.

[238] Denise Carvalho-Silva, Andrea Pierleoni, Miguel Pignatelli, Chuangkee Ong, Luca Fumis, Nikiforos Karamanis, Miguel Carmona, Adam Faulconbridge, Andrew Hercules, Elaine McAuley, Alfredo Miranda, Gareth Peat, Michaela Spitzer, Jeffrey Barrett, David G Hulcoop, Eliseo Papa, Gautier Koscielny, and Ian Dunham. Open targets platform: new developments and updates two years on. *Nucleic Acids Res.*, 47(D1):D1056–D1065, January 2019.

[239] Wytze J Vlietstra, Rein Vos, Anneke M Sijbers, Erik M van Mulligen, and Jan A Kors. Using predicate and provenance information from a knowledge graph for drug efficacy screening. *J. Biomed. Semantics*, 9(1):23, September 2018.

[240] Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher C Mader, Aravind Subramanian, and Todd R Golub. The drug repurposing hub: a next-generation drug library and information resource. *Nat. Med.*, 23(4):405–408, April 2017.

[241] Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-Alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res.*, 46(D1):D994–D1002, January 2018.

[242] Murat Iskar, Monica Campillos, Michael Kuhn, Lars Juhl Jensen, Vera van Noort, and Peer Bork. Drug-induced regulation of target expression. *PLoS Comput. Biol.*, 6(9), September 2010.

[243] Guangsheng Wu, Juan Liu, and Xiang Yue. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinformatics*, 20(Suppl 3):134, March 2019.

[244] F Kose, N E Kocer, A T Sumbul, A Sezer, and O Yilkan. Kaposi's sarcoma following chronic lymphocytic leukemia: A rare entity. *Case Rep. Oncol.*, 5(2):271–274, May 2012.

[245] Anuradha Avinash Belur, Arun Kumar Arumugam Raajasekar, Srikant Nannapaneni, and Thandavababu Chelliah. A case of kaposi's sarcoma in a HIV negative patient with CLL treated with rituximab. *Blood*, 124(21):4970–4970, December 2014.

[246] Damir Vučinić, Andrea Dekanić, Gordana Zamolo, Margita Belušić-Gobić, Ingrid Belac-Lovasić, and Tanja Batinac. Kaposi's sarcoma in an HIV-negative chronic lymphocytic leukemia patient without immunosuppressive therapy: A case report. *SAGE Open Med Case Rep*, 6:2050313X18799239, September 2018.

[247] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, 13(12):966–967, November 2016.

[248] Erdogan Taskesen and Marcel J T Reinders. 2D representation of transcriptomes by t-SNE exposes relatedness between human tissues. *PLoS One*, 11(2):e0149853, February 2016.

[249] Douglas M Hawkins. The problem of overfitting. *J. Chem. Inf. Comput. Sci.*, 44(1):1–12, January 2004.

[250] Adrià Fernández-Torras, Miquel Duran-Frigola, and Patrick Aloy. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med.*, 11(1):17, March 2019.

[251] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, 3rd, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, Kevin Hu, Alexander Y Andreev-Drakhlin, Jaegil Kim, Julian M Hess, Brian J Haas, François Aguet, Barbara A Weir, Michael V Rothberg, Brenton R Paolella, Michael S Lawrence, Rehan Akbani, Yiling Lu, Hong L Tiv, Prafulla C Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph

Lehar, Joshua M Korn, Dale A Porter, Michael D Jones, Javad Golji, Giordano Caponigro, Jordan E Taylor, Caitlin M Dunning, Amanda L Creech, Allison C Warren, James M McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E Maruvka, Andrew D Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D Jaffe, Andrew A Lane, David M Weinstock, Cory M Johannessen, Michael P Morrissey, Frank Stegmeier, Robert Schlegel, William C Hahn, Gad Getz, Gordon B Mills, Jesse S Boehm, Todd R Golub, Levi A Garraway, and William R Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, May 2019.

[252] Michael Costanzo, Jing Hou, Vincent Messier, Justin Nelson, Mahfuzur Rahman, Benjamin VanderSluis, Wen Wang, Carles Pons, Catherine Ross, Matej Ušaj, Bryan-Joseph San Luis, Emira Shuteriqi, Elizabeth N Koch, Patrick Aloy, Chad L Myers, Charles Boone, and Brenda Andrews. Environmental robustness of the global yeast genetic interaction network. *Science*, 372(6542), May 2021.

[253] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E Castel, Andrew R Hamel, Ana Viñuela, Amy L Roberts, Serghei Mangul, Xiaoquan Wen, Gao Wang, Alvaro N Barbeira, Diego Garrido-Martín, Brian B Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew Brown, Angel Martinez-Perez, José Manuel Soria, GTEx Consortium, Gad Getz, Emmanouil T Dermitzakis, Kerrin S Small, Matthew Stephens, Hualin S Xi, Hae Kyung Im, Roderic Guigó, Ayellet V Segrè, Barbara E Stranger, Kristin G Ardlie, and Tuuli Lappalainen. Cell type-specific genetic regulation of gene expression across human tissues. *Science*, 369(6509), September 2020.

[254] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*, 2(1):56–67, January 2020.

[255] Sehyoun Yoon, Nicolas H Piguel, Natalia Khalatyan, Leonardo E Dionisio, Jeffrey N Savas, and Peter Penzes. Homer1 promotes dendritic spine growth through ankyrin-g and its loss reshapes the synaptic proteome. *Mol. Psychiatry*, 26(6):1775–1789, June 2021.

[256] Sergey V Paushkin, Meenal Patel, Bansri S Furia, Stuart W Peltz, and Christopher R Trotta. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation. *Cell*, 117(3):311–321, April 2004.

[257] Hachung Chung, Jorg J A Calis, Xianfang Wu, Tony Sun, Yingpu Yu, Stephanie L Sarbanes, Viet Loan Dao Thi, Abigail R Shilvock, H-Heinrich Hoffmann, Brad R Rosenberg, and Charles M Rice. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. *Cell*, 172(4):811–824.e14, February 2018.

[258] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, Alan Norris, Philippe Sanseau, David Cavalla, and Munir Pirmohamed. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, 18(1):41–58, January 2019.

[259] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *J. Med. Chem.*, 63(16):8683–8694, August 2020.

[260] Jianzhu Ma, Samson H Fong, Yunan Luo, Christopher J Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F A Wessels, Marc Hafner, Roded Sharan, Jian Peng, and Trey Ideker. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer*, 2(2):233–244, February 2021.

[261] Zhuo Fu, Lina Zhao, Weidong Chai, Zhenhua Dong, Wenhong Cao, and Zhenqi Liu. Ranolazine recruits muscle microvasculature and enhances insulin action in rats. *J. Physiol.*, 591(20):5235–5249, October 2013.

[262] Robert H Eckel, Robert R Henry, Patrick Yue, Arvinder Dhalla, Pamela Wong, Philip Jochelson, Luiz Belardinelli, and Jay S Skyler.

Effect of ranolazine monotherapy on glycemic control in subjects with type 2 diabetes. *Diabetes Care*, 38(7):1189–1196, July 2015.

[263] Nishant R Shah, Michael K Cheezum, Vikas Veeranna, Stephen J Horgan, Viviany R Taqueti, Venkatesh L Murthy, Courtney Foster, Jon Hainer, Karla M Daniels, Jose Rivero, Amil M Shah, Peter H Stone, David A Morrow, Michael L Steigner, Sharmila Dorbala, Ron Blankstein, and Marcelo F Di Carli. Ranolazine in symptomatic diabetic patients without obstructive coronary artery disease: Impact on microvascular and diastolic function. *J. Am. Heart Assoc.*, 6(5), May 2017.

[264] Sofia K Forslund, Rima Chakaroun, Maria Zimmermann-Kogadeeva, Lajos Markó, Judith Aron-Wisnewsky, Trine Nielsen, Lucas Moitinho-Silva, Thomas S B Schmidt, Gwen Falony, Sara Vieira-Silva, Solia Adriouch, Renato J Alves, Karen Assmann, Jean-Philippe Bastard, Till Birkner, Robert Caesar, Julien Chilloux, Luis Pedro Coelho, Leopold Fezeu, Nathalie Galleron, Gerard Helft, Richard Isnard, Boyang Ji, Michael Kuhn, Emmanuelle Le Chatelier, Antonis Myridakis, Lisa Olsson, Nicolas Pons, Edi Prifti, Benoit Quinquis, Hugo Roume, Joe-Elie Salem, Nataliya Sokolovska, Valentina Tremaroli, Mireia Valles-Colomer, Christian Lewinter, Nadja B Søndertoft, Helle Krogh Pedersen, Tue H Hansen, MetaCardis Consortium*, Jens Peter Gøtze, Lars Køber, Henrik Vestergaard, Torben Hansen, Jean-Daniel Zucker, Serge Hercberg, Jean-Michel Oppert, Ivica Letunic, Jens Nielsen, Fredrik Bäckhed, S Dusko Ehrlich, Marc-Emmanuel Dumas, Jeroen Raes, Oluf Pedersen, Karine Clément, Michael Stumvoll, and Peer Bork. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature*, 600(7889):500–505, December 2021.

[265] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, 44(D1):D1214–9, January 2016.

[266] Amos Bairoch. The cellosaurus, a Cell-Line knowledge resource. *J. Biomol. Tech.*, 29(2):25–38, July 2018.

[267] The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47(D1):D330–D338, January 2019.

[268] Susan M Bello, Mary Shimoyama, Elvira Mitraka, Stanley J F Laulederkind, Cynthia L Smith, Janan T Eppig, and Lynn M Schriml. Disease ontology: improving and unifying disease annotations across species. *Dis. Model. Mech.*, 11(3), March 2018.

[269] Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, 49(D1):D344–D354, January 2021.

[270] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, January 2019.

[271] Daniel J Rigden and Xosé M Fernández. The 27th annual nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res.*, 48(D1):D1–D8, January 2020.

[272] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, 39(Database issue):D507–13, January 2011.

[273] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pages 1089–1090, NLD, August 2004. IOS Press.

[274] Daniel Domingo-Fernández, Charles Tapley Hoyt, Carlos Bobis-Álvarez, Josep Marín-Llaó, and Martin Hofmann-Apitius. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst Biol Appl*, 5:3, 2019.

[275] Frida Belinky, Noam Nativ, Gil Stelzer, Shahar Zimmerman, Tsippi Iny Stein, Marilyn Safran, and Doron Lancet. PathCards: multi-source consolidation of human biological pathways. *Database*, 2015, February 2015.

[276] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena M A Gendoo, Patrick Grossmann, Andrew H Beck, Hugo J W L Aerts, Mathieu Lupien, Anna Goldenberg, and Benjamin Haibe-Kains. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, April 2016.

[277] Yue Qin, Edward L Huttlin, Casper F Winsnes, Maya L Gosztyla, Ludivine Wacheul, Marcus R Kelly, Steven M Blue, Fan Zheng, Michael Chen, Leah V Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J Lee, Wei Ouyang, Sophie N Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F Kreisberg, Steven P Gygi, Jianzhu Ma, J Wade Harper, Gene W Yeo, Denis L J Lafontaine, Emma Lundberg, and Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature*, 600(7889):536–542, December 2021.

[278] Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.*, 10(1):5415, November 2019.

[279] Miquel Duran-Frigola, Eduardo Pauls, Oriol Guitart-Pla, Martino Bertoni, Víctor Alcalde, David Amat, Teresa Juan-Blanco, and Patrick Aloy. Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nat. Biotechnol.*, 38(9):1087–1096, September 2020.

[280] James Bergstra, Dan Yamins, and David Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*. SciPy, 2013.

[281] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, July 2021.

[282] Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for R and python. *J. Stat. Softw.*, 53:1–18, May 2013.

[283] G A Evans. Designer science and the "omic" revolution. *Nat. Biotechnol.*, 18(2):127, February 2000.

[284] Takeshi Obayashi, Shinpei Hayashi, Masayuki Shibaoka, Motoshi Saeki, Hiroyuki Ohta, and Kengo Kinoshita. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, 36(Database issue):D77–82, January 2008.

[285] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database issue):D901–6, January 2008.

[286] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35(Database issue):D198–201, January 2007.

[287] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database issue):D1100–7, January 2012.

[288] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and

Laura I Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028, April 2015.

[289] Mulin Jun Li, Panwen Wang, Xiaorong Liu, Ee Lyn Lim, Zhangyong Wang, Meredith Yeager, Maria P Wong, Pak Chung Sham, Stephen J Chanock, and Junwen Wang. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, 40(Database issue):D1047–54, January 2012.

[290] Adrià Fernández-Torras, Miquel Duran-Frigola, Martino Bertoni, Martina Locatelli, and Patrick Aloy. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nat. Commun.*, 13(1):1–18, September 2022.

[291] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. February 2017.

[292] A Gobbi, F Iorio, D Albanese, G Jurman, and J Saez-Rodriguez. BiRewire: High-performing routines for the randomization of a bipartite graph (or a binary event matrix), undirected and directed signed graph preserving degree distribution (or marginal totals), 2022.

[293] Torsten Hoffmann and Marcus Gastreich. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov. Today*, 24(5):1148–1156, May 2019.

[294] Teague Sterling and John J Irwin. ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Model.*, 55(11):2324–2337, November 2015.

[295] Oleksandr O Grygorenko, Dmytro S Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E Gubina, and Yurii S Moroz. Generating multibillion chemical space of readily accessible screening compounds. *iScience*, 23(11):101681, November 2020.

[296] Jean-Louis Reymond. The chemical space project. *Acc. Chem. Res.*, 48(3):722–730, March 2015.

[297] Steven M Corsello, Rohith T Nagari, Ryan D Spangler, Jordan Rossen, Mustafa Kocak, Jordan G Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A Tang, Vickie M Wang, Samantha A Bender, Evan Lemire, Rajiv Narayan, Philip Montgomery, Uri Ben-David, Colin W Garvie, Yejia Chen, Matthew G Rees, Nicholas J Lyons, James M McFarland, Bang T Wong, Li Wang, Nancy Dumont, Patrick J O'Hearn, Eric Stefan, John G Doench, Caitlin N Harrington, Heidi Greulich, Matthew Meyerson, Francisca Vazquez, Aravind Subramanian, Jennifer A Roth, Joshua A Bittker, Jesse S Boehm, Christopher C Mader, Aviad Tsherniak, and Todd R Golub. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer*, 1(2):235–248, February 2020.

[298] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P Overington, George Papadatos, Ines Smit, and Andrew R Leach. The ChEMBL database in 2017. *Nucleic Acids Res.*, 45(D1):D945–D954, January 2017.

[299] Yanli Wang, Stephen H Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A Shoemaker, Paul A Thiessen, Siqian He, and Jian Zhang. PubChem BioAssay: 2017 update. *Nucleic Acids Res.*, 45(D1):D955–D963, January 2017.

[300] Colm J Ryan, Peter Cimermančič, Zachary A Szpiech, Andrej Sali, Ryan D Hernandez, and Nevan J Krogan. High-resolution network biology: connecting sequence with function. *Nat. Rev. Genet.*, 14(12):865–879, December 2013.

[301] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules, 1988.

[302] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, November 2002.

[303] Oleg Devinyak, Dmytro Havrylyuk, and Roman Lesyk. 3D-MoRSE descriptors explained, 2014.

[304] M Pastor, G Cruciani, I McLay, S Pickett, and S Clementi. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.*, 43(17):3233–3243, August 2000.

[305] Sereina Riniker. Molecular dynamics fingerprints (MDFP): Machine learning from MD data to predict Free-Energy differences. *J. Chem. Inf. Model.*, 57(4):726–741, April 2017.

[306] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, May 2010.

[307] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.*, 12(1):56, September 2020.

[308] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical Graph-to-Graph translation for molecules. *arXiv*, June 2019.

[309] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, 9(2):513–530, January 2018.

[310] L M Kauvar, D L Higgins, H O Villar, J R Sportsman, A Engqvist-Goldstein, R Bukar, K E Bauer, H Dilley, and D M Rocke. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.*, 2(2):107–118, February 1995.

[311] Gaia V Paolini, Richard H B Shapland, Willem P van Hoorn, Jonathan S Mason, and Andrew L Hopkins. Global mapping of pharmacological space. *Nat. Biotechnol.*, 24(7):805–815, July 2006.

[312] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, Ryan Whaley, Richard A Glennon, Jérôme Hert, Kelan L H Thomas, Douglas D Edwards, Brian K Shoichet, and Bryan L Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, November 2009.

[313] Marnie L MacDonald, Jane Lamerdin, Stephen Owens, Brigitte H Keon, Graham K Bilter, Zhidi Shang, Zhengping Huang, Helen Yu, Jennifer Dias, Tomoe Minami, Stephen W Michnick, and John K Westwick. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.*, 2(6):329–337, June 2006.

[314] Daniel W Young, Andreas Bender, Jonathan Hoyt, Elizabeth McWhinnie, Gung-Wei Chirn, Charles Y Tao, John A Tallarico, Mark Labow, Jeremy L Jenkins, Timothy J Mitchison, and Yan Feng. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, 4(1):59–68, January 2008.

[315] Ryusuke Sawada, Michio Iwata, Yasuo Tabei, Haruka Yamato, and Yoshihiro Yamanishi. Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. *Sci. Rep.*, 8(1):156, January 2018.

[316] Susan L Holbeck, Jerry M Collins, and James H Doroshow. Analysis of food and drug administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol. Cancer Ther.*, 9(5):1451–1460, May 2010.

[317] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, July 2008.

[318] Martino Bertoni, Miquel Duran-Frigola, Pau Badia-I-Mompel, Eduardo Pauls, Modesto Orozco-Ruiz, Oriol Guitart-Pla, Víctor Alcalde, Víctor M Diaz, Antoni Berenguer-Llergo, Isabelle Brun-Heath, Núria Villegas, Antonio García de Herreros, and Patrick Aloy. Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.*, 12(1):3932, June 2021.

[319] Anne Mai Wassermann, Eugen Lounkine, John W Davies, Meir Glick, and L Miguel Camargo. The opportunities of mining historical and collective data in drug discovery. *Drug Discov. Today*, 20(4):422–434, April 2015.

[320] S Hellberg, M Sjöström, B Skagerberg, and S Wold. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.*, 30(7):1126–1135, July 1987.

[321] Nan Xiao, Dong-Sheng Cao, Min-Feng Zhu, and Qing-Song Xu. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31(11):1857–1859, June 2015.

[322] Yuting Xu, Deeptak Verma, Robert P Sheridan, Andy Liaw, Junshui Ma, Nicholas M Marshall, John McIntosh, Edward C Sherer, Vladimir Svetnik, and Jennifer M Johnston. Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model.*, 60(6):2773–2790, June 2020.

[323] Maxat Kulmanov and Robert Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, July 2019.

[324] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, 16(12):1315–1322, December 2019.

[325] Ehsaneddin Asgari and Mohammad R K Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, 10(11):e0141287, November 2015.

[326] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, December 2019.

[327] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.*, 32:9689–9701, December 2019.

[328] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and

Burkhard Rost. ProtTrans: Towards cracking the language of lifes code through Self-Supervised deep learning and high performance computing. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, July 2021.

[329] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv*, February 2019.

[330] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Syst*, 12(6):654–669.e3, June 2021.

[331] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, July 2021.

[332] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15), April 2021.

[333] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[334] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[335] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology meets biology: Interpreting attention in protein language models. *arXiv*, June 2020.

[336] V Joachim Haupt, Simone Daminelli, and Michael Schroeder. Drug promiscuity in PDB: Protein binding site similarity is key. *PLoS One*, 8(6):e65894, June 2013.

[337] Nathanaël Weill and Didier Rognan. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.*, 50(1):123–135, January 2010.

[338] Lydia Siragusa, Simon Cross, Massimo Baroni, Laura Goracci, and Gabriele Cruciani. BioGPS: navigating biological space to predict polypharmacology, off-targeting, and selectivity. *Proteins*, 83(3):517–532, March 2015.

[339] Alexander Stark, Shamil Sunyaev, and Robert B Russell. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326(5):1307–1316, March 2003.

[340] Miquel Duran-Frigola, Lydia Siragusa, Eytan Ruppin, Xavier Barril, Gabriele Cruciani, and Patrick Aloy. Detecting similar binding pockets to enable systems polypharmacology. *PLoS Comput. Biol.*, 13(6):e1005522, June 2017.

[341] Rajan Chaudhari, Long Wolf Fong, Zhi Tan, Beibei Huang, and Shuxing Zhang. An up-to-date overview of computational polypharmacology in modern drug discovery. *Expert Opin. Drug Discov.*, 15(9):1025–1044, September 2020.

[342] Rajiv Gandhi Govindaraj and Michal Brylinski. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinformatics*, 19(1):91, March 2018.

[343] Christiane Ehrt, Tobias Brinkjost, and Oliver Koch. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Comput. Biol.*, 14(11):e1006483, November 2018.

[344] Christiane Ehrt, Tobias Brinkjost, and Oliver Koch. Binding site characterization - similarity, promiscuity, and druggability. *Medchemcomm*, 10(7):1145–1159, July 2019.

[345] P Gainza, F Sverrisson, F Monti, E Rodolà, D Boscaini, M M Bronstein, and B E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods*, 17(2):184–192, February 2020.

[346] Martin Simonovsky and Joshua Meyers. DeeplyTough: Learning structural comparison of protein binding sites. *J. Chem. Inf. Model.*, 60(4):2356–2366, April 2020.

[347] Brandon J Bongers, Adriaan P IJzerman, and Gerard J P Van Westen. Proteochemometrics - recent developments in bioactivity and selectivity modeling. *Drug Discov. Today Technol.*, 32-33:89–98, December 2019.

[348] Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting Drug-Target interactions. *J. Chem. Inf. Model.*, 59(10):4131–4149, October 2019.

[349] Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J Dickson, Jose S Duca, Viktor Hornak, David R Koes, and Tom Kurtzman. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One*, 14(8):e0220113, August 2019.

[350] D L van der Velden, L R Hoes, H van der Wijngaart, J M van Berge Henegouwen, E van Werkhoven, P Roepman, R L Schilsky, W W J de Leng, A D R Huitema, B Nuijen, P M Nederlof, C M L van Herpen, D J A de Groot, L A Devriese, A Hoeben, M J A de Jonge, M Chalabi, E F Smit, A J de Langen, N Mehra, M Labots, E Kapiteijn, S Sleijfer, E Cuppen, H M W Verheul, H Gelderblom, and E E Voest. The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs. *Nature*, 574(7776):127–131, October 2019.

[351] Daniele Parisi, Melissa F Adasme, Anastasia Sveshnikova, Sarah Naomi Bolz, Yves Moreau, and Michael Schroeder. Drug repositioning or target repositioning: A structural perspective of

drug-target-indication relationship for available repurposed drugs. *Comput. Struct. Biotechnol. J.*, 18:1043–1055, April 2020.

[352] Pierre Bady, Sylvain Dolédec, Bernard Dumont, and Jean-François Fruget. Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *C. R. Biol.*, 327(1):29–36, January 2004.

[353] Xiaoshi Zhong, Rama Kaalia, and Jagath C Rajapakse. GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics*, 20(Suppl 9):918, December 2019.

[354] Sheng Wang, Emily R Flynn, and Russ B Altman. Gaussian embedding for large-scale gene set analysis. *Nat Mach Intell*, 2(7):387–395, July 2020.

[355] Philipp J Schubert, Sven Dorkenwald, Michał Januszewski, Viren Jain, and Joergen Kornfeld. Learning cellular morphology with neural networks. *Nat. Commun.*, 10(1):2736, June 2019.

[356] Eugene F Douglass, Jr, Robert J Allaway, Bence Szalai, Wenyu Wang, Tingzhong Tian, Adrià Fernandez-Torras, Ron Realubit, Charles Karan, Shuyu Zheng, Alberto Pessia, Ziaurrehman Tanoli, Mohieddin Jafari, Fangping Wan, Shuya Li, Yuanpeng Xiong, Miquel Duran-Frigola, Martino Bertoni, Pau Badia-I-Mompel, Lidia Mateo, Oriol Guitart-Pla, Verena Chung, DREAM CTD-squared Pancancer Drug Activity Challenge Consortium, Jing Tang, Jianyang Zeng, Patrick Aloy, Julio Saez-Rodriguez, Justin Guinney, Daniela S Gerhard, and Andrea Califano. A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. *Cell Rep Med*, 3(1):100492, January 2022.

[357] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform.*, 23(2), March 2022.

[358] U.S. Food and Drug Administration. FDA at a glance. https://www.fda.gov/about-fda/fda-basics/fact-sheet-fda-glance.

[359] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, 6(10):813–823, October 2006.

[360] Peter M Haverty, Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M Neve, Scott Martin, Jeff Settleman, Robert L Yauch, and Richard Bourgon. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533(7603):333–337, May 2016.

[361] Samuel W Brady, Alexander M Gout, and Jinghui Zhang. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. *Trends Genet.*, 38(2):194–208, February 2022.

[362] Christopher E Lietz, Erik T Newman, Andrew D Kelly, David H Xiang, Ziying Zhang, Caroline A Luscko, Santiago A Lozano-Calderon, David H Ebb, Kevin A Raskin, Gregory M Cote, Edwin Choy, G Petur Nielsen, Benjamin Haibe-Kains, Martin J Aryee, and Dimitrios Spentzos. Genome-wide DNA methylation patterns reveal clinically relevant predictive and prognostic subtypes in human osteosarcoma. *Commun Biol*, 5(1):213, March 2022.

[363] Martin Frejno, Chen Meng, Benjamin Ruprecht, Thomas Oellerich, Sebastian Scheich, Karin Kleigrewe, Enken Drecoll, Patroklos Samaras, Alexander Hogrebe, Dominic Helm, Julia Mergner, Jana Zecha, Stephanie Heinzlmeir, Mathias Wilhelm, Julia Dorn, Hans-Michael Kvasnicka, Hubert Serve, Wilko Weichert, and Bernhard Kuster. Proteome activity landscapes of tumor cell lines determine drug responses. *Nat. Commun.*, 11(1):3639, July 2020.

[364] Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Suppl_2):i911–i918, December 2020.

[365] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, July 2019.

[366] P Erridge. The pareto principle. *Br. Dent. J.*, 201(7):419, October 2006.

[367] Karl Weiss, Taghi M Khoshgoftaar, and Dingding Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):1–40, May 2016.

[368] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. May 2017.

[369] Amy Maxmen. One million coronavirus sequences: popular genome site hits mega milestone. http://dx.doi.org/10.1038/d41586-021-01069-w, April 2021.

[370] Yuan Huang, Chan Yang, Xin-Feng Xu, Wei Xu, and Shu-Wen Liu. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.*, 41(9):1141–1149, August 2020.

[371] Daniel Blanco-Melo, Benjamin E Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X Jordan, Kohei Oishi, Maryline Panis, David Sachs, Taia T Wang, Robert E Schwartz, Jean K Lim, Randy A Albrecht, and Benjamin R tenOever. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, 181(5):1036–1045.e9, May 2020.

[372] Mehdi Bouhaddou, Danish Memon, Bjoern Meyer, Kris M White, Veronica V Rezelj, Miguel Correa Marrero, Benjamin J Polacco, James E Melnyk, Svenja Ulferts, Robyn M Kaake, Jyoti Batra, Alicia L Richards, Erica Stevenson, David E Gordon, Ajda Rojc, Kirsten Obernier, Jacqueline M Fabius, Margaret Soucheray, Lisa Miorin, Elena Moreno, Cassandra Koh, Quang Dinh Tran, Alexandra Hardy, Rémy Robinot, Thomas Vallet, Benjamin E Nilsson-Payant, Claudia Hernandez-Armenta, Alistair Dunham, Sebastian Weigang, Julian Knerr, Maya Modak, Diego Quintero, Yuan Zhou, Aurelien Dugourd, Alberto Valdeolivas, Trupti Patil, Qiongyu Li, Ruth Hüttenhain, Merve Cakir, Monita Muralidharan, Minkyu Kim, Gwendolyn Jang, Beril Tutuncuoglu, Joseph Hiatt, Jeffrey Z Guo, Jiewei Xu, Sophia Bouhaddou, Christopher J P Mathy, Anna Gaulton, Emma J Manners, Eloy Félix, Ying Shi, Marisa Goff, Jean K Lim, Timothy McBride, Michael C O'Neal, Yiming Cai, Jason C J Chang, David J Broadhurst, Saker Klippsten, Emmie De Wit, Andrew R Leach, Tanja Kortemme, Brian Shoichet, Melanie Ott, Julio Saez-Rodriguez, Benjamin R tenOever, R Dyche Mullins, Elizabeth R Fischer, Georg Kochs, Robert

Grosse, Adolfo García-Sastre, Marco Vignuzzi, Jeffery R Johnson, Kevan M Shokat, Danielle L Swaney, Pedro Beltrao, and Nevan J Krogan. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell*, 182(3):685–712.e19, August 2020.

[373] Junmei Wang. Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. *J. Chem. Inf. Model.*, 60(6):3277–3286, June 2020.

[374] Miquel Duran-Frigola, Martino Bertoni, Roi Blanco, Víctor Martínez, Eduardo Pauls, Víctor Alcalde, Gemma Turon, Núria Villegas, Adrià Fernández-Torras, Carles Pons, Lídia Mateo, Oriol Guitart-Pla, Pau Badia-i Mompel, Aleix Gimeno, Nicolas Soler, Isabelle Brun-Heath, Hugo Zaragoza, and Patrick Aloy. Bioactivity profile similarities to expand the repertoire of COVID-19 drugs. *J. Chem. Inf. Model.*, 60(12):5730–5734, December 2020.

[375] Tara John and Isabelle Jani Friend. The unused "miracle". *CNN*, September 2021.

[376] Robin McKie. The vaccine miracle: how scientists waged the battle against covid-19. *The Guardian*, December 2020.

[377] Alice Park and Jamie Ducharme. Vaccine scientists are TIME's 2021 heroes of the year. *Time*, December 2021.

[378] Ewen Callaway. DeepMind's AI for protein structure is coming to the masses. http://dx.doi.org/10.1038/d41586-021-01968-y, July 2021.

[379] Ewen Callaway. What's next for AlphaFold and the AI protein-folding revolution. http://dx.doi.org/10.1038/d41586-022-00997-5, April 2022.

[380] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nat. Methods*, 19(6):679–682, June 2022.

[381] Mehmet Akdel, Douglas E V Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, Aditi Shenoy,

Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H M Rodrigues, Alistair S Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Kresten Lindorff-Larsen, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B Ascher, Janet M Thornton, Norman E Davey, Amelie Stein, Arne Elofsson, Tristan I Croll, and Pedro Beltrao. A structural biology community assessment of AlphaFold 2 applications. September 2021.

[382] Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, Frank DiMaio, Lauren Carter, Cameron M Chow, Gaetano T Montelione, and David Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, December 2021.

[383] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. February 2021.

[384] Miquel Duran-Frigola and Gemma Turon. Ersilia open source initiative. https://www.ersilia.io/.

[385] Yoson Park and Casey S Greene. A parasite's perspective on data sharing. *Gigascience*, 7(11), November 2018.

A

Supplementary Figures

A.1 Chapter 3.1

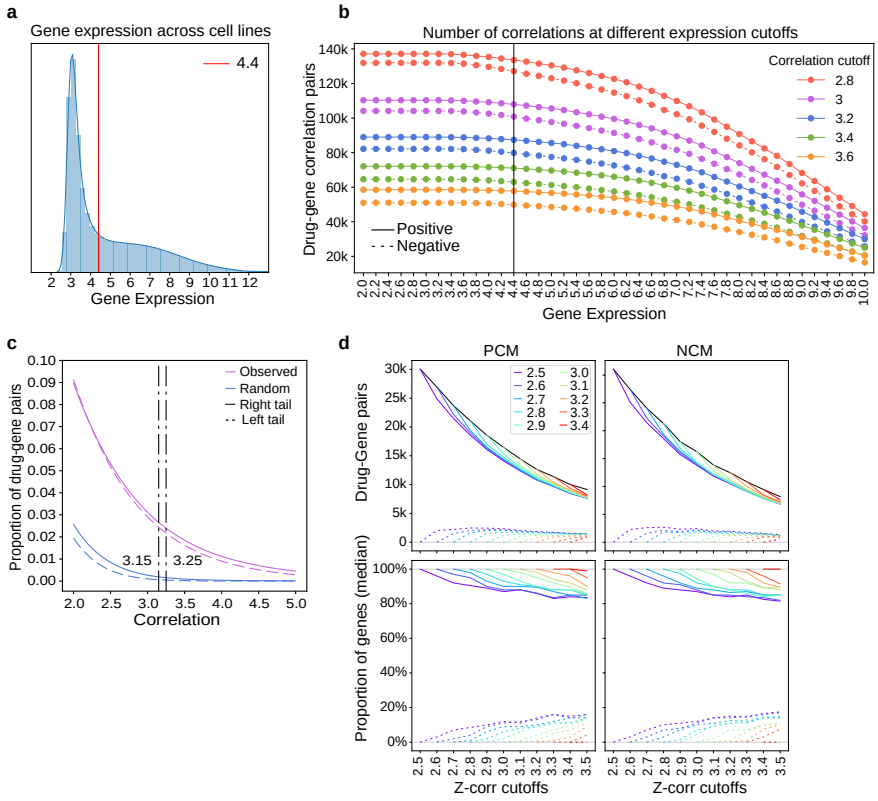


Figure A.1.1: Cutoffs exploration. **a** Distribution of gene expression values across the GDSC cell line panel. The chosen cutoff of 4.4 is shown in red. **b** Robustness of the previous cutoff, measured as number of drug-gene correlations found. Continuous lines correspond to positive correlations and dashed lines to negative correlations. The colors of the lines denote different possible zcor values; the chosen one was 3.2 (in blue) (see next panel). **c** Absolute zcor for the observed and randomized gene-drug correlations. We chose a cutoff of 3.2, as it corresponded to well-accepted *p* values of 0.05 and 0.001 in the observed and randomized distributions, respectively. **d** (Top panels) The black line denotes the number of drug-gene pairs encountered in modules (PCMs and NCMs) as a function of the zcor score cutoff [range 2.5-3.5]. The continuous colored lines show the number of drug-gene pairs 'conserved' in the modules as we move to higher zcor, with respect to the cutoff specified in the legend. On the contrary, dashed lines denote the genes that are added. (Bottom panels) Normalized version of the top panels, taking the total number of drug gene pairs (the black line) as a reference (100%).

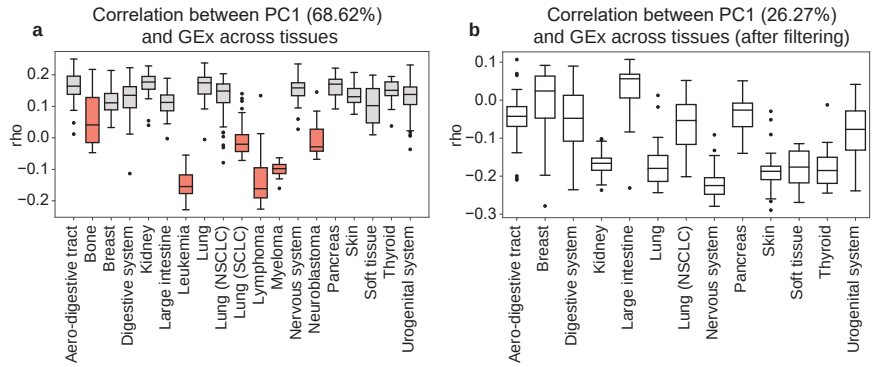


Figure A.1.2: Filtering tissue biases. a We performed a Principal Component Analysis (PCA) in the drug-gene correlation distribution and kept the first principal component (PC1, explaining 65.63% of the variance). Then, we correlated the PC1 loadings to basal gene expression of each CCL (rho). In light of the results, we removed CCLs derived from neuroblastomas, hematopoietic, bone and small cell lung cancer tissues (in red) due to their characteristic rho values. **b** Analysis of the tissue effect in drug-gene correlations after filtering the most influential tissues. After the filtering, PC1 explains only 26% of the total variance and none of the tissues is distinctively correlated.

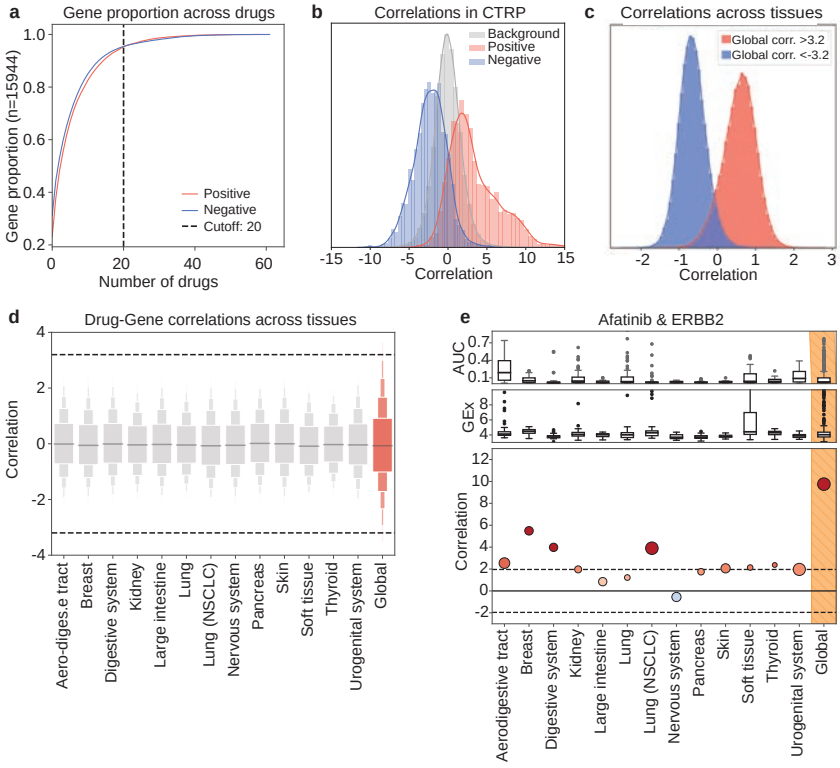


Figure A.1.3: Exploring the drug-gene correlations. **a** We counted the number times a positively (red) or negatively (blue) correlated gene appeared across drugs and plotted their cumulative distribution. The dashed black line shows the cutoff applied to identify frequently-correlated genes (5%). **b** We calculated drug-gene correlation in an external dataset (CTRP panel) and identified positive and negative correlations (+/- 3.2 zcor). When mapped CTRP-correlation pairs on the GDSC results, CTRP-positive and CTRP-negative correlations were also found positively (red) and negatively (blue) correlated in GDSC, respectively. **c** Distribution of drug-gene correlation medians across tissues for positively (red) and negatively (blue) correlated genes (zcor beyond +/- 3.2). **d** Drug-gene correlation distribution in each tissue. The right-most boxplot (in red) shows the correlations using all the tissues. **e** Afatinib-ERBB2 correlation across tissues.

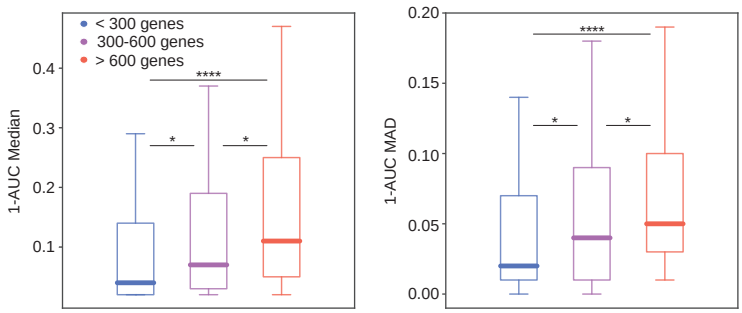


Figure A.1.4: The number of correlated genes depends on the drug response. Median (left) and median absolute deviation (MAD, right) of 1-AUC values per drug across CCLs. Results are split by the number of genes (<300, 300-600 and >600) correlated to each drug. Globally, higher (left) and more variable (right) 1-AUC values allow for more drug-gene correlations to be detected.

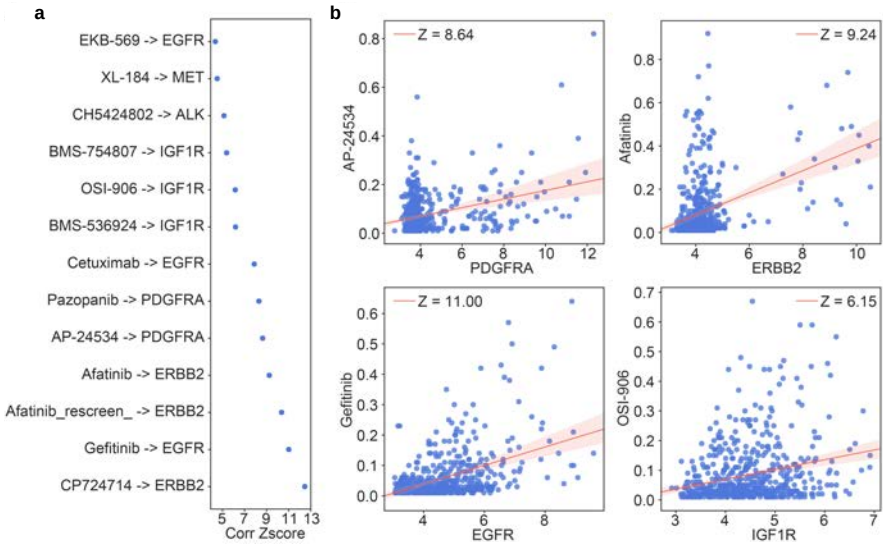


Figure A.1.5: Analyzing the correlation of the drugs' targets. **a** Correlation between drugs and cell surface receptor targets. **b** Four exemplary drugs whose nominal target gene expression correlates to cell line sensitivity.

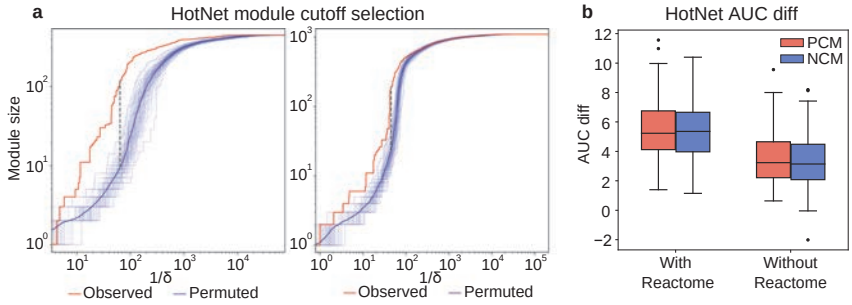


Figure A.1.6: HotNet module selection. **a** HotNet statistic for the positively-correlated module (PCM) detection in Afatinib. Red lines correspond to observed data and blue lines to random runs. In the right panel, HotNet was run without pre-filtering. In the left panel, a Reactome-based pre-filtering was applied. **b** Difference between observed and random HotNet curves (quantified as the subtraction of areas) across all drugs, with and without the Reactome filtering.

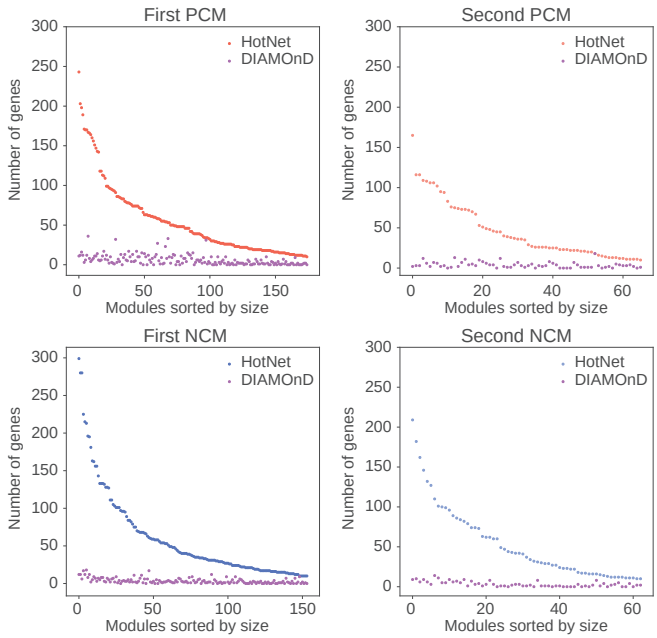


Figure A.1.7: Analyzing the impact of DIAMOnD compared to HotNet. Number of genes added by the DIAMOnD step in relation to genes added by HotNet.

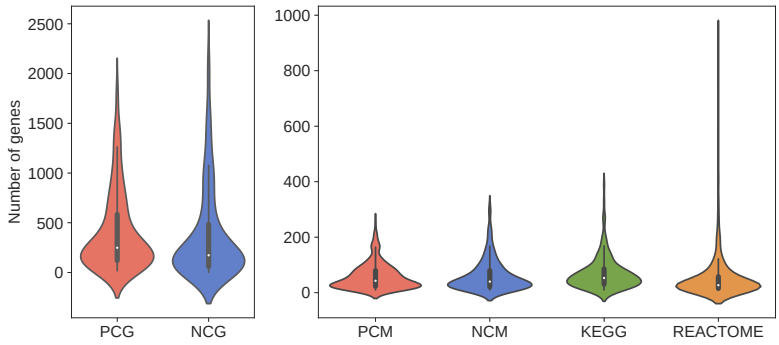


Figure A.1.8: Comparing modules' size with other gene sets. Number of positively and negatively correlated genes (PCGs, NCGs) per drug. Number of genes in positively and negatively correlated modules (PCMs, NCMs), compared to number of genes in KEGG and Reactome pathways.

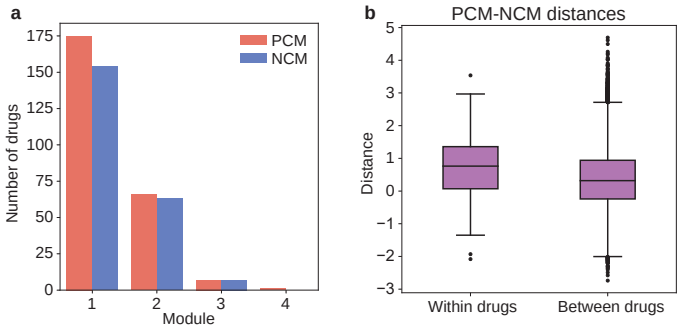


Figure A.1.9: Comparing positive and negative correlated gene modules. **a** Number of modules identified per drug (PCMs in red, NCMs in blue). **b** PCM vs NCM distances within drugs and between drugs.

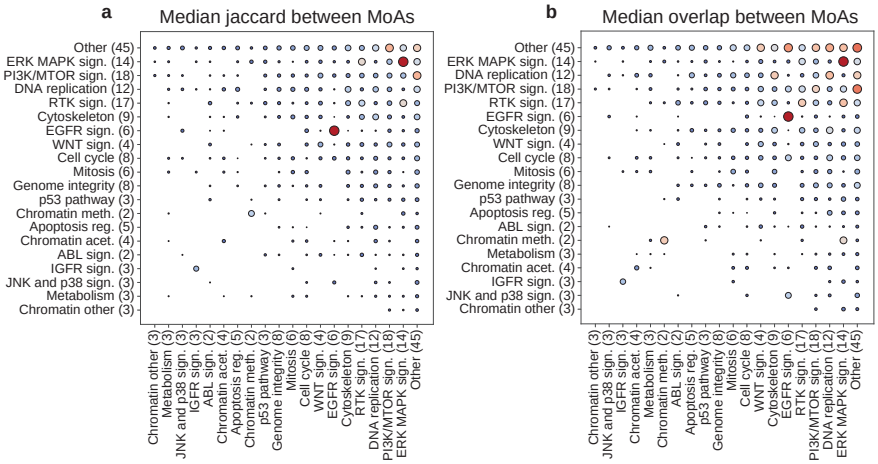


Figure A.1.10: Module-based drug similarity. 'Similarities' between drug modules of different drug classes. Larger and redder dots denote higher similarities. **a** Median Jaccard coefficient between genes [capped at 0.3 in the plot scale]. **b** Median overlap index (x-axis with respect to y-axis), capped at 0.5.

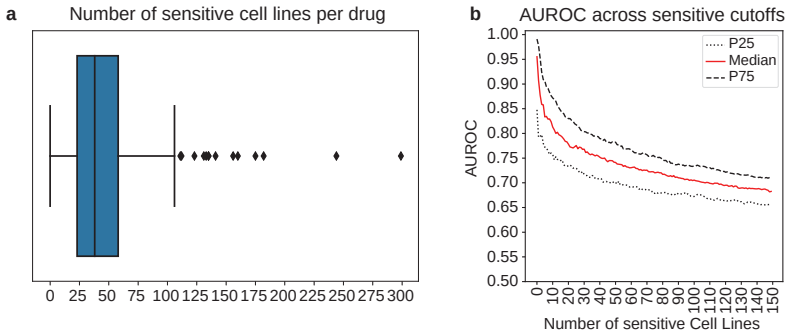


Figure A.1.11: Cell sensitivity calling. **a** Number of sensitive cell lines per drug according to the GDSC publication. **b** Predictive capability (AUROC) of the top-*n* sensitive cell lines, *n* ranging from 1 to 150.

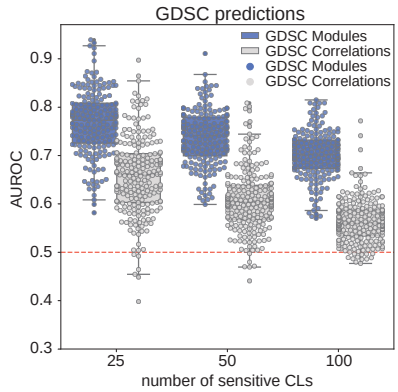


Figure A.1.12: Drug response prediction in the GDSC. AUROC for the GDSC drug predictions using drug modules (blue) and significant drug-gene correlations (i.e., full signatures, gray).

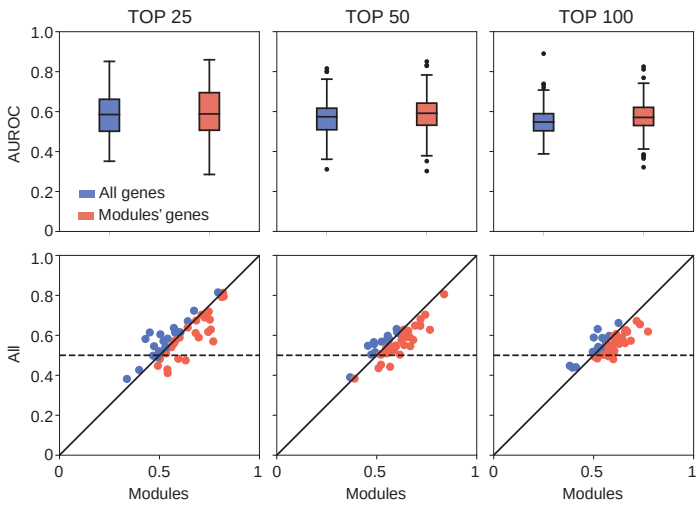


Figure A.1.13: Comparing module-driven prediction performances with the ones obtained with full gene expression signatures. Performance of random forest predictors of drug sensitivity (a predictor was built for each drug; predictions for the top 25, 50 and 100 most sensitive cell lines are shown). (Top) Distribution of AUROC for the predictors using full gene expression profiles (blue), and module-specific profiles (red). (Bottom) A paired view of the AUROC values.

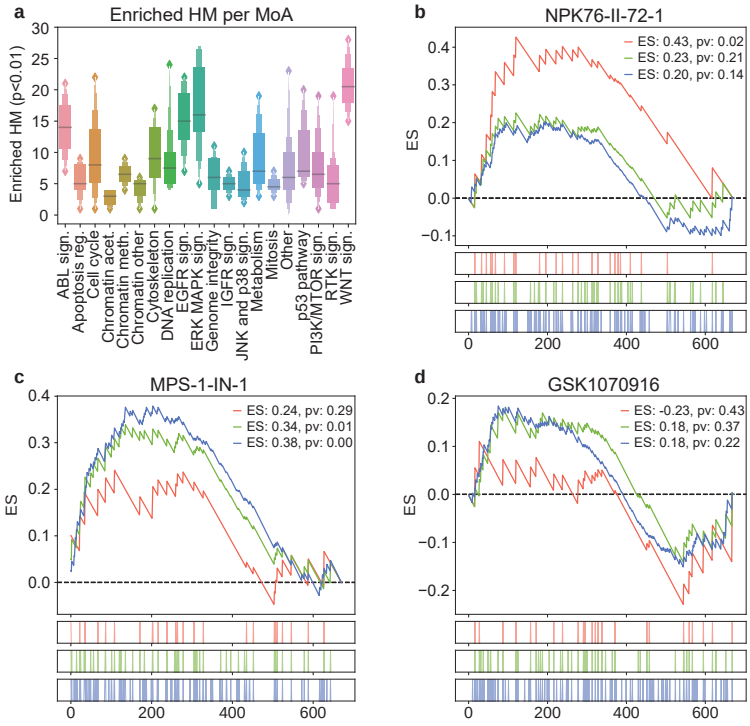


Figure A.1.14: Drug module-driven enrichments. **a** Number of enriched Hallmarks (HM) (p value < 0.01) across MoAs. **b,c,d** Myc gene expression enrichment in cell lines sensitive to NPK76-II-72-1, MPS-1-IN-1, and GSK1070916.

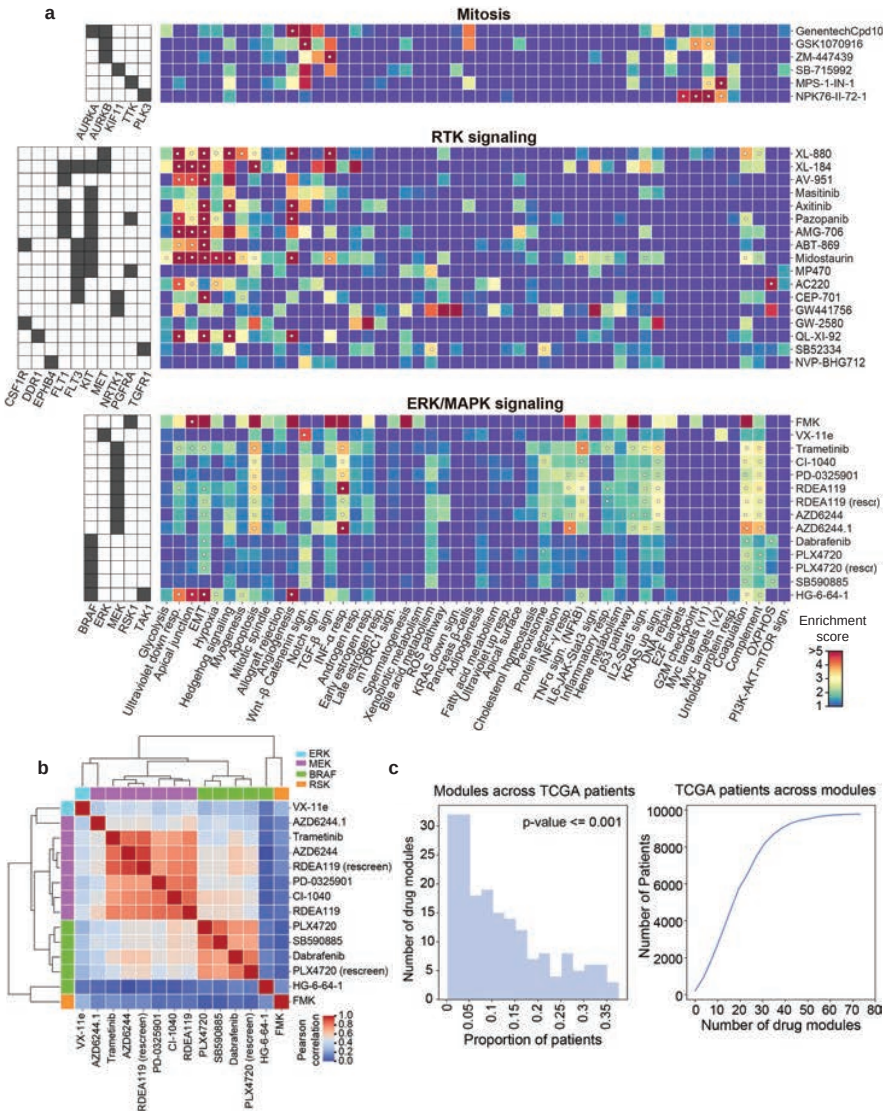


Figure A.1.15: Drug module characterization. **a** Heatmap showing enrichment scores (odds ratio) of the positively correlated genes (full signatures) in the Hallmark gene set collection for 3 different drug classes: Mitosis, RTK signaling, and ERK MAPK signaling. White dots denote significant enrichments (p value < 0.05). **b** Hierarchical clustering between ERK/MAPK inhibitors drugs according their cell line sensitivity correlations. **c** Number of enriched drug modules across the TCGA cohort (left plot) and cumulative distribution of the number of TCGA patients across the enriched drug modules (right plot).

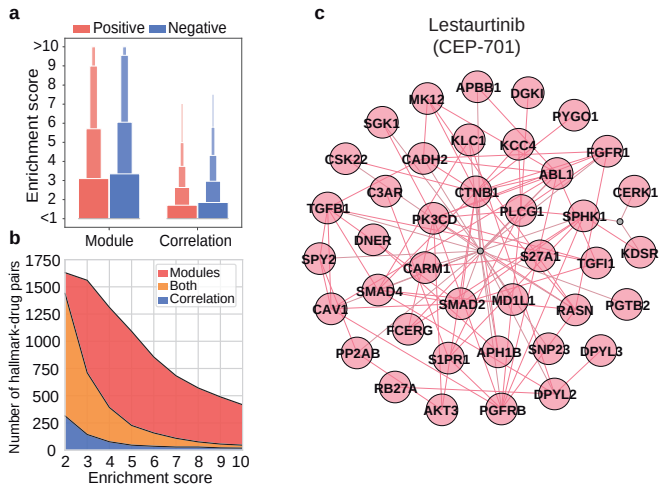


Figure A.1.16: Quantifyng enrichments associations. **a** Distribution of the enrichment scores in the Hallmark collection gene sets. Overall, higher enrichment scores are obtained using modules than using full signatures (PCGs and NCGs) (the gene universe used here is that of Reactome). **b** Similarly, number of Hallmark-drug pairs at different enrichment scores. We show the pairs found only with the modules (PCMs and NCMs, red), only with correlations (PCGs and NCGs, blue), or in both (orange). **c** A view of Lestaurtinib (CEP-701) module. For illustrative purposes, two out-of-the-module (non-correlated) proteins are shown (gray), one being very central and one being peripheral but acting as a 'bridging' node.

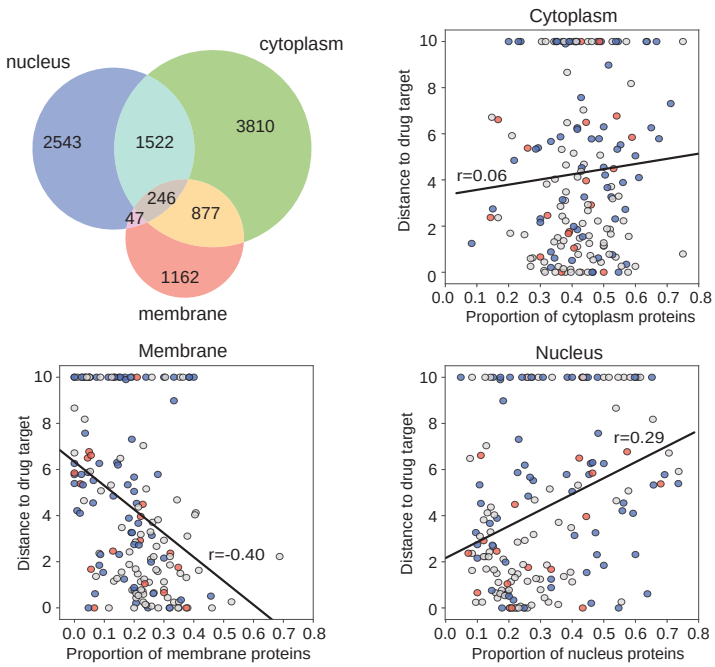


Figure A.1.17: Module distance to drug's target in different cell regions. Correlation between the module distance to the drug target and the proportion of membrane, cytoplasm and nucleus proteins in positively correlated modules (PCMs; first module). The Venn diagram shows the proportion of proteins found in each cellular component category. In each correlation plot, blue dots correspond to drugs targeting a nucleus protein whereas red dots correspond to drugs targeting a membrane protein. In gray, we show drugs which target is found in both the nucleus and the membrane. Drugs with higher proportion of membrane proteins in their modules tend to have their target 'nearby' the module (Spearman's r = -0.40), while modules with more nucleus proteins tend to have more distal targets (r = 0.29).

A.2 Chapter 3.2

Table A.2.1: Comparing pre-existing knowledge graphs. *Although our KG includes up to 150 datasets, we selected 66 as a reference to perform the embeddings.

KG dataset	Design use case	Entities	Edges	Entity Types	Relation Types	Datasets
Hetionet	Repurp.	47k	2.2M	11	24	29
DRKG	Repurp.	97k	5.7M	13	107	34
BioKG	General	105k	2M	10	17	13
PharmKG	Repurp.	7.6k	500k	3	29	7
OpenBioLink	Benchmark	184k	4.7M	7	30	17
Clinical KG	Personalized medicine	16M	220M	35	57	35
Bioteque (ours)	General	450k	30M	12	67	150 (66*)

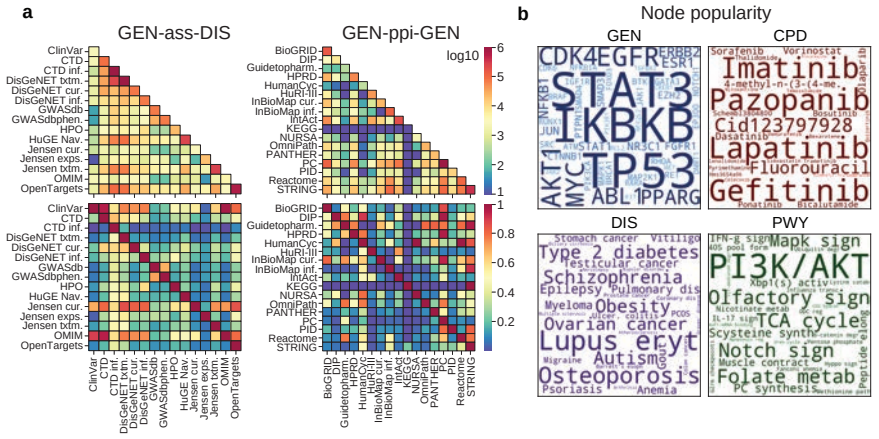


Figure A.2.1: Most popular data in the KG. a Number of edges (top row) and overlap (bottom row) between the datasets inside the 'gene associates with disease' (GEN-ass-DIS, left) and 'protein interacts protein' (GEN-ppi-GEN, right) associations. **b** Most popular nodes in the KG within the gene (GEN, blue), compound (CPD, red), disease (DIS, purple) and pathway (PWY, green) universe. Dataset associations were de-propagated across the corresponding ontologies (when possible) before computing the popularity of the nodes.

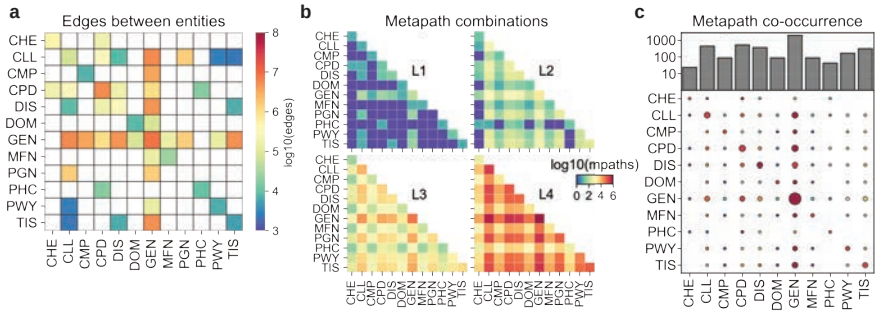


Figure A.2.2: Number of relations in the KG. a Number of possible edges between entities considering all associations and datasets available in the graph. **b** Theoretical number of metapaths of length 1, 2, 3, and 4. **c** Number of times that every entity (x-axis) co-occurs in a metapath with another entity (y-axis). The color scale illustrates the most used entities in each row, red and blue being the most and less used entities, respectively. The size is proportional to the number of times a given entity participates in a metapath.

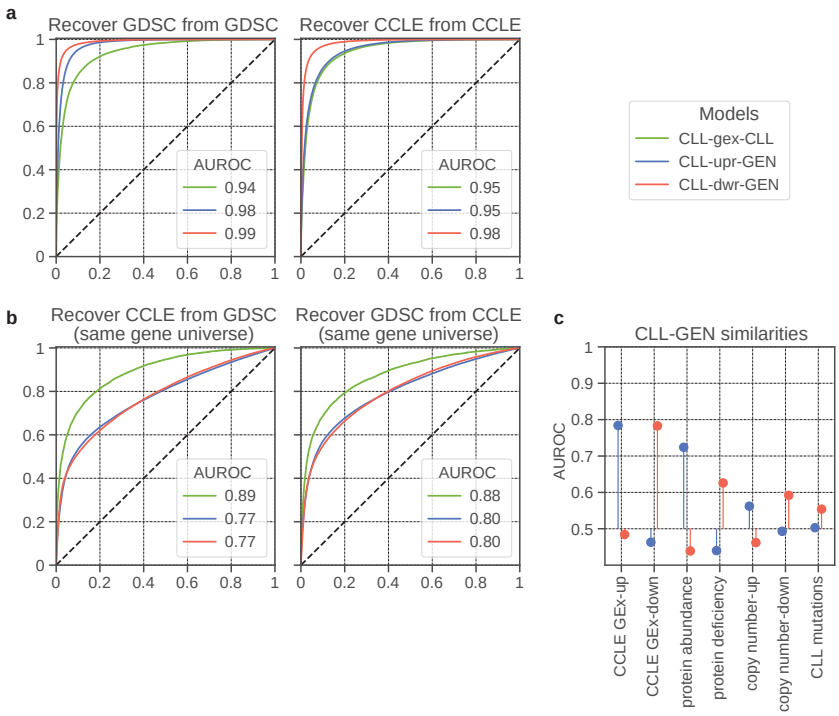


Figure A.2.3: Agreement between GDSC and CCLE embeddings. **a** Recovery of the original GDSC (left) and CCLE (right) network by their corresponding Bioteque embeddings. **b** Recovery of the CCLE (left) and GDSC (right) panels using the GDSC embeddings and CCLE embeddings, respectively. In contrast to Fig. 3.2.5d, embeddings were obtained from a GDSC and CCLE version in which we kept only those genes in common between both panels. **c** Characterization of the cell-gene (CLL-GEN) similarities for the 'cell upregulates gene' (CLL-upr-GEN) and 'cell downregulates gene' (CLL-dwr-GEN) metapath embeddings.

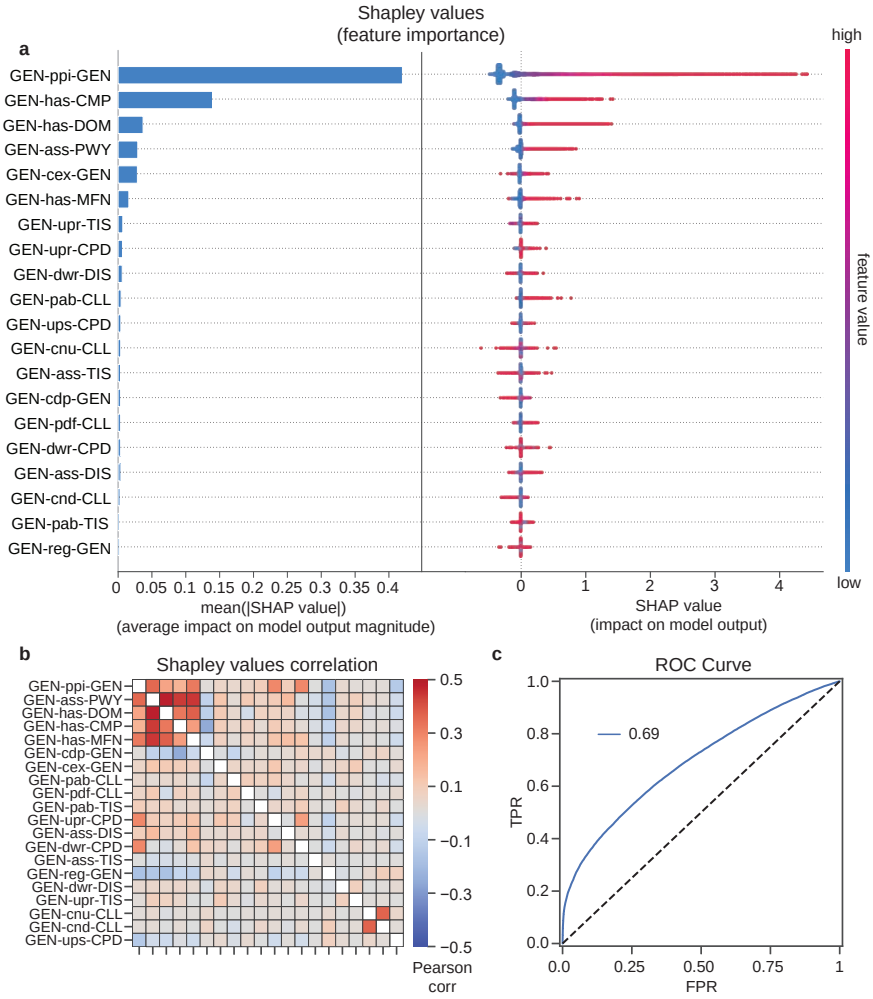


Figure A.2.4: Metapath importance when predicting new HuRI-III PPIs. **a** Feature importance measured as Shapley values (x-axis) for each metapath (y-axis) when predicting HuRI-III PPIs. The higher the Shapley value the higher the impact when predicting the correct class. (Left) Mean absolute Shapley value for each metapath. (Right) Individual Shapley values for each prediction (i.e., each dot corresponds to a predicted PPI). In the colour scale, red and blue indicate lower and higher P-values for the corresponding metapath, respectively. **b** Pairwise Pearson's correlation matrix of the Shapley value vectors for each metapath. The higher the correlation, the higher the agreement between metapaths when classifying a PPI. **c** ROC curve obtained from the model.

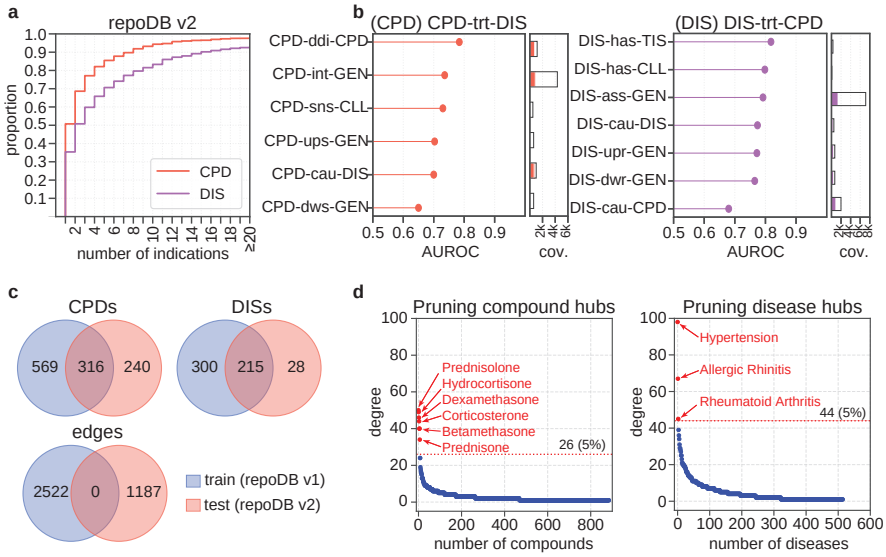


Figure A.2.5: Exploring the repoDB dataset. **a** Cumulative distribution of the number of treatment indications reported by repoDB v2 for the compounds (CPD, red) and diseases (DIS, purple). **b** Top L1 metapaths recapitulating CPD-CPD (left) and DIS-DIS (right) treatment similarities. This was measured by assessing (AUROC) how the different embeddings up-ranked CPD-CPD (or DIS-DIS) pairs associated with the same treatment. The bar plots show the number of nodes available by each metapath, colouring those that were covered by the 'compound treats disease' (CPD-trt-DIS) network. **c** Number of unique compounds (CPDs), diseases (DISs) and edges in the train (repoDB v1) and test (repoDB v2) splits after mapping the entities to the 'compound interacts protein' (CPD-int-GEN) and 'disease associates with gene' (DIS-ass-GEN) embedding universes. **d** Compounds (left) and diseases (right) of the train split were sorted according to their node degree in the repoDB CPD-trt-DIS network. The red line shows the degree corresponding to 5% of total possible associations. We highlight those compounds and diseases whose degree exceeded this limit and were, therefore, pruned.

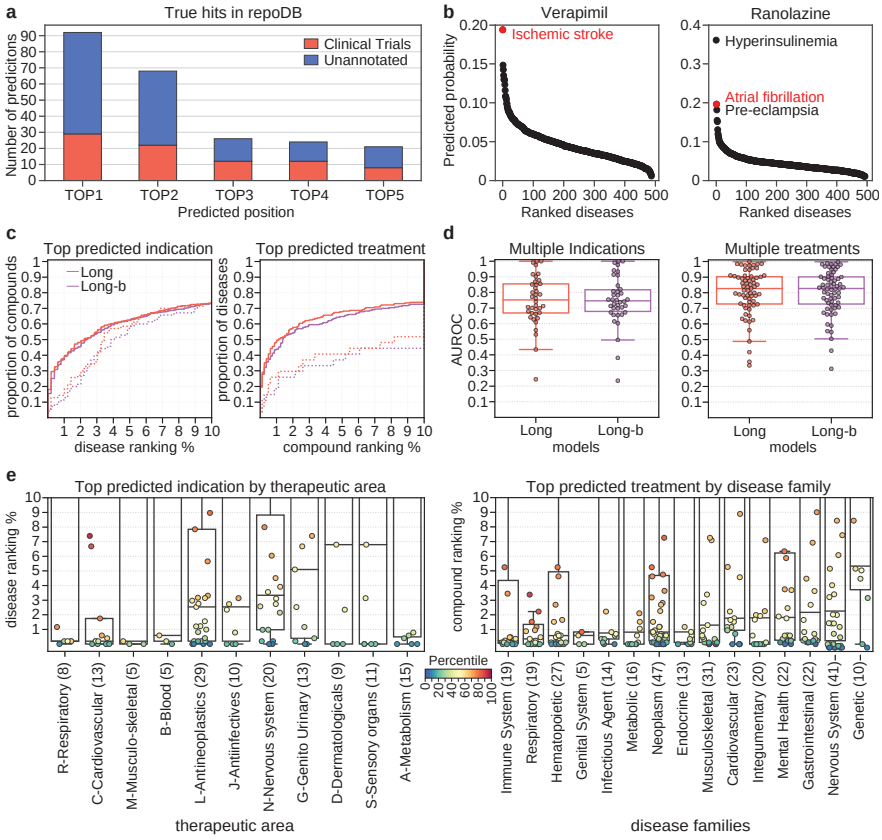


Figure A.2.6: Additional repoDB prediction results. **a** Number of compound-disease (CPD-DIS) repurposing pairs from repoDB (v2) (y-axis) correctly predicted by the *Long* model within the top 5 positions (x-axis). In red we color those predictions for which repoDB provides evidence of having been in clinical trials. **b** Predictions for all the screened new indications for Verapamil (left) and Ranolazine (right) drugs ranked (x-axis) according to the predicted probability given by the *Long* model (y-axis). In red we show those new indications validated in repoDB (v2). **c** Cumulative distribution (y-axis) of compounds (left) and diseases (right) according to the ranked position (x-axis) of the best predicted disease indication (left) or compound treatment (right) for the *Long* and *Long-b* models. The rankings are shown in percentages and only for the first 10% of compounds or disease predictions. Dotted line shows the distribution for those compounds or diseases with only one positive indication in repodDB (v1). **d** Classification performance obtained for each compound (n = 38, left plot) and disease (n = 67, right plot) with multiple (≥5) new indications reported in repoDB (v2). Box plots indicate median (middle line), 25^{th}, 75^{th} percentile (box), and max and min value within the $1.5*25^{th}$ and $1.5*75^{th}$ percentile range (whiskers). **e** We categorized the compounds (left) and diseases (right) according to their therapeutic area and disease family (x-axis) and showed the ranking of the best predicted indication and treatment (y-axis), respectively. Each dot corresponds to either a drug or a disease. The parenthesis in the x-axis indicates the total number of drugs or diseases in each class. Box plots indicate median (middle line), 25^{th}, 75^{th} percentile (box), and max and min value within the $1.5*25^{th}$ and $1.5*75^{th}$ percentile range (whiskers). Since the ranking was cut at the closest 10% of predictions, we coloured each drug or disease by the percentile it represents in the population of the corresponding group.

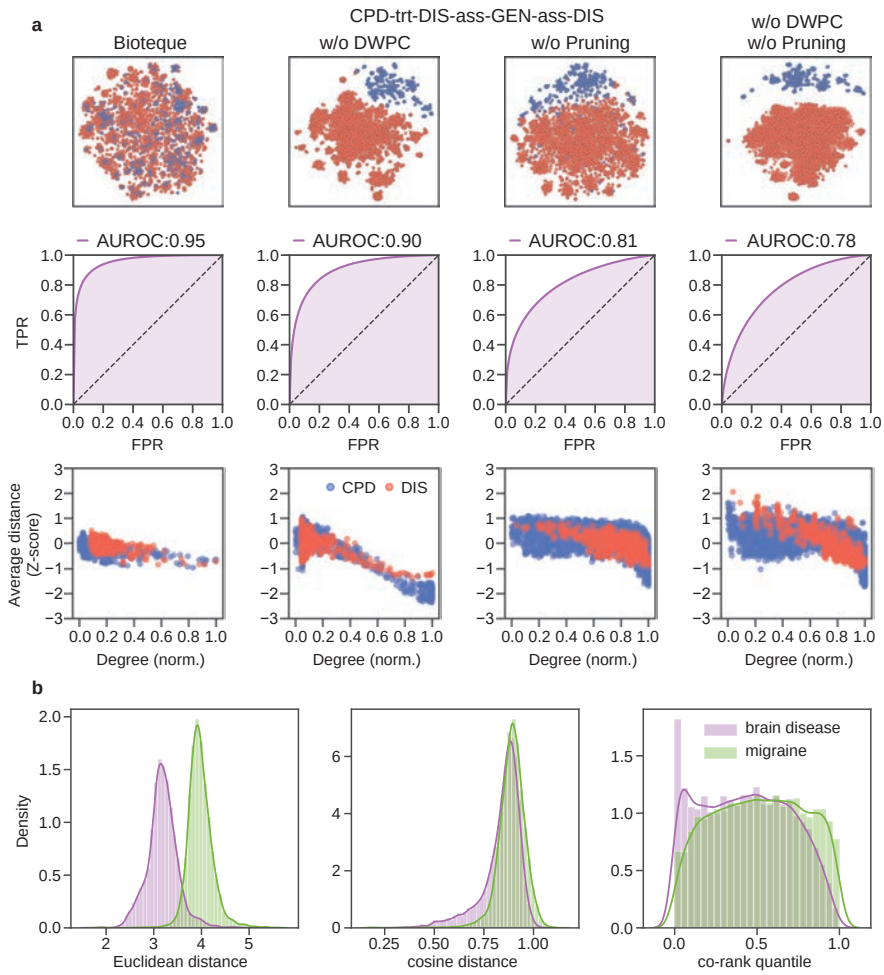


Figure A.2.7: Accounting for node degree biases. **a** Using the metapath embedding CPD-trt-DIS-ass-GEN-ass-DIS as reference we calculated 3 other embedding spaces where we removed the DWPC weights (w/o DWPC), the limitation in the number of edges (w/o pruning) or both (w/o DWPC, w/o pruning). Looking at the 2D representations (first row) we can see how removing either the DWPC or the pruning introduces biases in the space according to the node type, making them cluster separately in the 2D projection and affecting the ability of the space to recapitulate the original KG (second row). In the last row, we show the association between the average *z* score cosine distance of each node (y-axis) and their normalized degree (i.e., divided by the max degree within each node type) in the network (x-axis). Notice that, while it is expected that nodes with a higher degree will be, by definition, closer to more nodes, the average distance does not differ more than 1 standard deviation from the average (*z* scores between -1 and 1). However, removing either the DWPC or the pruning makes higher-degree nodes much closer, on average, to any other node in the space. **b** Distance distribution of the 'Brain disease' and 'Migraine' nodes to each of the genes available in the GEN-ass-DIS embedding space (obtained from DisGeNET). From left to right we show the distribution using Euclidean distances, cosine distances (1-cosine similarity), and co-rank quantiles (calculated as specified in the *Methods* section).

B

Publications

B.1 List of publications

1. **Fernández-Torras, A.**, Duran-Frigola, M. & Aloy, P. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med.* 11, 17 (2019). doi:10.1186/s13073-019-0626-x

2. Duran-Frigola, M., **Fernández-Torras, A.**, Bertoni, M. & Aloy, P. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 9, e1408 (2019). doi:10.1002/wcms.1408

3. Duran-Frigola, M., Bertoni, M., Blanco, R., Martínez, V., Pauls, E., Alcalde, V., Turon, G., Villegas, N., **Fernández-Torras, A.**, Pons, C., Mateo, L., Guitart-Pla, O., Badia-i-Mompel, P., Gimeno, A., Soler, N., Brun-Heath, I., Zaragoza, H. & Aloy, P. Bioactivity Profile Similarities to Expand the Repertoire of COVID-19 Drugs. *J. Chem. Inf. Model.* 60, 5730–5734 (2020). doi:10.1021/acs.jcim.0c00420

4. Douglass, E. F., Jr, Allaway, R. J., Szalai, B., Wang, W., Tian, T., **Fernández-Torras, A.**, Realubit, R., Karan, C., Zheng, S., Pessia, A., Tanoli, Z., Jafari, M., Wan, F., Li, S., Xiong, Y., Duran-Frigola, M., Bertoni, M., Badia-I-Mompel, P., Mateo, L., Guitart-Pla, O., Chung, V., DREAM CTD-squared Pancancer Drug Activity Challenge Consortium, Tang, J., Zeng, J., Aloy, P., Saez-Rodriguez, J., Guinney, J., Gerhard, D. S. & Califano, A. A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. *Cell Rep. Med.* 3, 100492 (2022). doi:10.1016/j.xcrm.2021.100492

5. **Fernández-Torras, A.**, Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr. Opin. Chem. Biol.* 66, 102090 (2022). doi:10.1016/j.cbpa.2021.09.001

6. **Fernández-Torras, A.**, Duran-Frigola, M., Bertoni, M., Locatelli, M. & Aloy, P. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat. Commun.* 13, 1–18 (2022). doi:10.1038/s41467-022-33026-0

7. Dan, Y., Radic, N., Gay, M., **Fernández-Torras, A.**, Arauz, G., Vilaseca, M., Aloy, P., Canovas, B. & Nebreda, R. A. Characterization of p38α signaling networks in cancer cells using quantitative proteomics and phosphoproteomics. *Mol. Cell. Proteomics (Submitted)*, *BioRxiv (preprint)* (2022). doi:10.1101/2022.09.09.507259

8. **Fernández-Torras, A.**, Locatelli M., Bertoni M., Duran-Frigola M. & Aloy P. BQsupports: Systematic annotation of biomedical support for binary data. *(In preparation)*, (2022).

B.2 Attachment of publications

Here we attach all the articles that have been published in the course of this thesis.

Fernández-Torras et al. Genome Medicine (2019) 11:17
https://doi.org/10.1186/s13073-019-0626-x

Genome Medicine

RESEARCH Open Access

Encircling the regions of the pharmacogenomic landscape that determine drug response

Adrià Fernández-Torras[1], Miquel Duran-Frigola[1*] and Patrick Aloy[1,2*]

Abstract

Background: The integration of large-scale drug sensitivity screens and genome-wide experiments is changing the field of pharmacogenomics, revealing molecular determinants of drug response without the need for previous knowledge about drug action. In particular, transcriptional signatures of drug sensitivity may guide drug repositioning, prioritize drug combinations, and point to new therapeutic biomarkers. However, the inherent complexity of transcriptional signatures, with thousands of differentially expressed genes, makes them hard to interpret, thus giving poor mechanistic insights and hampering translation to clinics.

Methods: To simplify drug signatures, we have developed a network-based methodology to identify functionally coherent gene modules. Our strategy starts with the calculation of drug-gene correlations and is followed by a pathway-oriented filtering and a network-diffusion analysis across the interactome.

Results: We apply our approach to 189 drugs tested in 671 cancer cell lines and observe a connection between gene expression levels of the modules and mechanisms of action of the drugs. Further, we characterize multiple aspects of the modules, including their functional categories, tissue-specificity, and prevalence in clinics. Finally, we prove the predictive capability of the modules and demonstrate how they can be used as gene sets in conventional enrichment analyses.

Conclusions: Network biology strategies like module detection are able to digest the outcome of large-scale pharmacogenomic initiatives, thereby contributing to their interpretability and improving the characterization of the drugs screened.

Background

Gene expression profiling has become a mainstay approach to characterize cell properties and status, unveiling links between gene activities and disease phenotypes. Early efforts were channeled into discovering transcriptional signatures that are specific to a disease state. This work involved the comparison of a relatively small number of diseased and healthy samples [1]. Although providing a rich account of disease biology, these studies have failed to yield better drug therapies, as causality and response to drug perturbations cannot be inferred directly from two-state (diseased vs. healthy) differential gene expression analysis [2, 3]. To address this issue, initiatives have flourished to profile the basal gene expression levels of hundreds of cell lines, together with their response to treatment over an array of drug molecules using a simple readout such as growth rate [4–7]. Provided that the panel of cell lines is large enough, this approach allows for a new type of gene expression analysis where basal expression levels are *correlated* to drug response phenotypes. A series of recent studies demonstrate the value of this strategy for target identification, biomarker discovery, and elucidation of mechanisms of action (MoA) and resistance [8–13].

The largest cell panels available today are derived from cancerous tissues, since a crucial step towards personalized cancer medicine is the identification of transcriptional signatures that can guide drug prescription.

* Correspondence: miquel.duran@irbbarcelona.org; patrick.aloy@irbbarcelona.org
[1]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

However, current signatures are composed of several hundred genes, thereby making them difficult to interpret, harmonize across platforms, and translate to clinical practice [14–16]. Recent assessment of sensitivity signatures for over 200 drugs [9] revealed that key genes include those involved in drug metabolism and transport. Intended therapeutic targets, though important, are detected in only a fraction of signatures, and cell line tissue of origin has been identified as a confounding factor throughout the signature detection procedure. In practice, the length of the signatures largely exceeds the number of sensitive cell lines available for each drug, which often yields inconsistent results between cell panels from different laboratories [14]. The current challenge is to filter and characterize transcriptional signatures so that they become robust, informative, and more homogeneous, while still retaining the complexity (hence the predictive power) of the original profiles [17].

Network biology offers means to integrate a large amount of omics data [18]. Most network biology capitalizes on the observation that genes whose function is altered in a particular phenotype tend to be co-expressed in common pathways and, therefore, co-localized in specific network regions [19]. Following this principle, it has been possible to convert genome-wide signatures to network signatures, or *modules*, that are less noisy and easier to interpret [20]. Raphael and co-workers, for instance, developed an algorithm to map cancer mutations on biological networks and identify "hot" regions that distinguish functional (driver) mutations from sporadic (passenger) ones [21]. Califano's group combined gene expression data with regulatory cellular networks to infer protein activity [22]. Overall, network-based methods come in many flavors and offer an effective framework to organize the results of omics experiments [23].

While many genes and proteins have enjoyed such a network-based annotation (being circumscribed within well-defined modules such as pathways and biological processes), drug molecules remain mostly uncharacterized in this regard. For a number of drugs, the mechanism of action is unclear [3] and off-targets are often discovered [24]. Recent publications of drug screens against cancer cell line panels, and the transcriptional signatures that can be derived from there, provide a broader view of drug activity and enable the full implementation of network biology techniques. Here we undertake the task of obtaining and annotating transcriptional modules related to 189 drugs. We show how these modules are able to capture meaningful aspects of drug biology, being robust to inherent biases caused by, for example, the cell's tissue of origin, and having a tight relationship to mechanisms of action and transportation events occurring at the membrane. Finally, we perform a series of functional enrichment analyses, which contribute to a better understanding of the molecular determinants of drug activity.

Methods

Data preparation and drug-gene correlations

We collected gene expression and drug response data from the GDSC resource (https://www.cancerrxgene.org). We first discarded those genes whose expression levels were low or stable across cell lines (Additional file 1: Figure S1A). To this end, we analyzed the distribution of basal expression of each gene in every CCL and filtered out those with an expression level below 4.4 (log2 units) across the panel (see Additional file 1: Figure S1B for a robustness analysis). Regarding drug response data, GDSC provides measurements of cell survival at a range of drug concentrations (area under the dose-response curve (AUC)). Since this measure is inversely proportional to drug sensitivity (i.e., the more sensitive the cell, the shorter its survival), we used the 1-AUC as a measure of potency. Thus, *positive* correlations denote drug sensitivity caused by gene overexpression while *negative* correlations indicate that sensitivity is associated with gene underexpression.

Recent studies report a confounding effect of certain tissues in the global analysis of drug-gene correlations [9]. In order to identify these potential biases in our dataset, we performed a principal component analysis (PCA) on the matrix of raw drug-gene correlations (Pearson's between 1-AUC and gene expression units). Then, we correlated the loadings of the first PC with gene expression values for each CCL. Finally, we filtered out CCLs belonging to tissues that were strongly correlated to the drug-gene correlation profiles (Additional file 1: Figure S2A). We removed leukemia, myeloma, lymphoma, neuroblastoma, small cell lung cancer (SCLC), and bone CCLs. In addition, we considered only drugs with sensitivity measurements available for at least 400 CCLs, as recommended by Rees et al. [9].

After this filtering process, we recalculated, for each drug-gene pair, the Pearson's correlation between basal gene expression and 1-AUC drug potencies across CCLs. We applied Fisher's z-transformation to the correlation coefficients in order to account for variation in the number of CCLs available for each drug [25]. Overall, we obtained positive and negative drug-gene correlations for 217 drugs and 15,944 genes across a total of 671 CCLs. Drug-gene correlations (z_{cor}) beyond ± 3.2 were considered to be significant (Additional file 1: Figures S1C and S1D shows that this cutoff is a robust choice).

Frequently correlated genes

For each gene, we counted the number of correlated drugs (z_{cor} beyond ± 3.2) and inspected the resulting

cumulative distribution (Additional file 1: Figure S3). Genes at the 5% end of the distribution were considered to be "frequently correlated genes" (FCGs). We found 869 positive and 799 negative FCGs, which were removed from further analyses. Finally, we performed enrichment analyses on those genes using the Gene Ontology database [26] and the DAVID toolbox (https://david.ncifcrf.gov/summary.jsp) (hypergeometric tests).

Tissue-specific correlations

First, we split the CCL panel into sets of CCLs belonging to the same tissue. We then calculated drug-gene correlations (z_{cor}) separately for each of the 13 tissues represented in our dataset. In order to verify that measures of *positively* correlated genes (PCGs) and *negatively* correlated genes (NCGs) were consistent across tissues, we calculated the median z_{cor} across tissues for each drug-PCG/NCG pair. In general, tissue-specific correlations had the same "direction" (i.e., same sign of z_{cor}) as the global correlation used throughout the study (Additional file 1: Figure S4A, left panel).

Drug-target correlations

We obtained drug targets from the GDSC resource (disambiguating them with DrugBank [27], when necessary). We assigned at least one target to 202 of the 217 drugs. We focused on the z_{cor} correlation of the targets to check whether target expression (positively) correlates with drug sensitivity. When more than one target was annotated per drug, we kept the maximum correlation. To validate the statistical significance of this measure, we randomly sampled genes (corresponding to the number of known targets per drugs; here again, we kept the maximum correlation). This process was repeated 1000 times for each drug. The mean and the standard deviation of this null distribution were used to derive a *z*-score, making results comparable between drugs.

Drug module detection

After removing frequently correlated genes from the list of drug-gene correlations, we kept 182 [median; Q1: 84, Q3: 372] positively and 122 [median; Q1: 41, Q3: 337] negatively correlated genes (PCGs, NCGs) per drug. Further, we used correlation values (z_{cor}) to run a gene-set enrichment analysis (GSEA) [28] for each drug and identify the genes that participate in enriched Reactome pathways [29, 30]. We only considered Reactome pathways composed of at least 5 genes. Then, for each drug, we kept the significantly correlated genes found in any of the enriched pathways (P value < 0.01). The resulting GSEA-filtered list of genes retained 100 [median; Q1: 49, Q3: 277] positive and 77 [median; Q1: 30, Q3: 221] negative correlations per drug. Then, taking the z_{cor} values as input scores, we submitted the GSEA-filtered list of genes to HotNet2 [31], using a high-confidence version of STRING [32] (confidence score > 700). We ran HotNet2 iteratively, keeping the largest module and removing its genes for the next iteration, until the modules had fewer than 5 genes or were not statistically significant (p value > 0.05). To recall strong drug-gene correlations "proximal" to the drug modules (missed, most likely, by the incomplete coverage of Reactome), we used the DIAMOnD module-expansion algorithm [29]. We considered only genes that (i) were correlated to the drug response, (ii) were not present in any of the Reactome pathways, and (iii) were in the top 200 closest genes to the module, according to DIAMOnD (this cut-off was proposed by the authors of DIAMOnD based on orthogonal functional analyses). Hence, we obtained at least one positively correlated module for 175 of the drugs (48 genes [median; Q1: 23, Q3: 83]) and one negatively correlated module for 154 of the drugs (40 genes [median; Q1: 21, Q3: 78]). Robustness analysis of this procedure is found in Additional file 1: Figure S1D. A GMT list of the drug modules can be found in Additional file 2. The correlation values of the genes in the drug modules are available in Additional file 3.

Distances between drug targets and modules

DIAMOnD [29] provides a list of genes sorted by their network-based proximity to the module. Accordingly, we retrieved from the STRING interactome the top closest 1450 genes (~ 10% of the largest connected component of the network) for every drug module. We then checked the ranking of drug targets in the resulting DIAMOnD lists, (conservatively) taking the median value when more than one target was available. To assess the proximity of drug targets to the modules, we measured distances to three different sets of random proteins. The first random set corresponded to the STRING proteome. For the second, we collected all genes defined as *Tclin* or *Tchem* in the Target Central Resource Database [33] (i.e., "druggable proteins"). Finally, the third random set included all pharmacologically active drug targets reported in DrugBank (https://www.drugbank.ca/).

Distances between modules

We calculated distances between positively and negatively correlated modules separately using the network distance proposed by Menche et al. [34]. This distance measure is sensitive to the number of genes (size) included in the modules. To normalize this measure, we devised the following procedure. First, we grouped drug modules on the basis of their size. Then, for each module, we calculated the distribution of shortest distances from each gene to the most central one [35]. We used this distribution to sample random modules from the

network. When the distribution constraint could not be fully met, we used the DIAMOnD algorithm [29] to retrieve the remaining genes (50% of the genes at maximum). We repeated this process to obtain 10 random modules of each size. Next, we distributed the random modules into ranges (intervals) of 5 (i.e., from 10 to 14 genes, from 15 to 19, etc.; 50 random modules per interval). Then, for each pair size, we randomly retrieved 100 pairs of modules and calculated the network-based distance between them. The mean and standard deviation of the distances at each pair size were used to normalize the observed distances, correspondingly (*z*-score normalization) (we checked that 100 random pairs were sufficient to approximate the mean and standard deviation of the population). The more negative the network distance (d_{net}), the more proximal the modules are. We provide the network distances as an Additional file 4.

Drug response prediction using drug modules

We performed drug response predictions in the GDSC dataset by using drug modules (only first PCMs and NCMs, to make results comparable between drugs). We devised a simple GSEA-like predictor in which CCLs were evaluated for their up-/downregulation of the modules, correspondingly. To this end, we first normalized the expression of each gene across the CCL panel (*z*-score). Then, for each drug, we ranked CCLs based on the GSEA enrichment scores (ES), taking drug modules as gene sets. To evaluate the ranking, we chose the top 25, 50, and 100 CCLs based on the *known* drug sensitivity profile. Performance was evaluated using the AUROC metric. Results were compared to those obtained with positively and negatively correlated genes (PCG, NCG) from the full signatures (z_{cor} beyond ± 3.2).

To check whether modules derived from GDSC generalize to other CCL panels, we applied the same procedure to the Cancer Therapeutics Response Portal (CTRP) (https://ocg.cancer.gov/programs/ctd2/data-portal). As done with the GDSC panel, we removed all CCLs derived from neuroblastomas, hematopoietic, bone, and small cell lung cancer tissues, leaving a total of 636 CCLs, 397 in common with our GDSC panel (67 drugs in common). Drug response predictions for CTRP were performed as detailed above. We used the best ES among all modules associated with the drug. In addition, we did the analysis using CCLs exclusive to CTRP (i.e., not shared with the GDSC panel).

Module enrichment in Hallmark gene sets

We downloaded the Hallmark gene set collection from the Molecular Signature Database (MSigDB) of the Broad Institute http://software.broadinstitute.org/gsea/index.jsp). We evaluated each gene set independently using a hypergeometric (Fisher's exact) test (first and second modules were merged, when applicable; the gene universe was that of GDSC). Enrichments can be found in the Additional file 5.

Drug module enrichments in the TCGA cohort

We downloaded gene expression data (median z-scores) for 9788 patients and 31 cancer tissues from the Pan-Cancer Atlas available in the cBioPortal resource (http://www.cbioportal.org). "Presence" or "expression" of the module in each patient was evaluated using GSEA (*P* value < 0.001), ensuring that the direction (up/down) of the enrichment score corresponded to the "direction" of the module (PCM/NCM). For a complete list of enrichment results, please see Additional file 6 (results are organized by tumor type). Further, to identify associations between drug modules and cancer driver genes, we checked whether patients "expressing module of drug X" (*P* value < 0.001) were "harboring a mutation in driver gene Y" (Fisher's exact test). We considered 113 driver genes (obtained as described in [36], using the "known" flag) (Additional file 7).

Characterization of drug modules

In order to characterize drug modules from different perspectives, we designed 21 features belonging to the following categories: (i) *General features* derived directly from the pharmacogenomics panel, (ii) *Network features* related to network measures such as topological properties, and (iii) *Biological features* encompassing a series of orthogonal analyses related to drug biology. For more information, please see Additional file 8 and its corresponding legend.

Results and discussion

The Genomics of Drug Sensitivity in Cancer (GDSC) is the largest cancer cell line (CCL) panel available to date [8]. This dataset contains drug sensitivity data (growth-inhibition, GI) for 265 drugs screened against 1001 cell lines derived from 29 tissues, together with *basal* transcriptional profiles of the cells (among other omics data). Aware of the work by Rees et al. [9], we first looked for the dominant effect of certain tissues in determining associations between drug response and gene expression. We found that CCLs derived from neuroblastoma, hematopoietic, bone, and small cell lung cancers may confound global studies of drug-gene correlations due to their unspecific sensitivity to drugs (Additional file 1: Figure S2A). These tissues were excluded from further analyses. We also excluded genes whose expression levels were low or constant across the CCL panel and drugs tested against fewer than 400 CCLs (see the "Methods" section for details). As a result, we obtained a pharmacogenomic dataset composed of 217 drugs, 15,944 genes, and 671 CCLs.

Following the conventional strategy to analyze pharmacogenomic datasets, we calculated *independent* drug-gene associations simply by correlating the expression level of each gene to the potency of each drug (area over the growth-inhibition curve; 1-AUC) across the CCL panel. We used a *z*-transformed version of Pearson's , as recommended elsewhere [25]. Figure 1a shows the pair-wise distribution of the *z*-correlation (z_{cor}) measures between the 15,944 genes and the 217 drugs. We validated the correlations identified in the GDSC panel on an independent set by applying the same protocol to the Cancer Therapeutic Response Portal (CTRP) panel [9] (Additional file 1: Figure S4B). To identify the strongest drug-gene associations, we set a cutoff of ± 3.2 z_{cor}, based on an empirical null distribution obtained from randomized data (see Additional file 1: Figure S1C and the "Methods" section). Please note that this is a widely adopted procedure that is not designed to detect *single* drug-gene associations (which would require multiple testing correction) [37]. Instead, and similar to signature identification in differential gene expression analysis, the goal is to identify sets of genes that are (mildly) correlated with drug response. For each drug, we obtained a median (Med) of 249 positively correlated genes [first quartile (Q1): 120, third quartile (Q3): 584], and Med of 173 negatively correlated genes [Q1: 59, Q3: 484] (Fig. 1b). Some drugs, like the BRAF inhibitor dabrafenib, or the EGFR inhibitor afatinib, had over 1500 positively and negatively correlated genes, while others, like the antiandrogen Bicalutamide or the p38 MAPK inhibitor Doramapimod, had hardly a dozen. We observed that the number of genes that correlate with drug response strongly depends on the drug class (Fig. 1c), EGFR and ERK-MAPK signaling inhibitors being the classes with the largest number of associated genes, and JNK/p38 signaling and chromatin histone acetylation inhibitors being those with the fewest correlations. This variation may be partially explained by the range of drug potency across the CCL panel, as it is "easier" to detect drug-gene correlations when the drug

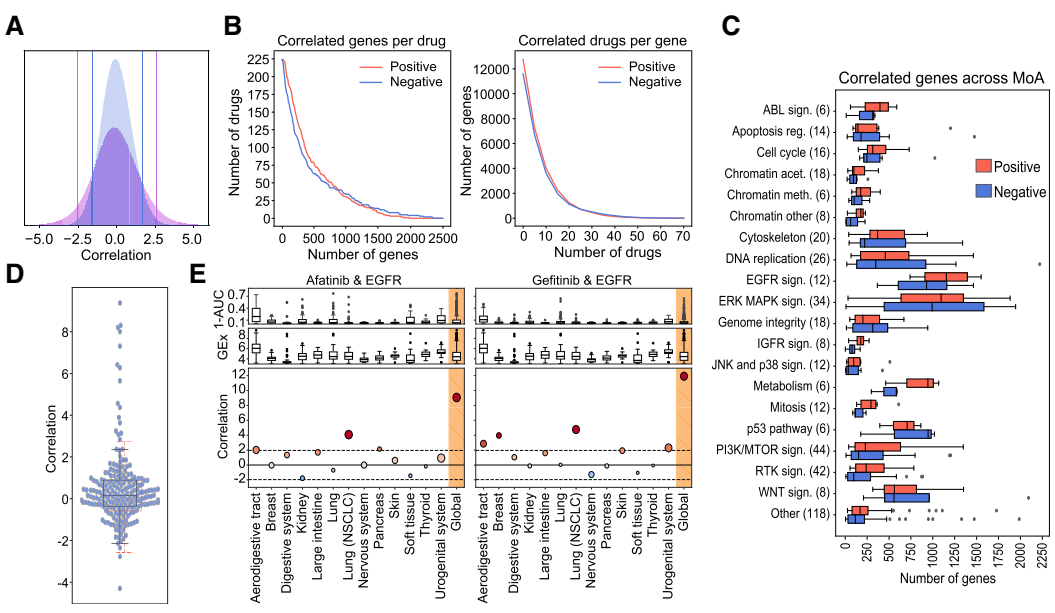


Fig. 1 Analysis of drug-gene correlations. **a** Observed drug-gene correlation distribution (purple) and randomized drug-gene correlation distribution (blue) (random permutation of expression values). Vertical lines denote the percentiles 5 and 95 of each distribution. **b** The left panel shows the "number of correlated genes per drug", while the right panel shows the "number of correlated drugs per gene". In the left panel, one can read, for example, that there are about 25 drugs (*y*-axis) with at least 1250 correlated genes (*x*-axis). Likewise, in the right panel, one can read that about 4000 genes (*y*-axis) are correlated to at least 10 drugs (*x*-axis). **c** Number of positively (red) and negatively (blue) correlated genes across drug classes. **d** Positively correlated targets (see the "Methods" section for details on the *z*-score normalization procedure of this correlation measure). Each dot represents one drug-target correlation. A full account of drug-target annotations is provided in Additional file 8. The red boxplot shows the background (random) distribution. **e** Drug-gene correlations (z_{cor}) between afatinib/gefitinib and the epidermal growth factor receptor (EGFR) across tissues. In the upper plots, we show the drug sensitivity (1-AUC) across tissues. In the middle plots, we show basal gene expression of EGFR across tissues. Bottom plots show the Afatinib/Gefitinib-EGFR correlation. The rightmost values refer to the correlation when all tissues are considered (Global). Size of the bubbles is proportional to the number of CCLs in each tissue

has a wide sensitivity spectrum (Additional file 1: Figure S5).

Similarly, analysis of independent drug-gene correlations suggests that some genes are positively correlated to many drugs. For instance, we found 5% of the genes to be associated with more than 10% of the drugs (Fig. 1b and Additional file 1: Figure S3). The transcripts of these "frequent positively correlated genes" are enriched in membrane processes, specifically focal adhesion (P value $< 5.2 \times 10^{-12}$) and extracellular matrix (ECM) organization (P value $< 5 \times 10^{-16}$), including subunits of integrin, caveolin, and platelet-derived growth factors (PDGFs). These genes determine, among others, the activation of Src kinases [38–41]. Overall, ECM proteins are known to play an important role in tumor proliferation, invasion, and angiogenesis [42, 43] and are often involved in the upstream regulation of cancer pathways [44] such as PI3K/mTOR [38–40], MAPK [39], and Wnt signaling [45], and in cell cycle and cytoskeleton regulation [46]. It is thus not surprising that ECM genes determine drug response in a rather unspecific manner.

On the other hand, "frequent negatively correlated genes" are associated with small molecule metabolism (xenobiotic metabolic processes, P value $< 3.2 \times 10^{-3}$). In this group, we found, among others, the cytochrome CYP2J2 and the GSTK1 and MGST glutathione transferases, which are highly expressed in cancers and known to confer drug resistance through their conjugating activity [47–50]. Following other studies that reported similar results [9], we checked for the presence of multidrug transporters (MDTs). Reassuringly, we found the efflux pump transporter ABCC3 and a total of 27 different solute carriers (SLCs) to be negatively correlated to the potency of many drugs. Of note, we also found the ABCA1 transporter and other 8 SLCs to be among the frequent positively correlated genes, thus emphasizing the key role of transporters and carriers in determining drug potency.

All of the above suggests that systematic analysis of independent drug-gene correlations is sufficient to highlight *unspecific* determinants of drug sensitivity and resistance (i.e., frequent positively and negatively correlated genes). However, while these determinants are recognized to play a crucial role, they do not inform targeted therapies, as they are usually unrelated to the mechanism of action of the drug. Thus, we assessed whether measuring drug-gene correlations would also be sufficient to elucidate drug targets, i.e., we tested whether the expression level of the target correlates with the potency of the drug. Since most drugs had more than one annotated target, to measure significance, we randomly sampled 1000 times an equal number of genes and derived an empirical z-score (see the "Methods" section). Figure 1d shows that the expression level of most drug targets did *not* correlate with drug response. In fact, only ~ 10% of the drugs had "positively correlated targets" (z-score > 1.9, P value ~ 0.05). Remarkably, the 6 EGF pathway inhibitors in our dataset were among these drugs, as were 3 of the 4 IGF pathway and 3 of the 21 RTK pathway inhibitors. We noticed that the molecular targets for these pathways were usually cell surface receptors, e.g., EGFR, IGFR, ALK, ERBB2, MET, and PDGFRA. Overall, of the 20 drugs with positively correlated targets, 13 bind to cell surface receptors, showing a propensity of drug-gene correlations to capture membrane targets (odds ratio = 15.13, P value = 1.9×10^{-7}). In Additional file 1: Figure S6, we show how this trend is driven mostly by the over-expression of the target on the cell surface.

The relatively small number of positively correlated targets illustrates how the analysis of expression levels alone is insufficient to reveal MoAs, especially when the drug target is located downstream of the cell surface receptors in a signaling pathway. Some authors have suggested that the tissue of origin of the cells might play a confounding role in defining drug response signatures. To address this notion, we repeated the calculation of Pearson's z_{cor} correlations separately for each of the 13 tissues in our dataset. In general, the trends observed at the tissue level were consistent with the global trends, although tissue-specific correlations were milder due to low statistical power (i.e., few cell lines per tissue) (Additional file 1: Figure S4A, right panel). Accordingly, we confirmed that none of the tissues had a globally dominant effect on the measures of drug-gene correlations (Additional file 1: Figure S2B) and verified that certain tissue-specific associations were still captured by the analysis. For instance, going back to the targeting of EGFR (which was positively correlated with Afatinib and Gefinitib), we show in Fig. 1e that the "global" correlation can be partly attributed to non-small cell lung cancer (NSCLC) cells (z_{cor} > 1.96, P value < 0.05). Indeed, afatinib and gefitinib have an approved indication for NSCLC. Both drugs correlate with EGFR also in the aerodigestive tract, an observation reported in an independent study dedicated to the discovery of drug-tissue/mutation associations (ACME) [7]. Moreover, and consistent with recent findings [51–54], gefitinib has a significant correlation to EGFR in breast cancers, whereas afatinib correlates with this target in pancreatic CCLs. Afatinib, in turn, is associated with ERBB2 in breast CCLs, as also confirmed by ACME analysis (Additional file 1: Figure S4C).

From drug-gene correlations to drug modules

The previous analysis demonstrates that conventional drug-gene correlations do *not* directly identify drug targets and suggests that standard transcriptional drug

signatures contain unspecific and indirect correlations that may mislead mechanistic interpretation. Recent advances in network biology precisely tackle these problems, as they can (i) filter signatures to make them more functionally homogeneous and (ii) allow for the measurement of network distances so that genes proximal to the target can be captured and connected to it, even if the expression of the target itself is not statistically associated with the drug.

Hence, we set to mapping drug-gene correlations onto a large protein-protein interaction (PPI) network, retaining only genes that could be grouped in network *modules* (i.e., strongly interconnected regions of the network). In the "Methods" section, we explain in detail the module detection procedure. In brief, starting from drug-gene correlations (Fig. 2A), we first filtered out those genes whose expression was frequently (and unspecifically) correlated to the potency of many drugs (Additional file 1: Figure S3). This reduced the number of associations to 182 [median; Q1: 84, Q3: 372] positively and 122 [median; Q1: 41, Q3: 337] negatively correlated genes per drug, respectively. Next, in order to identify genes acting in coordination (i.e., participating in enriched Reactome pathways [29, 30]), we adapted the gene set enrichment analysis (GSEA) algorithm [28] to handle drug-gene correlations (instead of gene expression fold-changes) (Fig. 2B). The resulting GSEA-filtered list of genes kept 100 [median; Q1: 49, Q3: 277] positive and 77 [median; Q1: 30, Q3: 221] negative correlations per drug. After this filtering, we submitted this list to HotNet2 [31], a module detection algorithm that was originally developed for the identification of recurrently mutated subnetworks in cancer patients (Fig. 2C; Additional file 1: Figure S7 shows the importance of the Reactome-based filtering previous to HotNet2). As a reference network (interactome) for HotNet2, we chose a high-confidence version of STRING [32], composed of 14,725 proteins and 300,686 interactions. HotNet2 further filtered the list of genes correlated to each drug, keeping only those that were part of the same network neighborhood. Finally, we used the DIAMOnD module expansion algorithm [29] to recover strong drug-gene correlations that had been discarded along the process. Although this step made a relatively minor contribution to the composition of the modules (less than 5% of the genes; Additional file 1: Figure S8), we did not want to lose any strong association caused by the limited coverage of the Reactome database (Fig. 2D).

Our pipeline yielded at least one "positively correlated module" (PCM) for 175 of the 217 drugs (48 genes [median; Q1: 23, Q3: 83]). Similarly, we obtained "negatively correlated modules" (NCMs) for 154 of the drugs (40 genes [median; Q1:21, Q3:78]). Thus, compared to the original signatures, drug modules are considerably smaller (80% reduction) (Fig. 3a) and are commensurate with manually annotated pathways in popular databases (Additional file 1: Figure S9). For roughly two thirds of the drugs, we obtained only one PCM and one NCM. For the remaining drugs, a second (usually smaller) module was also identified (Additional file 1: Figure S10A). The complete list of drug modules can be found in Additional file 2. Pair-wise drug-gene correlations of the modules are listed as Additional file 3. Additionally, the code of the module-detection pipeline is available at: https://github.com/sbnb-lab-irb-barcelona/GDSC-drug-modules.

Drug modules are tightly related to mechanisms of action

To assess the mechanistic relevance of drug modules, we measured their distance to the corresponding drug targets, i.e., we formulated the hypothesis that drug targets should be "proximal" to dysregulated network regions. To this end, we used the DIAMOnD algorithm again [29], this time to retrieve, for each drug, a list of genes ranked by their proximity to the corresponding drug module(s) (see the "Methods" section). Figure 3b shows that drug targets are remarkably up-ranked in these lists, making them closer to the drug modules than any other set of random proteins, including druggable genes and pharmacological receptors [33], which usually have prominent positions in the PPI network due to the abundant knowledge available for them. In 82% of the PCMs, the corresponding targets were among proximal proteins (top decile), which means a dramatic increase in mechanistic interpretability compared to the 12.25% of drugs that could be linked to their targets via conventional analysis of drug-gene correlations.

A unique feature of drug modules is that network-based distances can be natively measured between them [34]. We computed the distance between drug modules pair-wise (Additional file 4) and grouped them by drug class (Fig. 3c) (see the "Methods" section and alternative statistical treatments in Additional file 1: Figure S11). The diagonal of Fig. 3c clearly indicates that drugs belonging to the same category tend to have "proximal" modules (some of them in a highly specific manner, like in the case of ERK-MAPK signaling cascade inhibitors). Most interestingly, we could observe proximities between modules belonging to different drug classes. For instance, modules of drugs targeting RTK signaling were "located" near to those of drugs affecting genome integrity, in good agreement with recently reported cross-talk between these two processes [55, 56]. Likewise, and as proposed by some studies [57–59], IGFR-related drugs were "proximal" to drugs affecting cell replication events such as mitosis, cell cycle, and DNA replication.

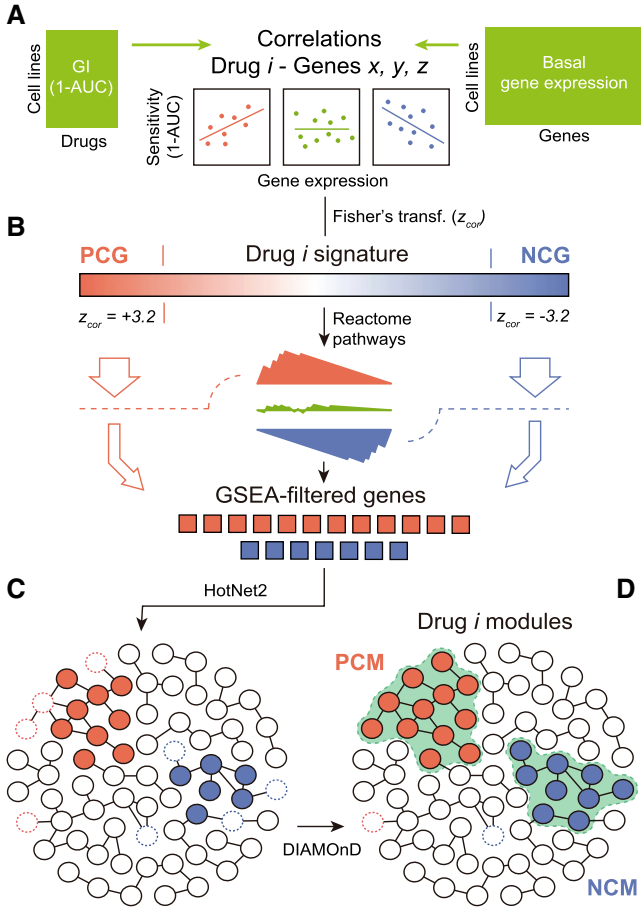


Fig. 2 Methodological pipeline to identify drug modules. (A) The process of obtaining modules starts with the calculation of z-transformed Pearson's correlation (z_{cor}) between gene expression and drug sensitivity data for each drug-gene pair. Correlations (z_{cor}) beyond ± 3.2 are considered to be significant. (B) We then run a gene-set enrichment analysis (GSEA) for each drug in order to identify genes that participate in enriched Reactome pathways. (C) GSEA-filtered genes are submitted to HotNet2 on the STRING interactome in order to identify drug modules. (D) Finally, modules are expanded (when possible) using the DIAMOnD algorithm to recall the few correlated genes that might have been excluded in step C as a result of the limited coverage of the Reactome database. This final step has a minor impact on the composition of the module

Drug modules retain the ability to predict drug response

We have shown that drug modules are related to the MoA of the drug, but the question remains as to the extent to which they retain the predictive capabilities of the full transcriptional profiles/signatures. In the CCL setting, gene expression profiles are valuable predictors of drug response [5, 11, 60] and crucially contribute to state-of-the-art pharmacogenomic models. To test whether our (much smaller) drug modules retained predictive power, we devised a simple drug sensitivity predictor based on the GSEA score (see the "Methods" section). In brief, given a drug, we tested whether cell lines *sensitive* to a certain drug were enriched in the corresponding drug modules. We expect genes in PCMs to be *over*-expressed in sensitive cell lines and those in NCMs to be *under*-expressed. Analogously, we took the positively and negatively correlated genes from the full drug-gene associations (signatures) and also performed a GSEA-based prediction. To nominate a cell "sensitive" to a certain drug, we ranked CCLs by their sensitivity and kept the top *n* CCLs, *n* being 25, 50, or 100, based on the distribution of sensitive cell lines provided by the authors of the GDSC (Additional file 1: Figure S12A). This simple binarization is, in practice, proportional to

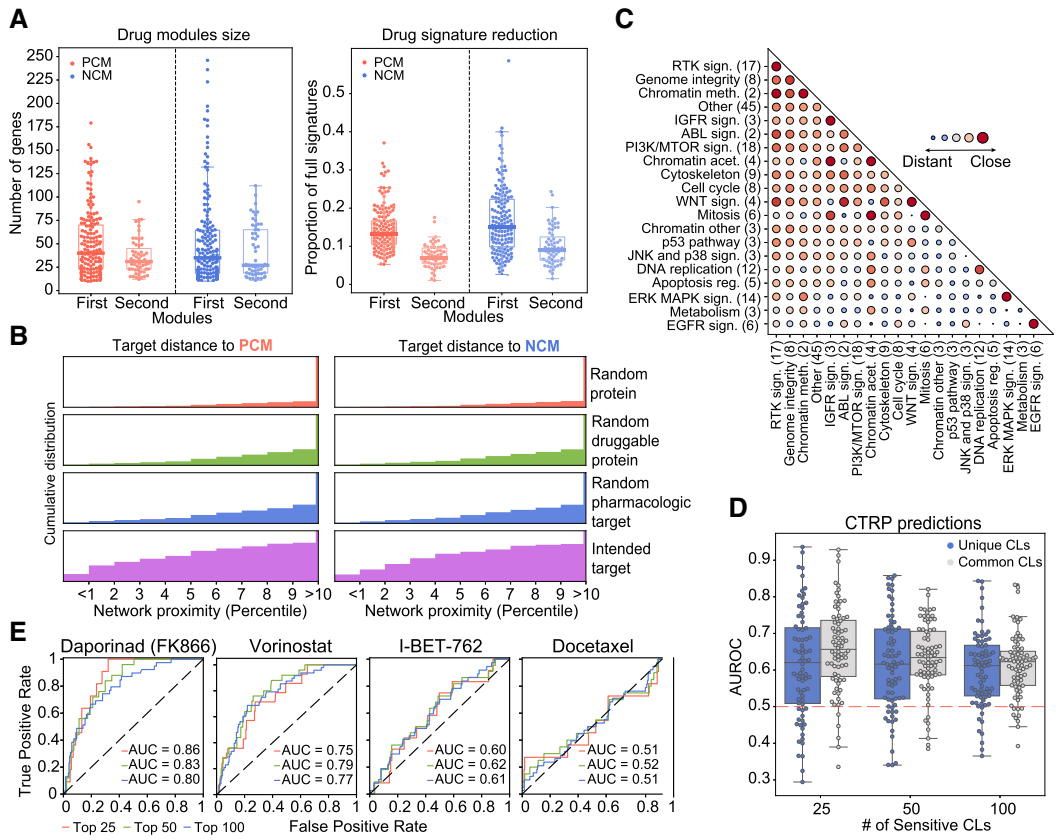


Fig. 3 Global drug module analysis. **a** Number of genes in positively and negatively correlated modules (PCMs and NCMs) (left). Proportion of genes in the modules with respect to PCGs and NCGs (i.e., full signature). **b** Distance between drug targets and PCMs/NCMs (purple cumulative distribution). Results are compared to random proteins from the STRING interactome (red), proteins sampled from the "druggable proteome" (Target Central Resource Database) (green), and proteins sampled from the pharmacological targets in DrugBank (blue). **c** Network-based distances between drug classes. The bigger the bubble, the closer the distance between drug classes. Drug classes are sorted by specificity in their proximity measures. Please note "distant" values in the diagonal are possible due to differences in drug modules belonging to the same class. The network-based distance calculation is detailed in the "Methods" section. **d** Predictive performances (AUROC) of the drug modules evaluated in the CTRP panel (top 25, 50, and 100 sensitive CCLs). Blue distributions correspond to results using unique CCLs (i.e., not shared with the GDSC panel). **e** Illustrative ROC curves for Daporinad (FMK866), Vorinostat, I-BET-762, and Docetaxel

more sophisticated "sensitive/resistant" categorizations such as the waterfall analysis [14], and it yields prediction performance metrics comparable between drugs.

Additional file 1: Figure S13 suggests that, when applied to the GDSC, drug module enrichment analysis can classify sensitive cell lines with high accuracy, especially for the top 25 sensitive cell lines (area under the ROC curve (AUROC) 0.77), which is a notable achievement considering that drug modules are 80% smaller than the original signatures. To assess the applicability of our modules outside the GDSC dataset, we performed an external validation with the CTRP panel of cell lines. About 37% of our drugs were also tested in this panel. In CTRP, drug sensitivity is measured independently of GDSC, which poses an additional challenge for prediction as a result of experimental inconsistencies [14]. Of the CCLs, 397 are shared between GDSC and CTRP, and gene expression data are also measured independently. We performed the GSEA-based sensitivity prediction for *all* CTRP CCLs. Figure 3d and e show the distribution of prediction performances for the 70 drugs, and illustrative ROC curves corresponding to four drugs (namely Daporinad, Vorinostat, I-BET-762 and Docetaxel), respectively. We found that, when focusing on the top 25 sensitive CCLs, over a quarter of the drugs had AUROC >0.7, including Daporinad. Acceptable

(AUROC > 0.6) predictions were achieved for half the cases (e.g., Vorinostat and I-BET-762), which is a comparable result to recent attempts to translate sensitivity predictors between different CCL panels [61]. For the remaining drugs, predictive performance did not differ to random expectation (AUROC < 0.6) (e.g., docetaxel). Notably, performance declined only slightly when considering CCLs that were *exclusive* to the CTRP panel (i.e., not part of the GDSC dataset) (Fig. 3d, blue boxes). The figure was comparable, if not better, to that obtained using full signatures (PCGs and NCGs) (Additional file 1: Figure S13, gray boxes). These observations support previous recommendations to pre-filter pharmacogenomic data based on prior knowledge [62] (Additional file 1: Figure S14).

Module-based characterization of drugs

Since drug modules are highly connected in biological networks, they are expected to be (at least to some extent) functionally coherent and easier to interpret. Accordingly, we tested the enrichment of drug modules in a collection of high-order biological processes (the Hallmark gene sets) available from the Molecular Signatures Database (MSigDB) [63]. Additional file 1: Figure S15A shows that the number of enriched Hallmark gene sets depends upon the MoA of the drug. The results of the enrichment analysis are given in Additional file 5 and as an interactive exploration tool based on the CLEAN methodology (Additional file 9; https://figshare.com/s/932dd94520d4a60f076d) [64]. We chose three drug classes to illustrate how to read these results, namely drugs targeting mitosis, RTK signaling inhibitors, and ERK-MAPK signaling inhibitors (Fig. 4a).

Drugs targeting mitosis have modules enriched in cell cycle and replication processes (Fig. 4a, top). Specifically, genes related to the Myc transcription factor are over-represented in three of the drug modules (NPK76-II-72-1, GSK1070916, and MPS-1-IN-1). The modules of these drugs have a rather distinct composition, NPK76-II-72-1 having the largest coverage of Myc-related genes and being, together with MPS-1-IN-1, related to both Myc1 and Myc2 processes. In Additional file 1: Figure S15B, we show how, for these two drugs, cell line sensitivity is dependent on Myc expression levels.

In contrast to mitosis inhibitors, drugs targeting the RTK pathway are enriched in biological processes outside the nucleus (Fig. 4a, middle), among these hypoxia and the epithelial-mesenchymal transition (EMT). Both mechanisms are known to be associated with tyrosine kinases [65, 66]. Interestingly, a subgroup of RTK inhibitors (namely ACC220, CEP-701, NVP-BHG712, and MP470) is characteristically associated with the PI3K-AKT-mTOR signaling cascade. With the exception of NVP-BHG712, these inhibitors have the tyrosine kinase FLT3 as a common target [67, 68]. Deeper inspection of FLT3 inhibitors reveals module proximities to certain PI3K inhibitors (e.g., GDC0941), and the PI3K-AKT-mTOR pathway is enriched in ERBB2 inhibitors as well (Additional files 4 and 5).

As for ERK-MAPK pathway inhibitors, we observed a total of 17 enriched Hallmarks, making this class of drugs the one with most variability in terms of enrichment signal of the modules (Fig. 4a, bottom; Additional file 1: Figure S15A). However, while some processes like apoptosis are detected in most of the drugs in this category, others are target-specific. Oxidative phosphorylation (OXPHOS), for example, is represented in 3 of the 4 BRAF inhibitors. It is known that, while BRAF inhibitors boost OXPHOS (leading to oncogene-induced senescence), activation of glycolytic metabolism followed by OXPHOS inactivation yields drug resistance [69, 70]. Similarly, VX11e (the only drug in our dataset targeting ERK2) shows a distinctive enrichment in Myc-regulated proteins, while FMK (the only drug targeting the Ribosomal S6 kinase) is enriched in p53 signaling pathway and inflammatory response processes. All these observations are consistent with previous studies [71–74], and Additional file 1: Figure S15C demonstrates that the variability observed between drugs in this class is driven mostly by differences in the sensitivity profiles of the drugs.

Overall, the enrichment signal (i.e., the functional coherence) of drug modules is substantially higher than that of full signatures (PCGs and NCGs) (Fig. 4b,c). This facilitates, in principle, the mechanistic interpretation of drug-gene correlation results (Additional file 1: Figure S15D). We show an illustrative module (CEP-701) in Fig. 4d.

We next examined whether our results could be extended beyond CCL panels. We found that drug modules are indeed identified (GSEA *P* value < 0.001) in the majority of patients in the TCGA clinical cohort (Additional file 1: Figure S15E; see the "Methods" section for details). Closer inspection by TCGA tumor type further supports the clinical relevance of our results (Additional file 6). For example, drugs affecting MAPK signaling (specifically, BRAF inhibitors, e.g., dabrafenib) have a tendency to "occur" in skin cutaneous melanomas (SKCM), as expected (Fig. 4e, blue). Of note, one PPAR inhibitor (FH535) was also found enriched in a high number of SKCM patients, in good agreement with work by others proposing the use PPAR inhibitors to treat skin cancer [75, 76]. Similarly, we observed an abundance of EGFR inhibitor modules among pancreatic cancers (PAAD) (Fig. 4e, green), in line with the known crucial role of EGFR in pancreatic tumorigenesis [77, 78]. As for gliobastomas (GBMs) (Fig. 4e, purple), we found two GSK3 inhibitors (CHIR-99021 and SB216763) and one TNKS inhibitor (XAV939), all of them targeting

Fig. 4 Drug module characterization. **a** Drug module enrichment analysis based on the Hallmark gene set (odds ratios in color, p values < 0.05 are marked with a white dot). For simplicity, three drug classes are shown: drugs affecting mitosis, RTK signaling, and ERK/MAPK signaling. **b** Distribution of the enrichment scores in the Hallmark collection gene sets. Overall, higher enrichment scores are obtained using modules than using full signatures (PCGs and NCGs) (the gene universe used here is that of Reactome). **c** Similarly, number of Hallmark-drug pairs at different enrichment scores. We show the pairs found only with the modules (PCMs and NCMs, red), only with correlations (PCGs and NCGs, blue), or in both (orange). **d** A view of Lestaurtinib (CEP-701) module. For illustrative purposes, two out-of-the-module (non-correlated) proteins are shown (gray), one being very central and one being peripheral but acting as a "bridging" node. **e** TCGA enrichment analysis of PCMs in four types of cancer: SKCM (Cutaneous Melanoma), PAAD (Pancreatic Adenocarcinoma), GBM (Glioblastoma Multiforme), and BRCA (Breast Carcinoma)

WNT signaling, which is a potential mechanism against this tumor type [79]. We also found one EGFR inhibitor (Gefitinib) and the PLK inhibitor NPK76-II-72-1 mentioned above in the context of Myc enrichment analysis. Both mechanisms have shown promise in EGFR- and Myc-activated gliomas, respectively [80, 81]. Finally, we encountered a more heterogeneous pattern in breast cancer patients (BRCA) (Fig. 4e, orange), including mechanisms supported by the literature, such as AKT, IRAK1, and PLK3 inhibition [82–84].

Beyond the tumor-type level, we looked for modules that were significantly enriched (odds ratio > 2, P value < 0.001) in patients harboring specific driver mutations (see the "Methods" section). A full account of this enrichment analysis is given in Additional file 7. We found, for instance, that modules of drugs targeting ERK/MAPK signaling are related to patients with mutations in HRAS and BRAF [85, 86] and that, in turn, BRAF is (together with KRAS) frequently mutated in patients "expressing" modules of EGFR signaling inhibitors [87]. Taken together, and although TCGA treatment response data is too scarce to allow for prediction assessment [88], these results indicate that the drug modules identified in CCLs hold promise for translation to clinical practice.

Conclusions

Two limitations of large-scale pharmacogenomic studies are the difficulty to reproduce results across screening

platforms and the eventual translation to clinics, as it remains unclear whether immortalized cells are able to model patient samples [89]. Another important limitation is the overwhelming number of drug-gene correlations that can be derived from these experiments, yielding signatures of drug sensitivity that are almost impossible to interpret. We have shown, for example, that (i) the number of correlated genes is highly variable across drugs, (ii) some genes are unspecifically correlated to many drugs, and (iii) not all drug-gene pairs are equally correlated in every tissue. We propose that converting transcriptional signatures to network modules may simplify the analysis, since network modules are smaller, more robust, and functionally coherent. We have validated this strategy by proving that drug response modules, which are enriched in biological processes of pharmacological relevance and exhibit comparable predictive power to the full signatures, are tightly related to the MoA. Further, we have characterized the modules extensively (Additional file 8 and e.g., Additional file 1: Figure S16) and confirmed their occurrence in the TCGA clinical cohort (Additional file 6 and Additional file 10).

However, our approach does have some of the limitations of ordinary transcriptomic analyses. Expression levels of mRNA do not perfectly match protein abundance, nor are they able to capture post-translational modifications such as phosphorylation events, which are key to some of the pathways studied here. Moreover, wide dynamic ranges in gene expression and drug sensitivity data are necessary for drug-gene correlations to be captured, thus requiring, in practice, considerably large panels of CCLs, which limits the throughput of the technique to a few hundred drugs. In particular, one cannot precisely measure correlations within poorly represented tissues, which in turn makes it difficult to disentangle tissue-specific transcriptional traits that may be irrelevant to drug response. Our module-based approach partially corrects for this confounding factor, although the integration of other CCL omics data (such as mutations, copy number variants and chromatin modifications) could further ameliorate these issues and also provide new mechanistic insights. In this context, systems biology tools that learn the relationships between different layers of biology are needed. Along this line, the release of CCL screens with readouts other than growth inhibition or proliferation rate [90, 91] will help unveil the connections between the genetic background of the cells and the phenotypic outcome of drug treatment.

All in all, transcriptomics is likely to remain the dominant genome-wide data type for drug discovery, as recent technical and statistical developments have drastically reduced its cost [92]. The L1000 Next-generation Connectivity Map, for instance, contains about one million post-treatment gene expression signatures for 20,000 molecules [90]. These signatures await to be interpreted and annotated, and more importantly, they have to be associated with pre-treatment signatures in order to identify therapeutic opportunities. We believe that network biology strategies like the one presented here will enable this connection, encircling relevant "regions" of the signatures and measuring the distances between them.

Additional files

Additional file 1: Contains supplementary figures 1–16. (PDF 2107 kb)

Additional file 2: Collection of drug modules in GMT format. The first column indicates the name of the drug while the second column indicates whether the module is a secondary module ("second_module") or not ("na"). From the third column onwards, there are the genes composing the module (gene names). (XLSX 219 kb)

Additional file 3: Drug module-gene correlations across tissues. (XLSX 2823 kb)

Additional file 4: Pair-wise distances between drug modules. Network distances (d_{net}) are normalized (*z*-scores): negative values denote proximity. Secondary modules receive with the suffix "_md2". See the "Methods" section for a detailed explanation of the network distance measurement. (XLSX 1742 kb)

Additional file 5: Enrichment scores and *p* values between drug modules (rows) and Hallmark gene sets (columns). For simplicity, secondary modules were merged with the main ones. (XLSX 453 kb)

Additional file 6: Enriched (p value < 0.001) drug module count across 31 TCGA cancer types, i.e., number of patients where each module is "expressed". (XLSX 79 kb)

Additional file 7: Cancer driver and drug module associations (OR > 2, p value < 0.001), based on patients "expressing/not-expressing" a module and "having/not-having" a driver mutation in the TCGA cohort. (XLSX 56 kb)

Additional file 8: We have chosen 21 features from network-based measures and other functional data: (i) General features *(columns 2–9)*. They illustrate basic and general features derived from the omics panel. We provide, for instance, the number of genes in each module, the average correlation among module genes and a measure of how "unique" are those genes with respect to other modules. Besides, we annotate drug classes and the AUROC predictions in both the GDSC and CTRP panels. (ii) Network features *(columns 10–12)*. These include distances between module genes and drug targets, "connectivity" within module genes (i.e., distance between them), and proximity to genes from other modules. (iii) Biological features *(columns 12–21)*. A series of biological features related to drug biology. We give most of them as simple proportions of genes/proteins. Among others, we provide the cellular compartmentalization of the genes, transcription factor specificity and the proportion of disease-related and "druggable" genes inside the module. Annotated drug targets are listed as well. (XLSX 156 kb)

Additional file 9: CLEAN cluster results using drug module genes and Hallmark gene sets. We provide an additional table with the significant associations between gene clusters and hallmark gene sets. Compressed folder (ZIP). The file can be found at https://figshare.com/s/932dd94520d4a60f076d (ZIP 3220 kb)

Additional file 10: Peer Review file. (PDF 1933 kb)

Acknowledgements
The authors would like to thank Prof. Ben Raphael and Dr. Matthew Reyna (Princeton University) for guidance in the use of HotNet2.

Funding
A.F-T. is a recipient of an FPI fellowship. P.A. acknowledges the support of the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R) and the European Research Council (SysPharmAD: 614944).

Availability of data and materials
Gene expression, drug response and drug target datasets are available in the cancerxxgene resource (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html). External validation data for drug response predictions are available in the ctd2 repository (https://ocg.cancer.gov/programs/ctd2/data-portal). Hallmark gene sets were obtained from the Molecular Signature Database (MSigDB) (http://software.broadinstitute.org/gsea/index.jsp). TCGA gene expression data were downloaded from cbioportal (http://www.cbioportal.org; "PanCancer Atlas" flag). TCGA cancer drivers are available in the OncoGenomic Landscapes resource (https://oglandscapes.irbbarcelona.org/). All data generated or analyzed during this study are included in this published article and its additional files. Code for the module detection pipeline is available at https://github.com/sbnb-lab-irb-barcelona/GDSC-drug-modules.

Authors' contributions
AF-T, MD-F, and PA conceived the study. AF-T and MD-F performed the analyses. AF-T, MD-F, and PA interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.

Received: 17 October 2018 Accepted: 5 March 2019
Published online: 26 March 2019

References

1. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. Nat Rev Genet. 2007;8(8):601.
2. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10(6):428–38.
3. Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov. 2011;10(7):507–19.
4. Monks A, Scudiero D, Skehan P, Shoemaker R, Paull K, Vistica D, Hose C, Langley J, Cronise P, Vaigro-Wolff A. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. J Natl Cancer Inst. 1991;83(11):757–66.
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
6. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(D1):D955–61.
7. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, Jones V, Bodycombe NE, Soule CK, Gould J, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. 2015;5(11):1210–23.
8. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54.
9. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12(2):109–16.
10. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, Bhutra S, Grossman RL, Stephanie Huang R. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Res. 2017;27(10):1743–51.
11. Lee S-I, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun. 2018;9(1):42.
12. Liu X, Yang J, Zhang Y, Fang Y, Wang F, Wang J, Zheng X, Yang J. A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. Sci Rep. 2016;6:22811.
13. Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, Chung M, Lu X, Stuart JM, Golub TR, Subramanian A, et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nat Commun. 2017;8(1):1186.
14. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, Quackenbush J. Inconsistency in large pharmacogenomic studies. Nature. 2013;504(7480):389–93.
15. Stransky NaG M, Kryukov GV, Garraway LA, Lehar J, Liu M, Sonkin D, Kauffmann A, Venkatesan K, Edelman EJ, Riester M, Barretina J, Caponigro G, Schlegel R, Sellers WR, Stegmeier F, Morrissey M, Amzallag A, Pruteanu-Malinici I, Haber DA, Ramaswamy S, Benes CH, Menden MP, Iorio F, Stratton MR, McDermott U, Garnett MJ, Saez-Rodriguez J. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528(7580):84–7.
16. Geeleher PaG ER, Seoighe C, Cox NJ, Huang RS. Consistency in large pharmacogenomic studies. Nature. 2016;540(7631):E1–2.
17. Garnett MJaE EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson JA, Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, Zhang T, O'Brien P, Boisvert JL, Price S, Hur W, Yang W, Deng X, Butler A, Choi HG, Chang JW, Baselga J, Stamenkovic I, Engelman JA, Sharma SV, Delattre O, Saez-Rodriguez J, Gray NS, Settleman J, Futreal PA, Haber DA, Stratton MR, Ramaswamy S, McDermott U, Benes CH. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483:570–5.
18. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009;461(7261):218–23.
19. Kim Y-A, Przytycka TM. Bridging the gap between genotype and phenotype via network approaches. Front Genet. 2012;3:227.
20. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761supp):C47–52.
21. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18(3):507–22.
22. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016;48(8):838–47.
23. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
24. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197–206.
25. Dančík V, Carrel H, Bodycombe NE, Seiler KP, Fomina-Yadlin D, Kubicek ST, Hartwell K, Shamji AF, Wagner BK, Clemons PA. Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. J Biomol Screen. 2014;19(5):771–81.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet 2000, 25(1):25–29.
27. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–72.
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.

29. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11(4):e1004120.
30. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.
31. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106–14.
32. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2016;45(D1):D362–8.
33. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. Nature Publishing Group. 2018;17:317.
34. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
35. Freeman LC. Centrality in social networks conceptual clarification. Soc Networks. 1978;1(3):215–39.
36. Mateo L, Guitart-Pla O, Duran-Frigola M, Aloy P. Exploring the OncoGenomic Landscape of cancer. Genome Medicine. 2018;10(1):61.
37. Smirnov PaK V, Maru A, Freeman M, Ho C, El-Hachem N, Adam GA, Ba-Alawi W, Safikhani Z, Haibe-Kains B. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res. 2018; 46(D1):D994–D1002.
38. Gilcrease MZ. Integrin signaling in epithelial cells. Cancer Lett. 2007;247(1):1–25.
39. Givant-Horwitz V, Davidson B, Reich R. Laminin-induced signaling in tumor cells. Cancer Lett. 2005;223(1):1–10.
40. Keely PJ, Westwick JK, Whitehead IP, Der CJ, Parise LV. Cdc42 and Rac1 induce integrin-mediated cell motility and invasiveness through PI(3)K. Nature. 1997;390(6660):632–6.
41. Gelderloos JA, Rosenkranz S, Bazenet C, Kazlauskas A. A role for Src in signal relay by the platelet-derived growth factor α receptor. J Biol Chem. 1998; 273(10):5908–15.
42. Patarroyo M, Tryggvason K, Virtanen I. Laminin isoforms in tumor invasion, angiogenesis and metastasis. Semin Cancer Biol. 2002;12(3):197–207.
43. Keely P, Parise L, Juliano R. Integrins and GTPases in tumour cell growth, motility and invasion. Trends Cell Biol. 1998;8(3):101–6.
44. Kim S-H, Turnbull J, Guimond S. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. J Endocrinol. 2011;209(2):139–51.
45. Du J, Zu Y, Li J, Du S, Xu Y, Zhang L, Jiang L, Wang Z, Chien S, Yang C. Extracellular matrix stiffness dictates Wnt expression through integrin pathway. Sci Rep. 2016;6:20395.
46. Moreno-Layseca P, Streuli CH. Signalling pathways linking integrins with cell cycle progression. Matrix Biol. 2014;34:144–53.
47. Lee CA, Neul D, Clouser-Roche A, Dalvie D, Wester MR, Jiang Y, Jones JP 3rd, Freiwald S, Zientek M, Totah RA. Identification of novel substrates for human cytochrome P450 2J2. Drug Metab Dispos. 2010;38(2):347–56.
48. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138(1):103–41.
49. Allocati N, Masulli M, Di Ilio C, Federici L. Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. Oncogenesis. 2018;7(1):8.
50. Hayes JD, Flanagan JU, Jowsey IR. Glutathione transferases. Annu Rev Pharmacol Toxicol. 2005;45:51–88.
51. Green MD, Francis PA, Gebski V, Harvey V, Karapetis C, Chan A, Snyder R, Fong A, Basser R, Forbes JF, et al. Gefitinib treatment in hormone-resistant and hormone receptor-negative advanced breast cancer. Ann Oncol. 2009; 20(11):1813–7.
52. Zhang X, Zhang B, Liu J, Liu J, Li C, Dong W, Fang S, Li M, Song B, Tang B, et al. Mechanisms of Gefitinib-mediated reversal of tamoxifen resistance in MCF-7 breast cancer cells by inducing ERα re-expression. Sci Rep. 2015;5:7835.
53. Huguet F, Fernet M, Giocanti N, Favaudon V, Larsen AK. Afatinib, an irreversible EGFR family inhibitor, shows activity toward pancreatic cancer cells, alone and in combination with radiotherapy, independent of KRAS status. Target Oncol. 2016;11(3):371–81.
54. Ioannou N, Dalgleish AG, Seddon AM, Mackintosh D, Guertler U, Solca F, Modjtahedi H. Anti-tumour activity of afatinib, an irreversible ErbB family blocker, in human pancreatic tumour cells. Br J Cancer. 2011;105(10):1554–62.
55. Mahajan K, Mahajan NP. Cross talk of tyrosine kinases with the DNA damage signaling pathways. Nucleic Acids Res. 2015;43(22):10588–601.
56. Chen M-K, Hung M-C. Regulation of therapeutic resistance in cancers by receptor tyrosine kinases. Am J Cancer Res. 2016;6(4):827–42.
57. Ish-Shalom D, Christoffersen CT, Vorwerk P, Sacerdoti-Sierra N, Shymko RM, Naor D, DeMeyts P. Mitogenic properties of insulin and insulin analogues mediated by the insulin receptor. Diabetologia. 1997;40(Suppl 2):25–31.
58. Borisov N, Aksamitiene E, Kiyatkin A, Legewie S, Berkhout J, Maiwald T, Kaimachnikov NP, Timmer J, Hoek JB, Kholodenko BN. Systems-level interactions between insulin–EGF networks amplify mitogenic signaling. Mol Syst Biol. 2009;5(1):256.
59. Mairet-Coello G, Tury A, DiCicco-Bloom E. Insulin-like growth factor-1 promotes G(1)/S cell cycle progression through bidirectional regulation of cyclins and cyclin-dependent kinase inhibitors via the phosphatidylinositol 3-kinase/Akt pathway in developing rat cerebral cortex. J Neurosci. 2009; 29(3):775.
60. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014;15(3):R47.
61. Juan-Blanco T, Duran-Frigola M, Aloy P. Rationalizing drug response in cancer cell lines. J Mol Biol. 2018.
62. Ferranti D, Krane D, Craft D. The value of prior knowledge in machine learning of complex network systems. Bioinformatics. 2017;33(22):3610–8.
63. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Systems. 2015;1(6):417–25.
64. Freudenberg JM, Joshi VK, Hu Z, Medvedovic M. CLEAN: CLustering Enrichment ANalysis. BMC Bioinformatics. 2009;10(1):234.
65. Glück AA, Aebersold DM, Zimmer Y, Medová M. Interplay between receptor tyrosine kinases and hypoxia signaling in cancer. Int J Biochem Cell Biol. 2015;62:101–14.
66. Thiery JP. Epithelial–mesenchymal transitions in development and pathologies. Curr Opin Cell Biol. 2003;15(6):740–6.
67. Lindblad O, Cordero E, Puissant A, Macaulay L, Ramos A, Kabir NN, Sun J, Vallon-Christersson J, Haraldsson K, Hemann MT, et al. Aberrant activation of the PI3K/mTOR pathway promotes resistance to sorafenib in AML. Oncogene. 2016;35(39):5119–31.
68. Nogami A, Oshikawa G, Okada K, Fukutake S, Umezawa Y, Nagao T, Kurosu T, Miura O. FLT3-ITD confers resistance to the PI3K/Akt pathway inhibitors by protecting the mTOR/4EBP1/Mcl-1 pathway through STAT5 activation in acute myeloid leukemia. Oncotarget. 2015;6(11):9189–205.
69. Haq R, Shoag J, Andreu-Perez P, Yokoyama S, Edelman H, Rowe GC, Frederick DT, Hurley AD, Nellore A, Kung AL, et al. Oncogenic BRAF regulates oxidative metabolism via PGC1α and MITF. Cancer Cell. 2013;23(3):302–15.
70. Haq R, Fisher DE, Widlund HR. Molecular pathways: BRAF induces bioenergetic adaptation by attenuating oxidative phosphorylation. Clin Cancer Res. 2014;20(9):2257–63.
71. Amatangelo MD, Goodyear S, Varma D, Stearns ME. c-Myc expression and MEK1-induced Erk2 nuclear localization are required for TGF-beta induced epithelial–mesenchymal transition and invasion in prostate cancer. Carcinogenesis. 2012;33(10):1965–75.
72. Marampon F, Ciccarelli C, Zani BM. Down-regulation of c-Myc following MEK/ERK inhibition halts the expression of malignant phenotype in rhabdomyosarcoma and in non muscle-derived human tumors. Mol Cancer. 2006;5:31.
73. Moens U, Kostenko S, Sveinbjørnsson B. The role of mitogen-activated protein kinase-activated protein kinases (MAPKAPKs) in inflammation. Genes. 2013;4(2):101–33.
74. Cho Y-Y, He Z, Zhang Y, Choi HS, Zhu F, Choi BY, Kang BS, Ma W-Y, Bode AM, Dong Z. The p53 protein is a novel substrate of ribosomal S6 kinase 2 and a critical intermediary for ribosomal S6 kinase 2 and histone H3 interaction. Cancer Res. 2005;65(9):3596–603.
75. Schadendorf D. Peroxisome proliferator-activating receptors: a new way to treat melanoma? J Investig Dermatol. 2009;129(5):1061–3.
76. Borland MG, Kehres EM, Lee C, Wagner AL, Shannon BE, Albrecht PP, Zhu B, Gonzalez FJ, Peters JM. Inhibition of tumorigenesis by peroxisome

proliferator-activated receptor (PPAR)-dependent cell cycle blocks in human skin carcinoma cells. Toxicology. 2018;404-405:25–32.
77. Ardito Christine M, Grüner Barbara M, Takeuchi Kenneth K, Lubeseder-Martellato C, Teichmann N, Mazur Pawel K, DelGiorno KE, Carpenter Eileen S, Halbrook Christopher J, Hall Jason C, et al. EGF receptor is required for KRAS-induced pancreatic tumorigenesis. Cancer Cell. 2012;22(3):304–17.
78. Tzeng C-WD, Frolov A, Frolova N, Jhala NC, Howard JH, Buchsbaum DJ, Vickers SM, Heslin MJ, Arnoletti JP. Epidermal growth factor receptor (EGFR) is highly conserved in pancreatic cancer. Surgery. 2007;141(4):464–9.
79. Lee Y, Lee J-K, Ahn SH, Lee J, Nam D-H. WNT signaling in glioblastoma and therapeutic opportunities. Lab Investig. 2015;96:137.
80. Rao SK, Edwards J, Joshi AD, Siu IM, Riggins GJ. A survey of glioblastoma genomic amplifications and deletions. J Neuro-Oncol. 2010;96(2):169–79.
81. Higuchi F, Fink AL, Kiyokawa J, Miller JJ, Koerner MVA, Cahill DP, Wakimoto H. PLK1 inhibition targets Myc-activated malignant glioma cells irrespective of mismatch repair deficiency–mediated acquired resistance to temozolomide. Mol Cancer Ther. 2018;17(12):2551.
82. Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. Ther Adv Med Oncol. 2014;6(4):154–66.
83. Wee ZN, Yatim SMJM, Kohlbauer VK, Feng M, Goh JY, Bao Y, Lee PL, Zhang S, Wang PP, Lim E, et al. IRAK1 is a therapeutic target that drives breast cancer metastasis and resistance to paclitaxel. Nat Commun. 2015;6:8746.
84. Fallah Y, Brundage J, Allegakoen P, Shajahan-Haq AN. MYC-Driven Pathways in Breast Cancer Subtypes. Biomolecules. Multidisciplinary Digital Publishing Institute. 2017;7:53.
85. Van Aelst L, Barr M, Marcus S, Polverino A, Wigler M. Complex formation between RAS and RAF and other protein kinases. Proc Natl Acad Sci. 1993; 90(13):6213.
86. Moodie SA, Willumsen BM, Weber MJ, Wolfman A. Complexes of Ras.GTP with Raf-1 and mitogen-activated protein kinase kinase. Science. 1993; 260(5114):1658.
87. Fitzgerald TL, Lertpiriyapong K, Cocco L, Martelli AM, Libra M, Candido S, Montalto G, Cervello M, Steelman L, Abrams SL, et al. Roles of EGFR and KRAS and their downstream signaling pathways in pancreatic cancer and pancreatic cancer stem cells. Adv Biol Regul. 2015;59:65–81.
88. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al. An integrated TCGA Pan-Cancer Clinical Data Resource to drive high-quality survival outcome analytics. Cell. 2018;173(2):400–416.e411.
89. Jaeger S, Duran-Frigola M, Aloy P. Drug sensitivity in cancer cell lines is not tissue-specific. Mol Cancer. 2015;14(1):40.
90. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–1452.e1417.
91. Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, Sokolnicki KL, Bray M-A, Kemp MM, Winchester E, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc Natl Acad Sci U S A. 2014;111(30):10911–6.
92. Frumkin I, Schirman D, Rotman A, Li F, Zahavi L, Mordret E, Asraf O, Wu S, Levy SF, Pilpel Y. Gene architectures that minimize cost of gene expression. Mol Cell. 2017;65(1):142–53.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

Check for updates

Received: 12 September 2018 | Revised: 30 November 2018 | Accepted: 13 December 2018

DOI: 10.1002/wcms.1408

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

ADVANCED REVIEW

Formatting biological big data for modern machine learning in drug discovery

Miquel Duran-Frigola[1] | Adrià Fernández-Torras[1] | Martino Bertoni[1] | Patrick Aloy[1,2]

[1]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Correspondence
Miquel Duran-Frigola, Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain.
Email: miquel.duran@irbbarcelona.org
Patrick Aloy, Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain.
Email: patrick.aloy@irbbarcelona.org

Funding information
H2020 European Research Council, Grant/Award Number: SysPharmAD: 614944; Spanish Ministerio de Economía y Competitividad, Grant/Award Number: BIO2016-77038-R

Biological data is accumulating at an unprecedented rate, escalating the role of data-driven methods in computational drug discovery. This scenario is favored by recent advances in machine learning algorithms, which are optimized for huge datasets and consistently beat the predictive performance of previous art, rapidly approaching human expert reasoning. The urge to couple biological data to cutting-edge machine learning has spurred developments in data integration and knowledge representation, especially in the form of heterogeneous, multiplex and semantically-rich biological networks. Today, thanks to the propitious rise in knowledge embedding techniques, these large and complex biological networks can be converted to a vector format that suits the majority of machine learning implementations. Here, we explain why this can be particularly transformative for drug discovery where, for decades, customary chemoinformatics methods have employed vector descriptors of compound structures as the standard input of their prediction tasks. A common vector format to represent biology and chemistry may push biological information into most of the existing steps of the drug discovery pipeline, boosting the accuracy of predictions and uncovering connections between small molecules and other biological entities such as targets or diseases.

This article is categorized under:
Computer and Information Science > Databases and Expert Systems
Computer and Information Science > Chemoinformatics

KEYWORDS
biological/chemical embeddings, biological signatures, biological similarity, machine learning

1 | INTRODUCTION

The deluge of molecular biology data that followed the sequencing of the human genome, almost two decades ago, has dramatically increased the complexity of biomedical research. The growth of biological databases is steeper than ever before,[1,2] being virtually every scientific paper supplemented with large data tables of experimental measurements. The cost of "omics" techniques such as exome sequencing outpaces the Moore's law,[3] and the repertoire of possible read-outs spans all levels of biology, from mutations in the DNA to epigenetics modifications, from mRNA expression to protein abundance, or from fluxes of metabolites to phosphorylation signaling cascades.[4]

Just like any other great technological breakthrough, "omics" platforms have trailed a hype cycle, first with inflated expectations, followed by disillusionment[5] and finally reaching mainstream adoption and realistic ambitions.[6,7] Systems biology (the main beneficiary of the post-genomic era) is now a mature discipline, with a solid community and a unique ability to interact with other scientific areas, ranging from evolutionary biology[8] to bed-side research.[9] Drug discovery, in particular,

 © 2018 Wiley Periodicals, Inc.

17509084, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

did put high hopes in the systems view of pharmacology. Disease etiology and treatment are extremely intricate processes, involving the interplay between a drug molecule and a very dynamic network of proteins, many times across several tissues with distinct characteristics and contexts. The promise of systems biology is, precisely, to connect phenotypes to convoluted molecular events, hence identifying the ideal intervention points to disrupt or ameliorate a disease process. In other words, systems approaches are expected to reconcile the two main traditions of drug discovery,[10] namely the phenotype-centered view that dominated pharmacology in the early days, and the target-centered view that took off after the molecular biology revolution in the 1980s and confides in the "one drug–one gene–one disease" paradigm.[11]

Despite its great potential, though, for the most part systems biology remains a *descriptive* discipline. Efforts so far have been put towards discovering associations between biological entities such as genes and diseases, drafting an architecture (a network) of statistical and physical interactions, but lacking awareness of causality events and dynamic response to perturbations. Constraint- and logic-based models[12,13] are committed to turning biological networks into *predictive* tools. However, these techniques only work well in controlled and relatively small biological systems such as genome-scale metabolic reconstructions[14,15] or certain signaling cascades.[16] The complexity of entire cells and organisms is still unattainable, requiring modular approaches on almost-complete and experimentally parametrized networks.[17] The biological information that is available in the databases does not adhere to these standards, as it comes from hundreds of different sources, each of them having peculiar data types and tackling concrete scientific questions with specific experimental conditions. The human protein–protein interactome, for example, is only ~10–30% complete[18] and contains interactions of various qualities, merged over a wide range of affinities and time-scales, and not being relevant to every cell type and tissue.[19]

Seduced by the impressive achievements of machine (deep) learning, especially in the fields of natural language processing and image recognition, some computational biologists are considering a shift towards less mechanistic, more data-driven predictors of biology.[20–22] Deep learning algorithms are data hungry, requiring millions of training samples and fair amounts of labeled data. It has been argued that, in many areas of biology, data is not "big enough" to fully exploit deep learning algorithms,[23] although previsions are that within the next decade sequencing data alone well equal, or even surpass, other big data archives such as social media or online videos.[24] This anticipates a central role of data-intensive algorithms in the near future of biomedicine, which poses a number of challenges, starting with the cost and infrastructure that is needed to store, process and share the information.[25] Another urgent matter is to correctly format biological data so that deep learning algorithms that were developed to handle text and image inputs can be smoothly transferred to systems biology tasks. The nature of biological data is considerably more complex than in the other big data fields. Dealing with diversity, inconsistency and incompleteness, among other issues, demands heavy specialist processing, hampering widespread adoption of deep learning by uninitiated researchers working on disease biology and drug discovery. Here, we discuss recent advances in knowledge representation of genuinely heterogeneous datasets, and explain how they can offer a generic and intuitive means to bridge the gap between biological big data repositories and state-of-the-art machine-learning tools.

2 | LESSONS FROM CHEMOINFORMATICS

Chemoinformatics is the branch of computer science devoted to the extraction and extrapolation of meaningful patterns from small molecule structures. Chemoinformatics was born shortly after bioinformatics, more than half a century ago, and the two fields have evolved rather independently.[26,27] While biologists primarily use computers to *understand* their systems, the major goal of chemoinformaticians is to *predict* active (hit) molecules from large collections of candidate compounds, and then optimize their properties to achieve increased therapeutic activities and reduced toxicity risks (hit-to-lead).[28] Hence, chemoinformatics is mainly concerned with the predictive power of virtual screening and the efficiency of molecular design. Compared to biology, this has made the field more welcoming to mathematical abstraction, since explicit knowledge representation and mechanistic understanding are not indispensable requirements to endow correct predictions.[11]

At the heart of chemoinformatics there is the "similarity principle," that is, the notion that similar compounds tend to have similar bioactivities. Thus, the basic chemoinformatics predictor is a simple similarity search where a new molecule is assigned the bioactivity of its closest analogs. Over the years, this rationale has motivated the invention of chemical "descriptors" of the compounds so that they can be compared, searched and classified at large.[29–31] The assortment of molecular descriptors includes numerical arrays of physicochemical properties such as logP and molar refractivity, topological properties that can be calculated from two-dimensional (2D) graphical representations of the molecules, and pharmacophoric features extracted from three-dimensional (3D) structures. A widespread modality of descriptors is the so-called molecular "fingerprint," which encodes small molecule structures as a binary (1/0) vector denoting the presence/absence of certain molecular substructures. Modern fingerprints are a multiple of 8 bits long, usually between 128 and 4,096 bits, and can be used along the drug discovery pipeline to infer targets and off-targets,[32] anticipate clinical side effects[33] or identify new therapeutic indications for clinically safe compounds.[34,35]

17590984, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The numerical vector format of small molecule descriptors makes them a natural input for machine learning, which is required when naïve similarity searches are not sufficient to produce acceptable predictions. Practically every wave of machine learning algorithms has flooded chemoinformatics,[28] starting with simple methods such as linear-discriminant analysis and decision trees, and continuing to support-vector machines, Bayesian classifiers and ensemble approaches like random forests.[36] Thus, it is not surprising that deep learning algorithms quickly caught the attention of chemoinformaticians, especially now that the scale, growth and variety of chemical data exceed the capacity of classical machine learning techniques.[37]

Deep learning comprises stacked layers of simple (but nonlinear) processing units that, starting with the input, each transform the representation at one layer into a representation at a deeper, more abstract layer. Thus, deep learning is a representation learning approach that yields an *embedding* of the raw data. This is a very appealing property to chemoinformatics, because it does not constrain predictive models to a predefined set of descriptors, and instead allows for descriptors (embeddings) to be learned automatically during the training.[38] As a result, SMILES strings,[39] 2D structural graphs[40] or even image drawings of the molecules[41] can be directly inputted to the neural networks, making the traditional feature selection process unnecessary[42] (Figure 1a). Using chemical embeddings obtained with a graph convolutional neural network, for example, the accuracy of predictions can be improved over using binary fingerprints and, more importantly, the influential substructures can be visualized to interpret and gain trust on the predictions.[43,44] Recently, it was shown that deep learning can be used in "low data" problems such as lead optimization, where enhanced analogs of hit compounds are sought with only a minimal amount of biological data available.[45] This was achieved by learning a refined similarity metric between the embeddings using a long short-term memory network. In the same vein, deep learning was used to predict drug–drug and drug–food interactions simply based on the names and structures of drugs and food constituents.[46]

Another remarkable application of deep learning in chemoinformatics is the generation of new chemical entities (Figure 1b).[47] Using variational autoencoders, it was possible to learn embeddings by simply reading the SMILES strings that are stored in a large compound repository (ZINC).[48] Then, these embeddings were used to reversibly generate novel and valid SMILES strings through the trained autoencoder.[49] Moreover, in a follow-up study, the autoencoder was coupled to another generative network to invent molecules with a desired anticancer activity.[50] Similarly, focused chemical libraries against *Plasmodium falciparum* (malaria) and *Staphylococcus aureus* were produced using a recurrent neural network pretrained on 1.4M molecules and fine-tuned only with ~1,000 compounds screened against each of the pathogens.[51] Other examples of de novo design of small molecules include the optimization for activity against DRD2[52] and JAK2.[53] Of note, this line of research gains yet more interest given the outstanding performance of a deep neural network trained on essentially every known

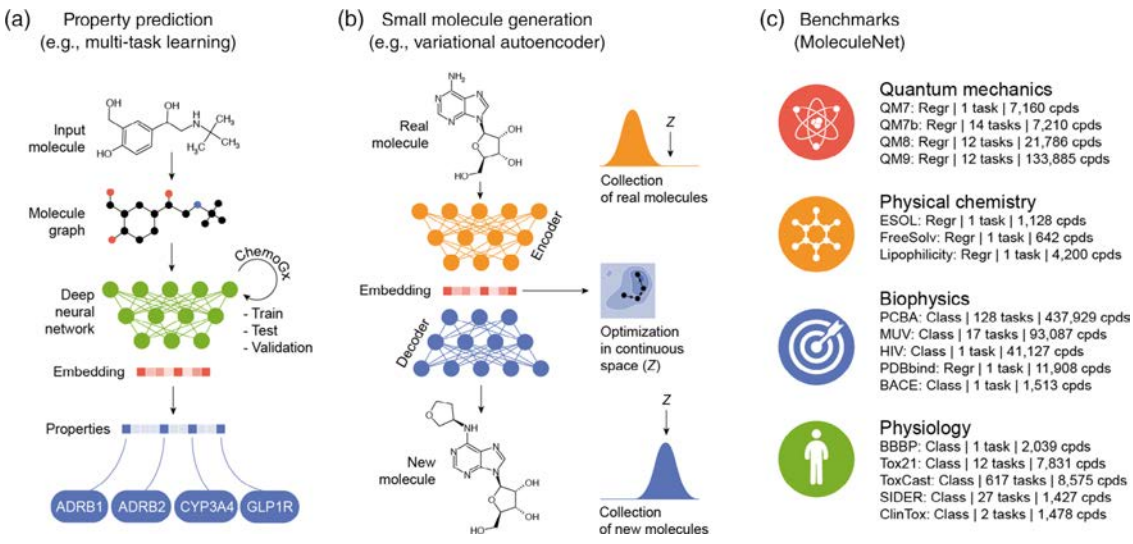


FIGURE 1 Deep learning in chemoinformatics. (a) A classical multitarget prediction exercise based on chemogenomics (ChemoGx) data. Deep neural networks can read a molecule structure as a graph (e.g., convolutional graph networks), and be trained to optimally perform a multitask classification. An inner (usually the last) layer of the network corresponds to the chemical embedding. (b) An autoencoder is a type of neural network that includes an encoder and a decoder, compressing and decompressing the data, respectively. The encoder maps the input to a latent space (embedding), and the decoder maps the embedding back to the original representation. The embedding is a continuous vector that can be optimized for a certain property of interest "*Z*". The interpolated vectors can be then decoded to generate new molecules. (c) MoleculeNet offers a number of benchmark datasets at different levels of resolution (from quantic properties to physiological properties of the molecules). For a brief explanation of the datasets, please visit http://moleculenet.ai/datasets-1

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

chemical reaction, being it able to automate retrosynthesis planning with a quality on par with peer-reviewed synthesis routes collected from the literature.[54]

The spectacular progress made in the fields of image recognition and natural language processing must be attributed not only to the advent of novel algorithms but also to the existence of benchmark datasets.[55] Well-curated and widely accepted gold standards constitute the best way to monitor progress, detect the limitations and, more importantly, identify significant improvements that will move the field in the right direction. Inspired by ImageNet[56] and WordNet,[57] MoleculeNet[58] was recently released, containing curated and diverse benchmark collections related to quantum mechanics, physicochemical, biophysical, and physiological properties of compounds (Figure 1c). In turn, MoleculeNet is integrated within DeepChem [deepchem.io], a toolchain that provides popular deep learning implementations with the aim of "democratizing" the use of high-quality algorithms in drug discovery.

3 | HETEROGENEOUS NETWORKS TO INTEGRATE ALL OF BIOLOGY

Computational biologists have to deal with datasets that are very different to chemoinformatics repositories. In chemical databases, there are millions of molecules with relatively poor annotations (i.e., the chemical structure and, eventually, some bioactivity records).[59,60] In contrast, biological databases annotate a relatively small set of biological entities (e.g., ~20,000 genes in human) with a comparatively large number of interactions between them[61,62] and associations to other biological entities such as diseases,[63,64] pathways,[65] molecular functions,[66] cells,[67] or tissues.[68] According to the 2018 report of the Molecular Biology Database Collection,[69] there are 1,737 online databases, spanning essentially every corner of biology.

Given the plethora of biological data sources, it would be useful to integrate "all" of them into a gigantic resource. While alluring, though, unifying the current biological knowledge implies a daunting effort, since data formats and identifiers need to be standardized,[70] and the process requires regular updates and is prone to legal tussles.[71] Recently, the Harmonizome was released[72] with the commitment of integrating datasets related to mammalian genes into a "harmonized" collection. As of August 2018, the Harmonizome centralizes 114 datasets provided by 66 online resources. About half of the repositories are from data-driven (high-throughput) studies, a third are from hypothesis-driven (low-throughput) studies, and the rest are from mixed sources. To build the Harmonizome, many choices were made concerning normalization methods and significance cut-offs, for example, of differential gene expression. In some cases, details had to be ignored such as the exact location of single nucleotide polymorphisms or binding sites proximal to a coding region, as well as the phosphorylation residues in a protein or the direct protein–protein contacts in a multimeric complex.[72] In practice, the Harmonizome publishes one processing script for each dataset, and simplifies the data to a list of relationships (a set of edges) denoting gene–gene and gene–attribute associations, where attributes are sequence features, cell lines, perturbation experiments, phenotypes, illnesses, drugs, etc. In total, the collection amounts to over 7 million edges. Thus, more than any other resource before, the Harmonizome testifies the original claim of network medicine, that is, that results of any biological experiment can be expressed as a graph, hence graphs are the best tool to obtain a "big picture" of disease biology.[73,74]

The Harmonizome also testifies that biomedical research is mostly gene-centric. Genes are connected between them and to many other biological entities (attributes), depending on the dataset. Ontologies provide a formal way of representing these biological entities, capturing their meaning with complicated hierarchies that consist of terms, relationships, and rules.[75,76] Again, the natural way of expressing these hierarchies is a graph, typically a directed acyclic graph that facilitates the browsing from specific ("leaf") terms to general ("root") concepts, and vice versa. In 2013, there were about 300 ontologies stored in the BioPortal,[77] and the number has more than doubled (722) ever since, amassing 95 billion direct annotations (bioportal.bioontology.org). Beyond the well-known Gene Ontology,[78] relevant controlled vocabularies for drug discovery are the disease,[79] the human phenotype,[80] the cell line,[81] the tissue,[82] the small molecule, and the bioassay[83] ontologies. The semantic knowledge contained in these ontologies has been complemented, in some cases, with further kinds of relationships between the terms, such as disease comorbidities,[84,85] pathway cross-talks[86] or genetic profile similarities between cell lines.[87,88]

Having every domain of biology expressed as a graph facilitates the interoperability between datasets and the merging of data from multiple sources. For example, gene–gene networks can be stacked in a multilayer (multiplex) network in which genes are connected through different types of pairwise edges such as mRNA co-expression, physical protein–protein interactions or cellular colocalization. This enables accurate assessment of the robustness[89] and redundancy[90] of biological systems, as well as detection of communities[91] and meaningful navigation across layers of regulation.[92] A successful and intuitive application of multilayer gene networks is PARADIGM,[93] a system that models the central dogma of biology (DNA–mRNA–protein) with multiple patient-specific "omics" measurements, and uses probabilistic inference to identify altered protein

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

activities in each patient. Another application regards the rewiring of protein–protein interactions in 107 human tissues by means of a multilevel interactome that was shown to capture tissue-specific functions of the proteins.[94]

Evidently, nodes other than genes can be conjoined with the above gene/protein-centric interactomes to obtain heterogeneous (multimodal) networks. A classical type of heterogeneous network in drug discovery is the bimodal graph comprising drug–drug similarities, protein–protein interactions and drug–target interactions, which have been widely used to identify the network-topological properties of successful drug targets[95,96] and discover new target classes.[97] A third type of node, namely diseases, is typically added to drug–protein networks, inserting disease genetics associations, drug indications and, occasionally, similarities between diseases based on, for example, shared phenotypes. Different flavors of the drug–protein–disease triad have shown power to pinpoint drug repositioning opportunities[34,98,99] and anticipate adverse drug events.[100] Recently, after a formidable knowledge integration effort, the Hetionet was presented,[101] building upon the previous networks and drastically augmenting them with transcriptomics, anatomical, and ontological knowledge. The Hetionet contains 2,250,197 bona fide connections between 47,031 nodes of 11 types, legitimating it as the largest (public) heterogeneous graph of biomedicine. Conveniently, the Hetionet is released as an easy-to-visualize graph database that offers seamless ease when querying several types of interactions (Box 1). To illustrate the features of heterogeneous networks, in Figure 2 we display an in-house version of the Hetionet, complemented with data from the Harmonizome.

BOX 1

GRAPH DATABASES

In classical relational databases, connectivity between two data tables is achieved using foreign-key references of columns, usually specified in a third pairwise table. The relational structure is suboptimal for biomedical applications, where relations between biological entities are the essential feature, and predefined rigid constraints such as column types are less necessary and often bothersome, given the diversity of data available. Instead, graph databases focus on the relationships (edges) between instances (nodes), and allow for flexible specification of node and edge attributes, which makes them a more suited data structure to store and operate on biomedical data.[102–104] The favorite graph database in biology is Neo4j, which has been shown to systematically outperform relational databases (e.g., MySQL) in a series of complex queries performed on heterogeneous data.[105]

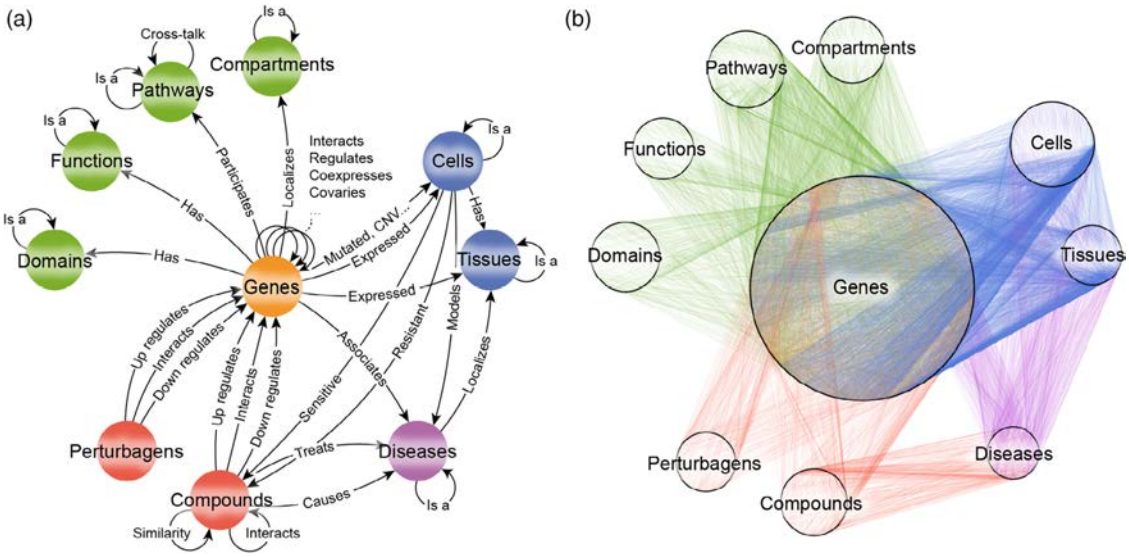


FIGURE 2 Heterogeneous network of biology. (a) A meta-graph of an in-house heterogeneous network, mostly inspired by Hetionet[101] and complemented with the Harmonizome.[72] For simplicity, only the most representative edge types are shown. "Is a" and "has" relationships typically refer to ontologies. (b) A view of the nodes and edges composing the network. To obtain a representative network, we sub-sampled 500 edges of each type. Different colors denote different types of edges, and size of the circles are proportional to the number of nodes

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 | TOWARDS BIOLOGICAL EMBEDDINGS

Heterogeneous networks are an excellent tool to represent biological knowledge explicitly. Querying these networks can help generate mechanistic hypotheses and extract rationale to *describe* observed phenomena. Perhaps more importantly, extensive experiments over the years have shown that large networks may also be exploited to *predict* unobserved phenomena,[106] especially when both the local and the global properties of the graph are utilized by the predictor. Like other big data graphs such as social networks, though, the scope of modern biomedical networks is computationally intractable by traditional graph analytics techniques,[107] which has fostered the development of graph *embedding* approaches that drastically reduce the dimensionality of the data while preserving the structural information and properties of the raw network.[108] In brief, network embedding algorithms learn to represent each node (biological entity) as a numerical vector, so that similar vectors correspond to "related" nodes in the original graph (Figure 3). In great resemblance to chemical embeddings, biological embeddings are an amenable input to subsequent machine learning tasks, and can be discovered automatically without the need for handcrafted design of features that "describe" the role of each node within the network.

Comprehensive surveys of network embedding algorithms can be found elsewhere.[107,108,110] There is an immense catalogue of algorithms, and code is distributed in a rushing pace (over 50 network embedding packages are available, many of them released during the last 2 years; https://github.com/chihming/awesome-network-embedding). Families of successful network embedding algorithms include adjacency matrix factorizations (e.g., graph Laplacian eigenmaps), local linear embeddings, isomaps, and a series of deep learning implementations that address several scenarios, such as the case of attributed networks or the preservation of network structure and properties. Below, we focus on a family of techniques defined by a two-step algorithm consisting of (a) the exploration of the network through random walks followed by (b) the learning of numerical vectors that represent the paths traveled by the random walker. This group of algorithms is uniquely flexible and scalable to huge networks. Of all the approaches to network embedding, this one is the most intuitive and the easiest to interpret and adapt to domain-specific needs, mainly thanks to the graphical, almost mechanistic simplicity of the random walk step (Figure 4).

4.1 | Efficient exploration of biological networks by random walks

Random walks are a popular tool to extract knowledge from biological networks. The algorithm simulates the behavior of a walker that moves from node to node stochastically (with a certain probability of restart). The intuition behind the method is that the paths traveled by the random walker will sample the vicinity of every node, hence providing a measure for node's relevance[111] and proximity to other nodes.[112] In computational biology, random walks were first applied to disease–gene prioritization, based on the proximity of candidate genes to disease-associated genes in a protein–protein interactome.[113] Further improvements of the algorithm enabled the weighting of edges in the network, acknowledging the fact that not all edges are equally important in an interactome, nor they are equally reliable.[114] Likewise, modern implementations can be parametrized to "encourage" the random walker to explore local or global regions of the graph.[115] In this line, a recent random walk scheme specifically designed to explore cancer-related regions of the interactome was able to stratify breast and glioblastoma tumors, discovering pathways in the network that were relevant to each tumor subtype.[116]

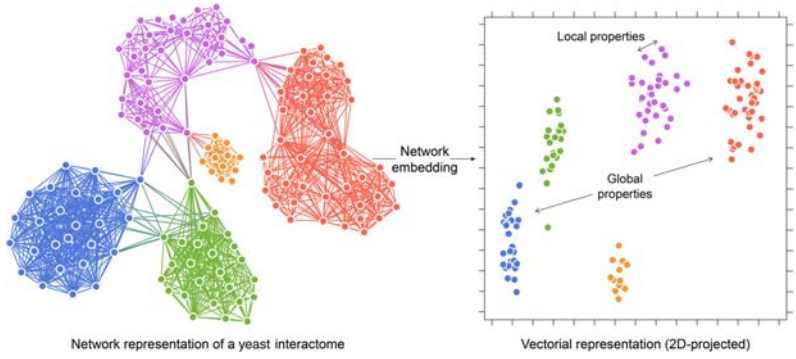


FIGURE 3 Network embedding example. The aim of network embedding is to represent graph entities (typically nodes) as numerical vectors (embeddings) that preserve graph properties, such as local distances, modularity and global organization. Here, we have embedded a fraction (~1%) of the yeast interactome[109] using a standard network embedding algorithm (node2vec; 128 dimensions), and projected the corresponding embeddings in a two-dimensional plane using t-Distributed Stochastic Neighbor Embedding (t-SNE)

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

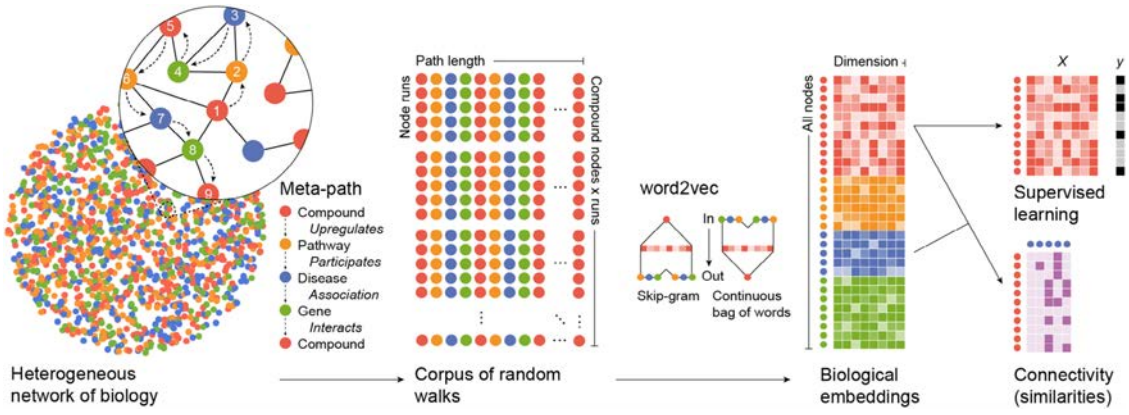


FIGURE 4 Biological embeddings. Given a heterogeneous network, the random walk algorithm can be run under the dictation of a certain meta-path. This will result in a "corpus" (text-like) output that can be apprehended with word2vec (using the skip-gram model or the continuous bag of words model). As a result, each node visited by the random walker will be mapped to an embedding space, that is, each node will be assigned a vector representation. Compound embeddings can be then used in subsequent supervised learning, for example, to predict a clinical property (y) of the molecules, given training data. Alternatively, embeddings of different types can be compared (connected) between them to discover, for example, compound-disease relationships

Applying the random walk algorithm to multilevel and multimodal networks is not straightforward, as naïve random walkers do not keep cognizance of the types of nodes and edges that they visit. Several adaptations of random walks to heterogeneous networks have been suggested recently,[117] refining for example, the search for disease-related genes, even in ill-studied conditions such as the Wiedemann–Rautenstrauch and the SHORT syndromes,[118] and contributing to the field of drug repositioning.[119] Most notably, the need for meaningful exploration of heterogeneous networks brought about the notion of "meta-paths." A meta-path is a sequence of edge types (e.g., drug–gene–disease) that guides the random walker throughout the network (Figure 4). Thus, meta-paths offer a means to capture numerous "semantic" relationships across one same reference biological network. In a series of studies conducted by the authors of the Hetionet, it was shown that different meta-paths can capture distinct aspects of the data, and a strategy was outlined to quantify what meta-paths are the most informative to ask a given "biological question." For example, in light of GWAS data, an association between *IRF1* and multiple sclerosis (MS) was justified by two meta-paths, namely the gene–tissue–disease meta-path ("*IRF1* is expressed in leukocytes, and leukocytes are relevant to MS") and the gene–gene–disease meta-path ("IRF1 interacts with IRF8, and *IRF8* is associated to MS").[120,121] Similarly, the *serendipitous* discovery that the antidepressant bupropion could be used for smoking cessation[122] was rationalized by several pieces of evidence such as the interaction between bupropion and CHRNA3, the fact that this drug causes insomnia, and the participation of CHRNA3 in nicotine-related pathways.[101]

4.2 | Embedding of random walk trajectories

The result of the random walk algorithm is a long list of paths (sequences of nodes) traveled by the random walker. In practice, this output can be seen as a "text corpus" where each node corresponds to a "word" and each path to a "phrase." This is a very convenient format given the technical revolution witnessed in the field of natural language recognition, especially through the set of methods known as word2vec,[123] which yield word embeddings that have an unusual ability to model semantic relationships between, for example, a noun and its gender ("man is to king as woman is to queen"). The word2vec framework offers two ways of training word embeddings, as given by a simple (one layer) neural network fed with a sliding window of words over the text (i.e., fixed-length chunks of sentences). In the continuous bag of words model, context words predict the current word; in the skip-gram model, the current word predicts its context words (Figure 4). Since semantically related words naturally occur in similar contexts, the resulting embeddings successfully capture the "meaning" of the words they represent. As a result, similar and semantically related words will have, correspondingly, numerically similar embeddings. Adapted to the network analysis field, word2vec-like methods such as DeepWalk[124] or node2vec[115] were soon developed to embed the behavior of random walks on homogeneous networks. These methods set solid grounds for the rapid move towards heterogeneous networks. For example, using the concept of meta-paths, metapath2vec maintains structural closeness among multiple types of nodes and edges.[125] A recent extension of the algorithm is even able to grasp

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

free-text attributed to the nodes, and to calculate embeddings for new (out-of-corpus) nodes that were not seen during the training process.[126]

4.3 | Biological embeddings to complement chemical embeddings

Many datasets of biology, including the Harmonizome and Hetionet, contain chemical entities. Hence, network embedding algorithms can be used to capture the "biological context" of compounds too. The resulting "biological embeddings" of small molecules would offer a complementary view to "chemical embeddings," which are dedicated to describing chemical structures. The idea of bringing together "chemical" and "biological" descriptors of small molecules is not new to drug discovery, and has been majorly exploited in the field of high-throughput screening[127,128] and high-content phenotypic screening[129] to optimize the hit rate of chemical libraries. Seminal studies, though, focused on one or few biological data types. The progress in data integration now allows for chemical traits to be combined to an arbitrary number of biological traits, including side-effect profiles,[130] cell-line sensitivity panels[131], and transcriptomic signatures.[132] This has shown to drastically improve the predictive power of daily chemoinformatics tasks such as target prediction[133,134] and anticipation of toxicity events,[135,136] sometimes by means of a simple aggregation of chemical and biological similarities. A recognized[137] and very restricting drawback of most of the current integrative drug predictors, especially of those that capitalize on the explicit links in the networks,[138] is that the accuracy drops sharply when the properties of new (unseen) drugs are to be predicted, compromising the practical interest of the strategy. Biological embeddings can, in principle, overcome this limitation, as they are less reliant on explicit relationships between entities, and sustain performance in notoriously incomplete datasets.[139,140] However, the extent to which biological embeddings of poorly characterized compounds remain informative needs to be systematically evaluated. This systematic analysis, we anticipate, shall determine if biological embeddings will be broadly accepted in the near future as a valid tool to enrich the chemoinformatics pipeline.

4.4 | Biological embeddings to connect small molecules to phenotypes

A singular feature of biological embeddings, compared to chemical embeddings, is that they can be *directly* compared ("connected") to the other biological entities in the network, without the need for previously existing data about the bioactivity of interest. The "connectivity" idea was popularized back in 2006 in the context of transcriptomics by the Connectivity Map initiative,[141] and has matured into the LINCS L1000 screening platform.[132] The LINCS L1000 measures gene expression signatures of ~20,000 compound treatments carried out in dozens of cell lines. In addition, ~7,000 genes are systematically "perturbed" through knock-down and over-expression experiments. This massive resource of mRNA expression profiles can be exploited to find gene expression signatures that "mimic" or "revert" a certain pattern of gene expression. For example, the inhibition/activation mode of action of an uncharacterized compound may be discovered by observing a similarity (mimicking) between the transcriptomic signature of the compound and the signature corresponding to the knock-down/over-expression of its actual target.[142] Likewise, new therapies may be proposed by identifying compounds that cancel out (revert) a disease-characteristic gene expression signature.[143–145] Interestingly, the Connectivity Map is expanding its portfolio of profiles beyond mRNA expression, and now includes cell-painting experiments of cell morphology, and proteomics P100 and GCP assays [clue.io].

As demonstrated by Hetionet and, especially, by the Harmonizome, "omics" signatures can be converted to a set of edges, hence the notion of "connectivity" may be generalized, in principle, to heterogeneous network analysis.[146] For example, using gene expression signatures, phenotype-specific gene regulatory networks were built and successfully "connected" to drugs through targets discovered at crucial points in the networks.[147] Besides, random walks have been successfully applied to the analysis of gene expression signatures overlaid on a protein–protein interactome,[148] advocating for the use of the two-step network embedding strategy presented herein. Reassuringly, it has been shown that gene expression profiles can be safely compressed to vectors of as few as 100 dimensions,[149] which is a typical size for network embeddings. Moreover, and suggestively, geometric operations between vectors in the embedding space have been formally associated to conjunctive logical queries on the graph,[150] setting the basis for the discovery of drugs that accomplish complex biological traits. This, we believe, may bring advancement in polypharmacy, multifactorial disease therapy, and precision medicine.

5 | CONCLUSIONS

Overall, expressing biological data as a huge heterogeneous network whose nodes can be embedded to numerical vectors opens a new avenue for computational drug discovery. First, because biological embeddings resemble in format the more established chemical embeddings, offering a complementary means to navigate the chemical space by virtue of similarity

17590084, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

searches that are more biologically relevant. Second, biological embeddings can be a natural input for most machine learning algorithms, which greatly facilitates the inheritance of methods developed in other fields such as text and image processing, as demonstrated by the swift incorporation of deep learning to the chemoinformatics toolbox via chemical embeddings. Finally, through distance measures within one same mathematical space, biological embeddings enable the long-sought connection of small molecules to other biological entities such as phenotypes or novel targets, in an unsupervised fashion that does not require previous bioactivity data.

While injecting dense (embedded) biological knowledge into the drug discovery pipeline may increase the efficiency of certain steps, some limitations have to be addressed before widespread acceptance of the approach by the community. The main barrier is the lack of interpretability (i.e., mechanistic understanding) of the models, which is crucial to gain confidence along the drug discovery process.[151] Whitening black-box predictions is an open challenge in machine learning. Attention is put to deciphering how a particular model relates its input to its output,[44] although generic solutions might not be sufficient to trace the interpretation back to the influential nodes in the biological network. Another limitation is the absence of benchmark datasets for "predictive" (perturbation-based) biology, making it difficult to optimize the network embedding protocol (an exception is the DREAM challenge, http://dreamchallenges.org). Without benchmark tests that refer to the phenotypic property of interest, traversal of the network by, for example, random walks may be erratic, exploring irrelevant regions of the graph while omitting the really predictive ones. Related to this, research is needed to identify meaningful meta-paths[152,153] and devise network sampling procedures that simulate complicated phenomena such as gene regulation,[154] spatial organization[155] or time-resolved dynamics.[156]

There is hope that deep neural network architectures will learn to overcome some of these limitations, much like they learn hidden patterns in images or the syntax and semantics in text phrases. However, this will chiefly depend on the availability of "big enough" data pertinent to the system of study,[20] and we agree with those who express "some healthy scepticism" about the prompt implementation of deep learning in biomedicine.[157] Evidently, the use of biological embeddings as inputs for machine learning is not restricted to deep learning, and other areas of artificial intelligence could be exploited as well (Box 2). Of note, automatic machine learning (AutoML) is showing outstanding progress,[158,159] freeing the user from the arduous tuning of hyper-parameters and the testing of different models and feature representations. AutoML holds promise for making machine learning accessible to nonexperts, which would promote acceptance of abstract (embedded) representations of the data by the community. Other interesting lines of research in artificial intelligence include semisupervised learning, especially in the absence of "negative" data (a common hurdle in computational biology[160,161]), and gradient-boosting methods,[162] which are dominating machine learning competitions for structured and tabular data. Coupled to this, there is a need for methods that estimate the uncertainty of machine-learning predictions, of which ensemble-based approaches are among the most practical and scalable, as opposed to classical Bayesian uncertainty estimates that require computation of probability distributions for every parameter in the model.[163,164]

BOX 2

FIVE MACHINE LEARNING KEYWORDS FOR DATA-DRIVEN DRUG DISCOVERY, SORTED ALPHABETICALLY

1. *Automated machine learning* (*AutoML*). The goal of AutoML is to provide off-the-shelf machine learning processes and methods that are accessible to nonmachine learning experts. This is achieved by the automatic determination of a well-performing machine learning pipeline, without the need for feature selection, choice of model, hyper-parameter optimization, and cross-validation.

2. *Feature learning.* The main motivation behind feature learning is to replace manual feature engineering by automatically detecting relevant patterns in the raw data, while dismissing noisy and noninformative traits. Feature learning can be supervised or unsupervised, and discovers mathematical representations of the data that are convenient to process by down-stream machine learning algorithms.

3. *Generative models.* Given samples, generative models try to learn the true data distribution so as to generate extra data points that resemble the observed samples but include some variations. Two popular generative neural network architectures are generative adversarial networks and generative autoencoders.

4. *One-shot learning.* Most classification algorithms require training on large datasets. Instead, one-shot learning aims to learn from one (or only a few) training samples. One-shot learning approaches human intelligence (which does not require huge amounts of examples to learn a concept) by incorporating "memory" and "comparisons" (metric learning) into neural network architectures.

5. *Positive-unlabeled (PU) learning.* PU learning handles the fact that, in many biological datasets, only a small portion of "positive" results/annotations are available, whereas a majority of "negative" results remain unreported or unknown.

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

All in all, as pharmaceutical research is moving towards precision medicine, there is a need to enrich computational methods with generic knowledge of biology as well as patient- and cohort-specific samples. Recent work on lung cancer patient selection demonstrates that the traditional "chemistry-first" approach can be sustained as long as biomarkers, genome-wide targets and genetic landscapes are included in the models.[165] Biological embeddings may help to generalize this approach, smoothing the transition from the blockbuster system of drugs to a personalized medicine one.[166]

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R) and the European Research Council (SysPharmAD: 614944).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLE

Systems biology-embedded target validation: Improving efficacy in drug discovery

REFERENCES

1. Marx V. Biology: The big challenges of big data. *Nature*. 2013;*498*:255–260.
2. Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016. Data growth and integration. *Nucleic Acids Res*. 2016;*44*:D20–D26.
3. Muir P, Li S, Lou S, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol*. 2016;*17*:53.
4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;*18*:83.
5. Barash CI. Omics challenges and unmet translational needs. *Appl Transl Genom*. 2016;*10*:1.
6. Butcher EC. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov*. 2005;*4*:461–467.
7. Tsigkinopoulou A, Baker SM, Breitling R. Respectful modeling: Addressing uncertainty in dynamic system models for molecular biology. *Trends Biotechnol*. 2017;*35*:518–529.
8. Papp B, Notebaart RA, Pál C. Systems-biology approaches for predicting genomic evolution. *Nat Rev Genet*. 2011;*12*:591–602.
9. Apweiler R, Beissbarth T, Berthold MR, et al. Whither systems medicine? *Exp Mol Med*. 2018;*50*:e453.
10. Swinney DC, Anthony J. How were new medicines discovered? *Nat Rev Drug Discov*. 2011;*10*:507–519.
11. Keiser MJ, Irwin JJ, Shoichet BK. The chemical basis of pharmacology. *Biochemistry*. 2010;*49*:10267–10276.
12. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet*. 2014;*15*:107–120.
13. Wynn ML, Consul N, Merajver SD, Schnell S. Logic-based models in systems biology: A predictive and parameter-free network analysis method. *Integr Biol (Camb)*. 2012;*4*:1323–1337.
14. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotechnol*. 2010;*28*:245–248.
15. Yizhak K, Gaude E, Le Devedec S, et al. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife*. 2014;*3*:e03641.
16. Sharan R, Karp RM. Reconstructing Boolean models of signaling. *J Comput Biol*. 2013;*20*:249–257.
17. Le Novere N. Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet*. 2015;*16*:146–158.
18. Vidal M. How much of the human protein interactome remains to be mapped? *Sci Signal*. 2016;*9*:eg7.
19. Washburn MP. There is no human interactome. *Genome Biol*. 2016;*17*:48.
20. Rampasek L, Goldenberg A. TensorFlow: biology's gateway to deep learning? *Cell Syst*. 2016;*2*:12–14.
21. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;*15*.
22. Webb S. Deep learning for biology. *Nature*. 2018;*554*:555–557.
23. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. *Cell*. 2018;*173*:1581–1592.
24. Stephens ZD, Lee SY, Faghri F, et al. Big data: Astronomical or genomical? *PLoS Biol*. 2015;*13*:e1002195.
25. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med*. 2010;*2*:84.
26. Gasteiger J. Chemoinformatics: A new field with a long tradition. *Anal Bioanal Chem*. 2006;*384*:57–64.
27. Oprea TI, May EE, Leitao A, Tropsha A. Computational systems chemical biology. *Methods Mol Biol*. 2011;*672*:459–488.
28. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. *Drug Discov Today*. 2017;*22*:1680–1685.
29. Livingstone DJ. The characterization of chemical structures using molecular properties. A survey. *J Chem Inf Comput Sci*. 2000;*40*:195–209.
30. Rajarshi G, Egon W. A survey of quantitative descriptions of molecular structure. *Curr Top Med Chem*. 2012;*12*:1946–1956.
31. Sagarika S, Chandana A, Minati K, Bijay KM. A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. *Curr Comput Aided Drug Des*. 2016;*12*:181–205.
32. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007;*25*:197–206.
33. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: A chemical fragment-based approach. *BMC Bioinformatics*. 2011;*12*:169.
34. Duran-Frigola M, Mateo L, Aloy P. Drug repositioning beyond the low-hanging fruits. *Curr Opin Syst Biol*. 2017;*3*:95–102.
35. Luo H, Wang J, Li M, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*. 2016;*32*:2664–2671.
36. Mitchell JB. Machine learning methods in chemoinformatics. *WIREs Comput Mol Sci*. 2014;*4*:468–481.

17590984, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

37. Tetko IV, Engkvist O, Koch U, Reymond J-L, Chen H. BIGCHEM: Challenges and opportunities for big data analysis in chemistry. *Mol Inform*. 2016;*35*: 615–621.
38. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des*. 2016;*30*:595–608.
39. Kwon S, Yoon S. DeepCCI: End-to-end deep learning for chemical-chemical interaction prediction. arXiv:1704.08432; 2017.
40. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. arXiv:1704.01212; 2017.
41. Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv:1706.06689; 2017.
42. Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. How much chemistry does a deep neural network need to know to make accurate predictions? arXiv: 1710.02238; 2017.
43. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP. Convolutional networks on graphs for learning molecular fingerprints. arXiv:1509.09292; 2015.
44. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. arXiv:1602.04938; 2016.
45. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. *ACS Cent Sci*. 2017;*3*:283–293.
46. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc Natl Acad Sci*. 2018;*115*:E4304–E4311.
47. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*. 2018;*361*:360–365.
48. Sterling T, Irwin JJ. ZINC 15—Ligand discovery for everyone. *J Chem Inf Model*. 2015;*55*:2324–2337.
49. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. arXiv:1610.02415; 2016.
50. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm*. 2017;*14*:3098–3104.
51. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Centr Sci*. 2018;*4*: 120–131.
52. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. Application of generative autoencoder in de novo molecular design. arXiv:1711.07839; 2017.
53. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de-novo drug design. arXiv:1711.10907; 2017.
54. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. 2018;*555*:604–610.
55. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;*23*:1241–1250.
56. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 I.E. Conference on Computer Vision and Pattern Recognition, Miami, FL; 2009, p. 248–255.
57. Miller GA. WordNet: A lexical database for English. *Commun ACM*. 1995;*38*:39–41.
58. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;*9*:513–530.
59. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;*45*:D945–D954.
60. Wang Y, Bryant SH, Cheng T, et al. PubChem BioAssay: 2017 update. *Nucleic Acids Res*. 2017;*45*:D955–D963.
61. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;*45*:D362–D368.
62. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;*159*:1212–1226.
63. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017;*45*:D833–D839.
64. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: A platform for therapeutic target identification and validation. *Nucleic Acids Res*. 2017;*45*: D985–D994.
65. Fabregat A, Korninger F, Viteri G, et al. Reactome graph database: Efficient access to complex pathway data. *PLoS Comput Biol*. 2018;*14*:e1005968.
66. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: Expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 2016;*44*:D336–D342.
67. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;*483*: 603–607.
68. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;*45*:580–585.
69. Rigden DJ, Fernández XM. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res*. 2018;*46*: D1–D7.
70. Aranda B, Blankenburg H, Kerrien S, et al. PSICQUIC and PSISCORE: Accessing and scoring molecular interactions. *Nat Methods*. 2011;*8*:528–529.
71. Oxenham S. Legal confusion threatens to slow data science. *Nature*. 2016;*536*:16–17.
72. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)*. 2016;*2016*. pii: baw100. https://doi.org/10.1093/database/baw100
73. Jacunski A, Tatonetti NP. Connecting the dots: Applications of network medicine in pharmacology and disease. *Clin Pharmacol Ther*. 2013;*94*:659–669.
74. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nat Rev Genet*. 2010;*12*:56.
75. Bard JBL, Rhee SY. Ontologies in biology: Design, applications and future challenges. *Nat Rev Genet*. 2004;*5*:213–222.
76. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: A functional perspective. *Brief Bioinform*. 2015;*16*: 1069–1080.
77. Salvadores M, Alexander PR, Musen MA, Noy NF. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semant Web*. 2013;*4*: 277–284.
78. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. *Nat Genet*. 2000;*25*:25–29.
79. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 2015;*43*:D1071–D1078.
80. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008;*83*:610–615.
81. Bairoch A. The Cellosaurus: A cell-line knowledge resource. *J Biomol Tech*. 2018;*29*:25–38.
82. Gremse M, Chang A, Schomburg I, et al. The BRENDA Tissue Ontology (BTO): The first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res*. 2011;*39*:D507–D513.
83. Schurer SC, Vempati U, Smith R, Southern M, Lemmon V. BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J Biomol Screen*. 2011;*16*:415–426.
84. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nat Rev Genet*. 2016;*17*:615.
85. Duran-Frigola M, Rossell D, Aloy P. A chemo-centric view of human health and disease. *Nat Commun*. 2014;*5*:5676.
86. Sam SA, Teel J, Tegge AN, Bharadwaj A, Murali TM. XTalkDB: A database of signaling pathway crosstalk. *Nucleic Acids Res*. 2017;*45*:D432–D439.

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

87. Lee Y-F, Lee C-Y, Lai L-C, Tsai M-H, Lu T-P, Chuang EY. CellExpress: A comprehensive microarray-based cancer cell line and clinical sample gene expression analysis online system. *Database (Oxford)*. 2018;*2018*. pii: bax101. https://doi.org/10.1093/database/bax101
88. Wang H, Huang S, Shou J, et al. Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. *BMC Genomics*. 2006;*7*:166.
89. Osat S, Faqeeh A, Radicchi F. Optimal percolation on multiplex networks. *Nat Commun*. 2017;*8*:1540.
90. De Domenico M, Nicosia V, Arenas A, Latora V. Structural reducibility of multilayer networks. *Nat Commun*. 2015;*6*:6864.
91. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science*. 2010;*328*: 876–878.
92. Kleineberg K-K, Boguñá M, Ángeles Serrano M, Papadopoulos F. Hidden geometric correlations in real multiplex networks. *Nat Phys*. 2016;*12*:1076–1081.
93. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;*26*:i237–i245.
94. Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*. 2017;*33*:i190–i198.
95. Jeon J, Nim S, Teyra J, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med*. 2014;*6*:57.
96. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol Ther*. 2013;*138*:333–408.
97. Duran-Frigola M, Mosca R, Aloy P. Structural systems pharmacology: The role of 3D structures in next-generation drug development. *Chem Biol*. 2013;*20*: 674–684.
98. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*. 2011;*7*:496.
99. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun*. 2017;*8*:573.
100. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. *Nat Commun*. 2016;*7*:10331.
101. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*. 2017;*6*:e26726.
102. Have CT, Jensen LJ. Are graph databases ready for bioinformatics? *Bioinformatics*. 2013;*29*:3107–3108.
103. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. *BioData Mining*. 2016;*9*:23.
104. Mungall CJ, McMurry JA, Köhler S, et al. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2017;*45*:D712–D722.
105. Yoon BH, Kim SK, Kim SY. Use of Graph Database for the integration of heterogeneous biological data. *Genomics Inform*. 2017;*15*:19–27.
106. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. arXiv:1807.00123; 2018.
107. Cai H, Zheng VW, Chang KC-C. A comprehensive survey of graph embedding: Problems, techniques and applications. arXiv:1709.07604; 2017.
108. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. arXiv:1705.02801; 2017.
109. Collins SR, Kemmeren P, Zhao X-C, et al. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics*. 2007;*6*: 439–450.
110. Cui P, Wang X, Pei J, Zhu W. A survey on network embedding. arXiv:1711.08752; 2017.
111. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th International World-Wide Web Conference (WWW 1998); 1998, p. 107–117.
112. Pan J-Y, Yang H-J, Faloutsos C, Duygulu P. Automatic multimedia cross-modal correlation discovery. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2004, p. 653–658.
113. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*. 2008;*82*:949–958.
114. Li T, Wernersson R, Hansen RB, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat Methods*. 2016;*14*:61.
115. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. arXiv:1607.00653; 2016.
116. Zhang W, Ma J, Ideker T. Classifying tumors by supervised network propagation. *Bioinformatics*. 2018;*34*:i484–i493.
117. Navarro C, Martínez V, Blanco A, Cano C. ProphTools: General prioritization tools for heterogeneous biological networks. *GigaScience*. 2017;*6*:1–8.
118. Valdeolivas A, Tichit L, Navarro C, et al. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*. 2018. https://doi.org/10.1093/bioinformatics/bty637
119. Luo H, Wang J, Li M, et al. Computational drug repositioning with random walk on a heterogeneous network. *IEEE/ACM Trans Comput Biol Bioinform*. 2018;1.
120. Himmelstein DS, Baranzini SE. Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *PLoS Comput Biol*. 2015;*11*:e1004259.
121. Greene CS, Himmelstein DS. Genetic association-guided analysis of gene networks for the study of complex traits. *Circ Cardiovasc Genet*. 2016;*9*:179–184.
122. Harmey D, Griffin PR, Kenny PJ. Development of novel pharmacotherapeutics for tobacco dependence: Progress and future directions. *Nicotine Tob Res*. 2012; *14*:1300–1318.
123. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv: 1301.3781; 2013.
124. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. arXiv: 1403.6652; 2014.
125. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017, p. 135–144.
126. Zhang C, Swami A, Chawla NV. CARL: Content-aware representation learning for heterogeneous networks. arXiv:1805.04983; 2018.
127. Helal KY, Maciejewski M, Gregori-Puigjané E, Glick M, Wassermann AM. Public Domain HTS Fingerprints: Design and evaluation of compound bioactivity profiles from PubChem's Bioassay Repository. *J Chem Inf Model*. 2016;*56*:390–398.
128. Cortes Cabrera A, Petrone PM. Optimal HTS fingerprint definitions by using a desirability function and a genetic algorithm. *J Chem Inf Model*. 2018;*58*:641–646.
129. Wawer MJ, Li K, Gustafsdottir SM, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc Natl Acad Sci*. 2014;*111*:10911–10916.
130. Baker NC, Fourches D, Tropsha A. Drug side effect profiles as molecular descriptors for predictive modeling of target bioactivity. *Mol Inform*. 2015;*34*:160–170.
131. Chabner BA. NCI-60 cell line screening: A radical departure in its time. *J Natl Cancer Inst*. 2016;*108*:djv388.
132. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;*171*:1437–1452. e1417.
133. Madhukar NS, Khade P, Huang L, Gayvert K, Galletti G, Stogniew M, Allen JE, Giannakakou P, Elemento O. A new big-data paradigm for target identification and drug discovery. bioRxiv 2017.
134. Zhu S, Bing J, Min X, Lin C, Zeng X. Prediction of drug–gene interaction by using Metapath2vec. *Front Genet*. 2018;*9*:1–10.

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

135. Gayvert Kaitlyn M, Madhukar Neel S, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol*. 2016;*23*: 1294–1301.

136. Madhukar NS, Gayvert K, Gilvary C, Elemento O. A machine learning approach predicts tissue-specific drug adverse events. bioRxiv 2018.

137. Guney E. Reproducible drug repurposing: when similarity does not suffice. In: Altman RB, Keith Dunker A, Hunter L, Ritchie MD, Murray T, Klein TE, eds. *Pacific Symposium on Biocomputing 2017*. Singapore: World Scientific, 2016; p. 132–143.

138. Vilar S, Hripcsak G. The role of drug profiles as similarity metrics: Applications to repurposing, adverse effects detection and drug–drug interactions. *Brief Bioinform*. 2017;*18*(4):670–681.

139. Yang D, Wang S, Li C, Zhang X, Li Z. From properties to links: Deep network embedding on incomplete graphs. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017, p. 367–376.

140. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. 2018;*34*:i457–i466.

141. Lamb J, Crawford ED, Peck D, et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;*313*: 1929–1935.

142. Sawada R, Iwata M, Tabei Y, Yamato H, Yamanishi Y. Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. *Sci Rep*. 2018;*8*:156.

143. Chen B, Ma L, Paik H, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat Commun*. 2017;*8*:16022.

144. Duan Q, Reid SP, Clark NR, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl*. 2016;*2*:16015.

145. Wu H, Huang J, Zhong Y, Huang Q. DrugSig: A resource for computational drug repositioning utilizing gene expression signatures. *PLoS One*. 2017;*12*: e0177743.

146. Li L, He X, Borgwardt K. Multi-target drug repositioning by bipartite block-wise sparse multi-task learning. *BMC Syst Biol*. 2018;*12*:55.

147. Zickenrott S, Angarica VE, Upadhyaya BB, del Sol A. Prediction of disease–gene–drug relationships following a differential network analysis. *Cell Death Dis*. 2016;*7*:e2040.

148. Soul J, Hardingham TE, Boot-Handford RP, Schwartz J-M. PhenomeExpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes. *Sci Rep*. 2015;*5*:8117.

149. Filzen TM, Kutchukian PS, Hermes JD, Li J, Tudor M. Representing high throughput expression profiles via perturbation barcodes reveals compound targets. *PLoS Comput Biol*. 2017;*13*:e1005335.

150. Hamilton WL, Bajaj P, Zitnik M, Jurafsky D, Leskovec J. Querying complex networks in vector space. arXiv:1806.01445; 2018.

151. Plenge RM. Disciplined approach to drug discovery and early development. *Sci Transl Med*. 2016;*8*:349ps315.

152. Shakibian H, Moghadam CN. Mutual information model for link prediction in heterogeneous complex networks. *Sci Rep*. 2017;*7*:44981.

153. Meng C, Cheng R, Maniu S, Senellart P, Zhang W. Discovering meta-paths in large heterogeneous information networks. Proceedings of the 24th International Conference on World Wide Web; 2015, p. 754–764.

154. Kittas A, Delobelle A, Schmitt S, Breuhahn K, Guziolowski C, Grabe N. Directed random walks and constraint programming reveal active pathways in hepatocyte growth factor signaling. *FEBS J*. 2015;*283*:350–360.

155. Xu X, Lu L, He P, Chen L. Protein localization prediction using random walks on graphs. *BMC Bioinform*. 2013;*14*:S4.

156. Weng T, Zhao Y, Small M, Huang D. Time-series analysis of networks: Exploring the structure with random walks. *Phys Rev E*. 2014;*90*:022804.

157. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform*. 2016;*35*:3–14.

158. Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Efficient and robust automated machine learning. Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume. 2; 2015, p. 2755–2763.

159. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J Mach Learn Res*. 2017;*18*:826–830.

160. Hameed PN, Verspoor K, Kusljic S, Halgamuge S. Positive-unlabeled learning for inferring drug interactions based on heterogeneous attributes. *BMC Bioinform*. 2017;*18*:140.

161. Yang P, Li X-L, Mei J-P, Kwoh C-K, Ng S-K. Positive-unlabeled learning for disease gene identification. *Bioinformatics*. 2012;*28*:2640–2647.

162. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. arXiv:1603.02754; 2016.

163. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv:1612.01474; 2016.

164. Pearce T, Zaki M, Brintrup A, Neel A. Uncertainty in neural networks: Bayesian ensembling. arXiv:1810.05546; 2018.

165. McMillan EA, Ryu M-J, Diep CH, et al. Chemistry-first approach for nomination of personalized treatment in lung cancer. *Cell*. 2018;*173*:864–878.e829.

166. Pavelić K, Martinović T, Kraljević PS. Do we understand the personalized medicine paradigm? *EMBO Rep*. 2014;*16*:133–136.

How to cite this article: Duran-Frigola M, Fernández-Torras A, Bertoni M, Aloy P. Formatting biological big data for modern machine learning in drug discovery. *WIREs Comput Mol Sci*. 2019;9:e1408. https://doi.org/10.1002/wcms.1408

17590884, 2019, 6, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1408 by Consorci De Serveis Universitaris De Catalunya, Wiley Online Library on [24/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

This article is made available via the ACS COVID-19 subset for unrestricted RESEARCH re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for the duration of the World Health Organization (WHO) declaration of COVID-19 as a global pandemic.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Application Note

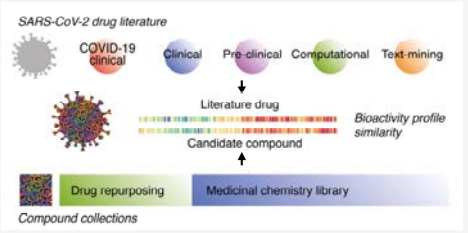
Bioactivity Profile Similarities to Expand the Repertoire of COVID-19 Drugs

Miquel Duran-Frigola,* Martino Bertoni, Roi Blanco, Víctor Martínez, Eduardo Pauls, Víctor Alcalde, Gemma Turon, Núria Villegas, Adrià Fernández-Torras, Carles Pons, Lídia Mateo, Oriol Guitart-Pla, Pau Badia-i-Mompel, Aleix Gimeno, Nicolas Soler, Isabelle Brun-Heath, Hugo Zaragoza, and Patrick Aloy*

Cite This: J. Chem. Inf. Model. 2020, 60, 5730−5734 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Until a vaccine becomes available, the current repertoire of drugs is our only therapeutic asset to fight the SARS-CoV-2 outbreak. Indeed, emergency clinical trials have been launched to assess the effectiveness of many marketed drugs, tackling the decrease of viral load through several mechanisms. Here, we present an online resource, based on small-molecule bioactivity signatures and natural language processing, to expand the portfolio of compounds with potential to treat COVID-19. By comparing the set of drugs reported to be potentially active against SARS-CoV-2 to a universe of 1 million bioactive molecules, we identify compounds that display analogous chemical and functional features to the current COVID-19 candidates. Searches can be filtered by level of evidence and mechanism of action, and results can be restricted to drug molecules or include the much broader space of bioactive compounds. Moreover, we allow users to contribute COVID-19 drug candidates, which are automatically incorporated to the pipeline once per day. The computational platform, as well as the source code, is available at https://sbnb.irbbarcelona.org/covid19.



INTRODUCTION

A new coronavirus, named SARS-CoV-2, is the responsible agent for the current 2019−2020 viral pneumonia (COVID-19) outbreak,[1,2] which is already affecting millions of people worldwide and causing hundreds of thousands of deaths. The COVID-19 pandemic has prompted an unprecedented effort by the scientific community to understand its molecular constituents and find an effective treatment to mitigate viral infectiveness and symptoms. This is reflected in the over 6000 COVID-related publications that appeared in the past few weeks.[3] Huge efforts are being invested in the discovery of an effective vaccine, but even the most optimistic scenarios suggest that it will not be available until 2021. Other drug discovery projects have been launched to target specific viral proteins, particularly its main protease (Mpro).[4] However, these initiatives, even if successful, could take even longer to deliver an approved drug. Thus, the repurposing of existing drugs is our best chance to face the current outbreak therapeutically, since approved drugs have known safety profiles and are ready to be tested in humans. For instance, several compounds initially developed to treat HIV (e.g., lopinavir/ritonavir)[5] or Ebola (e.g., remdesivir),[6] as well as antimalarial drugs (e.g., hydroxychloroquine),[7] are being tested against COVID-19. Indeed, we conducted a limited review of the most relevant scientific literature and identified over 200 compounds that are potentially active against COVID-19 with different levels of experimental support, from purely computational predictions to preclinical and drugs already in clinical trials.

We now exploit this literature mining effort to identify other compounds with the potential to be effective against COVID-19. To this aim, we use the Chemical Checker (CC), a resource that provides processed, harmonized, and integrated bioactivity data for about 1 million small molecules.[8] In the CC, bioactivity data are expressed in a vector format, which naturally extends the notion of chemical similarity between compounds to similarities between bioactivity profiles. The CC organizes data into five levels of increasing complexity, ranging from drug binding profiles to clinical outcomes, and thus enables similarity searches that should be mechanistically and clinically relevant.

Special Issue: COVID19 - Computational Chemists Meet the Moment

Received: April 22, 2020
Published: July 16, 2020

Downloaded via 79.153.122.3 on October 24, 2022 at 15:04:03 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© 2020 American Chemical Society

https://dx.doi.org/10.1021/acs.jcim.0c00420
J. Chem. Inf. Model. 2020, 60, 5730−5734

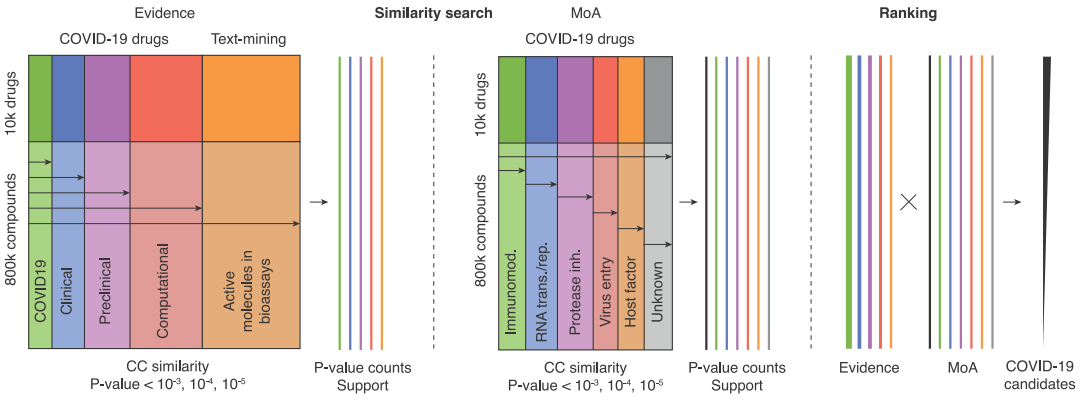


Figure 1. Methodological strategy. We use the list of COVID-19 compounds extracted from the literature, with different levels of experimental evidence, as bait to search for compounds with similar bioactivity or chemical features among the 800,000 molecules contained in the CC. We also include compounds that are positive in relevant bioassays, identified through automatic mining of the COVID-19 literature, and for which we find further bioactivity support in the CC. We keep and rank the top 10,000 most similar molecules to bait compounds and weight them to favor molecules with similar properties to those with higher levels of experimental evidence.

In the current resource, we use CC signatures to identify similarities between bioactive compounds and the list of current COVID-19 drug candidates (i.e., *bait* compounds). The similarity search is performed systematically across the large chemical space encompassed by the CC, thereby substantially expanding the portfolio of potential molecules effective against SARS-CoV-2. Results are stratified between drug molecules and a broader medicinal chemistry space, thus offering ranked lists of compounds that should be of value for drug repurposing endeavors as well as preclinical screening campaigns.

METHODOLOGICAL STRATEGY

Our resource capitalizes on an ongoing literature curation effort done by our group. Additionally, we welcome contributions from the broader scientific community via web form, allowing users to include compounds under investigation in their laboratories, or to update the evidence level as new COVID-19 experiments accumulate. The scientific evidence supporting COVID-19 drug candidates is variable: some compounds come from computational predictions, some have proven their value in preclinical tests, others are approved drugs with a therapeutic indication unrelated to infectious diseases, and, finally, some are drugs currently used to fight SARS-CoV-2-related pathogens. The mechanisms of action (MoA) suggested to confer efficacy are also variable, ranging from immunomodulators to protease inhibitors. During curation, we classify literature COVID-19 candidates by their level of evidence and MoA (Figure 1). By the 18th of April, 2020, we have found that 230 small molecules have been suggested as potential treatments for COVID-19.

Starting from the SMILES representation of a compound, we derive CC bioactivity signatures for each COVID-19 literature bait compound. We then run bioactivity similarity searches against the ∼1 million bioactive molecules characterized in the CC and keep the top 10,000 most similar compounds for each search type. Likewise, we conduct conventional similarity searches solely based on 2D representations of the compounds (2048-bit Morgan fingerprints, radius 2). Similarities are expressed as empirical P-values (−log 10 scale) derived from the expected similarity distribution across the full search space. A simple *support* measure is provided for each compound by adding up the number of similar COVID-19 drugs (weighted by −log 10 P-value and level of evidence, as shown in Figure 1).

In addition, we complement our literature curation effort with a further level of evidence, namely, text-mining, based on the automatic detection of experiments (bioassays) that could be relevant to COVID-19. More specifically, we process the text description of the ∼1.2 million bioassays catalogued in the ChEMBL database and rank them according to their relevance to the current corpus of about 30,000 articles related to COVID-19 and other coronavirus infections.[9] ChEMBL bioassays[10] are ranked using two complementary approaches: (i) We construct a retrieval query from the bioassay descriptions and use it to score each of the paragraphs and abstracts contained in the articles collection. We then use statistics of the score distribution of top scoring documents to rank the bioassays. And (ii), we manually labeled a set of (*seed*) molecules that tested positive in ∼100 bioassays relevant to COVID-19. We then automatically identify compounds from all the bioassay descriptions and compute their contextual embeddings. Finally, we rank the bioassays according to their cosine similarity to the seed molecules. We then keep the 1000 most relevant COVID-19 literature bioassays, as ranked by either text-mining approach and identify those bioactive molecules within the CC universe that tested positive (<10 μM) in at least one of them. Finally, we cross these results with the 10,000 compounds obtained from the similarity searches described above and assign an extra literature-evidence level (text-mining) to those in common, which are then used as bait compounds.

The pipeline runs automatically every day, so that we always provide the most updated results. Searches are precomputed for each evidence strength and MoA.

THE RESOURCE

Results of the large-scale similarity search are made available as a web-resource at https://sbnb.irbbarcelona.org/covid19. The interface contains five tabs:

https://dx.doi.org/10.1021/acs.jcim.0c00420
J. Chem. Inf. Model. 2020, 60, 5730−5734

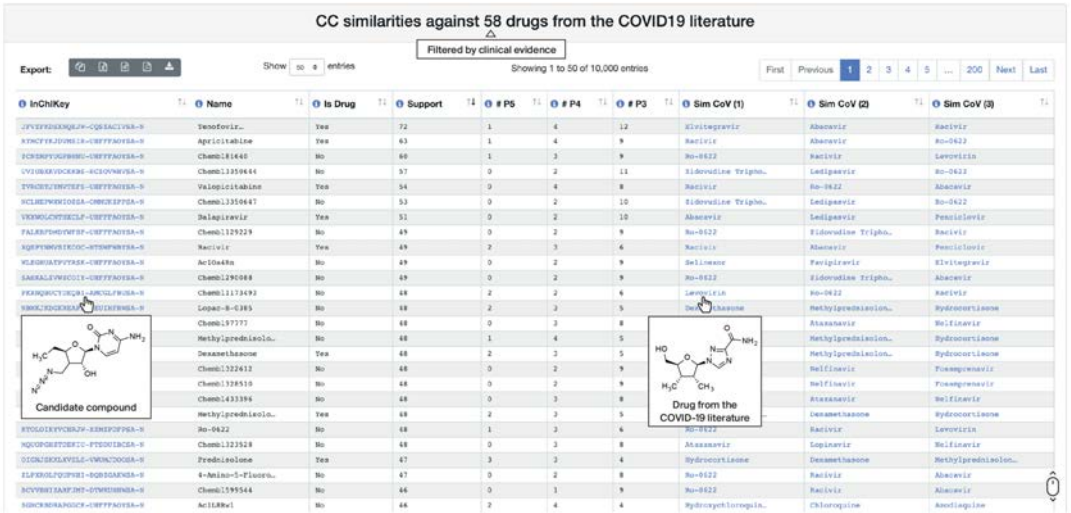


Figure 2. Querying the compound similarity matrix. The pre-computed similarity matrices can be queried to extract candidates with the properties of interest. The dynamic tables show information about each candidate compound: InChIKey, name, whether it is an approved drug, its level of support, number of COVID-19 bait compounds to which it is similar to different P-values (10^{-5}, 10^{-4}, and 10^{-3}), and the three most similar bait compounds. Additionally, for each molecule, we provide its structure and links to the corresponding CC page. Figure produced on the 18th of April, 2020.

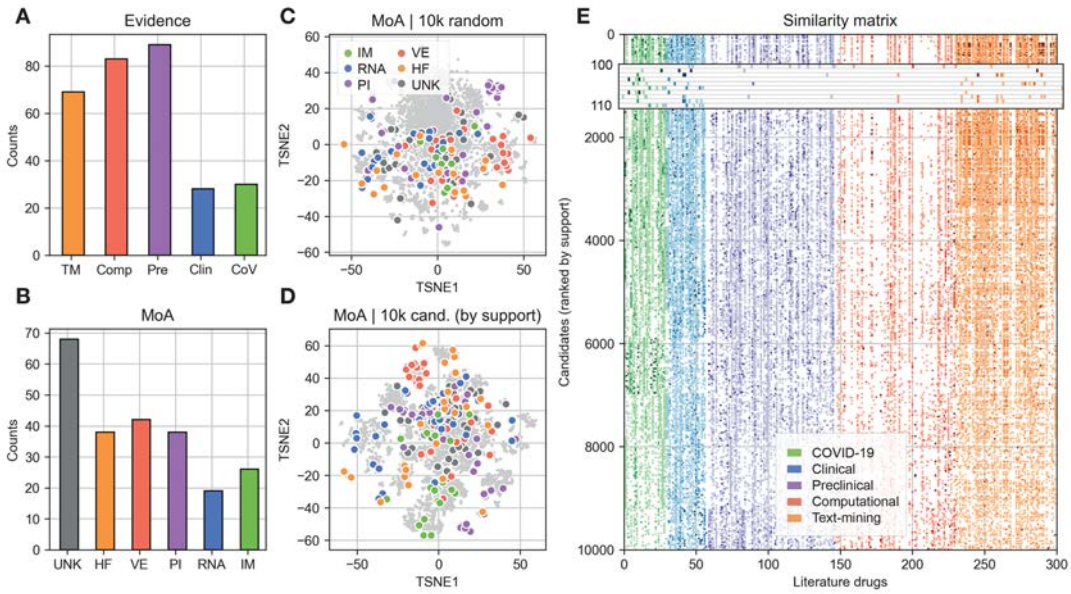


Figure 3. COVID-19 literature bait compounds' composition and functional diversity. Number of literature bait compounds split according to their (A) level of experimental evidence or (B) MoA. (C) t-SNE projections of the bait compounds on the global space of bioactive CC molecules and on the top 10,000 candidate compounds (D), coloured by MoA. (E) A global view on the similarity matrix, stratified by level of evidence. Figure produced on the 18th of April, 2020.

Candidates. We provide the 10,000 molecules, within the CC universe of 1 M bioactive compounds, that are more similar to the COVID-19 bait compounds collected from the literature (Figure 2). The precomputed similarity matrix can be queried to extract candidates that fulfill properties of interest by selecting among the levels of evidence for the bait

https://dx.doi.org/10.1021/acs.jcim.0c00420
J. Chem. Inf. Model. 2020, 60, 5730−5734

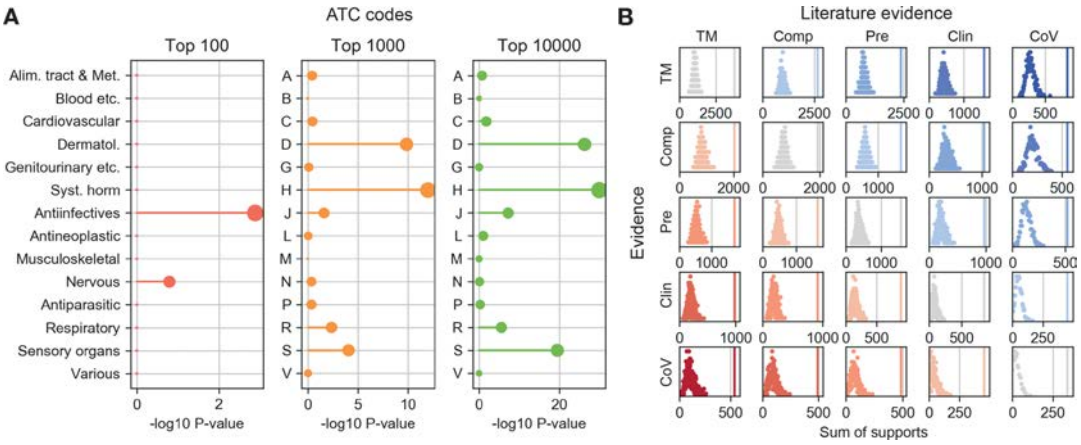


Figure 4. Benchmark of the strategy. (A) Enrichment analysis of therapeutic areas (ATC categories) among the top ranked candidate compounds. (B) Leave-one-out cross-validation to assess whether compounds at different levels of evidence (rows) are retrieved by our similarity search using the COVID-19 bait literature drugs (columns). The vertical line indicates the sum of support for observed candidates, and distributions represent the background expectation of the search. Figure produced on the 18th of April, 2020.

compounds as well as their MoA. In addition, the resulting list of molecules can be sorted following different criteria, including whether they are approved/experimental drugs, the cumulative level of support, or their similarity to specific COVID-19 literature drugs. Full and partial tables can be downloaded and exported to several formats, including the SMILES string representation for all the compounds.

Literature. This tab lists the COVID-19 bait compounds extracted from the literature, together with their level of experimental evidence and, if known, the MoA that confers efficacy against SARS-CoV-2.

Documentation. Here, we present a brief description of the methodological strategy, and more importantly, we offer updated statistics and benchmarks of the resource. In particular, we quantify the number of literature bait compounds available at each level of evidence and MoA (Figure 3A,B) and project CC signatures on a 2D plane to offer a global view of the chemical space explored by our resource (Figure 3C,D). We see that, while significantly diverse, COVID-19 bait compounds cluster in certain regions of the chemical space, and we find new candidate molecules in their vicinity. Reassuringly, when we analyze the therapeutic categories of the top-ranked candidates, as expected, we retrieve a significant number of anti-infective drugs (Figure 4A). Other therapeutic categories such as hormonal treatments are enriched after the highest-ranking compounds. Note that, for this enrichment analysis, only drug molecules could be considered since ATC annotations are not available for most of the compounds in the CC. Finally, we perform a leave-one-out cross-validation to assess whether bait compounds can be retrieved by our similarity search. Figure 4B shows that known COVID-19 drugs are significantly up-ranked when using and evaluating all levels of evidence (Figure 4B).

Contribute. Through this form, users can contribute to the resource by including their molecules of interest. We require the name and SMILES representation of the molecules as well as their level of experimental evidence, MoA, and references, if available. After each submission, we manually check the data and incorporate it in the next daily update.

Code. This links to the Gitlab repository containing the complete code to run the pipeline and analyze results.

Overall, we believe that the tool presented herein explores regions of the bioactive chemical space that could be relevant to COVID-19 treatment. Our web-based resource is updated daily and can be used to dynamically search for candidates related to COVID-19 drugs with varying levels of evidence and MoA. Therefore, our resource will be useful to a broad range of COVID-19 drug discovery approaches, ranging from those seeking a repurposing opportunity to those departing from the *in vitro* screening of compounds.

AUTHOR INFORMATION

Corresponding Authors

Miquel Duran-Frigola − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*; orcid.org/0000-0002-9906-6936; Email: miquel.duran@irbbarcelona.org

Patrick Aloy − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain*; Email: patrick.aloy@irbbarcelona.org

Authors

Martino Bertoni − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Roi Blanco − *Amazon Search Science and AI, 08018 Barcelona, Catalonia, Spain*

Víctor Martínez − *Amazon Search Science and AI, 08018 Barcelona, Catalonia, Spain*

Eduardo Pauls − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine*

https://dx.doi.org/10.1021/acs.jcim.0c00420
J. Chem. Inf. Model. 2020, 60, 5730−5734

(IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain

Víctor Alcalde − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Gemma Turon − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Núria Villegas − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Adrià Fernández-Torras − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Carles Pons − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Lídia Mateo − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Oriol Guitart-Pla − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Pau Badia-i-Mompel − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Aleix Gimeno − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Nicolas Soler − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Isabelle Brun-Heath − *Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08020 Barcelona, Catalonia, Spain*

Hugo Zaragoza − *Amazon Search Science and AI, 08018 Barcelona, Catalonia, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00420

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101003633 (RiPCoN).

■ REFERENCES

(1) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727−733.

(2) Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265−269.

(3) Search | COVID-19. https://search.bvsalud.org/global-research-on-novel-coronavirus-2019-ncov/.

(4) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289.

(5) Cao, B.; Wang, Y.; Wen, D.; Liu, W.; Wang, J.; Fan, G.; Ruan, L.; Song, B.; Cai, Y.; Wei, M.; Li, X.; Xia, J.; Chen, N.; Xiang, J.; Yu, T.; Bai, T.; Xie, X.; Zhang, L.; Li, C.; Yuan, Y.; Chen, H.; Li, H.; Huang, H.; Tu, S.; Gong, F.; Liu, Y.; Wei, Y.; Dong, C.; Zhou, F.; Gu, X.; Xu, J.; Liu, Z.; Zhang, Y.; Li, H.; Shang, L.; Wang, K.; Li, K.; Zhou, X.; Dong, X.; Qu, Z.; Lu, S.; Hu, X.; Ruan, S.; Luo, S.; Wu, J.; Peng, L.; Cheng, F.; Pan, L.; Zou, J.; Jia, C.; Wang, J.; Liu, X.; Wang, S.; Wu, X.; Ge, Q.; He, J.; Zhan, H.; Qiu, F.; Guo, L.; Huang, C.; Jaki, T.; Hayden, F. G.; Horby, P. W.; Zhang, D.; Wang, C. A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. *N. Engl. J. Med.* **2020**, *382*, 1787.

(6) Grein, J.; Ohmagari, N.; Shin, D.; Diaz, G.; Asperges, E. Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N. Engl. J. Med.* **2020**, *382*, 2327.

(7) Lover, A. A. Quantifying treatment effects of hydroxychoroquine and azithromycin for COVID-19: a secondary analysis of an open label non-randomized clinical trial. *medRxiv* **2020**.

(8) Duran-Frigola, M.; Pauls, E.; Guitart-Pla, O.; Bertoni, M.; Alcalde, V.; Amat, D.; Juan-Blanco, T.; Aloy, P. Extending the small molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.* **2020**, in press, DOI: 10.1038/s41587-020-0502-7

(9) https://allenai.org/data/cord-19.

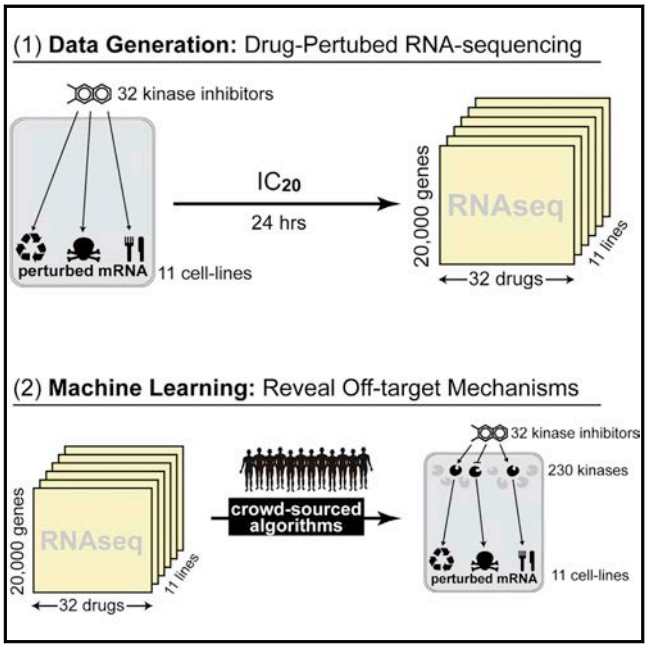
(10) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

Cell Reports
Medicine

Article

A community challenge for a pancancer drug mechanism of action inference from perturbational profile data

Graphical abstract



Highlights

- Drug-perturbed RNA sequencing data can be used to identify drug targets
- Technology-based drug-target definitions often subsume literature definitions
- Literature and screening datasets provide complementary information on drug mechanisms

Authors

Eugene F. Douglass, Jr., Robert J. Allaway, Bence Szalai, ..., Justin Guinney, Daniela S. Gerhard, Andrea Califano

Correspondence

ac2248@cumc.columbia.edu

In brief

Douglass et al. report the results of a crowdsourced challenge to develop machine-learning algorithms that use drug-perturbed transcriptome data to rapidly predict drug targets on a proteomic scale. Winning methods effectively predicted off-target binding of clinical kinase inhibitors and clarified disparate literature on these drugs' mechanisms of action.

Douglass et al., 2022, Cell Reports Medicine *3*, 100492
January 18, 2022 © 2021
https://doi.org/10.1016/j.xcrm.2021.100492

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

A community challenge for a pancancer drug mechanism of action inference from perturbational profile data

Eugene F. Douglass, Jr.,[1,2,17] Robert J. Allaway,[3,17] Bence Szalai,[4,17] Wenyu Wang,[5] Tingzhong Tian,[6] Adrià Fernández-Torras,[7] Ron Realubit,[1] Charles Karan,[1] Shuyu Zheng,[5] Alberto Pessia,[5] Ziaurrehman Tanoli,[5] Mohieddin Jafari,[5] Fangping Wan,[6] Shuya Li,[6] Yuanpeng Xiong,[8] Miquel Duran-Frigola,[7] Martino Bertoni,[7] Pau Badia-i-Mompel,[7] Lídia Mateo,[7] Oriol Guitart-Pla,[7] Verena Chung,[3] DREAM CTD-squared Pancancer Drug Activity Challenge Consortium, Jing Tang,[5] Jianyang Zeng,[6,9] Patrick Aloy,[7,10] Julio Saez-Rodriguez,[11,18] Justin Guinney,[3,18] Daniela S. Gerhard,[12,18] and Andrea Califano[1,13,14,15,16,18,19,*]

[1]Department of Systems Biology, Columbia University Irving Medical Center, 1130 Saint Nicholas Ave., New York, NY 10032, USA
[2]Pharmaceutical and Biomedical Sciences, University of Georgia, 250 W. Green Street, Athens, GA 30602, USA
[3]Computational Oncology Group, Sage Bionetworks, 2901 Third Ave., Ste 330, Seattle, WA 98121, USA
[4]Semmelweis University, Faculty of Medicine, Department of Physiology, Budapest, Hungary
[5]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[6]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[7]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[8]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[9]MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China
[10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[11]Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany
[12]Office of Cancer Genomics, National Cancer Institute, NIH, Bethesda, MD 20892, USA
[13]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, 1130 Saint Nicholas Ave., New York, NY 10032, USA
[14]Department of Medicine, Columbia University Irving Medical Center, 630 W 168th Street, New York, NY 10032, USA
[15]Department of Biochemistry & Molecular Biophysics, Columbia University Irving Medical Center, 701 W 168th Street, New York, NY 10032, USA
[16]Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W 168th Street, New York, NY 10032, USA
[17]These authors contributed equally
[18]Senior author
[19]Lead contact
*Correspondence: ac2248@cumc.columbia.edu
https://doi.org/10.1016/j.xcrm.2021.100492

SUMMARY

The Columbia Cancer Target Discovery and Development (CTD2) Center is developing PANACEA, a resource comprising dose-responses and RNA sequencing (RNA-seq) profiles of 25 cell lines perturbed with ~400 clinical oncology drugs, to study a tumor-specific drug mechanism of action. Here, this resource serves as the basis for a DREAM Challenge assessing the accuracy and sensitivity of computational algorithms for *de novo* drug polypharmacology predictions. Dose-response and perturbational profiles for 32 kinase inhibitors are provided to 21 teams who are blind to the identity of the compounds. The teams are asked to predict high-affinity binding targets of each compound among ~1,300 targets cataloged in DrugBank. The best performing methods leverage gene expression profile similarity analysis as well as deep-learning methodologies trained on individual datasets. This study lays the foundation for future integrative analyses of pharmacogenomic data, reconciliation of polypharmacology effects in different tumor contexts, and insights into network-based assessments of drug mechanisms of action.

INTRODUCTION

Non-canonical drug targets are known to contribute to clinical toxicity due to off-target effects. More recent work, however, suggests that off targets may contribute to clinical efficacy.[1,2] Systematic, *de novo* elucidation of compound mechanisms of action (MoAs), including polypharmacology, is thus emerging as a critical, yet still highly elusive, problem in clinical oncology.

This is an open access article under the CC BY-NC-ND IGO license (http://creativecommons.org/licenses/by-nc-nd/3.0/igo/).

Availability of methodologies for the comprehensive assessment of on- and off-target drug binding could help discriminate between targets driving efficacy or toxicity and those producing non-relevant clinical effects.[3]

Traditionally, the molecular targets of a drug that comprise its MoA have been defined by detailed thermodynamic (binding strength) and crystallographic (binding structure) characterization of a drug's interaction with individual proteins.[4] This approach is quite effective, as it directly facilitates structure-based drug design. Unfortunately, such a "one-drug/one-target" paradigm is often insufficient to mechanistically elucidate clinical phenotypes induced by even classical drugs, such as aspirin.[5,6] As a result, there is an urgent need to systematically re-assess drug MoAs in terms of their proteome-wide polypharmacology, which is defined as their ability to inhibit or activate proteins across a comprehensive, proteome-wide landscape.[7]

An increasing number of efforts have emerged to leverage large-scale perturbational profiles—e.g., mRNA profiles of cell lines and tissues before and after perturbation with a small compound—to predict both high-affinity binding targets and context-specific effectors.[8–11] The key assumption behind the use of perturbational profiles for this purpose is that differential gene expression is controlled by transcription factors and co-factors that represent the key downstream effectors of a compound's high-affinity binding targets (Figure 1A).[12,13] For example, the drug lapatinib inhibits EGFR (Epidermal Growth Factor Receptor), which induces gene expression changes via downstream transcription factors, including MYC and E2F family proteins (effectors).[14,15] As a result, drug-induced differential expression of MYC and E2F transcriptional targets may help distinguish EGFR inhibitors from inhibitors with a different downstream effector repertoire (Figure 1A). By extension, compounds targeting the same proteins should induce similar transcriptional signatures, which in turn can shed insight into its MoA (Figure S1).

The availability of compound- and tissue-specific dose-response curves (DRCs) further improves target assessments. First, it allows perturbational profile generation at high, yet sub-lethal, concentrations, thus preventing an emergence of cell-mediated responses, such as apoptosis or cellular stress, that would confound the true MoA. Second, the availability of differential cell viability in multiple molecularly distinct tissues further informs on compound activity based on distinct cellular and pathway architectures.[16]

Protein kinases represent one of the most thoroughly studied drug target classes. Protein kinase inhibitors are designed to target some of the most frequently mutated oncogenes, whose inhibition has been the hallmark of the oncogene addiction hypothesis.[17] Moreover, ATP-competitive pull-down assays enable effective and systematic binding affinity measurements across comprehensive protein kinase repertoires. The most comprehensive such evaluation to date, the Kinome-Binding Resource (KBR), measured the affinity of 230 clinically relevant kinase inhibitors across 255 kinases.[18] While restricted to this protein class, this dataset is well-suited to benchmarking methods aimed at predicting drug polypharmacology by providing criteria for the evaluation of systems pharmacology approaches (Figure S2).

To assess the research community's ability to predict kinase inhibitors' MoAs from drug perturbation profiles, we designed a DREAM Challenge[19,20] using KBR to provide ground-truth compound MoAs and PANACEA (Pan-cancer Analysis of Chemical Entity Activity)—a large-scale resource comprising genome-wide RNA sequencing (RNA-seq) profiles and matched DRCs of multiple cell lines following perturbation with hundreds of clinically relevant compounds—to provide data that may be used to predict compound MoAs. This significantly extends previous computational and systems pharmacology DREAM challenges by shifting the question from drug sensitivity to MoA prediction. The PANACEA data used in this challenge includes matched DRCs and perturbational RNA-seq profiles representing 11 cell lines following perturbation with approximately 400 clinical oncology drugs in replicate—including US Food and Drug Administration (FDA)-approved and late-stage (phase 2 and 3) compounds (Figure 1B). From the challenge, we specifically selected a subset of 32 kinase inhibitors that were also represented in the KBR (Figures 1C and 1D).

Challenge participants were provided with perturbational profiles and DRCs for each drug (blinded) and cell line (Figure 1E) and were asked to predict high-affinity binding targets for the 32 selected drugs by developing and training machine-learning algorithms using these data. Teams were further encouraged to use public data sources, such as the Cancer Cell Line Encyclopedia,[21] the Genomics of Drug Sensitivity in Cancer database,[22] and the CMap L1000 database,[23] and to leverage insights and models developed in previous DREAM Challenges.[8,19,24,25]

Previous projects, such as the IDG-DREAM Drug Kinase Binding Challenge[25] and the Multi-targeting Drug DREAM Challenge,[26] challenged the community to develop computational methods that leveraged publicly available chemical (e.g., chemical fingerprints, protein structures) and kinase binding data to predict drug-target interactions without using compound treatment data in a biological context. In contrast, this challenge asked the community to develop methods that could rank the proteins most affected by a compound using publicly available biochemical data. In order to make the challenge realistic, participants were blinded to the compound identity and to the fact that they were selected from the KBR collection. The challenge operated from December 2019 to February 2020 and led to the development and assessment of state-of-the-art approaches for inferring drug MoAs from perturbational profiles, as described herein.

RESULTS

Challenge requirements and data

In the CTD2 Pancancer Drug Activity DREAM Challenge, participants were asked to use DREAM-provided and publicly available pharmacogenomic datasets—including cell-line-matched DRCs and gene expression profiles of drug-naive and -perturbed cells (perturbational profiles)—to predict compound binding proteins (high-affinity targets) of 32 anonymized drugs (Figure 1A, Table S1). The DREAM-provided dataset comprised 704 DRCs and matched perturbational profiles of these 32 drugs (Table S2) in 11 cell lines representing molecularly distinct tumor

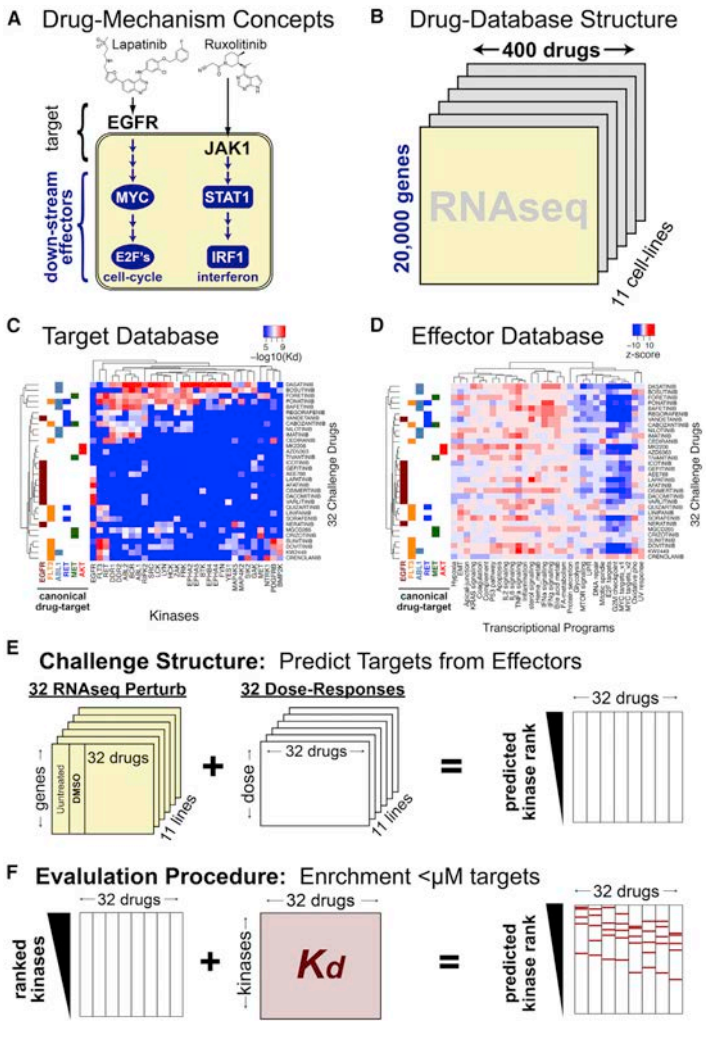


Figure 1. Underlying data and structure of polypharmacology community challenge

(A) Drug mechanism can be divided into direct binding targets and downstream effectors.

(B) The PANACEA-database-given transcriptional profiles of cell lines perturbed by clinical oncology drugs.

(C) Kinome-binding profiles of 32 kinase inhibitors.

(D) Transcriptional Hallmark programs induced by 32 kinase inhibitors (this data represents the average of two technical replicates where the same cell line was perturbed and sequenced on 2 different days).

(E) Challenge structure: participants are given perturbed RNA-seq and dose response data and asked to predict protein targets.

(F) Challenge evaluation: participant predictions are evaluated based on the enrichment of <μM binders within each drug target prediction vector.

subtypes in replicate (Figure 1E). All drugs used in the challenge had perturbational profiling in PANACEA and high-affinity binding characterization in the Kinome-Binding Resource (KBR).[18] While the full PANACEA manuscript is being published independently, all data related to this challenge is made contextually available with the publication of this manuscript (see Data and code availability).

Participants were encouraged to combine these data with additional publicly available resources to infer high-affinity binding targets of the 32 drugs from a repertoire of ~1,300 potential drug targets, defined as the union of all DrugBank-reported targets and the 255 kinases profiled in the KBR. Drug names were obfuscated to prevent trivial training of the algorithm on the KBR data (Figure 1F), and participants were not aware that the KBR data would be used as a gold standard for performance assessment.

Consistent with past DREAM studies, the challenge included a *leaderboard round* followed by a final *validation round*.[20] During the former, teams were allowed to submit up to five predictions for the 32 compounds, which were scored and posted to a public leaderboard. The purpose of this round was to enable experimentation and conceptual flexibility in model development by providing rapid feedback on the accuracy of the model while also encouraging competition among participants. A limit of 5 submissions was chosen to allow model refinement without compromising the statistical independence of the training and testing model, thus minimizing the potential for over-fitting. In the final validation round, participants were asked to submit their final model's predictions with the accompanying source code, thus allowing for objective validation of their methodology.

Model performance was evaluated according to each team's ability to prioritize bona fide targets of the 32 drugs, with the latter defined as having a dissociation constant $K_d < 1$ μM in the KBR, according to two complementary metrics, which were summarized by two sub-challenges:

Sub-challenge 1 (SC1) was designed to assess the ability of each submitted prediction to identify high-affinity binding targets ($K_d < 1$ μM) of each of the 32 compounds among the top 10 highest-scoring predicted targets. The rationale for selecting the top 10 targets was to represent the number of predictions that could be realistically validated using experimental assays. For each submitted drug prediction, a p value was calculated by filtering the prediction list to consider only

Table 1. Number of additional datasets used by participants for training and algorithm class

Team	SC1	SC2	No. of drug-AUC datasets[a]	No. of drug-mRNA datasets[b]	No. of drug-target datasets[c]	Total training datasets	Algorithm class
Netphar	12.6	70.9	6	1	4	11	similarity
SBNB	11.7	59.2	6	3	2	11	similarity
Xielab	13.8	50.3	6	2	1	9	similarity
Atom	17.4	49.3	–	2	4	6	NN
DMIS_PDA	13.8	35.2	–	2	1	3	NN
Theragen	15.1	17.3	–	2	1	3	similarity
Signal	6.3	6.1	–	1	2	4	regression
TeamAxolotl	6.2	1.1	–	–	2	3	NN
AMbeRland	3.3	1.1	–	–	–	0	unsup.
SenthamizhaV	7.4	0.9	–	–	–	0	unsup.

[a]Drug sensitivity (AUC) databases include: NCI60,[27] GDSC,[22] CTRP,[28] gCSI,[29] CCLE,[21] and other manually curated data.
[b]Drug mRNA perturbation databases include: L1000-drugs, L1000-shRNA,[23] and CREEDS.[30]
[c]Drug target datasets include: DrugBank,[31] ChEMBL,[32] KEGG,[33] and MATADOR.[34]

targets in the KBR and comparing the number of bona fide targets (K_d < 1 μM in the KBR) in the top 10 predicted targets to a null model generated from all possible targets and was similarly filtered to consider only targets in the KBR. A final integrated score was computed by averaging the -$\log_2$(p value) for each drug across all 32 drugs.

Sub-challenge 2 (SC2) was designed to assess the ability of each submitted prediction to accurately rank all the (for the participants) unknown bona fide targets (K_d < 1 μM in the KBR) of each of the 32 compounds by computing their enrichment—and associated p value—within the ranked list of predicted targets. The rationale for this second metric was to provide a more comprehensive and fine-grained comparison of the different methodologies (Figure 1F). Similar to SC1, a final integrated score was computed by averaging the -$\log_2$(p value) for each drug across all 32 drugs.

Challenge results

During the leaderboard phase, 21 teams contributed 86 prediction matrices of which 39 (45%) showed a geometric mean (across drugs for each team) p value of <0.01 for both SC1 and SC2. Interestingly, SC1 and SC2 scores revealed distinct distribution profiles: on average, most predictions were statistically significantly enriched on the top 10 target metric (SC1) but not on the entire list enrichment (SC2) (Figures S3A and S3B)

Consistent with previous DREAM Challenges, we assessed whether the performances across teams were statistically different for both sub-challenges by estimating a Bayes factor using a bootstrap analysis (see STAR Methods). The Bayes factor is a metric used to compare two (or more) statistical models; a model with a Bayes factor BF $\leq$3 indicates that the model is statistically indistinguishable from the top-ranked model. Figures S3C and 3D summarize the results of this analysis, with each box showing a team's bootstrapped scores, and the color of the box indicating the Bayes factor relative to the top performer. Using this criteria, Team Atom and Team Netphar were confirmed as the top performers in SC1 and SC2, respectively (Figures S3C and S3D), while team SBNB was a close second in SC2 (Bayes factor 3–5). A description of the algorithms from teams Atom, Netphar, and SBNB is provided in the STAR Methods.

When scoring the algorithms for the challenge, we filtered the predictions to the 255 kinases in the gold-standard dataset (i.e., KBR compendium). However, it would be possible in principle for challenge participants to rank kinase targets in the correct order but below incorrect targets not included in the KBR, such that this filtering step would boost their performance relative to participants who had ranked kinase targets in the correct order and above non-kinase targets. To address this issue, SC1 and SC2 results were re-scored considering the full list of 1,259 targets. As with the scoring for the main challenge, SC1 evaluated the top 10 predictions to assess whether any given top 10 overlapped with the gold-standard dataset. SC2 looked at the rank of the gold-standard targets within the submitted predictions. This analysis (Figures S3E and S3F) changed the ranking of some teams in SC1, most notably Team Theragen, whose original second place fell to ninth place when non-KBR targets were not prefiltered. In addition, Team Netphar's SC1 performance substantially increased, moving from 5th to 1st position.

To better understand the models and the difference in their performances, we examined sub-challenge scores on an individual drug basis (Figures S3G and S3H). Two clusters emerged, which separated teams based on whether they had used additional training datasets to train their algorithms. (Table 1). In general, consistent with prior results on the value of evidence integration,[25] overall performance was positively correlated with the number of additional databases utilized in the analysis, accounting for 27% of the variance in SC1 and a remarkable 82% of the variance in SC2.

Training data source contribution to model performances

Both SC2 winning teams, Netphar and SBNB, employed multiple highly curated datasets for training their algorithm. Netphar relied on the multi-database resources DrugComb (cytotoxicity)[35] and DrugTargetCommons (drug targets),[36] and SBNB relied on the multi-modality ChemicalChecker database.[37] Figure 2 provides a high-level conceptual summary of the types of

CellPress
OPEN ACCESS

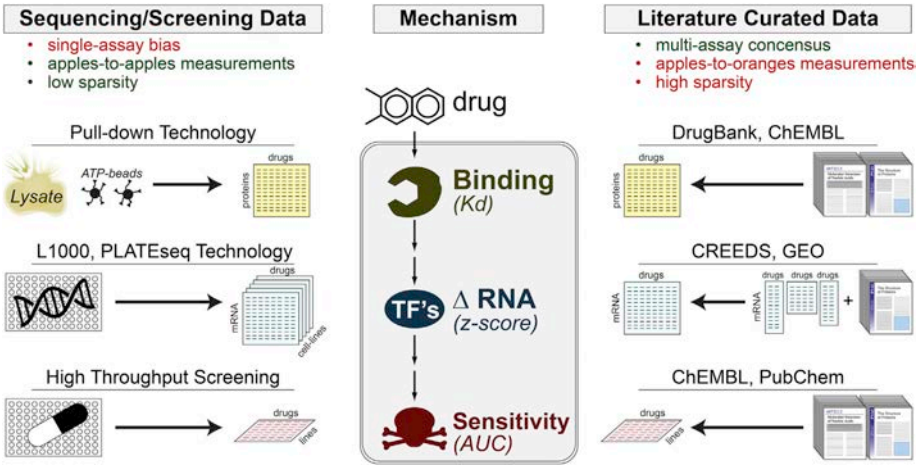


Figure 2. The universe of training data used in this challenge

Drug-perturbation datasets can be divided into two major categories: technology-based and literature-based, each with distinct limitations.

data sources included in these meta-databases organized by data type and source.

Overall, the datasets used to train the algorithms could be divided into two main categories: experimental screening-based and literature curation-based (Figure 2). Screening approaches have the advantages of providing measurements that are quantitative, directly comparable, and systematic (i.e., low sparsity). However, they may suffer from technological platform bias. Literature curation has the advantage of reflecting a multi-laboratory consensus but suffers from the disparate, *ad hoc* nature of the measurements and from lack of systematic assessment (high sparsity) (Figure 2). Team performance was further stratified based on whether they relied on (1) drug-target databases, (2) drug-perturbational databases, and/or (3) cytotoxicity databases. As further discussed below, drug-target and -perturbational databases provided the greatest accuracy boost across all drugs.

Critically, all teams chose to use literature-based datasets for identifying candidate drug targets (Figure 2). This is an important detail because while methods were trained on literature-based "drug-target" definitions, they were eventually evaluated based on objective, high-accuracy ATP-competitive assays (Figure 1F). To better understand the overlap between literature- and ATP-based drug targets, we evaluated the overlap between DrugBank and KBR targets (Figure 3). Specifically, we measured the number of DrugBank-reported protein kinase targets that were recovered across a range of affinity thresholds from 1 nM to 10 μM in the KBR (Figure 3A). Encouragingly, almost 80% of them were identified in the KBR using a $K_d < 1$ μM threshold (Figure 3A), consistent with a common "rule-of-thumb" for drug-lead development.[4]

Interestingly, while a 1 μM threshold identified the majority of DrugBank kinase targets, it also revealed the presence of a significant number of new targets not reported in DrugBank (Figure 3B). Overall, this shows that while DrugBank is mostly recapitulated by the KBR, the reverse is not true, suggesting that DrugBank may not contain all high-affinity targets of a drug. A key question raised by this comparison is whether the winning method's performance may have been driven entirely by canonical DrugBank targets. To address this question, we evaluated the ratio between the scores of the top three winning teams when either DrugBank or KBR targets were used as bona fide high-affinity targets of the 32 drugs used in the challenge (Figure S4). While the scores based on DrugBank targets were consistently higher (Netphar, 3:2; SBNB, 4:2; and Atom, 1.7:1.4), they all showed positive enrichment within the prediction vector (Figure S4). This result implies that literature-curated drug targets can be successfully used to bootstrap the polypharmacology analysis of otherwise uncharacterized drugs, thus further supporting the value of these resources. However, this may reduce algorithm performance for new compounds that are not yet included in any database.

In addition to DrugBank, two additional drug target databases—ChEMBL[32] and DrugTargetCommons[36]—were used by the top-performing teams. Plotting the overlap of all drug-target pairs across all four drug-target databases, only 121 targets (34%) were found to be unique to the KBR (Figure 3C). Taken together, these databases provided up to 2,386 additional drug-target interactions, of which 520 (21%) were evaluated in the KBR but were found to have affinities >1 μM, suggesting that they are false positive drug-target interactions (Figure 3D).

We compared the ranked performance of each prediction using these various databases as gold standards (KBR/Kinome, ChEMBL, DrugTargetCommons, KinomeScan, and DrugBank) to evaluate the stability of the predictions with different ground truths (Figure S5). As is expected, different ground-truth datasets have a substantial effect on team ranking, though the SC2 metric (rank across all targets for which data are available in

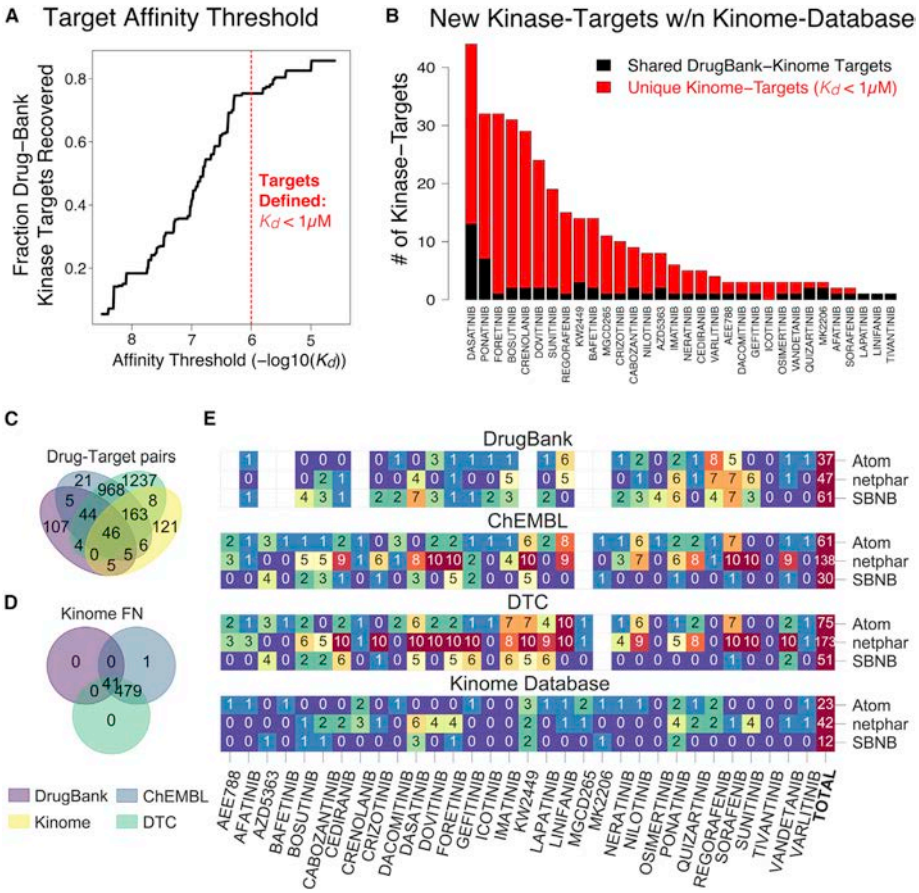


Figure 3. Comparison of DrugBank and kinome drug target definitions

(A) An affinity threshold of 1 μM within the kinome database successfully recovered almost 80% of the kinase targets within DrugBank.

(B) The kinome-defined drug targets appear to reveal a large number of new drug-targets (in red) in addition to the canonical drug targets (in black).

(C) Drug target pairs overlap across four drug target universes.

(D) Drug target pairs not detected in the kinome database used for PANACEA evaluation.

(E) Number of successful top 10 predictions for each drug and team across the different drug target universes.

the gold-standard dataset) is more stable than the SC1 metric (rank across top ten targets only).

Interestingly, when comparing the overlap of the top 10 targets predicted by the winning teams in each database, the observed differences strongly reflect the training datasets used by each team (Figure 3E). For instance, as one would expect, SBNB and Netphar results were biased toward DrugBank and DrugTargetCommons targets, respectively.

Kinase groups have distinct transcriptional programs

We next explored drivers of model performance by examining prediction accuracy for individual kinase inhibitor groups (Figures 4A and 4B).[38] Significant heterogeneity in methods performance across individual drugs was observed, suggesting that differences in modeling strategies (see the next section) may be leveraged to predict different drug classes. For instance, all winning methods performed better on the tyrosine kinase inhibitors group than on any other kinase group (Figure 4C).

Based on this observation, we hypothesized that specific kinase groups and families may be associated with distinct transcriptional programs. To evaluate the general relationships between kinase targets and mRNA programs, we assessed the correlation of the KBR-reported K_d with transcriptional hallmarks (as detailed in Supplemental information) across 84 drugs present in both databases (Figure 4D). This correlation matrix is plotted with phylogenetic tree-based kinase groups annotated

CellPress
OPEN ACCESS

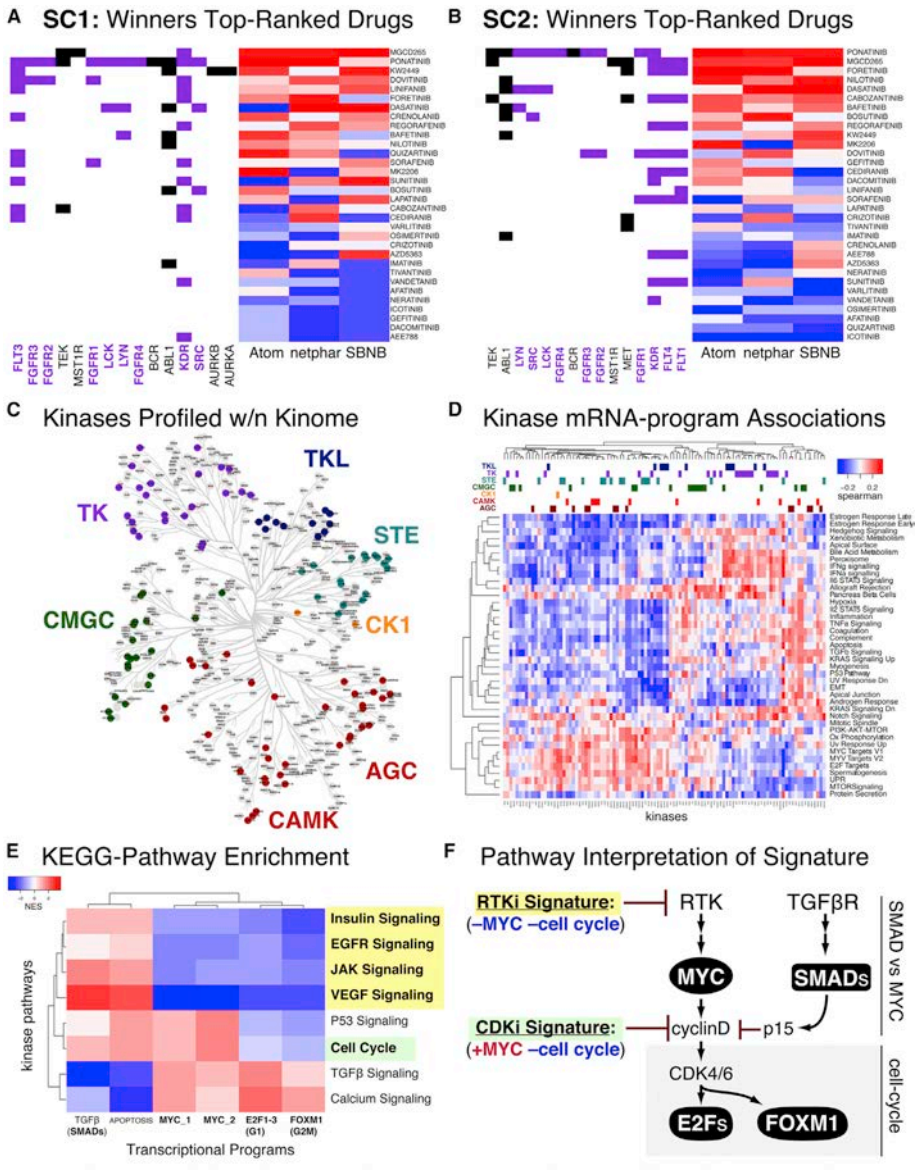


Figure 4. Different kinase pathways show distinct mRNA signatures when inhibited

(A and B) Across all models, tyrosine kinase (TK)-targeting drugs performed the best.

(C) Distribution of kinases profiled across the Human Kinome annotated by kinase group.

(D) Correlation of kinase-binding data with transcriptional program.

(E and F) KEGG pathway transformation of kinase space from (C) revealed pathway-specific transcriptional signatures

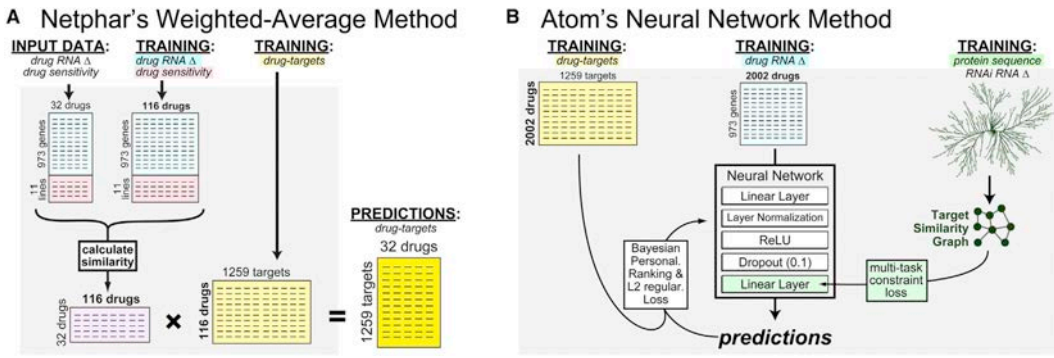


Figure 5. Comparison of the two winning strategies: weighted similarity and neural networks

(A) Team Netphar (who won SC2) used a simple matrix manipulation procedure to predict drug targets.

(B) Team Atom (who won SC1) used a protein-sequence-trained neural network.

on the top bars. Examining the protein kinase mRNA program matrix, no strong kinase group clustering was observed, indicating that kinase class is not generally sufficient to predict downstream transcriptional effects (Figure 4D, columns), although tyrosine kinases showed weak clustering with proliferation programs (Figure 4D, rows, bottom cluster): E2F targets, MYC targets, G2M checkpoint, oxidative phosphorylation, and mTOR-signaling.

To better understand the nature of the biological pathways underlying this association, we projected kinases into the KEGG pathway space[39] (see Method details for figures), which yielded a matrix of associations between kinase-signaling pathways and downstream transcriptional programs (Figure 4E). This analysis revealed a distinct pattern of transcriptional signatures that distinguished tyrosine kinase inhibitors from cell-cycle inhibitors and TGFβ inhibitors (Figure 4E), consistent with the known hierarchical structure of these signaling pathways, e.g., MYC and cell-cycle suppression via RTK-inhibition, in contrast to cell-cycle but not MYC suppression via CDK-inhibition (Figure 4F).[14]

Methodological summary

Overall, the methods submitted to the final validation round could be broken into three general categories:

1. Methods relying on a weighted average of differential gene expression and area under the curve (AUC)-based DRC similarity across drugs and drug targets. These included Netphar, SBNB, Xielab, and Theragen.
2. Methods relying on neural networks trained on prior information relating differential gene expression to drug-targets. These included Atom, DMIS_PDA, and TeamAxolotl
3. Methods based on fully unsupervised data transformation combining differential gene expression and DRC data. These included AMbeRland, SenthamizhamV, and Signal.

Generally, similarity-weighted average methodologies performed best in SC2 (Netphar 1st, SBNB 2nd, and Theragen 3rd)—i.e., they were better at predicting the entire range of targets— while Neural Network-based methodologies performed best in SC1 (Atom 1st and DMIS_PDA 3rd)—i.e., they were better at predicting targets in the range that could lead to realistic experimental validation. Fully unsupervised methods showed the worst performance. Nonetheless, they achieved statistical significance without leveraging any prior knowledge, suggesting the potential for mechanistic insight that could be combined with prior knowledge in future approaches.

In addition, there were differences in the training datasets used by algorithms in the first two categories. While weighted similarity methods used both transcriptional and cytotoxicity data (Figures 1E and 5A), neural network methods were trained exclusively with transcriptional profile data (see Contribution of drug sensitivity data). Intriguingly, the winning neural network method (Atom) used protein sequence data to further train their neural network (Figure 5B). This particular prior knowledge is worth noting, as it underlies several traditional approaches to structure-based drug design (e.g., ligand docking to homology models) and off-target discovery (e.g., BLAST searches in DrugBank[31]). Unfortunately, while such an approach may eventually help distinguish high-affinity binding targets from key downstream effectors, the use of protein sequence information improved Atom's performance only by a small, non-statistically significant amount.

Contribution of drug sensitivity data

Previous work has shown that training on drug sensitivity profile data can provide a comparable prediction performance to training on transcriptional signatures.[40] As such, we sought to investigate the contributions of drug sensitivity and drug transcriptional data to the performance of the winning Netphar model (which utilized both). Drug sensitivity training data was obtained from DrugComb, a curated database that includes batch-corrected drug sensitivities for both single drugs and drug combinations.[35] In addition to the commonly used IC_{50} (half maximum inhibitory concenration), DrugComb provides an AUC-based relative inhibition (RI) metric,[41] which captures both the potency and efficacy of drug responses (Figure S6A).

Examining correlations between predicted and gold-standard targets, we found that adding drug sensitivity data significantly improved prediction accuracy relative to transcriptional data alone (Figure S6B). Performance improvements were driven by several individual drugs whose targets were poorly predicted based on perturbational profile data only, including sunitinib, crizotinib, and crenolanib (Figure S6C). Finally, we tested whether the additional efficacy information provided by the RI metric improved model performance. Indeed, use of the RI metric in the predictive algorithms produced statistically significant, albeit marginal, overall improvement (median 0.18 compared to 0.19, paired Wilcoxon test p value = 0.025), highlighting the potential value of this metric in modeling drug properties.

DISCUSSION

MoA elucidation is a critical, yet time-consuming, step in the drug development process,[42] as it helps to identify on- and off-target effects supporting the activity of the compound (polypharmacology) as well as off-target effects that may cause unwanted toxicity. This addresses two major reasons for clinical trial failures: lack of safety and efficacy.[43,44] Failure rates may be substantially reduced if compound MoAs could be assessed more accurately and comprehensively (Figures S1A and S1B).

A drug MoA is defined as the set of biochemical interactors and effectors through which the drug produces its pharmacological effects, both positive and negative. These are almost invariably cell-context-specific. Despite its relevance, MoA characterization still represents a significant challenge, which is only partially addressed by experimental and computational strategies. Most of the experimental approaches rely on direct binding assays, such as ATP competitive pull-down,[18] affinity purification[45,46] or affinity chromatography assays.[47] These labor-intensive methods are generally limited to the identification of high-affinity binding targets rather than the full protein repertoire responsible for compound activity in a tissue and are often restricted to a specific protein family, such as protein kinases (Figure S1A). Thus, critically relevant targets outside of these relatively narrow confines may be missed, as shown by the recent reclassification of the MET tyrosine receptor kinase inhibitor tivantinib as a microtubule inhibitor.[48] Indeed, drug polypharmacology is emerging as a critical concept that increasingly impacts the mechanistic understanding of a drug's disease-specific impact, for instance via a field effect mediated by multiple targets rather than by their primary, high-affinity binding target (Figure S1B). For example, OTS964 is a compound originally developed as a MELK inhibitor and was recently shown to manifest its antitumoral activity via an entirely different target, CDK11, which had originally been missed in its MoA characterization.[1]

A few computational approaches have also been developed to infer MoA,[49–51] including using structural and/or genomic information,[52] text-mining algorithms,[53] or data mining.[54,55] As such, they rely on detailed three-dimensional structures of both the drug molecules and the target proteins or on prior knowledge of related compounds. More recently, systematic gene expression profiling (GEP) following compound perturbations in cell lines[8,11,23,56] has furthered the development of computational methods for MoA analysis (Figure S1B).

To address these issues, we hosted a DREAM community challenge to assess computational approaches for drug MoA inference from drug perturbational profiles using a comprehensive, experimental protein kinase binding affinity benchmark. The objective benchmark used in this challenge is the Kinome-Binding Resource (KBR), a systematic set of ATP-competitive binding assays assessing the ability of 230 candidate kinase inhibitor molecules to bind to one of 255 protein kinases (Figure S2).

A critical issue emerging from the evaluation of individual prediction performance and individual databases is that the concept of *drug target* is still poorly defined and inconsistent (Figure S2). For instance, even when restricting the comparison strictly to protein kinases, the comparison of targets defined in DrugBank versus KBR shows that the former may be missing data and may contain false positive targets whose binding affinity is >1 μM (Figure 3). Yet, it is unclear whether there may be false negatives in the KBR, for example, if allosteric binding or protein degradation occurred upon drug binding, as it would be missed by an ATP-competitive binding assay. More critically, it is unclear whether the targets reported in one database but not the other may play a relevant pharmacological role either in disease treatment or in the emergence of undesirable side effects.

While this was not the main objective of the DREAM Challenge, the study also provides significant insights on the network of effector proteins downstream of high-affinity binding targets. Indeed, the fact that the perturbational signature significantly contributed to correct target inference suggests that downstream transcriptional regulators represent a valuable reporter assay that can distinguish the MoA of different compounds (Figures 4 and S6). Furthermore, the analysis shows that availability of matched DRCs and perturbational profile data for each drug provided a significant contribution to the quality of the prediction. For instance, drugs such as sunitinib, crizotinib, and crenolanib produced significantly worse performances when the analysis was restricted to perturbational profiles but performed significantly better when DRCs and perturbational profile data were integrated.

An interesting observation that emerged from this challenge is that tyrosine kinase inhibitors were predicted with higher accuracy by all methods (Figures 4A–4C). Examining correlations between binding constants and transcriptional profiles, we found that tyrosine kinases inhibitors were mostly associated with suppression of proliferation signatures (Figure 4D). This is perhaps unsurprising, as growth factor control of the cell cycle is typically mediated by receptor tyrosine kinases. Looking at enrichment of KEGG pathways within Figure 4D's correlation matrix, we were able to identify a decoupling in the effects of MYC and the cell cycle (Figure 4E) that was consistent with the hierarchy of known proliferation pathways (Figure 4F). These results provide evidence that drug-perturbed transcriptional signatures can retain information on the signaling pathways directly downstream of molecular drug targets.

While we did not observe major differences between model performances based on modeling strategy, generally, similarity-weighted average methodologies performed best in SC2 (Figures S7 and S8), while neural-network-based methodologies performed best in SC1 (Figure S9). An important insight arising

from the challenge is that computational methods for MoA inference are best at identifying similarities between unknown compounds and compounds already reported in existing databases rather than at elucidating compound MoAs *de novo*. Indeed, all of the methodologies that did not rely on prior databases underperformed when compared with those that did. The fact that all the proposed methodologies produced statistically significant results suggests that genome-wide perturbational profiles bring *de novo* predictions of compound MoAs a step closer to being effectively useful in drug discovery.

For the best-performing drug classes, differential transcriptional signals could be traced to specific patterns of co-regulated transcriptional gene sets or hallmarks (Figures 1D, 4D, and 4E). These patterns can be directly explained by the hierarchical structure of kinase signaling cascades in canonical pathways (Figure 4F). Critically, this insight highlights the strengths and weaknesses of mRNA-based target inference where:

- Targets within the same pathway can be difficult to differentiate (e.g., EGFR, RAS, RAF, and MEK inhibitors) due to transcriptional phenocopying.
- Targets at pathway branch points are easier to predict due to the differential transcriptional effects they induce (e.g., RTK versus CDK versus TGFβ inhibitors).

For example, while RTK inhibitors could be effectively distinguished from CDK inhibitors, distinguishing the more subtle differences between drugs within each class should prove more challenging.

Overall, this work suggests that predictive models can leverage perturbational data to effectively infer the MoA of small molecules and to reveal biological and clinical insights about druggable pathways. Future studies using computational modeling to tackle this problem will be critical to the successful application of these methods. Specifically, developing a more systematic knowledge of drug targets, particularly for non-kinase targets, may improve the ability of the community to develop accurate models. Additional development and benchmarking of unsupervised prediction methods may also be required for the accurate prediction of targets of novel molecules. Finally, future work will be necessary to elucidate the best practices, limitations, and general applicability of these methods as a step in the drug discovery pipeline.

Limitations of the study

It is important to note that these finding are most applicable to kinases due to the focus of our drug library on kinase inhibitors. While kinase inhibitors form the largest class of targeted therapy (which often assume specificity), care should be taken in extending the results to other oncology drugs such as cell-cycle inhibitors and DNA-damaging agents. In addition, it should be noted that our perturbation data are collected on only 11 cell lines and so may not recapitulate the transcriptional effects of these 32 kinase inhibitors across all cancer types. Finally, while several different methods were evaluated in this challenge, other methods not evaluated in the present study may also be performant when applied to this machine-learning problem.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell line viability
- METHOD DETAILS
  - Collaborative methods overview
  - Compound titration curves
  - Perturbational profile generation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Profile normalization
  - Kinome and PANACEA data formatting
  - Baseline model
  - Scoring algorithms
  - Alternate gold standard evaluation
  - Determination of top performers and data leak
  - Detailed computational procedure Figures 1C and 1D
  - Detailed computational procedure Figure 3A
  - Detailed computational procedure Figures 4A and 4B
  - Detailed computational procedure Figure S4
  - Detailed algorithm for winning team "Netphar"
  - Detailed algorithm for winning team "SBNB" (Figure S8):
  - Detailed algorithm for winning team "ATOM" (Figure S9):

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2021.100492.

CONSORTIA

The members of the CTD2 Drug Activity DREAM Challenge Community Consortium are: Renata Retkute, Alidivinas Prusokas, Augustinas Prusokas, Andrea Degasperi, Yasin Memari, João M. L. Dias, Guillermo de Anda-Jáuregui, Santiago Castro-Dau, Cristóbal Fresno, Laura Gómez-Romero, Humberto Gutiérrez-González, Enrique Hernández-Lemus], Soledad Ochoa, José María Zamora-Fuentes, Yue Qiu, Di He, Lei Xie, Gwanghoon Jang, Jungsoo Park, Sungjoon Park, Buru Chang, Sunkyu Kim, Jaewoo Kang, Eugene F. Douglass Jr., Robert Allaway, Bence Szalai, Ron Realubit, Charles Karan, Wenyu Wang, Tingzhong Tian, Adrià Fernández-Torras, Jing Tang, Shuyu Zheng, Alberto Pessia, Ziaurrehman Tanoli, Mohieddin Jafari, Fangping Wan, Shuya Li, Yuanpeng Xiong, Jianyang Zeng, Miquel Duran-Frigola, Martino Bertoni, Pau Badia-i-Mompel, Lídia Mateo, Oriol Guitart-Pla, Patrick Aloy, Verena Chung, Julio Saez-Rodriguez, Justin Guinney, Daniela Gerhard, and Andrea Califano (see Data S1 for consortium author affiliations).

ACKNOWLEDGMENTS

This work was supported by the National Cancer Institute's Office of Cancer Genomics Cancer Target Discovery and Development (CTD^2) initiative. The results published here are based in whole or in part upon data generated by CTD^2 Network (https://ocg.cancer.gov/programs/ctd2/data-portal), established by the National Cancer Institute's Office of Cancer Genomics. P.A.

acknowledges the support of the Generalitat de Catalunya (RIS3CAT Emergents CECH: 001-P-001682 and VEIS: 001-P-001647), the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R), the European Research Council (SysPharmAD: 614944), and the European Commission (RiPCoN: 101003633). B.S. was supported by the Premium Postdoctoral Fellowship Program of the Hungarian Academy of Sciences (460044). A.C. was supported by a Cancer Target Discovery and Development Center grant (U01-CA168426) and by NIH shared instrumentation grants S10-OD012351 and S10-OD021764. A.C., J.G., R.J.A., and E.F.D. were supported through supplemental funding from the CTD2 program (U01-CA217862). J.T. acknowledges the support of European Research Council (ERC) starting grant DrugComb (no. 716063) and the Academy of Finland (no. 317680). J.T. and Z.T. were supported by European Commission H2020 EOSC-life (no. 824087). W.W. was funded by the Doctoral Program of Biomedicine, University of Helsinki, and K. Albin Johanssons stiftelse. M.J. was supported by the Academy of Finland (no. 332454). The Challenge team NetPhar acknowledges their colleagues Bulat Zagidullin and Jehad Aldahdooh for technical support, as well as the CSC – IT Center for Science, Finland, for computational resources.

AUTHOR CONTRIBUTIONS

E.F.D., R.J.A., and B.S. contributed equally to this work and have the right to list their names first on their CVs. Conceptualization, E.F.D., A.C., D.S.G., A.F.-T., M.D.-F., P.A., W.W., J.T., and T.T.; data curation, E.F.D., R.R., C.K., A.F.-T., M.D.-F., L.M., W.W., S.Z., A.P., Z.T., M.J., T.T., and S.L.; formal analysis, E.F.D., A.C., A.F.-T., M.D.-F., M.B., P.B.-i.-M., L.M., W.W., J.T., S.Z., M.J., A.P., T.T., S.L., R.J.A., and B.S.; funding acquisition, A.C., P.A., J.T., W.W., J.Z., and J.G.; investigation, E.F.D., A.C., D.G., A.F.-T., M.D.-F., M.B., P.B.-i.-M., L.M., P.A., J.T., W.W., A.P., M.J., T.T., F.W., S.L., Y.X., J.Z., R.J.A., B.S., and J.S.-R.; methodology, E.F.D., A.C., A.F.-T., M.D.-F., M.B., P.B.-i.-M., L.M., J.T., W.W., A.P., M.J., T.T., F.W., S.L., R.J.A., B.S., and J.S.-R.; project administration, E.F.D., A.C., D.G., M.D.-F., P.A., J.T., W.W., T.T., J.Z., R.J.A., and J.G.; resources, P.A., J.T., J.Z., V.C., and R.J.A., software, E.F.D., A.F.-T., M.D.-F., M.B., P.B.-i.-M., L.M., O.G.-P., W.W., J.T., S.Z., M.J., A.P., T.T., F.W., S.L., R.J.A., V.C., and B.S.; supervision, A.C., D.S.G., M.D.-F., P.A., J.T., J.Z., J.G., and J.S.-R.; validation, E.F.D., A.F.-T., M.D.-F., M.B., P.B.-i.-M., W.W., J.T., S.Z., T.T., F.W., R.J.A., and B.S.; visualization, E.F.D., A.F.-T., M.D.-F., W.W., F.W., T.T., R.J.A, and V.C.; writing – original draft, E.F.D., A.C., D.S.G., A.F.-T., M.D.-F., P.A., W.W., T.T., J.G., R.J.A., and B.S.; writing – review & editing, E.F.D., A.C., D.S.G., A.F.-T., M.D.-F., P.A., W.W., J.T., S.Z., M.J., Z.T., A.P., T.T., F.W., J.G., R.J.A., V.C., and J.S.-R.

DECLARATION OF INTERESTS

A.C. is founder, equity holder, and consultant of DarwinHealth, Inc., a company that has licensed the PANACEA database used in this manuscript from Columbia University. Columbia University is also an equity holder in DarwinHealth, Inc. J.S.-R. has received funding from GSK and Sanofi and personal fees from Travere Therapeutics. B.S. received consultation fees from Turbine, Ltd. All other authors declare no competing interests.

Received: February 5, 2021
Revised: August 8, 2021
Accepted: December 15, 2021
Published: January 18, 2022

REFERENCES

1. Lin, A., Giuliano, C.J., Palladino, A., John, K.M., Abramowicz, C., Yuan, M.L., Sausville, E.L., Lukow, D.A., Liu, L., Chait, A.R., et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci. Transl. Med. *11*, eaaw8412.
2. Dar, A.C., Das, T.K., Shokat, K.M., and Cagan, R.L. (2012). Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. Nature *486*, 80–84.
3. Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. *4*, 682–690.
4. Anderson, A.C. (2003). The process of structure-based drug design. Chem. Biol. *10*, 787–797.
5. Bedard, P.L., Hyman, D.M., Davids, M.S., and Siu, L.L. (2020). Small molecules, big impact: 20 years of targeted therapy in oncology. Lancet *395*, 1078–1088.
6. Proschak, E., Stark, H., and Merk, D. (2019). Polypharmacology by Design: A Medicinal Chemist's Perspective on Multitargeting Compounds. J. Med. Chem. *62*, 420–444.
7. Milletti, F., and Vulpetti, A. (2010). Predicting polypharmacology by binding site similarity: from kinases to the protein universe. J. Chem. Inf. Model. *50*, 1418–1431.
8. Bansal, M., Yang, J., Karan, C., Menden, M.P., Costello, J.C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R., et al.; NCI-DREAM Community; NCI-DREAM Community (2014). A community computational challenge to predict the activity of pairs of compounds. Nat. Biotechnol. *32*, 1213–1222.
9. Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., and di Bernardo, D. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc. Natl. Acad. Sci. USA *107*, 14621–14626.
10. Shen, Y., Alvarez, M.J., Bisikirska, B., Lachmann, A., Realubit, R., Pampou, S., Coku, J., Karan, C., and Califano, A. (2017). Systematic, network-based characterization of therapeutic target inhibitors. PLoS Comput. Biol. *13*, e1005599.
11. Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R., et al. (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441–451.
12. Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. *48*, 838–847.
13. Alvarez, M.J., Subramaniam, P.S., Tang, L.H., Grunn, A., Aburi, M., Rieckhof, G., Komissarova, E.V., Hagan, E.A., Bodei, L., Clemons, P.A., et al. (2018). A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. Nat. Genet. *50*, 979–989.
14. Kolch, W., Halasz, M., Granovskaya, M., and Kholodenko, B.N. (2015). The dynamic control of signal transduction networks in cancer cells. Nat. Rev. Cancer *15*, 515–527.
15. Blumer, K.J., and Johnson, G.L. (1994). Diversity in function and regulation of MAP kinase pathways. Trends Biochem. Sci. *19*, 236–240.
16. Schenone, M., Dančík, V., Wagner, B.K., and Clemons, P.A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. Nat. Chem. Biol. *9*, 232–240.
17. Weinstein, I.B. (2002). Cancer. Addiction to oncogenes–the Achilles heel of cancer. Science *297*, 63–64.
18. Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P.-A., Reinecke, M., Ruprecht, B., Petzoldt, S., Meng, C., et al. (2017). The target landscape of clinical kinase drugs. Science *358*, eaan4368.
19. Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al.; NCI DREAM Community (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212.
20. Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., and Stolovitzky, G. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. *17*, 470–486.
21. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The

Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

22. Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740–754.
23. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437–1452.e17.
24. Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., et al.; AstraZeneca-Sanger Drug Combination DREAM Consortium (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat. Commun. *10*, 2674.
25. Cichońska, A., Ravikumar, B., Allaway, R.J., Wan, F., Park, S., Isayev, O., Li, S., Mason, M., Lamb, A., Tanoli, Z., et al.; IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium (2021). Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat. Commun. *12*, 3307.
26. Xiong, Z., Jeon, M., Allaway, R.J., Kang, J., Park, D., Lee, J., Jeon, H., Ko, M., Jiang, H., Zheng, M., et al. (2021). Crowdsourced identification of multi-target kinase inhibitors for RET- and TAU-based disease: the Multi-Targeting Drug DREAM Challenge. PLoS Comput. Biol. *17*, e1009302.
27. Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.
28. Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. *5*, 1210–1223.
29. Haverty, P.M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R.M., Martin, S., Settleman, J., Yauch, R.L., and Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature *533*, 333–337.
30. Wang, Z., Monteiro, C.D., Jagodnik, K.M., Fernandez, N.F., Gundersen, G.W., Rouillard, A.D., Jenkins, S.L., Feldmann, A.S., Hu, K.S., McDermott, M.G., et al. (2016). Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat. Commun. *7*, 12846.
31. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46* (*D1*), D1074–D1082.
32. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. *47* (*D1*), D930–D940.
33. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Res. *47* (*D1*), D590–D595.
34. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G., Gewiess, A., Jensen, L.J., et al. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. *36*, D919–D922.
35. Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., and Tang, J. (2019). DrugComb: an integrative cancer drug combination data portal. Nucleic Acids Res. *47* (*W1*), W43–W51.
36. Tang, J., Tanoli, Z.-U.-R., Ravikumar, B., Alam, Z., Rebane, A., Vähä-Koskela, M., Peddinti, G., van Adrichem, A.J., Wakkinen, J., Jaiswal, A., et al. (2018). Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. Cell Chem. Biol. *25*, 224–229.e2.
37. Duran-Frigola, M., Pauls, E., Guitart-Pla, O., Bertoni, M., Alcalde, V., Amat, D., Juan-Blanco, T., and Aloy, P. (2020). Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat. Biotechnol. *38*, 1087–1096.
38. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934.
39. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45* (*D1*), D353–D361.
40. Szalai, B., Subramanian, V., Holland, C.H., Alföldi, R., Puskás, L.G., and Saez-Rodriguez, J. (2019). Signatures of cell death and proliferation in perturbation transcriptomics data-from confounding factor to effective prediction. Nucleic Acids Res. *47*, 10010–10026.
41. Malyutina, A., Majumder, M.M., Wang, W., Pessia, A., Heckman, C.A., and Tang, J. (2019). Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer. PLoS Comput. Biol. *15*, e1006752.
42. Scannell, J.W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Nat. Rev. Drug Discov. *11*, 191–200.
43. Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. *3*, 711–715.
44. Wehling, M. (2009). Assessing the translatability of drug projects: what needs to be scored to predict success? Nat. Rev. Drug Discov. *8*, 541–546.
45. Hirota, T., Lee, J.W., St John, P.C., Sawa, M., Iwaisako, K., Noguchi, T., Pongsawakul, P.Y., Sonntag, T., Welsh, D.K., Brenner, D.A., et al. (2012). Identification of small molecule activators of cryptochrome. Science *337*, 1094–1097.
46. Ito, T., Ando, H., Suzuki, T., Ogura, T., Hotta, K., Imamura, Y., Yamaguchi, Y., and Handa, H. (2010). Identification of a primary target of thalidomide teratogenicity. Science *327*, 1345–1350.
47. Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.
48. Basilico, C., Pennacchietti, S., Vigna, E., Chiriaco, C., Arena, S., Bardelli, A., Valdembri, D., Serini, G., and Michieli, P. (2013). Tivantinib (ARQ197) displays cytotoxic activity that is independent of its ability to bind MET. Clin. Cancer Res. *19*, 2381–2392.
49. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., et al. (2009). Predicting new molecular targets for known drugs. Nature *462*, 175–181.
50. Lomenick, B., Hao, R., Jonai, N., Chin, R.M., Aghajan, M., Warburton, S., Wang, J., Wu, R.P., Gomez, F., Loo, J.A., et al. (2009). Target identification using drug affinity responsive target stability (DARTS). Proc. Natl. Acad. Sci. USA *106*, 21984–21989.
51. Miller, W.H., Jr., Schipper, H.M., Lee, J.S., Singer, J., and Waxman, S. (2002). Mechanisms of action of arsenic trioxide. Cancer Res. *62*, 3893–3903.
52. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics *24*, i232–i240.
53. Li, J., Zhu, X., and Chen, J.Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput. Biol. *5*, e1000450.
54. Hansen, N.T., Brunak, S., and Altman, R.B. (2009). Generating genome-scale candidate gene lists for pharmacogenomics. Clin. Pharmacol. Ther. *86*, 183–189.
55. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. *18*, 133–145.
56. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., et al. (2006). The

CellPress
OPEN ACCESS

Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935.

57. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

58. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

59. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

60. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

61. Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat. Med. *23*, 405–408.

62. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Softw. *4*, 1686.

63. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

64. Metz, K.S., Deoudes, E.M., Berginski, M.E., Jimenez-Ruiz, I., Aksoy, B.A., Hammerbacher, J., Gomez, S.M., and Phanstiel, D.H. (2018). Coral: Clear and Customizable Visualization of Human Kinome Data. Cell Syst. *7*, 347–350.e1.

65. Smirnov, P., Kofia, V., Maru, A., Freeman, M., Ho, C., El-Hachem, N., Adam, G.-A., Ba-Alawi, W., Safikhani, Z., and Haibe-Kains, B. (2018). PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res. *46* (*D1*), D994–D1002.

66. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45* (*D1*), D945–D954.

67. Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2012). BPR: Bayesian Personalized Ranking from Implicit Feedback. arXiv, 1205.2618v1. https://arxiv.org/abs/1205.2618.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
AEE788	SelleckChem	S1486
Afatinib	SelleckChem	S1011
AZD5363	SelleckChem	S8019
Bafetinib	SelleckChem	S1369
Bosutinib	SelleckChem	S1014
Cabozantinib	SelleckChem	S1119
Cediranib	SelleckChem	S1017
Crenolanib	SelleckChem	S2730
Crizotinib	SelleckChem	S1068
Dacomitinib	SelleckChem	S2727
Dasatinib	SelleckChem	S1021
Dovitinib	SelleckChem	S1018
Foretinib	SelleckChem	S1111
Gefitinib	SelleckChem	S1025
Icotinib	SelleckChem	S2922
Imatinib	SelleckChem	S2475
KW2449	SelleckChem	S2158
Lapatinib	SelleckChem	S2111
Linifanib	SelleckChem	S1003
MGCD365	SelleckChem	S1361
MK2206	SelleckChem	S1078
Neratinib	SelleckChem	S2150
Nilotinib	SelleckChem	S1033
Osimertinib	SelleckChem	S7297
Ponatinib	SelleckChem	S1490
Quizartinib	SelleckChem	S1526
Regorafenib	SelleckChem	S1178
Sorafenib	SelleckChem	S1040
Sunitinib	SelleckChem	S1042
Tivantinib	SelleckChem	S2753
Vandetanib	SelleckChem	S1046
Varlitinib	SelleckChem	S2755
Critical commercial assays		
CellTiter-Glo Luminescent Viability Assay	Promega	G7570
Deposited data		
PANACEA gene expression profiles.	This paper	GEO: GSE186341
Experimental models: Cell lines		
AsPC-1	ATCC	ATCC Cat# CRL-1682; RRID:CVCL_0152
DU 145	ATCC	ATCC Cat# HTB-81; RRID:CVCL_0105
EFO-21	DSMZ	DSMZ Cat# ACC-235; RRID:CVCL_0029
HCC1143	ATCC	ATCC Cat# CRL-2321; RRID:CVCL_1245

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HF2597	Henry Ford	N/A
HSTS	Broad	RRID:CVCL_L296
KRJ1	Califano Lab	RRID:CVCL_8886
LNCaP	ATCC	ATCC Cat# CRL-1740; RRID:CVCL_1379
NCI-H1793	ATCC	ATCC Cat# CRL-5896; RRID:CVCL_1496
PANC-1	ATCC	ATCC Cat# CRL-1469; RRID:CVCL_0480
U-87 MG	ATCC	ATCC Cat# HTB-14; RRID:CVCL_0022
Software and algorithms		
STAR aligner, 2.5.2b	Dobin et al.[57]	https://github.com/alexdobin/STAR
Limma 3.48.1	Ritchie et al.[58]	https://bioconductor.org/packages/release/bioc/html/limma.html
DESeq2	Love et al.[59]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ComBat	Johnson et al.[60]	https://bioconductor.org/packages/release/bioc/html/sva.html
Analysis code	This paper	https://github.com/Sage-Bionetworks-Challenges/CTD2-Panacea-Challenge

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Andrea Califano (ac2248@cumc.columbia.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data used in the challenge, submission writeups, and other Challenge resources can be found at https://www.doi.org/10.7303/syn20968331.

Raw data is also available through the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under accession number GEO: GSE186341. Code for scoring the predictions and for generating the null models is available here: https://github.com/Sage-Bionetworks-Challenges/CTD2-Panacea-Challenge, and a Docker container that was used to deploy the scoring algorithm in this challenge is available to all registered Synapse users via the Synapse Docker registry (https://www.synapse.org/#!Synapse:syn20968331/wiki/597042). Links to all submitted writeups, Docker containers containing method source code, and Docker documentation for this challenge can be found on the Challenge wiki: https://www.synapse.org/#!Synapse:syn20968331/wiki/607259.

Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell line viability

Cell-lines were obtained from ATCC and cultured using prescribed conditions. To determine optimal seeding density for compound titrations (i.e., cell-growth is linear for the duration of experiment), 3.2 million cells of each cell line were plated and viability measured using CelTiter Glo (Promega Corp.) at 24, 48, 72 and 96 hours. Briefly, 10 mL of 320,000 cells/mL cell-solution was added to column 11 of a 12w deep-well plate. 5mL from column 11 was then serially diluted 1:1 from column 11 through column 2. The Hamilton MicroLab automated liquid handling system's Cell Line Optimization protocol was used to split the 12 w plates between 4 384 well plates for incubation. 384 well plates were stored in the incubator and at 24, 48, 72 and 96 hours 1 plate was removed and allowed to sit for 15 minutes at room temperature. 25 uL of Cell Titer Glo was added to each well and shaken at 800rpm for 5 min. Finally luminescence was read using the EnVision Multi-Label Reader (Perkin Elmer Inc.). The seeding density which resulted in linear increase of the cells was used for the perturbation experiments.

METHOD DETAILS

Collaborative methods overview

The PANACEA database was developed in collaboration between Columbia University Irving Medical Centers (CUIMC)'s High Throughput Screening Center (HTS), Sulzberger Genome Center and the Califano Laboratory in the Department of Systems Biology. Briefly, HTS handled cell-culture, cell-perturbation experiments and RNA extraction; the Genome Center performed RNA sequencing and the Califano laboratory performed data normalization, quality control, benchmarking and scientific and statistical analysis.

Compound titration curves

To determine the 48h ED_{20} of each drug, cell lines were plated into 96-well tissue culture plates, in 100 μL total volume, and incubated at 37°C. After 16 hours the plates were removed from the incubator and compounds were transferred into assay wells (1 μL) in triplicate. Plates were then returned to the incubator. After 48 hours the assay plates were removed from the incubator and allowed to cool to room temperature prior to the addition of 100 μL of CellTiter-Glo (Promega Inc.) per well. The plates were then mechanically shaken for 5 minutes prior to readout on the EnVision Multi-Label Reader (Perkin Elmer Inc.) using the enhanced luminescence module. Relative cell viability was computed using matched DMSO control wells as reference. ED_{20} was estimated by fitting a four-parameter sigmoid model to the titration results.

Perturbational profile generation

Using the previously described plating and perturbation procedure we perturbed each cell-line with each drug at its 48h ED_{20} value (measured above) or its CMax concentration. In order to optimize the clinical translation potential of the perturbation databases, we used the CMax, defined as the maximum plasma concentration after the administration of the drug at the maximum tolerated dose in patients, (whenever available from published pharmacokinetic studies), as an upper bound for the perturbation studies (Table S1). The mRNA from these cells was isolated and profiled by PLATESeq (Nat. Commun. 2017, 8, 105) at 24h after each perturbation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Profile normalization

RNASeq reads were mapped for each well to the human reference genome assembly 38 using the STAR aligner,[57] version 2.5.2b. Individual plates counts files were then combined, normalized and corrected for batch effects. First, individual counts files were combined across genes and ERCC2 spike-in counts removed, yielding the raw counts file for each cell-line experiment. Second, raw counts were quantile normalized and variance stabilized based on the negative binomial distribution with the DESeq R system package.[59] To account for plate-based batch effects (which are common with drug-perturbed transcriptomic data) normalized expression was batch corrected using ComBat.[60]

Kinome and PANACEA data formatting

Kinome-binding data from Klaeger et al.[18] was downloaded at https://www.proteomicsdb.org/#projects/4257 via "Supplementary Table 3 Drug Matrices." Raw data was transformed to -log10 scale and NA's replaced with the matrix maximum -log10(Kd) of ~4.3 to represent the limit of detection of the technology. PANACEA differential gene expression data was calculated using a moderated Student's t test as implemented in the limma package[58] from Bioconductor (version 3.48.1) with respect to pooled DMSO controls across all cell-line plates.

Baseline model

For the baseline model, we used drug perturbation gene expression data from the LINCS-L1000 project[23] and drug-target information from the Drug Repurposing Hub.[61] We calculated consensus signatures[40] for each drug with known target molecules. The DREAM-PANACEA gene expression dataset was standardized using the control measurements, and consensus signature (average across cell lines) was calculated for each DREAM-PANACEA drug. We calculated the similarity (Spearman's correlation) matrix between the LINCS and DREAM-PANACEA drug signatures, using only the measured (landmark) genes of LINCS-L1000. For each DREAM-PANACEA drug, we performed target enrichment (including the mode of action (i.e., activation or inhibitor), using the *viper* R package[12]) using the drug similarity vector and the known targets of the LINCS drugs. The normalized enrichment scores from target enrichment were further rank transformed for each drug, and submitted as baseline prediction.

Scoring algorithms

Participants submitted predictions for a list of 1259 "druggable" targets and 30 drugs, with each prediction being a confidence score between 0 and 1 (where one is most confident that the target is a true target of a drug). We then filtered each submission to only consider the 255 targets in the gold standard dataset. For the purposes of calculating p values, we created 1000 null models by generating 1000 random prediction sets. These random predictions were generated by sampling (without replacement) the 255 gold standard targets using the dplyr "sample_frac" function to obtain a randomly-ranked set of targets (this procedure was repeated 1000 times).[62]

CellPress
OPEN ACCESS

For each submission, we filtered the predictions to the 255 kinases being evaluated. For SC1, we scored teams by evaluating the enrichment of their top 10 predictions for each drug in the gold standard dataset, as well as for one null model prediction, performing a paired Wilcoxon rank sum test (Mann-Whitney test) to generate a p value for each prediction. We repeated this each null model to generate a distribution of 1000 p values for each submission, and calculated the mean p value as the participants' score. For SC2, the methodology and null models were identical, but instead of evaluating the enrichment of the top 10 predicted targets in the gold standard dataset, we assessed the ranks of the true targets within the full vector of 255 predicted targets for each drug. We again performed a paired Wilcoxon rank sum test (Mann-Whitney test) to generate a p value for each submission. We repeated this for each null model to generate a distribution of 1000 p values for each submission and calculated the mean p value as the participants' score. In the post-challenge phase of this study, we re-evaluated the performance of each team (Figure S3) by repeating this analysis but omitting the kinase filtering step described above.

Alternate gold standard evaluation

Data from ChEMBL, DrugTargetCommons, KinomeScan (generated by HMS LINCS consortium), and DrugBank for the 1259 "druggable" targets used in this challenge were collected and formatted in the same manner as the KBR dataset used in the challenge. A set of 1000 null models was generated using the 1259 "druggable" targets. The scoring was performed for the SC1 and SC2 metrics as described in the previous section. Due to the very different target universes and completeness of each dataset, we converted the absolute scores to ranks to make it easier to compare the relative differences between the different datasets.

Determination of top performers and data leak

Winners were determined by calculating a Bayes factor relative to the top-ranked submission in each category. In this context, we used the Bayes factor, a likelihood ratio, to compare the difference between the top-ranked model and all other models in each subchallenge. The Bayes factor indicates the relative difference between the predictive power of the two models; with larger Bayes factor values corresponding an larger difference between the models. Ties were defined as models with a Bayes factor ≤ 3 relative to the top-ranked model.

We calculated Bayes factors by bootstrapping all of the submissions that qualified for final scoring by performing 10000 iterations of sampling with replacement for each submission. For each bootstrap, we calculated the p values as described above to generate a distribution of scores for each submission. Using this distribution of p values, Bayes factors were calculated for each submission relative to the top-scoring team using the challengescoring R package (https://github.com/sage-bionetworks/challengescoring). Ties were defined as submissions with a Bayes factor ≤ 3 relative to the top submission. During the scoring of the final round, we discovered that a portion of the Challenge dose-response data had been revealed to the public via a preprint. Upon reviewing the writeups, we saw that Team netphar (not knowing that this was the challenge data) described using this information to fine-tune some of the compound predictions for better performance. To ensure a level playing field and to ensure that this team's model was generalizable and did not use the preprint data, we worked with Team netphar to remove this fine-tuning step and rescore the prediction. Importantly, the analyses presented in this manuscript to determine the top performers used the new prediction file that omits the fine-tuning step and leaked data.

Detailed computational procedure Figures 1C and 1D

PANACEA differential gene expression data were transformed into "Transcriptional Hallmarks" based on definitions of 50 transcriptional signatures defined in[63]. Briefly, an average z-score was calculated for each signature by averaging the z-scores of the individual genes for each signature. PanACEA cell-lines were then averaged to yield a single 32x50 matrix reflecting the relationships of 32 drugs and 50 transcriptional hallmarks. PANACEA and Kinome-binding matrices were then processed for visualization by (1) filtering for the top 30 kinases and signatures by variance and (2) clustering rows and column based on pearson correlation. Filtered data were then visualized using the heatmap.2 function of the gplots package in R. Sidebar annotations of canonical drug-targets were defined based on the DrugBank definitions of drug-targets as detailed in Table S1.

Detailed computational procedure Figure 3A

To assess the agreement between DrugBank-Literature and Kinome-binding data, we first defined our reference as: all the kinase-targets defined by DrugBank for our 32-drug library. As the Kinome-binding data give continuous measurements, it is necessary to define a Kd-threshold to binarize the Kinome-data to compare with DrugBank. For each Kd-threshold, we then calculated the coverage of DrugBank by counting the number of drug-kinase edges identified in the Kinome data and divided by the total number of drug-kinase edges in DrugBank. **(B)** To visualize the new-targets defined in the Kinome-data (but not in DrugBank) we plotted the number of overlapping drug-targets in black (defined in Figure 3A) and newly identified drug-targets in red (defined as NOT being present within DrugBank) for each drug. We then sorted based on total number of targets to aid assessment of polypharmacology.

Detailed computational procedure Figures 4A and 4B

To better understand the performance of the three winning models on individual drugs we recalculated team-scores for each drug as z-scores for enrichment (in red) or depletion (in blue) for < uM targets within each drug-vector for both SC1 and SC2. Recalculated scores were sorted by the rank of the average performance across all three teams to identify the drugs which all models performed

well on. To better understand the type of inhibitors that models performed the best on we calculated the enrichment of each drugbank target kinases (as defined in Table S1) over the ranked 32-drug vector in Figures 4A and 4B using the aREA algorithm in the viper package in R.[12] **(C)** To visualize the kinases sampled by the Klaeger et. al[18] definitions of drug-targets we color-coded individual kinase-nodes within the Human Kinome phylogenetic tree obtained from the CORAL tool.[64] Kinases measured in the Kinome-dataset were color coded based the Kinase group that they were a member of as defined in Manning et al.[38] **(D)** To better understand the relationships between individual kinases and down-stream transcriptional programs we calculated the correlation matrix between the Kinome Kd's (250 kinases x 84 drugs) and the pan-cancer transcriptional signature PANACEA-data (50 signatures x 84 drugs) across 84 overlapping drugs that occurred in both datasets. Correlation matrix was then clustered based on correlation and visualized using the heatmap.2 function in the gplots package in R. Top side-bars were color-coded by kinase groups as defined in Manning et al.[38] and colors were chosen to match the Kinome-coverage-phylogenetic tree in Figure 4C. **(E)** To better understand the signaling pathways involved in the kinase-mRNA correlations in Figure 4D, we transformed the individual kinase columns in Figure 4D's correlation matrix into KEGG-defined signaling pathways. This was done by calculating the enrichment of pathway-specific kinases within each transcriptional program vector using the aREA algorithm in the viper package in R. This generated a normalized enrichment score (NES) for each kinase-pathway/mRNA-program pair that is equivalent to a z-score. The visualization on Figure 4E was obtained from the raw pathway-program matrix by slicing top columns associated with receptor tyrosine kinase controlled pathways.

Detailed computational procedure Figure S4

To compare the relative scores of DrugBank-defined targets and New-Kinome-Data-set defined targets within the winning teams predictions we first normalized all scores by the average-score. The purpose of this was to assure that a random-selection of drug-targets would have a normalized score of 1. For each winning team, and for each drug, we then calculated the average scores of Drug-Bank defined targets and Kinome-defined targets. Encouraging, while DrugBank Targets were consistently higher than Kinome-defined targets, both sets consistently scored better than random sets of drug-targets.

Detailed algorithm for winning team "Netphar"

The Netphar team collected three types of data related to the compounds: 1) Drug sensitivity data; 2) Drug induced gene signature data and 3) Drug target interaction data. For drug sensitivity data, we utilized the DrugComb database, which is a crowd-sourcing database to collect comprehensive drug sensitivity screen data, including both monotherapy drug screens and drug combination screens.[35] DrugComb currently consists of drug sensitivity data for 466k combination and 710k monotherapy drug screenings. From DrugComb, we found n = 116 drugs that have dose-response data on at least 7 of the 11 cell lines. Furthermore, for each compound-cell pair, we determined IC20 and RI (relative inhibition, which is based on area under the log10-scaled dose-response curves[41]) score, as more robust measures for drug sensitivity.

For drug-target interaction data, we utilized the DrugTargetCommons which is a crowdsourcing-based database to collectively and manually curate the comprehensive drug-target bioactivity values.[36] The bioactivity values were transformed into a confidence score between 0 and 1 to indicate the binding affinity potential.

To determine the best machine learning models to predict the drug targets, we considered two classes of methods including weighted averaging and regression (Figure below). For weighted averaging, the prediction was made based on the multiplication of the Pearson correlation matrix and the drug-target interaction matrix; while for regression, we considered standard machine learning algorithms including ElasticNet, RandomForest and GBM (Gradient Boosting Machine), for which the model was trained on the n = 116 compounds that were found in DrugComb, and then tested on the n = 32 Challenge compounds. We have utilized the LINCS-L1000 data[23] to evaluate the methods, and determined the weighted averaging approach that performed better than regression based on 10-fold cross validation.

Detailed algorithm for winning team "SBNB" (Figure S8):

As SBNB team, we approached the challenge as a data integration exercise, where we first adapted the transcriptional and sensitivity signatures of the DREAM Challenge compounds to the format of the Chemical Checker (CC).[37] The CC is a resource that provides processed, harmonized, and ready-to-use bioactivity signatures for about 1M compounds, offering a rich portrait of the small molecule data available in the public domain, and opening an opportunity for making queries that would be otherwise impossible using chemical information alone. The CC expresses bioactivity data as numerical vectors, making them suitable for similarity measurements, clustering, visualization and prediction tasks. Among others, the CC contains cell line sensitivity (Sens) and differential gene expression (DGEx) bioactivity signatures for tens of thousands of compounds (CC compounds), being thus possible to relate this data to the DREAM compounds. To integrate DREAM compounds with CC compounds, we built six different signature types from those bioactivity spaces similar to the ones provided by the DREAM challenge. In three of them, we used growth-inhibition (GI) data of eight cell lines common to the Cancer Therapeutics Response Portal (CTRP[28]) and the DREAM panel. We then used GI data as features to train a classifier to infer the expected CTRP sensitivity (Sens) profile of DREAM compounds, as well as biomarkers and annotations from the PharmacoDB resource.[65] Thus, we could connect the DREAM compounds to the hundreds of drugs available in the public drug sensitivity panels. Likewise, DREAM DGEx data were integrated with LINCS DGEx (Level 5) signatures,[23] along with the Touchstone reference collection of perturbational profiles. Additionally, we mapped the DREAM and LINCS

DGEx to a collection of manually curated expression signatures from Gene Expression Omnibus (CREEDS[30]), in order to capture cell-unspecific profiles, since only one cell line was shared between the DREAM and LINCS L1000 resources. We moved from individual gene expression to global expression signatures with the aim of capturing possible transcriptional regulatory programs shared among the compounds, enabling thus a more comprehensive integration of the DREAM and LINCS datasets.

Once we contextualised DREAM compounds within the larger CC compounds collection, we used the Sens/DiffGEx signatures as input for conventional target prediction methods, based on previously known ligand-binding profiles. In brief, to prevent overfitting due to the limited number of CC-exp compounds, we first used CC signatures to train a k-nearest neighbors (kNN) classifier to identify the most probable targets for each DREAM compound.

We simply looked in DrugBank for CC compounds having similar signatures to the DREAM compounds, and suggested the CC annotated targets as putative targets for the DREAM compounds. Then, in a second step, we used a much larger set of over 100k bioactive compounds in the Chemical Checker, for which we inferred their gene expression and cell line sensitivity signatures to train a multitask, quality-aware artificial neural network (ANN) primarily based on chemogenomics data (i.e., compound interactions) from ChEMBL[66] and refined it with DrugBank drug-target data.[31] More specifically, we trained a deep neural network (implemented with Tensorflow v1.12.) with 2 hidden layers (of 256 and 128 units, both using RELU activation and 20% dropout for regularization) and a last multitask classification output layer (with sigmoid activation and 1 unit for each annotated target). Given the gene expression and cell line sensitivity signatures of a drug, this architecture returns a vector of probabilities for each annotated target. We first trained the model using the ChEMBL universe of targets (456 proteins, 87904 different compounds) for 50 epochs with a high learning rate (1e-3). We used the trained network as a starting point to fine-tune the network (transfer learning on the whole network with a low learning rate of 1e-5) with Drugbank targets (456 proteins, 3409 compounds). In both cases we used the normalized average of compound signature confidence (obtained from the CC pipeline) to weight the sample, hence, making the network predictions aware of the input quality (quality-aware). To obtain the final ranking we first computed the closest 1, 5 and 10 nearest neighbors, assembling the results and using them to rank each target accordingly, as previous attempts showed a good performance for the challenge SC2. Then, to improve the challenge SC1, we reordered the top 10 targets for each drug according to the ANN prediction (i.e., we placed in the top 10 the top 10 targets with higher probability scores according to the ANN). Finally, those protein targets of the challenge not annotated in Drugbank were placed at the end, ranked according to the drug counts in ChEMBL (thus, sorted by their prior probability of being a target).

Detailed algorithm for winning team "ATOM" (Figure S9):

For each compound, its compound-perturbed gene expression feature was calculated from Level 5 data of the LINCS L1000 platform of phase I (GSE92742) and phase II (GSE70138). To obtain a consensus feature for each compound without considering other conditions like cell line, dose and time, all the Level 5 signatures corresponding to the same compound were selected and averaged using MODZ algorithm introduced in L1000 paper.[23] In order to suit for the challenge, we compared the RNA-seq data with the L1000 data and selected 973 overlapping genes as input features.

During the model training, a graph-based multi-task constraint was used to train our model (described below). The target similarity graph was constructed by using two types of metrics, including a sequence similarity from protein primary sequences as well as a genomic similarity from gene knockdown perturbed gene expression profiles. The protein primary sequences were first obtained from UniProt database according to their gene IDs. Then, the Smith-Waterman sequence alignment scores were computed by an alignment tool (https://github.com/mengyao/Complete-Striped-Smith-Waterman-Library). The sequence similarity between two proteins was then defined as the normalized alignment scores, that is, $\frac{sw(s_1,s_2)}{\sqrt{sw(s_1,s_1)\times sw(s_2,s_2)}}$, where $sw(s_1,s_2)$ stands for the alignment score between protein sequences s_1 and s_2. The gene knockdown perturbed gene expression profiles were obtained from the L1000 database and processed using the same protocol as drug features described above. The genomic similarity between two targets was defined as $max\{0,r(e_1,e_2)\}$, where $r(e_1,e_2)$ stands for Pearson's correlation coefficient between gene expression profiles e_1 and e_2. Finally, we averaged these two matrices and constructed a K-nearest neighbor (KNN) (K = 10) graph as our final target similarity graph.

As the problem is to predict the potential targets for a compound/drug of interest, we formulate this problem as a multi-label classification problem, where the input of a compound is the compound-perturbed gene expression feature $x \in R^{973}$ derived from LINCS L1000 platform, and the output is a binary vector $y \in R^{769}$ indicating the binding probabilities to 769 pre-defined protein targets. We used an ensemble of neural networks to make predictions.

We use an ensemble of single-layer neural networks to model the relationship between x and y. The model architecture of each base learner (i.e., a single-layer neural network) is shown in Figure 5. For each base learner, three losses are used to train its parameters. The first one is Bayesian Personalized Ranking (BPR) loss.[67] Specifically, let $S_{i,j}$ denote the predicted score between drug i and protein j produced by our model. Then, BPR loss is defined as: $BPR\_Loss = -loglog(S_{i,j} - S_{i,k})$, where protein j is the known target of drug i while protein k is not. During neural network training, we sampled a batch (batch size = 256) of drugs, and for each drug i, we sampled pairs (i, j) and (i, k) to perform forward and backward propagation. The second loss is a multi-task constraint loss (Zhou et al., 2011). The multi-task constraint loss is defined as: $Multitask\_Loss = trace(WLW^T)$, where W is the learnable parameter of the last layer of the neural network, L is the normalized graph laplacian of the target similarity graph defined above. This loss encourages the similar targets to have similar classifiers. The last loss is the weight decay (i.e., L2_regularization) for controlling the model

complexity. The combined loss is defined as BPR_Loss + λ_1*Multitask_Loss* + λ_2 L2_regularization, where λ_1 and λ_2 are used to balance different losses. This combined loss was optimized by Adam optimizer with learning rate = 0.001.

We used 10-fold cross validation to train models. For each fold, 1/10 of the drugs were used as test data. Among the remaining 9/10 drugs, 1/10 of drugs were left out as validation data and the rest drugs were used as training data. This strategy was used to perform hyperparameter selection (i.e., dropout rate, λ_1, λ_2, hidden size of the neural network and training epoch). During training, early stopping was used to prevent overfitting. For each epoch, we compared the model performance on validation data with the best performance. The training process would be stopped as long as the performance on the validation data no longer improves in consecutive 100 epochs.

We used ensemble learning approach to further boost the performance. We constructed 100 different neural network models from { λ_1 = 0.0001, 0.00001} × { λ_2 = 0.0001} × {hidden size of neural network = 256, 512, 1024, 2048, 4096} × {10 different folds}. These hyperparameter ranges produced decent prediction performance during our hyperparameter selection. We then averaged the prediction scores from these models to produce the final scores.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Connecting chemistry and biology through molecular descriptors

Adrià Fernández-Torras[1], Arnau Comajuncosa-Creus[1], Miquel Duran-Frigola[1,2] and Patrick Aloy[1,3]

Abstract

Through the representation of small molecule structures as numerical descriptors and the exploitation of the similarity principle, chemoinformatics has made paramount contributions to drug discovery, from unveiling mechanisms of action and repurposing approved drugs to *de novo* crafting of molecules with desired properties and tailored targets. Yet, the inherent complexity of biological systems has fostered the implementation of large-scale experimental screenings seeking a deeper understanding of the targeted proteins, the disrupted biological processes and the systemic responses of cells to chemical perturbations. After this wealth of data, a new generation of data-driven descriptors has arisen providing a rich portrait of small molecule characteristics that goes beyond chemical properties. Here, we give an overview of biologically relevant descriptors, covering chemical compounds, proteins and other biological entities, such as diseases and cell lines, while aligning them to the major contributions in the field from disciplines, such as natural language processing or computer vision. We now envision a new scenario for chemical and biological entities where they both are translated into a common numerical format. In this computational framework, complex connections between entities can be unveiled by means of simple arithmetic operations, such as distance measures, additions, and subtractions.

Addresses

[1] Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain

[2] Ersilia Open Source Initiative, Cambridge, United Kingdom

[3] Institució Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

Corresponding author: Aloy, Patrick. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain (patrick.aloy@irbbarcelona.org)

Current Opinion in Chemical Biology 2022, **66**:102090

This review comes from a themed issue on **Omics (2022)**

Edited by **Mohan Babu**

For complete overview of the section, please refer to the article collection Omics (2022)

Available online 6 October 2021

https://doi.org/10.1016/j.cbpa.2021.09.001

1367-5931/© 2021 Elsevier Ltd. All rights reserved.

Keywords

Molecular descriptors, Bioactivity signatures, Biological embeddings.

Introduction

Small molecules are an excellent tool to probe biological functions and the primary choice of pharmaceutical companies, as they are easy to manufacture, store, and distribute, and synthetic chemists can conceive a broad variety of them [1]. Some commercial and public chemical collections include up to 10^9 compounds, with the number increasing to 10^{20} for proprietary libraries, which means that the chemical space available to researchers is essentially infinite [2]. Moreover, new strategies based solely on the combination of two- or three-step reaction sequences estimate that it would be possible to readily synthesize ~29 billion compounds [3*]. The size of the accessible chemical space easily explodes if fewer constraints are applied, with some plausible estimates exceeding 10^{60} compounds for molecules under 500 Da [4]. In addition, and perhaps more importantly, in the last years high-throughput screening (HTS) assays have penetrated the public research sector (e.g. the study by Subramanian et al. [5] and Corsello et al. [6*]), providing depth of annotation to the compound collections. This is reflected in the increasing number of bioactive small molecules catalogued in open databases, which already amount to over two million entries [7,8].

Querying compounds in these databases differ greatly from querying proteins or genes. Biological sequences are richly annotated, and even when they are not, evolutionary and structural domains help link them to molecular functions, which, in turn, contributes to our understanding of higher-order biological processes [9]. Compared to biological sequences, small molecules spell a much more complicated code which, for the most part, has not been explored by the rules of natural evolution. In consequence, there is no clear and continuous connection between structure and function, which converts an apparently simple task, such as measuring similarity between two molecules into an open problem driving a whole field of research.

In practice, representing chemical compounds in a meaningful way (for compound similarity measures or other computational chemistry calculations) requires the selection of a small molecule descriptor. Among the classical chemical notations, we find the simplified molecular input line entry system (SMILES) that, although it might be ambiguous (i.e. one molecule can be described with multiple SMILES), it is very intuitive and widely used [10]. Other popular molecular descriptors encode the structural, topological and/or physicochemical properties of the compounds. These descriptors can account for the presence or absence of a specific set of pre-defined chemical groups, like in the case of the molecular access system keys [11], defined dynamically by listing the 2D structural elements encountered in a molecule. For example, in the extended connectivity fingerprints atoms are enumerated, and neighboring elements and bonds are captured. Other complex descriptors broaden the structural information by capturing the spatial 3D coordinates of the atoms [12] or go beyond molecular geometry and consider environment-dependent properties, such as the active site of the receptor [13] or those derived from molecular simulations [14], within a given radius [15]. These and other similar descriptors have been at the core of chemoinformatics and are still the first choice in most applications (see the study by David et al. [16] for a recent and very comprehensive review). However, the last years have witnessed the expansion of a new generation of molecular descriptors, deemed to be 'data-driven' and based on deep learning approaches, that are engineered on the basis of large-scale chemistry databases and are thus adaptable to a given task or region of the chemical space [17]. In particular, graph and text-based autoencoders are able to embed the information provided by 2D structures and SMILES strings, respectively, into a dense numerical vector belonging to a 'latent space' [18]. Simple measures such as Euclidean distances within the latent space are able to capture chemical similarity and, when coupled to machine learning algorithms, these descriptors have shown state-of-the-art performance in several biophysics and physiological benchmark datasets [19].

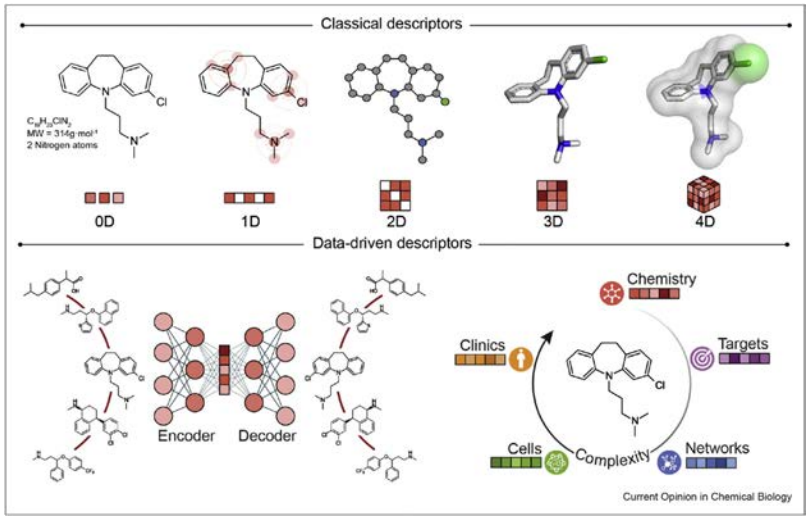
A natural extension of this first generation of data-driven descriptors is to include the wealth of bioactivity information available in the databases, to encapsulate, in the form of 'bioactivity descriptors', the experimental evidence gathered over years of research. Here, we review some recent attempts to provide these biologically relevant molecular descriptors and discuss how a descriptor-based approach may help integrate small molecules with larger biomolecules in a common framework able to capture several layers of biological complexity encompassing protein targets to cellular pathways and disease phenotypes.

Extending the similarity principle beyond chemical structures

Chemical descriptors, in their different flavors, encode the physicochemical and structural properties of small molecules and provide a computer-friendly format to represent and compare them (Fig. 1). However, these descriptors do not incorporate bioactivity information explicitly, which handicaps the discovery of links between small molecules and other entities, such as proteins or cells. In pioneering work, instead of focusing on chemical structures, Kauvar et al. [20] characterized a set of compounds according to their ability to bind a panel of 18 receptors and used these affinity profiles to assess similarities between them. The idea of relating small molecules based on their target profiles was further developed over the next years [21,22], enhancing the performance in classical chemoinformatics tasks (e.g. target prediction). In a more complex attempt to capture phenotypic effects induced by drug activity in cells, MacDonald et al. [23] used a protein complementation assay to monitor the status of several cellular pathways after compound perturbation. Then, they derived pathway activity fingerprints for over a hundred compounds and found that pathway-based similarities strongly correlated with known structure−activity relationships. Similarly, Young et al. [24] combined automated microscopy with image analysis to profile the biological effects of a compound library. They integrated the resulting phenotypic profiles with the chemical structure of the compounds and their predicted targets and found that the combination of the three features had a substantially higher capacity to identify mechanisms of action than either one in isolation.

Indeed, the popularity of HTS assays has revealed that it is possible to establish relationships between compounds based on their functional activity rather than their chemical structure. For instance, it was suggested that molecules triggering similar transcriptional responses in cell lines might share mechanisms of action, an observation that inspired the implementation of the connectivity map [25] and the following library of integrated network-based cellular signatures (LINCS L1000) [5] initiatives. These libraries provide a catalogue of transcriptional signatures in different cell lines, measured as a result of a systematic screening of genetic (CRISPR or shRNA) and pharmacological perturbations, which has been exploited, for instance, to suggest potential targets for a given compound [26]. Likewise, molecules that inhibit the growth of a similar subset of cell lines (i.e. that have similar sensitivity profiles) [27] or drugs that elicit similar side effects, also tend to share mechanisms of action [28], even if their 2D or 3D structures appear to be unrelated.

Figure 1



Encoding chemical molecules through their chemistry and bioactivity. Molecular descriptors allow for the mathematical treatment of chemical and structural features of molecules. There is a wide range of strategies to generate such descriptors. Simple approaches account for global molecular properties (0D, e.g. molecular weight) or the presence of particular structural features (1D, e.g. encoding circular environment of each atom up to a specific radius). The molecular topology (2D, e.g. distance matrices between atoms) or the spatial information of the atoms (3D, e.g. cartesian coordinates) can be encapsulated by conveniently representing molecules as chemical graphs. In addition, there are sophisticated methods that capture environment-dependent properties, such as functional regions or intramolecular interactions (4D, e.g. energetically favorable binding sites or multiple conformational states). Driven by the bloom of high-throughput assays and the following population of compound libraries, a new generation of data-driven descriptors based on deep learning strategies encode molecules into abstract latent spaces, representing molecular similarities as simple distance measures between numerical vectors. Furthermore, molecular descriptors have expanded beyond chemistry, integrating relevant biological data from heterogeneous bioactivity assays and providing a complementary framework to assess molecular similarity.

Building upon these seminal works, we recently presented the chemical checker (CC), a resource that integrates the major chemogenomics and drug activity repositories and represents the largest collection of small molecule bioactivity signatures available to date [29**]. The CC gathers experimentally determined bioactivity data for about 1M small molecules in the medicinal chemistry space and provides bioactivity descriptors in five levels of increasing biological complexity. The first level of descriptors characterizes the chemical properties of the compounds, including their 2D and 3D structures, scaffolds, functional groups, and physicochemical properties. The second level captures information on the protein receptors of the molecules, including known mechanisms of action, metabolizing enzymes and HTS binding assays. Descriptors in the third level of complexity address the propagation of the target perturbations triggered by the small molecules, including protein–protein interactions and pathways provided by several types of biological networks. The fourth level of signatures captures the bioactivity of the compounds measured at the cellular level, with assays including differential gene expression and sensitivity profiles in cancer cell-line panels. Finally, for the few compounds that reached clinical stages, the fifth level of CC signatures encodes details on their therapeutic areas, adverse side effects and drug–drug interactions. A known limitation of the CC was that the number of molecules with reported bioactivities diminished at each level of complexity, and thus, we could only derive a limited set of bioactivity descriptors corresponding to a minority of well-characterized compounds. To extend the coverage of bioactivity descriptors to uncharacterized molecules, we trained a collection of deep neural networks (i.e. 'signaturizers') that are able to infer bioactivity signatures for any compound of interest, even when only its chemical structure is available. We were able to assign a confidence score to the predictions of the signaturizers and systematically apply them to sets of compounds beyond drug molecules, including plant metabolites and food ingredients [30*].

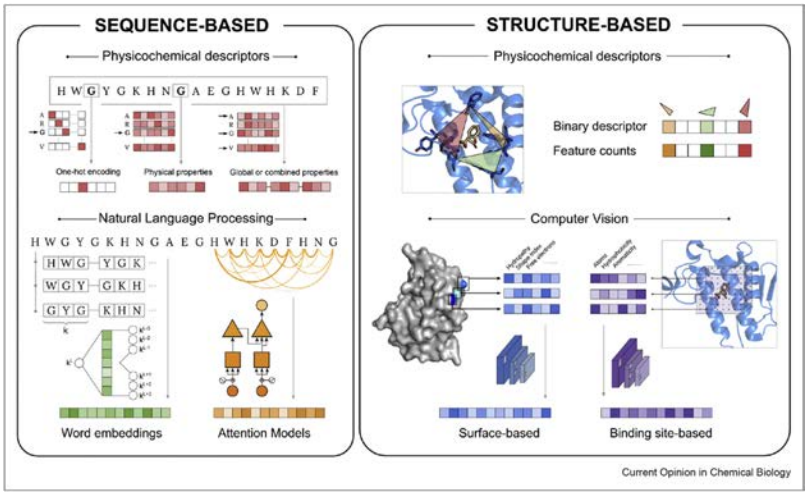
Overall, bioactivity signatures provide a complementary means to describe small molecules, focusing on the integration of multiple types of experimental data [31].

Indeed, these descriptors have proven useful to navigate the chemical space in a biologically relevant manner and boost the performance in many drug discovery tasks that typically rely on chemical descriptors, for example, target identification or toxicity prediction [30*].

Target descriptors to complement small molecule bioactivity signatures

In the quest to predict small-molecule bioactivities, often through machine learning approaches, the chemical compounds represent only one part of the equation. To match the rich chemical representations described previously, researchers are also developing methods to encapsulate information available for the biomolecular targets (Fig. 2). Protein sequence descriptors, for example, annotate the identity and the physicochemical properties of each amino-acid (e.g. the study by Hellberg et al. [32]) or measure general features of the full-length sequence, such as global residue composition and distribution (e.g. the study by Xiao et al. [33]). In any case, these relatively simple representations have been used in a battery of bioinformatics tasks, including protein engineering [34] or function prediction [35]. Like in the case of 'data-driven' descriptors for small molecules, deep learning is providing new ways to describe biological sequences. For instance, in a recent study, Alley et al. [36*] applied deep neural networks to a vast set of unlabeled sequences, yielding semantics-rich descriptors that capture structural, evolutionary and biophysical properties of proteins. These descriptors have proven their value to predict the stability of *de novo* designed proteins, but their agnostic nature and versatile format make them a suitable input for almost any machine learning task involving proteins. In general, protein sequences are treated as text data, which allows for borrowing techniques from natural language processing, a discipline that has made extraordinary progress for knowledge representation [37,38]. In a first attempt to systematically benchmark language models (LMs) for protein modeling, Rao et al. [39] designed a set of tasks assessing protein embeddings and reported promising results for a variety of models involving evolutionary understanding and protein engineering. Earlier this year, Elnaggar et al. [40**] explored the limits of up-scaling LMs trained on protein sequences achieving, for the first time, performances competitive with evolutionary models, but requiring much less time to compute. Just recently, while reviewing the new advances in language modelling for protein sequences, Bepler and Berger [41] extended their previous work and pretrained a protein LM conditioned to structure prediction tasks (e.g. the

Figure 2



Target and binding pocket descriptors. The simplest way to represent a target protein sequence is by encoding the identity or the physicochemical properties of its amino-acids, either individually (i.e. one-hot encoding) or using sliding windows to capture their short-range environment. To account for more distant amino-acid relationships, proteins can be encoded using techniques borrowed from natural language processing (i.e. word embeddings or attention models), where sequences are often treated as a set of constant-length overlapping fragments or k-mers. Whenever high-resolution models of target proteins are available, these can be used to derive structure-based descriptors. The classical ones consider the geometry and physicochemical properties of the binding pockets by calculating distances between pharmacophoric points and transforming them into high−dimensional profiles, accounting for the presence or absence of a given pharmacophoric geometry. More recently, computer vision and deep learning techniques have been adapted to embed structural properties of protein surfaces and specific binding pocket features.

model was forced to predict residue contacts and structural similarity during training) [42**]. By including evolutionary and structural information, they not only showed improvements in downstream tasks (e.g. protein function prediction) but also evidenced that hybrid approaches leveraging both data-driven sequences and physics-based domains can help LM to better embrace the sequence-structure—function paradigm. In another fresh work, Rao et al. [43] trained an LM taking multiple sequence alignments as input, conversely to the single sequence approach. Their model showed a better recapitulation of evolutionary variation and set a new state-of-the-art on unsupervised protein structure prediction [44]. It is worth noting that learning from both the multiple sequence alignments and the interplay between protein sequence and structure has been paramount to AlphaFold2 success in achieving outstanding accurate 3D protein structure predictions [45**]. Most of these successful models are based on transformers, such as the bidirectional encoder representations from transformers, a widely used architecture in text recognition [46]. However, as with almost any method involving deep learning, the interpretability of these protein LMs is very limited. In a remarkable attempt to shed light on the biological and biophysical information captured by bidirectional encoder representations from transformers -based descriptors, Vig et al. [47*] thoroughly analyzed the inner layers of the deep neural network and found that they uncovered relevant associations in the 3D space, such as residues that were far apart in the sequence but spatially close in the structure or those constituting the protein binding sites. We refer the reader to the study by Bepler and Berger[42**] for an insightful review of LMs in protein biology.

Binding between targets and ligands is determined by the biophysical properties of protein 3D structures and, in particular, the surface residues where potentially druggable pockets are found. Indeed, while a study exploring the binding promiscuity of over 160 drugs could not identify correlations between drug promiscuity and their chemical features (e.g. hydrophobicity), it did reveal structural similarities amongst their protein targets, highlighting the need to study binding site similarity across the proteome [48]. Thus, whenever high-resolution structures of the target proteins are available, more specific descriptors can be developed. Classic pocket descriptors measure the geometrical and electrostatic features of small molecule binding sites and translate them into binary fingerprints that just account for the presence or absence of a given structural motif (e.g. the study by Weill and Rognan [49], Siragusa et al. [50]), in the same way, that extended connectivity fingerprint or molecular access system descriptors do for chemical compounds. Cavity similarities based on these binding pocket fingerprints have unveiled interesting cases of remote homology between proteins [51] and are the basis for several polypharmacology strategies [52,53]. The popularity of methods to compare druggable pockets prompted the creation of thorough benchmark datasets, such as TOUGH-M1 [54] and the protein site pairs for the evaluation of cavity comparison tools [55], which pointed out the strengths and weaknesses of a variety of descriptor types and approaches, and provided a gold standard to validate pocket comparison strategies to come. Systematic evaluation has revealed that some descriptors are better suited than active sites of related proteins, while others perform better to describe macromolecular binding interfaces, being the latter more appropriate for drug polypharmacology and repurposing studies [56]. If progress in natural language processing has enabled sequence-based descriptors, progress in image analysis and computer vision has prompted the development of 3D structure-based descriptors. For instance, Gainza et al. [57**] devised a novel strategy to segment high-resolution protein surfaces into overlapping radial patches, mapping chemical, and geometrical features onto them. These data are then transferred into a convolutional neural network (CNN) to generate the descriptors, which can be fine-tuned for specific tasks, such as ligand-binding pocket similarity or protein—protein interaction interface comparisons. DeeplyTough is another recent method that also uses CNNs to encode 3D characteristics of protein binding pockets [58*]. The peculiarity of DeeplyTough is that it has been trained to ensure that similar pockets are encoded into similar descriptors, while retaining the ability to account for small structural variations and differentiate closely related binding sites. In a recent protein site pairs for the evaluation of cavity comparison tools benchmark, pocket comparisons based on these descriptors scored among the best [55].

The significant improvement of both chemical and protein descriptors has prompted the development of proteochemometric strategies, where machine learning models are trained on a combination of ligand and target representations [59*]. Indeed, these kinds of approaches have already shown superior performances in multi-target bioactivity prediction compared with classical methods [60], although some results may be over-optimistic due to bias in the training datasets as pointed out in the study by Chen et al. [61]. Moreover, Bongers et al. [59*] showed that structure-based descriptors are often superior when a detailed definition of the target is needed (i.e. to distinguish drug selectivity among members of the same protein family), while sequence-based ones are better suited for more generic models, especially when key structural details are lacking.

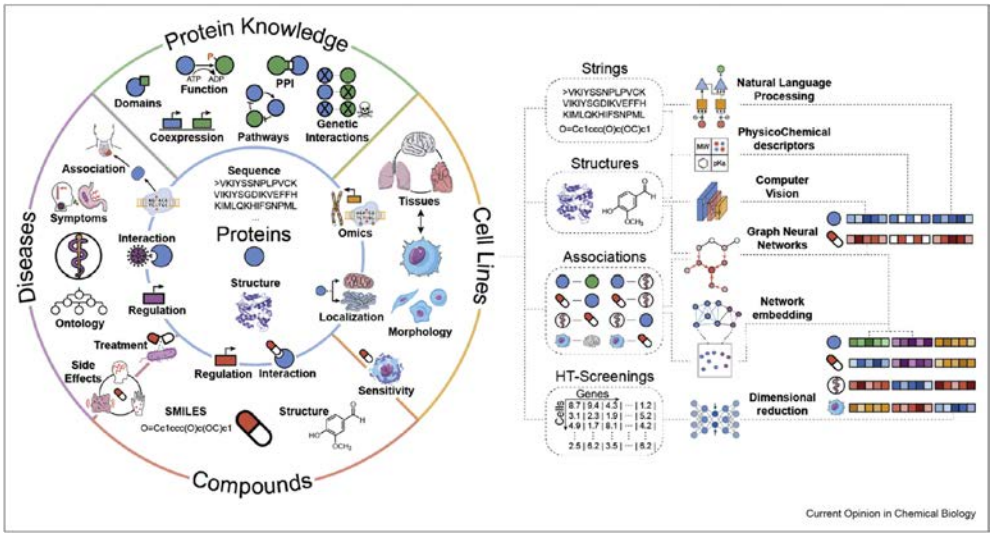
Capturing biological complexity in biomolecular descriptors

From a drug discovery perspective, genomic initiatives are providing new target opportunities [62,63], but many of these correspond to gene products thought to be undruggable, and the avalanche of data has not spurred the development of truly personalized, or even precision, therapies based on the exquisite interaction between a drug and an optimal target [64]. In fact, whole-cell phenotypic screenings continue to be the approach that contributes the most to the discovery of first-in-class medicines, while target-centric approaches appear more useful only for the development of follow-on products [65,66]. Thus, to tackle complex phenotypes, we need to move away from the 'one disease, one target, one drug' paradigm and consider the complexity of human pathologies from the early stages of the drug development process. Indeed, a growing fraction of recently approved drugs is associated with pharmacological biomarkers at the genomic scale [67], meaning that omics experiments are able to identify links between biomolecular profiles and drug action. This evidence is often complementary to the modulation of the intended therapeutic target and thus offer a more systemic view of drug activity.

In an attempt to capture this systemic complexity, it is increasingly common for HTS experiments to simultaneously characterize multiple omics profiles (i.e. trans-omics analyses) [68,69] so that several views of small molecule action can be analyzed in parallel. New methodologies are flourishing to deal with such data (e.g. the study by Argelaguet et al. [70]) and yet, these methods mainly adapt existing strategies developed in the past for single omics experiments, and often draw conclusions from the most informative data type, while the rest are used as support. It is, thus, fundamental to come up with strategies able to capture the coordinated interplay of the many regulatory layers present in biological systems (Fig. 3).

Integrating many levels of biology into a single resource is a daunting task because one needs to standardize data formats and identifiers, normalize records across different resources and categorize the observations by applying significance cutoffs (e.g. of differential gene

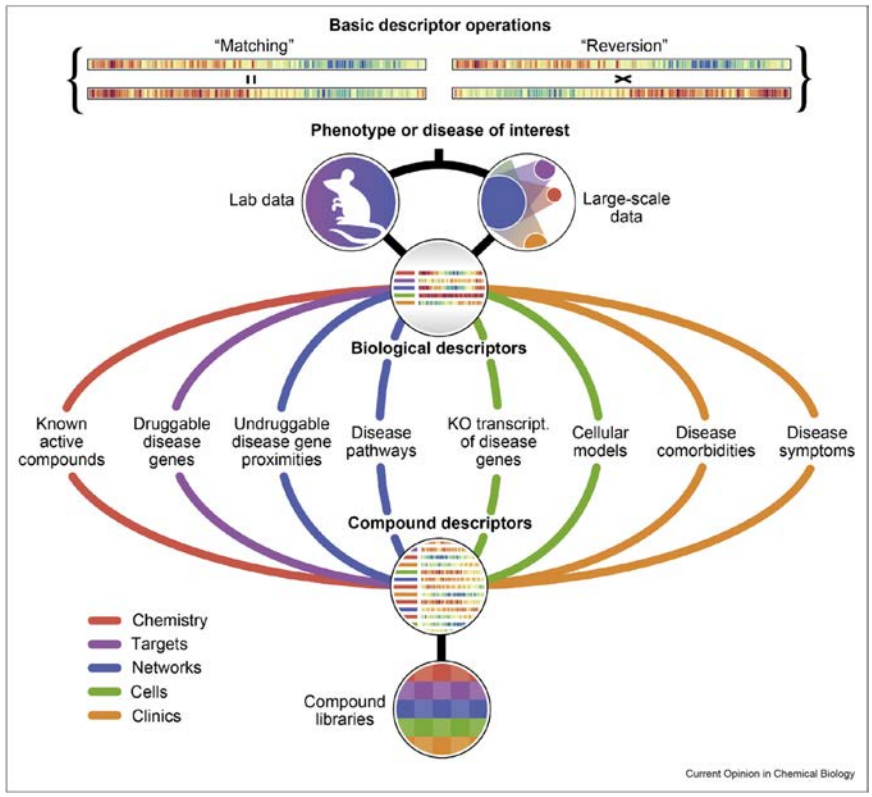
Figure 3



Capturing biological complexity in the form of descriptors. Bioactive chemical compounds often interact with their molecular targets to exert their function. However biological complexity spans far beyond protein targets, and long-range effects have a clear impact on drug action. At a molecular level, genes and proteins interact forming complex networks that regulate the physiology. Many of these physical or functional connections and their effects can be captured by individual biology experiments, while the integration of multi-omic unmasks the interrelations between different regulation layers. However, there is a resolution gap where we lose causality and all we can measure are somehow vague associations between molecules and higher-order phenotypic observations, such as a disease state. Depending on the nature of each experimental readout, different encoding strategies have been optimized to condense such complex biological data in the form of vector-like descriptors suitable for modern machine learning. String-like data, such as gene sequences or compound SMILES, are often encoded through the use of natural language models. Structural data, like the one representing protein and chemical structures or cellular morphology, is better suited for convolutional or graph neural networks. Alternatively, if the data to be encoded represent relationships between different biological entities, such as protein networks or compound–gene associations, network embedding techniques seem to yield the best results. Finally, as the readout of high-throughput screening experiments, such as drug sensitivity or cell transcriptomics, yields big numerical matrices, they are best condensed through the use of autoencoders.

expression). Unlike chemical data, where we often have millions of molecules with relatively poor annotations, biological databases annotate a relatively small set of biomolecules with a large number of interactions between them and associations with other biological entities, such as diseases, pathways, molecular functions, cells, and tissues. According to the 2020 report of the Molecular Biology Database Collection [71], there are 1637 active online databases, spanning every corner of biology. The first successful attempts to organize multiple databases into a single resource (e.g. Harmonizome [72] and Hetionet [73]) have structured the information in the form of a network, or knowledge graph, focused on the relationships (edges) between biological entities (nodes). However, the magnitude of biological networks is computationally intractable by traditional graph analysis techniques [74] which, also, in this case, has boosted the development of graph embedding approaches to reduce the dimensionality of the data while preserving the structural information and properties of the network [75**]. Thanks to these advances, we have been able to release the Bioteque, a resource of biological network embeddings of unprecedented size and scope [76*]. Bioteque descriptors are derived from a gigantic heterogeneous network (more than 550k nodes and 30M edges) that harmonizes data extracted from >200 data sources, including 12 different biological entities (e.g. genes, diseases, drugs) linked through 67 types of relationships (e.g. 'drug *treats* disease', 'gene *interacts with* gene'). We have shown that this concise representation of the data can be used to evaluate and characterize a wide array of experimental observations (e.g. drug sensitivity assays), and have illustrated how these omics-based descriptors can be plugged into machine learning tasks, similar to what is done with their counterparts centered on proteins and chemical compounds. Also recently, Cantini et al. [77*] evaluated the performance of several embedding methodologies to

Figure 4



Connecting biology and chemistry through molecular descriptors. A common framework for small molecule and biological descriptors will enable a direct comparison between compound structures, bioactivity data and biological entities such as protein targets, cell lines or disease symptoms.

integrate continuous multi-omics data (e.g. gene expression, copy number variation, methylation and miRNA expression). In addition to evaluating the preservation of the original (raw data) structure, the authors also assessed their performance in predicting clinical outcomes in a cancer cohort, as well as classifying multi-omics single-cell data from cancer cell lines. They found that, while the performance of each method significantly changed depending on the task, a concomitant analysis of multiple datasets (i.e. multiple co-inertia analysis) [78] was the most consistent across different benchmarks.

While omics data has provided us with a broad understanding of biological phenomena, there are biological entities that are not easy to describe from a molecular perspective, as they usually involve ontological concepts or high-order functions. Biological pathways, often represented by gene ontology terms, are commonly embedded by grouping genes that participate in similar biological processes or have related functional categories [79]. Recently, Wang et al. [80*] introduced an approach in which multiple gene sets are represented together in the embedding space, using a protein–protein interaction network as a measure of proximity between genes. This type of gene set descriptors has shown an improved capacity to identify new functionally related gene set members and reveal subnetworks with clinical prognostic capacity in sarcoma samples. At a cellular level, Schubert et al. [81] trained a CNN to learn embeddings of neuron images, where each embedding represented a fragment of the cell thus capturing the neuron morphology. They proved the power of these embeddings to identify subcellular compartments, cell types and, more importantly, detect neuron reconstruction errors. Going one step up in the hierarchy of the biological organization, Zitnik and Leskovec [82] developed OhmNet, a set of protein descriptors that take into consideration the specific protein–protein interactions within each human tissue, as well as the inter-tissue relationships, so that proteins with similar network neighborhoods in similar tissues are placed proximally in the embedding space. Then, they showed that these tissue-aware protein descriptors provide more accurate

Box 1. Most used machine learning methods in the development of chemical and biological descriptors.

Autoencoders	An autoencoder is a type of artificial neural network used to derive compressed representations of input data through an unsupervised learning strategy. Autoencoders are composed of an encoder and a decoder that compress and reconstruct the data, respectively. Autoencoders have been used, for example, to map large collections of compounds to the latent space defined by the encoder component, which provides a more suitable representation for machine learning pipelines.
Attention-based encoders (Transformers)	Transformers are a timely family of deep learning models based on attention mechanisms that have been especially successful at language modeling. Qualitatively speaking, attention refers to the upweighting of relevant parts of the input sequence, usually those that confer 'meaning' to it. A direct analogy can be established with protein sequences, where some amino-acids are more functionally relevant than others. Thus, when large protein sequence databases are processed with attention-based encoders, relevant descriptors can be extracted from the inner layers of the model.
Convolutional neural networks (CNN)	Convolutional neural networks are most commonly applied to image data as they naturally extract high- and low-order features from, for example, spatial data through the successful implementation of convolutional and pooling layers. Similarly, 2D and 3D structures of proteins or small molecules can be processed with these kinds of networks, typically by taking graph representations as input.
Network embedding and graph neural networks (GNN)	Network embedding comprises the set of techniques aimed at representing networks entities (typically nodes) in a vector format. Plausible results of a network embedding will assign similar vectors to neighbors in the original network, being able to capture higher-order organizations such as clusters of strongly connected nodes. A classical way of deriving network embeddings consists of an initial exploration of the network by a 'random walker', followed by a conventional sequence embedding based on the registered node-to-node trajectories. More recently, by involving graph neural networks these techniques can now jointly embed node and edge features (e.g. chemical properties) together with the network structure, enabling inductive learning. Large-scale biological networks are usually processed with network embedding techniques.

predictions of tissue-specific protein functions than alternative approaches, making them a powerful tool to transfer these learned functions to the lesser characterized tissues. In related work, the same authors have embedded different networks (i.e. protein−protein, drug-target and disease-gene interactions) to explore the mechanisms of action of drugs [83*]. Here, they modeled how drug effects spread through a hierarchy of biological functions coordinated by the underlying protein−protein interaction network. Thus, for each drug and disease, they learnt a diffusion profile to identify the key proteins and biological functions involved in treatment providing a transparent interpretation of the drug therapy.

Overall, these embedding-based descriptors provide a scalable and intuitive means to capture complex relationships between biological entities, and they represent an excellent strategy to integrate the deluge of biological data in a format that is readily amenable for downstream machine learning applications.

Concluding remarks

In this article, we have provided an overview of methods to represent chemical and biological entities in a common framework based on numerical descriptors. Although the approach may strike as too abstract to researchers uninitiated in data science, it has the unique advantage of capturing a number of data points that would otherwise be intractable. On top of that, this type of representation helps uncover links between entities by means of simple arithmetic calculations, such as similarity and distance measures between descriptors or additions to represent higher−order processes. The strategy can be applied at the atomistic level (e.g. compound similarity), as well as the phenotypic level, as first demonstrated by the connectivity map and LINCS L1000 [5,25] in the context of gene expression data. Indeed, dissimilarities between chemical and disease perturbation signatures can be leveraged to find small molecules that potentially revert a specific disease gene expression profile, hence providing support for drug-disease indications [84]. We have recently exploited connectivities between bioactivity descriptors based on pathways, biological processes or interactome networks to identify compounds that revert Alzheimer's disease signatures *in vitro* and *in vivo* [85], mimic the phenotypic effects of biodrugs (e.g. daclizumab, ustekinumab and cetuximab) [29**] and indirectly target cancer proteins thought to be undruggable [29**].

We envisage a scenario for computational chemistry and biology where drug candidates and biological entities will be first described with numerical vectors in the light of the available data, coming either from public repositories or in-house experiments (Fig. 4). These data would include structural features of the molecules and the targets, together with omics profiles, such as gene expression data, as well as large-scale biological networks and ontologies. Data will be linked at different levels with relatively simple operations, allowing for ultra-large, unbiased and systematic identification of the existing connections between the chemical space and the intricate biological space defined by disease biology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank M. Miñarro-Lleonar for their helpful suggestions about structural visualization. P.A. acknowledges the support of the Generalitat de Catalunya (RIS3CAT Emergents CECH: 001-P-001682 and VEIS: 001-P-001647), the Spanish Ministerio de Ciencia, Innovación y Universidades (PID2020-119535RB-I00) and the European Commission (RiPCoN: 101003633). AF-T is a recipient of an FPI fellowship (BES-2017-083053) and AC is supported through an FI fellowship from the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the European Social Fund.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Sterling T, Irwin JJ: **ZINC 15–ligand discovery for everyone**. *J Chem Inf Model* 2015, **55**:2324−2337.

2. Hoffmann T, Gastreich M: **The next level in chemical space navigation: going far beyond enumerable compound libraries**. *Drug Discov Today* 2019, **24**:1148−1156.

3. * Grygorenko OO, Radchenko DS, Dziuba I, Chuprina A, Gubina KE, Moroz YS: **Generating multibillion chemical space of readily accessible screening compounds**. *iScience* 2020, **23**:101681.
Novel synthetic strategy that allows the creation of a multi-billion compound library based on a two- or three-step three-component reactions of pre-validated building blocks. Initial validations show an ~80% of synthesis success, opening the door to the construction of ultra-large chemical libraries.

4. Reymond JL: **The chemical space project**. *Acc Chem Res* 2015, **48**:722−730.

5. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, *et al.*: **A next generation connectivity Map: L1000 platform and the first 1,000,000 profiles**. *Cell* 2017, **171**:1437−1452. e1417.

6. * Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, Humeidi R, Peck D, Wu X, Tang AA, *et al.*: **Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling**. *Nat Can (Que)* 2020, **1**:235−248.
Drug repurposing exercise of 4518 compounds against 578 human cancer cell lines. The authors use a barcoding method to screen the drugs against cell lines in pools, finding that a surprisingly large number of non-oncology drugs are able to selectively inhibit growth of subsets of cell lines.

7. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, *et al.*: **The ChEMBL database in 2017**. *Nucleic Acids Res* 2017, **45**:D945−D954.

8. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, Thiessen PA, He S, Zhang J: **PubChem**

BioAssay: 2017 update. *Nucleic Acids Res* 2017, **45**: D955–D963.

9. Ryan CJ, Cimermančič P, Szpiech ZA, Sali A, Hernandez RD, Krogan NJ: **High-resolution network biology: connecting sequence with function**. *Nat Rev Genet* 2013, **14**:865.

10. Weiniger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *J Chem Inf Comput Sci* 1988, **28**:31–36.

11. Durant JL, Leland BA, Henry DR, Nourse JG: **Reoptimization of MDL keys for use in drug discovery**. *J Chem Inf Comput Sci* 2002, **42**:1273–1280.

12. Devinyak O, Havrylyuk D, Lesyk R: **3D-MoRSE descriptors explained**. *J Mol Graph Model* 2014, **54**:194–203.

13. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S: **GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors**. *J Med Chem* 2000, **43**:3233–3243.

14. Riniker S: **Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences**. *J Chem Inf Model* 2017, **57**:726–741.

15. Rogers D, Hahn M: **Extended-connectivity fingerprints**. *J Chem Inf Model* 2010, **50**:742–754.

16. David L, Thakkar A, Mercado R, Engkvist O: **Molecular representations in AI-driven drug discovery: a review and practical guide**. *J Cheminf* 2020, **12**:56.

17. Sanchez-Lengeling B, Aspuru-Guzik A: **Inverse molecular design using machine learning: generative models for matter engineering**. *Science* 2018, **361**:360–365.

18. Jin W, Barzilay R, Jaakkola TS. *Hierarchical graph-to-graph translation for molecules*. arXiv; 2019. 1907.11223.

19. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V: **MoleculeNet: a benchmark for molecular machine learning**. *Chem Sci* 2018, **9**:513–530.

20. Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Goldstein A, Bukar R, Bauer KE, Dilley H, Rocke DM: **Predicting ligand binding to proteins by affinity fingerprinting**. *Chem Biol* 1995, **2**:107–118.

21. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space**. *Nat Biotechnol* 2006, **24**:805–815.

22. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, *et al.*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**: 175–181.

23. MacDonald ML, Lamerdin J, Owens S, Keon BH, Bilter GK, Shang Z, Huang Z, Yu H, Dias J, Minami T, *et al.*: **Identifying off-target effects and hidden phenotypes of drugs in human cells**. *Nat Chem Biol* 2006, **2**:329–337.

24. Young DW, Bender A, Hoyt J, McWhinnie E, Chirn GW, Tao CY, Tallarico JA, Labow M, Jenkins JL, Mitchison TJ, *et al.*: **Integrating high-content screening and ligand-target prediction to identify mechanism of action**. *Nat Chem Biol* 2008, **4**:59–68.

25. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, *et al.*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**: 1929–1935.

26. Sawada R, Iwata M, Tabei Y, Yamato H, Yamanishi Y: **Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures**. *Sci Rep* 2018, **8**:156.

27. Holbeck SL, Collins JM, Doroshow JH: **Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines**. *Mol Canc Therapeut* 2010, **9**:1451–1460.

28. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity**. *Science* 2008, **321**: 263–266.

29. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, Juan-Blanco T, Aloy P: **Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker**. *Nat Biotechnol* 2020, **38**:1087–1096.
**
Presentation of the Chemical Checker (CC), a resource that provides processed, harmonized and ready-to-use bioactivity signatures of 1M small molecules, and offers a rich portrait of the compounds available in the public domain. The CC divides data into five levels of increasing complexity following the way we think of drug activity, including chemistry, targets, networks, cells and clinics.

30. Bertoni M, Duran-Frigola M, Badia-i-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llergo A, Brun-Heath I, *et al.*: **Bioactivity descriptors for uncharacterized chemical compounds**. *Nat Commun* 2021, **12**:3932.
*
Collection of deep neural networks to infer bioactivity signatures for any compound of interest, even when little or no experimental information is available. These 'signaturizers' relate to the 25 types of bioactivities present in the Chemical Checker [29], including target profiles, cellular response and clinical outcomes, and can be used as drop-in replacements for chemical descriptors in chemoinformatics tasks.

31. Wassermann AM, Lounkine E, Davies JW, Glick M, Camargo LM: **The opportunities of mining historical and collective data in drug discovery**. *Drug Discov Today* 2015, **20**:422–434.

32. Hellberg S, Sjostrom M, Skagerberg B, Wold S: **Peptide quantitative structure-activity relationships, a multivariate approach**. *J Med Chem* 1987, **30**:1126–1135.

33. Xiao N, Cao DS, Zhu MF, Xu QS: **protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences**. *Bioinformatics* 2015, **31**: 1857–1859.

34. Xu Y, Verma D, Sheridan RP, Liaw A, Ma J, Marshall NM, McIntosh J, Sherer EC, Svetnik V, Johnston JM: **Deep dive into machine learning models for protein engineering**. *J Chem Inf Model* 2020, **60**:2773–2790.

35. Kulmanov M, Hoehndorf R: **DeepGOPlus: improved protein function prediction from sequence**. *Bioinformatics* 2020, **36**: 422–429.

36. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM: **Unified rational protein engineering with sequence-based deep representation learning**. *Nat Methods* 2019, **16**: 1315–1322.
*
Application of deep learning to derive semantically-rich protein descriptors from their amino-acid sequences and show that, since they preserve the structural, evolutionary and physicochemical information encoded in the sequences, are useful in protein engineering.

37. Asgari E, Mofrad MR: **Continuous distributed representation of biological sequences for deep proteomics and genomics**. *PloS One* 2015, **10**, e0141287.

38. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B: **Modeling aspects of the language of life through transfer-learning protein sequences**. *BMC Bioinf* 2019, **20**:723.

39. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS: **Evaluating protein transfer learning with TAPE**. *Adv Neural Inf Process Syst* 2019, **32**:9689–9701.

40. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, *et al.*: *ProtTrans: towards cracking the language of life's code through self-supervised learning*. bioRxiv; 2021.
**
The authors present different pretrained LMs and compare them to state-of-the-art solutions using evolutionary information. They show, for the first time, that LMs can perform equally well than evolutionary baseline models, while requiring much less time to compute. Additionally, they provide a hub for protein sequence embedding models where their own benchmarks and comparisons are available (https://github.com/agemagician/ProtTrans).

41. Bepler T, Berger B: *Learning protein sequence embeddings using information from structure*. aRxiv; 2019. arXiv:1902.08661vol. 2.

42. Bepler T, Berger B: **Learning the protein language: evolution, structure, and function**. *Cell Syst* 2021, **12**:654–669. e653.
**
Review of the major breakthroughs in the field of LMs and an illustration of how these models can be tailored to protein down-stream tasks

by incorporating evolutionary and physics-based inductive biases during pre-training.

43. Rao R, Iiu J, Verkuil R, Meier J, Canny JF, Abbeel P, Sercu T, Rives A: *MSA transformer*. bioRxiv; 2021.

44. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, *et al.*: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci U S A* 2021:118.

45. ** Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589, https://doi.org/10.1038/s41586-021-03819-2.
ALphaFold2 is a novel machine learning approach that uses physical and biological knowledge encoded in the protein structure and multi-sequence alignments to train a deep learning algorithm (transformer) able to predic the three-dimensional structure of new proteins at an unprecedented accuracy.

46. Devlin J, Chang M, Lee K, Toutanova K: *BERT: pre-training of deep bidirectional transformers for language understanding*. arXiv; 2018. arXiv:1810.0480.

47. * Vig J, Madani A, Varshney L, Xiong C, Socher R, Rajani R: *BERTology meets biology: interpreting attention in protein language models*. arXiv; 2020. arXiv:2006.15222.
Thorough analysis of transformer (BERT) models to study the features that the models learn from the protein sequences. This article complements previous works by correlating the attention weights of the BERT models to known biological associations, such as the evolutionary or spatial proximity of amino-acids or the composition of functional sites.

48. Haupt VJ, Daminelli S, Schroeder M: **Drug promiscuity in PDB: protein binding site similarity is key**. *PloS One* 2013, **8**, e65894.

49. Weill N, Rognan D: **Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites**. *J Chem Inf Model* 2010, **50**:123–135.

50. Siragusa L, Cross S, Baroni M, Goracci L, Cruciani G: **BioGPS: navigating biological space to predict polypharmacology, off-targeting, and selectivity**. *Proteins* 2015, **83**:517–532.

51. Stark A, Sunyaev S, Russell RB: **A model for statistical significance of local similarities in structure**. *J Mol Biol* 2003, **326**:1307–1316.

52. Duran-Frigola M, Siragusa L, Ruppin E, Barril X, Cruciani G, Aloy P: **Detecting similar binding pockets to enable systems polypharmacology**. *PLoS Comput Biol* 2017, **13**, e1005522.

53. Chaudhari R, Fong LW, Tan Z, Huang B, Zhang S: **An up-to-date overview of computational polypharmacology in modern drug discovery**. *Expet Opin Drug Discov* 2020, **15**:1025–1044.

54. Govindaraj RG, Brylinski M: **Comparative assessment of strategies to identify similar ligand-binding pockets in proteins**. *BMC Bioinf* 2018, **19**:91.

55. Ehrt C, Brinkjost T, Koch O: **A benchmark driven guide to binding site comparison: an exhaustive evaluation using tailor-made data sets (ProSPECCTs)**. *PLoS Comput Biol* 2018, **14**, e1006483.

56. Ehrt C, Brinkjost T, Koch O: **Binding site characterization - similarity, promiscuity, and druggability**. *Medchemcomm* 2019, **10**:1145–1159.

57. ** Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein MM, Correia BE: **Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning**. *Nat Methods* 2020, **17**:184–192.
Novel strategy to generate descriptors based on a set of chemical and geometrical features extracted from protein surfaces and transferred to a convolutional neural network. The generated descriptors are application-specific, including pocket similarity comparison, interaction sites prediction and ultrafast scanning of surfaces.

58. * Simonovsky M, Meyers J: **DeeplyTough: learning structural comparison of protein binding sites**. *J Chem Inf Model* 2020, **60**:2356–2366.
Convolutional neural network to generate descriptors of protein binding pockets based on their three-dimensional representations. The network is trained so that slightly dissimilar pockets can be distinguished, achieving robustness to nuisance variations. DeeplyTough is indeed one of the best pocket comparison methods according its results in the ProSPECCT benchmark.

59. * Bongers BJ, AP IJ, Van Westen GJP: **Proteochemometrics-recent developments in bioactivity and selectivity modeling**. *Drug Discov Today Technol* 2019, **32–33**:89–98.
General overview on recent proteochemometric approaches, with special emphasis on sequence- and structure-based target pocket descriptors and when to better use them depending on the task.

60. Torng W, Altman RB: **Graph convolutional neural networks for predicting drug-target interactions**. *J Chem Inf Model* 2019, **59**:4131–4149.

61. Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, Koes DR, Kurtzman T: **Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening**. *PloS One* 2019, **14**, e0220113.

62. Behan FM, Iorio F, Picco G, Goncalves E, Beaver CM, Migliardi G, Santos R, Rao Y, Sassi F, Pinnelli M, *et al.*: **Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens**. *Nature* 2019, **568**:511–516.

63. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, *et al.*: **Defining a cancer dependency Map**. *Cell* 2017, **170**:564–576. e516.

64. van der Velden DL, Hoes LR, van der Wijngaart H, van Berge Henegouwen JM, van Werkhoven E, Roepman P, Schilsky RL, de Leng WWJ, Huitema ADR, Nuijen B, *et al.*: **The Drug Rediscovery protocol facilitates the expanded use of existing anticancer drugs**. *Nature* 2019.

65. Swinney DC, Anthony J: **How were new medicines discovered?** *Nat Rev Drug Discov* 2011, **10**:507–519.

66. Parisi D, Adasme MF, Sveshnikova A, Bolz SN, Moreau Y, Schroeder M: **Drug repositioning or target repositioning: a structural perspective of drug-target-indication relationship for available repurposed drugs**. *Comput Struct Biotechnol J* 2020, **18**:1043–1055.

67. https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling.

68. Kawata K, Hatano A, Yugi K, Kubota H, Sano T, Fujii M, Tomizawa Y, Kokaji T, Tanaka KY, Uda S, *et al.*: **Trans-omic analysis reveals selective responses to induced and basal insulin across signaling, transcriptional, and metabolic networks**. *iScience* 2018, **7**:212–229.

69. Vitrinel B, Koh HWL, Mujgan Kar F, Maity S, Rendleman J, Choi H, Vogel C: **Exploiting interdata relationships in next-generation proteomics analysis**. *Mol Cell Proteomics* 2019, **18**:S5–S14.

70. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O: **Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets**. *Mol Syst Biol* 2018, **14**, e8124.

71. Rigden DJ, Fernandez XM: **The 27th annual Nucleic Acids Research database issue and molecular biology database collection**. *Nucleic Acids Res* 2020, **48**:D1–D8.

72. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A: **The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins**. *Database* 2016:2016.

73. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE: **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**. *eLife* 2017, **6**, e26726.

74. Cai H, Zheng VW, Chang KC-C: *A comprehensive survey of graph embedding: problems, techniques and applications*. arXiv; 2017. *1709.07604*.

75. ** Li M, Huang K, Zitnik M: *Representation learning for networks in biology and medicine: advancements, challenges, and opportunities*. arXiv; 2021. arXiv:2104.04883.
Comprehensive review on network embedding approaches used in biology, providing a detailed overview of the different techniques that emerged in the last years, together with illustrative examples of their applicability on different biological entities such as proteins, small

molecules, diseases, omics experiments, protein interactions and even health records.

76. * Fernández-Torras A, Duran-Frigola M, Aloy P: *Integrating and formatting biological knowledge in the Bioteque, a comprehensive repository of biomolecular descriptors.* bioRxiv; 2021.
Presentation of the 'Bioteque', a gigantic knowledge graph centered on 12 types of biological entities and containing over 550k nodes and 30M edges from >200 data sources. The authors then use network embedding techniques to systematically provide node descriptors capturing a whole variety of biological contexts.

77. * Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A: **Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer**. *Nat Commun* 2021, **12**:124.
Thorough benchmark of different joint dimensionality reduction methods focusing on continuous omics data, including gene expression, copy number variation, miRNAs and methylation analyses.

78. Bady P, Doledec S, Dumont B, Fruget JF: **Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities**. *C R Biol* 2004, **327**:29–36.

79. Zhong X, Kaalia R, Rajapakse JC: **GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings**. *BMC Genom* 2019, **20**:918.

80. * Wang S, Flynn ER, Altman RB: **Gaussian embedding for large-scale gene set analysis**. *Nat Mach Intell* 2020, **2**:387–395.
Interesting application of network-based gene set embedding approach, where each gene set is represented as a multivariate Gaussian distribution rather than a single point in the embedding space, according to the proximity of these genes in a protein–protein interaction network.

81. Schubert PJ, Dorkenwald S, Januszewski M, Jain V, Kornfeld J: **Learning cellular morphology with neural networks**. *Nat Commun* 2019, **10**:2736.

82. Zitnik M, Leskovec J: **Predicting multicellular function through multi-layer tissue networks**. *Bioinformatics* 2017, **33**:i190–i198.

83. * Ruiz C, Zitnik M, Leskovec J: **Identification of disease treatment mechanisms through the multiscale interactome**. *Nat Commun* 2021, **12**:1796.
Strategy to identify potential treatment mechanisms for disease by comparing diffusion states (embeddings) from a random walker exploration of protein-interaction, drug-target and disease-gene networks. By comparing drug and disease diffusion profiles, the multiscale interactome provides an interpretable basis to identify the proteins and biological functions that explain successful treatments.

84. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua MS, So S, Butte AJ: **Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets**. *Nat Commun* 2017, **8**:16022.

85. Pauls E, Bayod S, Mateo L, Alcalde V, Juan-Blanco T, Saido T, Saito T, Berebguer Llergo A, Stephan Otto Attolini C, Gay M, *et al.*: *Identification and drug-induced reversion of molecular signatures of Alzheimer's disease onset and progression in AppNL-G-F, AppNL-F and 3xTg-AD mouse models.* bioRxiv; 2021.

nature communications

Article

https://doi.org/10.1038/s41467-022-33026-0

Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque

Received: 6 May 2022

Accepted: 30 August 2022

Published online: 09 September 2022

Check for updates

Adrià Fernández-Torras[1], Miquel Duran-Frigola[1,2], Martino Bertoni[1], Martina Locatelli[1] & Patrick Aloy[1,3]

Biomedical data is accumulating at a fast pace and integrating it into a unified framework is a major challenge, so that multiple views of a given biological event can be considered simultaneously. Here we present the Bioteque, a resource of unprecedented size and scope that contains pre-calculated biomedical descriptors derived from a gigantic knowledge graph, displaying more than 450 thousand biological entities and 30 million relationships between them. The Bioteque integrates, harmonizes, and formats data collected from over 150 data sources, including 12 biological entities (e.g., genes, diseases, drugs) linked by 67 types of associations (e.g., 'drug treats disease', 'gene interacts with gene'). We show how Bioteque descriptors facilitate the assessment of high-throughput protein-protein interactome data, the prediction of drug response and new repurposing opportunities, and demonstrate that they can be used off-the-shelf in downstream machine learning tasks without loss of performance with respect to using original data. The Bioteque thus offers a thoroughly processed, tractable, and highly optimized assembly of the biomedical knowledge available in the public domain.

Systematic measurements of biological samples through omics technologies, together with efforts to distil the scientific literature into structured databases, are providing an ever-growing corpus of biomedical and biomolecular information[1]. Indeed, the data stored in the EMBL-EBI has increased sixfold in the last few years, from 40 petabytes in 2014 to over 250 in 2021[2]. Associated with this phenomenon, a variety of nomenclatures have been proposed, along with identifiers, levels of resolution (e.g., protein isoforms or gene splice variants) and experimental conditions, making data integration and harmonization across platforms a challenging step[3]. As a result, even though as many as 1641 resources were listed in the 2021 Online Molecular Biology Database Collection[4], only a small portion are broadly used, and hundreds remain isolated with their own particular formats[5,6]. Aware of the situation, several initiatives have emerged to standardize biological data by establishing common vocabularies and formats. For instance, the pioneering Harmonizome[7] was able to integrate knowledge from several gene-centric databases by representing data (e.g., gene expression, disease genetics, etc.) in a simple discretized format that was applicable to each type of data.

Nowadays, in an attempt to capture the complexity of biological systems, multiple omics profiles are often measured simultaneously (i.e., trans-omics analyses)[8,9] so that complementary views of a given phenotype or event can be considered in parallel and as a whole[10]. However, current methods mainly adapt and combine existing strategies developed to analyse individual omics data, and often the net result is that most conclusions are drawn from the most informative single data type, while the rest are used as support. It is thus fundamental to devise strategies able to capture the coordinated interplay of the many regulatory layers present in biological systems. Himmelstein et al. suggested the use of knowledge graphs (KG) as a tool to integrate heterogeneous biomolecular data[11,12]. In a biomedical KG, nodes represent biological or chemical entities (e.g., genes, cell lines, diseases, drugs,

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. [2]Ersilia Open Source Initiative, Cambridge, UK. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. e-mail: patrick.aloy@irbbarcelona.org

etc.), and edges capture the interactions or relationships between them (e.g., 'drug treats disease' or 'cell upregulates gene'). This concept has recently been expanded to include clinical entities[13].

However, large biomedical networks are intractable by conventional graph analytics techniques[14], thus prompting the development of dimensionality reduction techniques that learn numerical feature representations of nodes and links in a low dimensional space (aka network embeddings). As a result, network embeddings reduce the dimensionality of the data while preserving the topological information and the connectivity of the original network[15]. Moreover, the vectorial format of the nodes resulting from network embedding approaches is better suited as an input for machine learning algorithms. For instance, Zitnik and Leskovek presented a set of protein embeddings that consider the protein interactions within each human tissue, as well as inter-tissue relationships, and showed their potential to predict tissue-specific protein functions[16]. Later on, the same authors embedded several networks (i.e., protein–protein, drug–target and disease–gene interactions) to explore the mechanisms of drug action[17]. Recently, Cantini et al. evaluated the capacity of several dimensionality reduction methods to integrate continuous multi-omics data (e.g., gene expression, copy number variation, miRNAs and methylation)[18], assessing their ability to preserve the structure of the original data and their prediction performance in different tasks. Overall, embedding-based descriptors provide a scalable and standard means to capture complex relationships between biological entities and they integrate the myriad of omics experiments associated with them[19,20].

To make biomedical knowledge embeddings available to the broad scientific community, we have developed the Bioteque, a resource of unprecedented size and scope that contains pre-calculated embeddings derived from a gigantic heterogeneous network (more than 450k nodes and 30M edges). The Bioteque harmonizes data extracted from over 150 data sources, including 12 distinct biological entities (e.g., genes, diseases, compounds) linked through 67 types of relationships (e.g., 'compound treats disease', 'gene interacts with gene'). We demonstrate that Bioteque embeddings retain the information contained in the large biological network and illustrate with examples how this concise representation of the data can be used to evaluate, characterize and predict a wide set of experimental observations. Finally, we offer an online resource to facilitate access and exploration of the pre-calculated embeddings (https://bioteque.irbbarcelona.org).

Results

A comprehensive biomedical knowledge graph (KG)

To build a KG that integrates biological and biomedical knowledge available in the public domain, we first defined the basic entities (nodes) of the network and the relationships between them (edges). As shown in Fig. 1a, the resource is gene-centric.

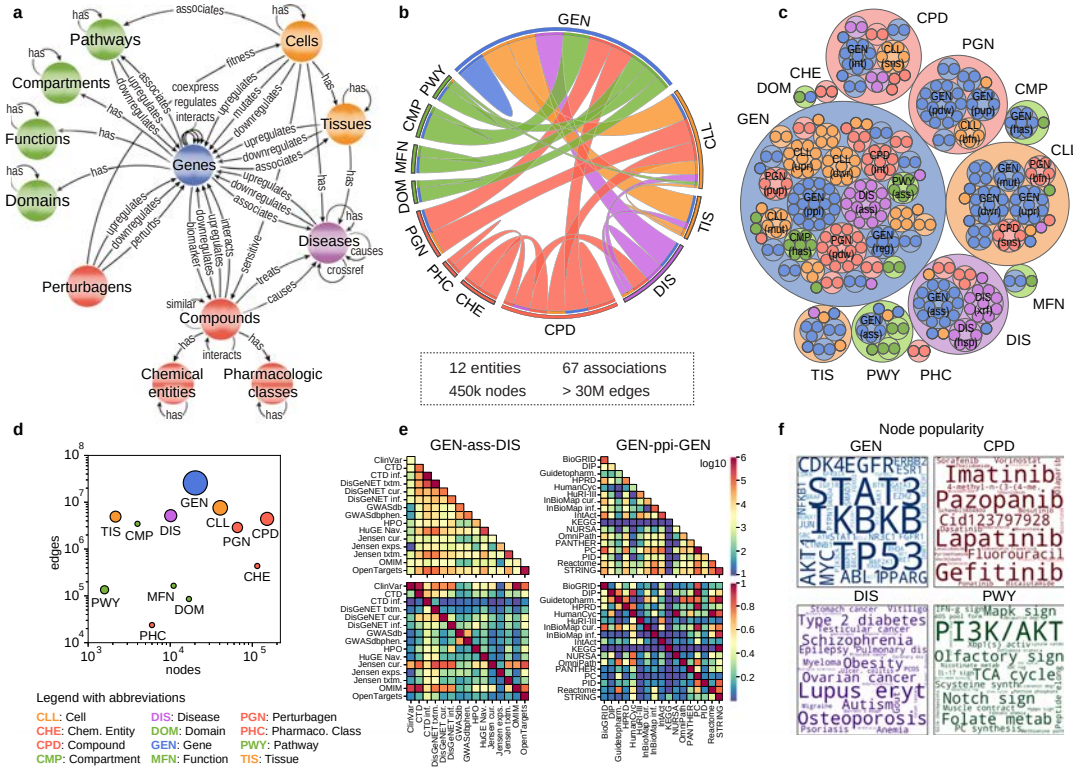


Fig. 1 | Building the Bioteque knowledge graph (KG). a Metagraph of the Bioteque, showing all the entities and the most representative associations (metaedges) between them. **b** Circos plot representation of the KG, showing the relationships between nodes. **c** Treeplot showing the number of datasets used to construct each metaedge. **d** Total number of nodes (x-axis) and edges (y-axis) available for each entity type. The size of the circles is proportional to the number of metaedges in which the entities participate. **e** Number of edges (top row) and overlap (bottom row) between the datasets inside the 'gene associates with disease' (GEN-ass-DIS, left) and 'protein interacts protein' (GEN-ppi-GEN, right) associations. **f** Most popular nodes in the KG within the gene (GEN, blue), compound (CPD, red), disease (DIS, purple) and pathway (PWY, green) universe. Dataset associations were de-propagated across the corresponding ontologies (when possible) before computing the popularity of the nodes. A propagated version of this plot is shown in Supplementary Fig. 1.

Table 1 | Biological and chemical entities in the knowledge graph (KG)

Metanode	Abbreviation	Nodes	Metaedges	Edges	Example 1	Example 2
Cell	CLL	40,681	15	7,512,366	CLL-upr-GEN	CLL-mut-GEN
Cellular component	CMP	3992	2	3,461,731	GEN-has-CMP	CPD-hsp-CMP
Chemical entity	CHE	115,002	2	435,011	CHE-hsp-CHE	CHE-hsp-CPD
Compound	CPD	153,279	12	5,713,785	CPD-int-GEN	CPD-trt-DIS
Disease	DIS	10,144	10	5,037,293	GEN-ass-DIS	CPD-cau-DIS
Domain	DOM	16,913	2	85,747	GEN-has-DOM	DOM-hsp-DOM
Gene	GEN	20,229	42	25,788,255	GEN-ppi-GEN	GEN-pho-GEN
Molecular function	MFN	11,006	2	164,447	GEN-has-MFN	MFN-hsp-MFN
Pathway	PWY	1585	4	133,851	GEN-ass-PWY	PWY-hsp-PWY
Perturbagen	PGN	66,988	7	2,889,047	PGN-bfn-CLL	PGN-gfn-CLL
Pharmacological class	PHC	6072	2	31,004	CPD-has-PHC	PHC-hsp-PHC
Tissue	TIS	2157	8	4,928,112	GEN-ass-TIS	TIS-upr-GEN

We show the number of nodes, metaedges and edges contained in the KG for each metanode, as well as some examples of metaedges.

Thus, genes and gene products (GEN) are represented in the centre of the KG scheme and are involved in most associations. To better characterize genes and proteins, we collected their molecular function (MFN), cellular component localization (CMP), functional structure or domains (DOM), and biological processes or pathways (PWY). Additionally, we included information on cell lines (CLL), one of the most studied entities in biology, as well as their anatomical ensembles, namely the tissues (TIS). Analogously, chemical compounds (CPD) are depicted together with pharmacological classes (PHC) and chemical entities (CHE), two common vocabularies for medicinal compounds. Diseases (DIS) are abnormal conditions that have been widely studied in various fields, giving rise to a wide diversity of interactions between different nodes. Furthermore, although CPD and DIS are two of the major perturbational agents found in repositories like GEO[21] and LINCS[22], we also considered other biological entities such as miRNA, shRNA and overexpression vectors that can also act as perturbagens (PGN). To connect the entities in the Bioteque, we defined 67 types of associations reflecting biological relationships between them. An example of such an association would be a gene that is associated with a given pathway (GEN-ass-PWY) and might be downregulated in a certain cell (GEN-dwr-CLL) or tissue type (GEN-dwr-TIS), or a drug compound that is used to treat a disease (CPD-trt-DIS). A comprehensive list of all the biological and chemical entities included in the Bioteque, as well as the different associations, are summarized in Fig. 1a and Table 1 and provided in Supplementary Data 1 and 2.

Having defined the biological entities and their interactions, we populated the Bioteque with data collected from representative datasets and resources. We first incorporated data from the Harmonizome[7], the most complete compendium of biological datasets to date, and added data from another 100 reference datasets. Each dataset was mapped to the KG scheme (or metagraph) depicted in Fig. 1a. Inspired by the Harmonizome strategy, we processed each dataset separately following author guidelines, when possible ("Methods"). In brief, we binarized continuous data so that it could be represented in a network format, and we standardized identifiers from multiple sources.

The current version of the KG contains over 450k nodes, belonging to 12 types of biological entities (metanodes), and over 30M edges, representing 67 types of relationships (metaedges) (Fig. 1b). In general, the size of our KG is comparable to other recently published biomedical KGs[13,23–25]. In fact, taking as a reference the comparison made by Bonnet et al.[26], our KG is the most comprehensive in the number of processed datasets, the second most comprehensive with respect to entities, edges, and relation types, and the third regarding entity types (Supplementary Table 1). Not surprisingly, genes and proteins account for most of the edges (25M) and metaedges (42) in the graph (Fig. 1c, d). In terms of the number of reference datasets, protein interactions (GEN-ppi-GEN) and gene-disease associations (GEN-ass-DIS) are the most represented metaedges, supported by 17 and 15 datasets, respectively (Fig. 1c). A comparison of data extracted from each dataset revealed that, although there is some overlap, most sets cover distinct associations, probably due to differences in the focus of the underlying experiments (i.e., physical[27] vs. functional[28] PPIs or drug-driven[29] vs. genomics-driven[30] gene associations) (Fig. 1e).

Calculation of network embeddings across the KG

To integrate the biological knowledge gathered, we devised an approach to obtain, for a given node in the KG, a set of embeddings capturing different contexts defined by one or more types of relationships between this node and other entities (Fig. 2a). For example, the pharmacological context of a certain compound can be captured by 'compound interacts with protein' associations, while its clinical context may be captured by 'compound treats disease' links. The embedding procedure is as follows. We first define the types of biological entities (metanodes) to be connected and the sequence of relationships (metaedges) between them that we wish to explore. This sequence of relationships is called metapath. We then systematically examined all possible paths from the source and target nodes of the metapath, downweighting highly connected nodes to ensure exhaustive exploration of the network[11]. This step yields a simplified homogeneous (when source and target metanodes belong to the same type) or bipartite (when source and target metanodes belong to different types) graph that can be explored with conventional network embedding techniques. We chose to use a random walk method, where the trajectories of an agent that explores the network are retained and eventually fed into a text-embedding algorithm[31]. As a result, for each node in the network, a 128-dimensional vector (i.e., an embedding) is obtained, so that similar vectors are given to nodes that are proximal in the network. During this process, we mostly keep different datasets separately (i.e., without merging equivalent networks in different sources) to preserve the original information captured in them[32]. A more detailed description of the protocol is provided in the "Methods" section.

We have created a resource of pre-calculated biomedical embeddings, the Bioteque, where we have exhaustively considered most metapaths of length 1 and 2 extracted from the KG (i.e., direct connections between source and target nodes, or with one intermediate node between them). In addition, we have curated a collection of 135 metapaths of length ≥3. Overall, the Bioteque currently holds a total of 81, 785, and 175 embeddings of length 1, 2, and ≥3,

https://doi.org/10.1038/s41467-022-33026-0

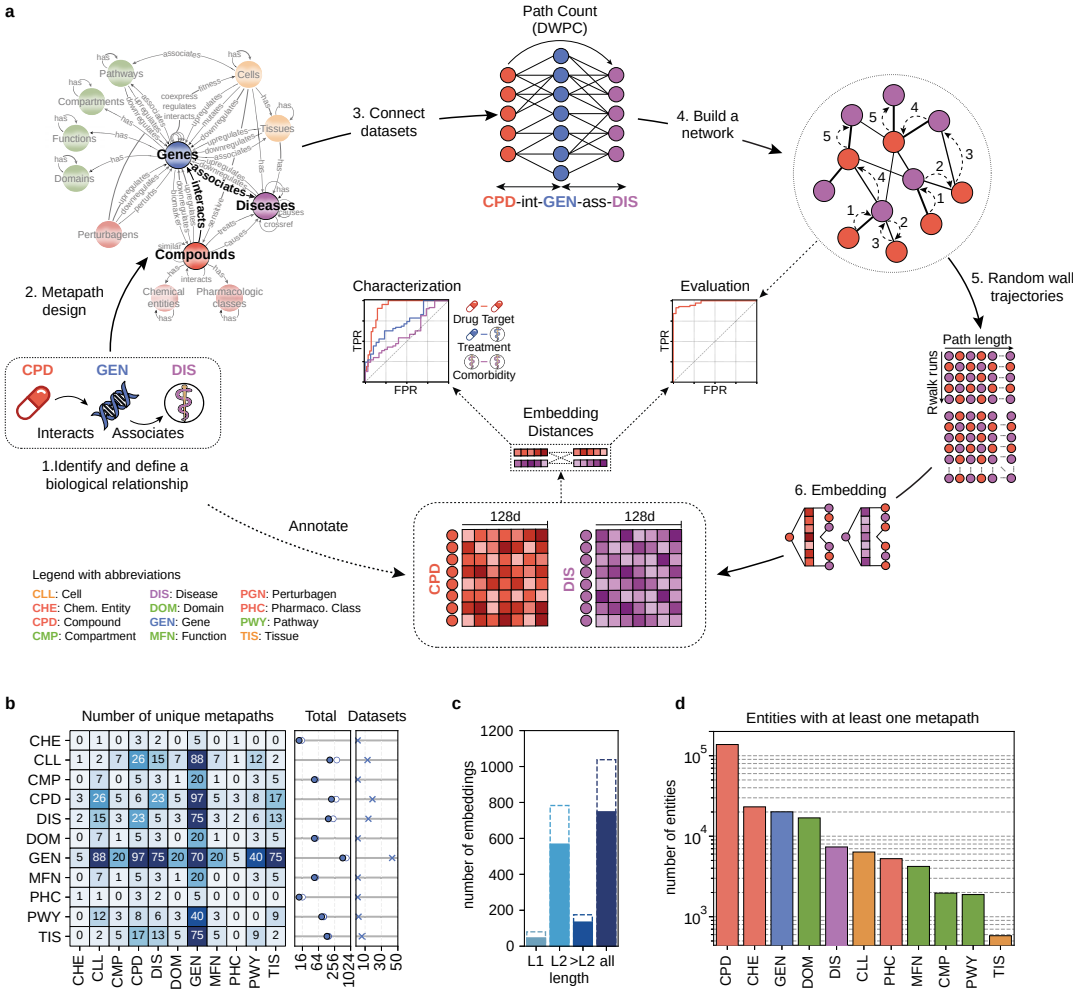


Fig. 2 | Generating the Bioteque embeddings. a Scheme of the methodology. First, we define the biological entities to be connected and the specific context to be explored. Then a source-target network is derived by traversing all the paths available from the source to the target nodes of a given metapath. The vicinity of each node in the network is then explored by a random walker and embedded in a 128-dimensional space. Finally, embeddings are evaluated and characterized. **b** Number of unique metapath embeddings linking each entity. In the middle plot, the filled dots indicate the total number of unique metapaths while the empty dots show the total number of metapath-dataset combinations. In the rightmost plot, we show the number of entity-specific datasets used in the metapaths. **c** Number of metapath-dataset embedding combinations obtained at each metapath length. Solid bars highlight the number of unique metpaths. **d** Number of nodes within each entity with at least one embedding in the Bioteque resource. Note that during metapath construction, perturbagen (PGN) entities are always mapped to the corresponding perturbed genes. Thus, although used to construct several metapaths, PGN nodes are not explicitly embedded, i.e., they are not the first or last nodes in the metapaths.

respectively (Fig. 2c and Supplementary Data 3). Length 1 (L1) metapaths correspond to direct associations in the knowledge graph and provide the simplest domain knowledge representations of the entities. Larger metapaths (>L1), on the other hand, are either dedicated to connecting different entities through a third one (i.e., CPD-int-GEN-ass-DIS) or extend L1 associations to similar entities (i.e., CPD-int-GEN-ppi-GEN or CPD-trt-DIS-ass-GEN-ass-DIS), allowing the identification of more complex relationships between biological entities (i.e., two compounds may target different proteins yet affect the same pathway, or CPD-int-GEN-ass-PWY).

Given that the constructed KG is gene-centric, genes (GEN) are the most frequently embedded biological entity in the resource (515 unique metapaths from 43 different datasets) followed by compounds (CPD), cell lines (CLL), and diseases (DIS) (198, 168 and 150 unique metapaths, respectively) (Fig. 2b). Furthermore, most of the metapaths used gene entities, such as those derived from omics experiments or literature curated annotations, as bridges to connect distinct entities (Supplementary Fig. 2). Compounds also play an important role, connecting pharmacological classes and chemical entities to the rest of the graph and being a major source of metapaths embedding cell lines, diseases and tissues.

Overall, the Bioteque provides a collection of 1041 embeddings obtained from 746 unique metapaths, covering all entities defined in the biological KG (Fig. 2d).

https://doi.org/10.1038/s41467-022-33026-0

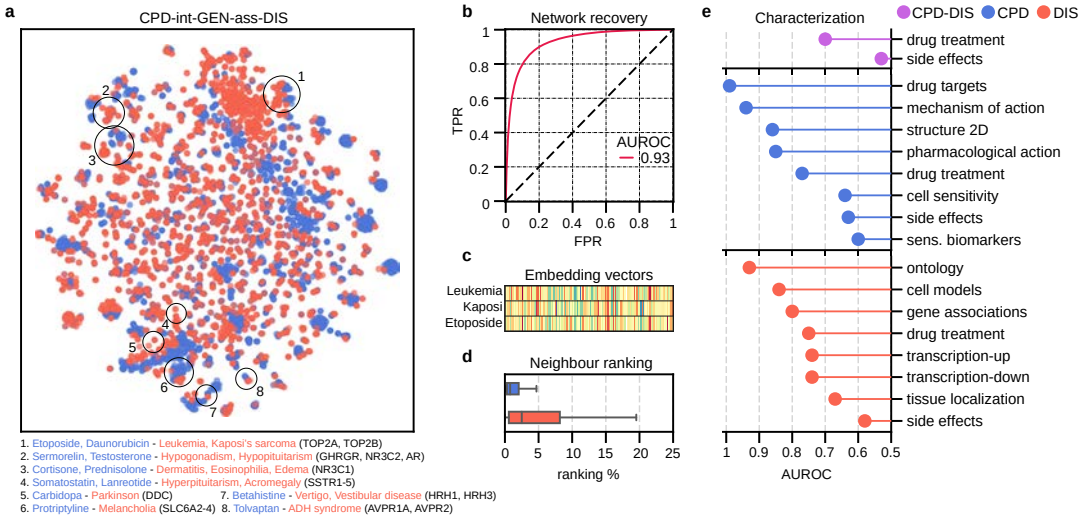


Fig. 3 | A Bioteque embedding summary card. a 2D projection (opt-SNE) of the compound (CPD, blue) and disease (DIS, red) embeddings from the metapath 'compound interacts protein associates with disease' (CPD-int-GEN-ass-DIS). We highlight clusters of compounds and diseases sharing treatment evidence. We highlight some representative compounds and diseases found in these clusters, together with the drug targets associated with the diseases. **b** ROC curve validation when reconstructing the original network with the corresponding embeddings. **c** Visual representation of the embedding vectors of leukaemia (top) and Kaposi's sarcoma (middle), together with the drug Etoposide (bottom). **d** Ranking proportion in which the putative CPD ($n = 131{,}648$) and DIS ($n = 134{,}997$) neighbours are found. Box plots indicate median (middle line), 25th, 75th percentile (box) and max value within the 1.5*75th percentile (whiskers). **e** Recapitulation of orthogonal associations by using embedding distances. The AUROC (x-axis) summarizes the performance obtained when ranking the orthogonal associations. Drug targets are collected from Drugbank[85], the Drug Repurposing Hub[86] and PharmacoDB[87], and gene-disease associations are obtained from Open Targets[29].

Embeddings retain the interactions in the original KG, and reveal relationships between biological entities depending on the scope and type of data

Having obtained embeddings for all nodes in the KG, we performed a set of analyses to, on the one hand, validate that the embeddings retained the connectivity observed in the KG and, on the other, to characterize each embedding space in the light of other (orthogonal) datasets in the Bioteque. As an illustrative example, Fig. 3 shows the analysis of the metapath CPD-int-GEN-ass-DIS, corresponding to compounds that interact with genes, which are, in turn, associated with a disease.

To validate the embeddings, we calculated their cosine distances pairwise, and checked that proximal embeddings corresponded to edges in the KG (Fig. 3b), measured with the Area Under the Receiver Operating Characteristic (AUROC) metric. Similarly, when we used the embedding distances to rank entity pairs, we found their known neighbours in the closest 10% of possible nodes (Fig. 3d). Note that the goal of this study is not to benchmark the embedding method (which is already a well-accepted implementation in the field[31]), but to provide an assessment of the approach across a comprehensive set of cases.

Analogously, distances between embeddings can be used to measure whether the dimensional space preserves similarities among entities that share biological traits (i.e., cell lines sharing tissue of origin or genes sharing molecular functions). Following this rationale, we can characterize the type of biological signal captured by a given metapath by comparing its embeddings to a battery of reference biological traits, an approach already used to benchmark drug-drug similarities on the basis of shared chemical features[33]. The use of embeddings allows for straightforward comparison of entities of the same type (for example, similarity of cell lines according to their upregulated genes can be measured by computing distances of CLL entities in the CLL-upr-GEN embedding). Likewise, it is easy to compare and uncover correlations between different types of associations. For instance, the correlation between copy number amplification and upregulation can be assessed by considering similarities in the CLL-cnu-GEN and CLL-upr-GEN embedding spaces. In the CPD-int-GEN-ass-DIS example, drug targets and gene-disease associations are among the biomedical traits that are better recapitulated by the compound and disease embeddings (Fig. 3e). Accordingly, we see how compounds and diseases associated with similar treatments are close in the embedding space. We also observe that compound-disease treatment similarity is achieved at the edge level (AUROC: 0.7), suggesting that not only compounds and diseases with similar treatments are close in the embedding space, but also that compound-disease treatment pairs are often found in the same vicinity. Indeed, compound and disease-associated genes have proven useful in drug treatment prediction exercises[12,34].

A projection of the 128-dimensional embeddings onto a 2D space reveals clusters of drugs and treatments which, by the definition of the metapath, have identifiable targets (Fig. 3a). We find, for instance, drug-disease groups associated with the treatment of leukaemia (e.g., Etoposide and Daunorubicin), hormonal disorders (e.g., Somatostatin and Sermorelin), nervous system disorders (e.g., Carbidopa, Betahistine, and Protriptyline), and inflammatory conditions (e.g., Cortisone and Prednisolone). We observe that most of these drugs target a small subset of proteins or protein families directly related to the diseases, such as the growth hormone-releasing hormone receptor (GHRHR) for hypogonadism treatment, the somatostatin receptor (SSTR) for acromegaly treatment, and the DOPA decarboxylase to prevent dopamine formation in the treatment of Parkinson's disease. Additionally, the analysis reveals that drugs approved to treat either leukaemia or Kaposi's sarcoma cluster, share the topoisomerase II alpha (TOP2A) enzyme as target (Fig. 3c). Indeed, comorbidity between these two diseases has been reported in several studies[35–37], although, to the best of our knowledge, the role of TOP2A in this comorbidity has not been yet described.

https://doi.org/10.1038/s41467-022-33026-0

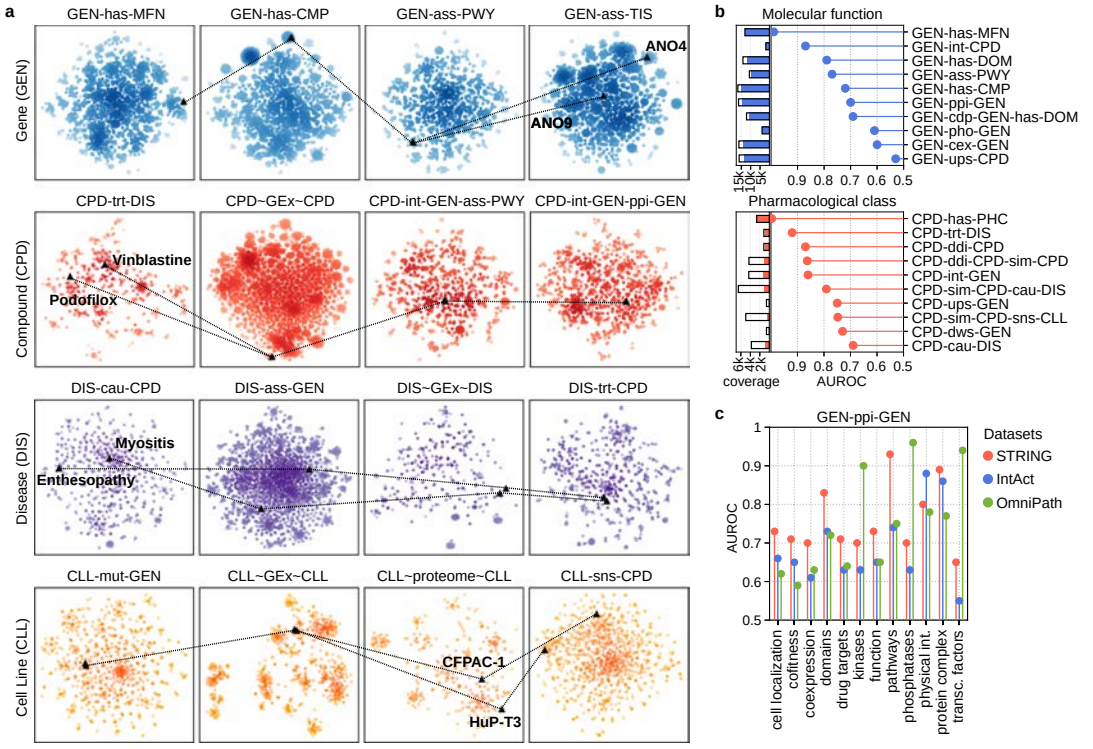


Fig. 4 | Comparison of embeddings built from different metapaths and datasets. a Four illustrative examples showing pairs of genes (GEN), compounds (CPD), diseases (DIS) and cell lines (CLL) with similarities or differences depending on the metapaths. The extended nomenclature of each metapath can be found in Supplementary Data 2. **b** Top metapaths (y-axis) recapitulating (AUROC, x-axis) gene molecular function (MFN, blue) and compound pharmacological class (PHC, red). The coloured bars indicate the proportion of nodes in the metapath that could be assessed (i.e., with annotated molecular function or pharmacological classes). **c** Gene embedding characterization of three reference PPI datasets, namely STRING, IntAct and OmniPath. We limited the analysis to the common gene universe (9395 genes) between the three sources.

The repertoire of embeddings encoded in the Bioteque enables exploration of a given biomedical entity from multiple perspectives, often corresponding to different biological contexts, such as genes with the same biological role yet expressed in different tissues, or cell lines with similar transcriptional profiles but dissimilar at the proteome and drug response levels (Fig. 4a). When performed systematically, this analysis quantifies the relationship of a certain metapath with the other metapaths in our collection, which in turn helps assessing the types of biological traits that it captures. Figure 4b shows ten of the top metapaths recapitulating gene molecular function and compound pharmacological class. We see that genes targeted by the same compounds or having similar domains tend to share molecular function while, as expected, sets of interacting compounds, or those with similar binding profiles, tend to belong to the same pharmacological class.

Additionally, one can explore differences among datasets within a single metapath. In Fig. 4c, we embedded three well-known protein-protein interaction (PPI) networks, representing functional interactions (STRING[28]), physical interactions (IntAct[27]), and protein-signalling interactions (OmniPath[38]), and quantified the capacity of these networks to capture a variety of biological features, from cellular localization to protein complexes. The diversity of functional interactions contained in STRING favours recapitulation of most of the features explored, especially those involving similar biological pathways (AUROC: 0.93), protein complexes (AUROC: 0.89) and protein domains (AUROC: 0.83). Not surprisingly, IntAct better preserves physical interactions (AUROC: 0.88) and shows good performance with protein complexes (AUROC: 0.86). Finally, OmniPath shows an enrichment in signalling processes such as kinase-substrate (AUROC: 0.9), phosphatase-substrate (AUROC: 0.96) and transcription factor interactions (AUROC: 0.94), in good agreement with the type of resources used to build this network.

In general, the different considerations followed to populate these networks may favour some domains of knowledge, hence suiting different tasks, which can be efficiently and systematically revealed by transforming them into embeddings. In the next sections, we present three illustrative examples on how these biological embeddings can be used off-the-shelf in a variety of tasks.

Gene expression embeddings as biological descriptors of cell lines

Gene Expression (GEx) experiments have been widely used to characterize cellular identity and state, as they broadly recapitulate tissues of origin[39] and they are notable genomic biomarkers for anticipating drug response[40]. However, these experiments typically measure the expression of 15–20k genes, yielding numerical profiles that are computationally demanding and prone to overfitting problems when used as input in machine learning approaches with limited data[41,42].

We thus explored whether our more succinct 128-dimensional vectors were able to retain the information contained within the full GEx profile. Taking the Genomics of Drug Sensitivity in Cancer (GDSC)[40] panel as a reference, we collected, for each cell line, the basal

https://doi.org/10.1038/s41467-022-33026-0

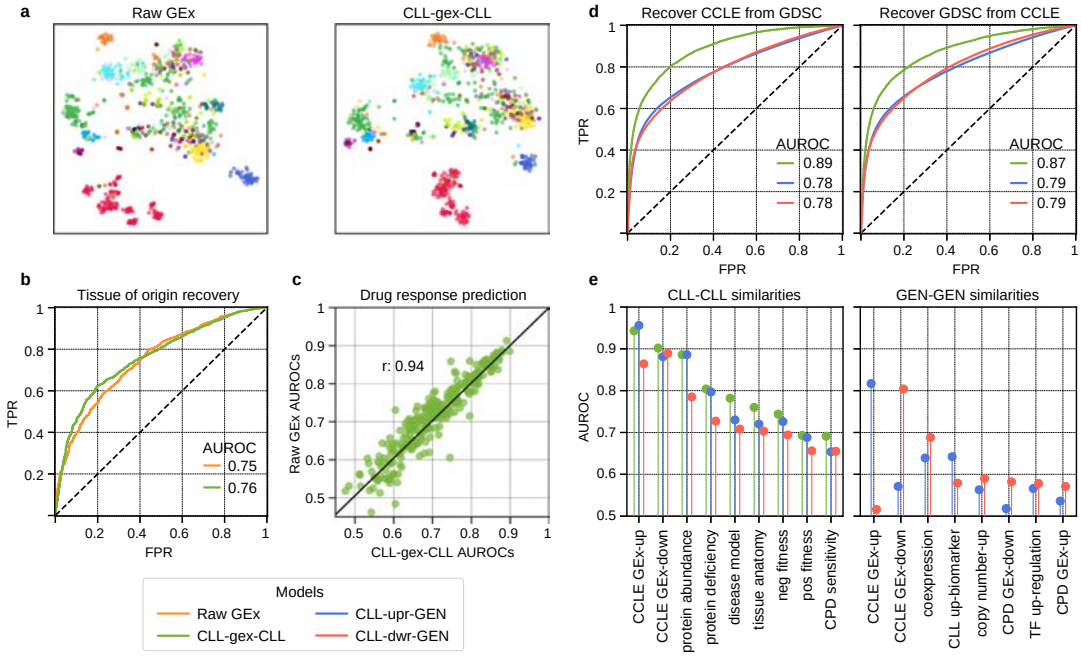


Fig. 5 | Analysis of gene expression (GEx) embeddings. a 2D projection of the raw GEx (left) and the corresponding Bioteque 'cell has similar gex cell' (CLL-gex-CLL) embedding (right). Each dot corresponds to one cell line and is coloured by tissue of origin. **b** Tissue recovery by the raw GEx and the CLL-gex-CLL embedding. **c** Drug response prediction performance (AUROC) for each drug in the GDSC resource from models trained with either the raw GEx (y-axis) or the CLL-gex-CLL embeddings (x-axis). **d** Recovering CCLE (left) and GDSC (right) cell-cell (CLL-CLL) similarities (green), cell-gene (CLL-GEN) upregulation (upr) similarities (blue) and CLL-GEN downregulation (dwr) similarities (red) using embedding distances from the GDSC and the CCLE embedding spaces, respectively. **e** Characterization of the CLL-CLL (left) and GEN–GEN (right) embedding similarities for three metapaths: CLL-gex-CLL (green), CLL-upr-GEN (blue) and CLL-dwr-GEN (red).

(raw) GEx (17.7 K Genes) and the corresponding Bioteque metapath embedding CLL-dwr+upr-GEN-dwr+upr-CLL (hereafter CLL-gex-CLL), aimed at capturing gene expression similarities between cell lines.

We first examined the similarity landscape of the cell lines by performing a 2D projection of the raw and embedded GEx. By colouring the cell lines according to their tissue of origin, we visually verified the capacity of the CLL-gex-CLL embedding to resemble the raw GEx data (Fig. 5a). Indeed, cosine similarities between CLL-gex-CLL vectors up-ranked CLLs sharing tissue of origin with a similar rate as when using correlations between raw GEx vectors (AUROC: 0.75 and 0.76, respectively) (Fig. 5b).

Next, we assessed the capacity of our embeddings to predict the drug response of each cell line. To this end, we trained a standard machine learning model (a random forest classifier) for each of the 262 drugs in the panel and predicted sensitive/resistant responses using the raw GEx and our embeddings independently ("Methods"). Indeed, we found that the capacity of the CLL-gex-CLL embedding to recapitulate drug response is equivalent to that observed when the raw GEx data is used (average AUROC: 0.70 and 0.71, respectively). Moreover, the models based on embeddings had strong concordance with the raw GEx model (0.94 Pearson correlation) (Fig. 5c). This level of agreement is remarkable and represents a clear advantage for the embeddings since they are smaller, easier to handle and do not require expert knowledge to pre-process the raw data. A disadvantage of the embedding approach is the less obvious interpretability of predictions.

After verifying that the Bioteque GEx embeddings retain the basal transcriptional information from the cell lines, we used them to compare profiles obtained from different cell line panels. Specifically, we compared the GDSC with the Cancer Cell Line Encyclopaedia (CCLE)[43]. In agreement with previous reports, we observed a strong correspondence between the two panels, measured as CLL-gex-CLL similarities in the embedding space (AUROC: 0.89) (Fig. 5d). To assess whether these similarities were driven by the up- or downregulation of the same genes, we repeated the analysis focusing on the CLL-upr-GEN and CLL-dwr-GEN embeddings and checked whether the CLL-GEN similarities in the GDSC panel were also preserved in the CCLE. In general, the recovery score of cell line-specific up-/downregulated genes (i.e., CLL-GEN pairs) was lower (AUROC: 0.78) (Fig. 5d). We obtained similar results when we reversed the exercise and used CCLE embeddings to recapitulate GDSC similarities (Supplementary Fig. 3). This finding suggests that, while cell line similarities between panels are robust (i.e., cell lines sharing similar transcriptional signatures in one panel also share similar ones in the other), the specific transcriptional changes of a given cell line may differ. The characterization of the CLL-CLL and GEN–GEN distances further confirmed the better recapitulation of cell line similarity in comparison to gene similarity between panels (AUROC: 0.9 and 0.8 for the CLL-CLL and GEN–GEN similarities, respectively) (Fig. 5e). Furthermore, the CLL-CLL similarity characterization revealed a strong concordance between protein and transcript levels (AUROCs: 0.9 and 0.8 for protein abundance and deficiency, respectively), which was partially driven by the same CLL-GEN pairs (AUROC: 0.72 and 0.63 for the protein abundance and protein deficiency CLL-GEN pairs, respectively) (Supplementary Fig. 3). In addition to tissue of origin, we also observed resemblances between cell lines used to model a given disease (AUROC: 0.78), sharing fitness profiles (AUROC: 0.72 for negative and 0.69 for positive fitness profiles) and similar drug responses (AUROC: 0.7). Finally, the GEN–GEN

similarities also revealed a mild recapitulation of known co-expressed gene pairs (AUROC: 0.64 and 0.69, for the up- and downregulated gene similarities, respectively), thereby suggesting that some of the genes commonly up- or downregulated in the same cell lines from different panels may share the same transcriptional regulatory programmes.

On the whole, our approach retains meaningful information from the original data into a reduced number of dimensions (128 vs ~20k), even when the data comes from a much noisier source such as transcriptomic technologies. We believe that the standardized and dense format of our embeddings provides a by-default way to integrate and compare omics datasets.

Assessing the uniqueness of new omics datasets

Since the consolidation of high-throughput omics technologies, several long-term initiatives have been established to comprehensively characterize certain levels of biological systems (i.e., genetic interactions in yeast[44] or the transcriptomes of cell line panels and human tissues[43,45]). After several years running, all these efforts have had to balance a potential decrease in novelty and an increase in costs as the screens approach saturation. The Bioteque provides a corpus of biological data that is cast to a single format and, as such, it offers a means to quantify the degree of novelty of new data releases of omics experiments. As an illustrative example, we analyse the systematic charting of the Human Reference Interactome (HuRI) with the yeast two-hybrid methodology, which has already identified over 50,000 protein-protein interactions (PPIs) of high quality over the last 15 years[46–48].

To estimate the level of support from different experiments and assess the novelty of the latest HuRI release (HuRI-III[48]), we used the embedding space of relevant metapaths to determine the biological context of each pair of interacting proteins. In brief, for each gene-gene pair, we calculated an empirical *P* value corresponding to the measured similarity in the embedding space, which allowed for commensurate comparison of distance/similarity measures performed in different embedding spaces (see "Methods"). Note that, to have a fair representation of the known physical interactions, we embedded an older version of the protein-interaction network, without including any of the entries from HuRI-III. We then categorized each interaction in HuRI-III into four groups, depending on the level of support contained in the Bioteque embeddings. In this regard, we labelled them as (i) known and supported interactions (covered by GEN-ppi-GEN and at least another metapath), (ii) known interactions (only covered by GEN-ppi-GEN), (iii) supported interactions (covered by other metapaths but not GEN-ppi-GEN) and (iv) potentially novel interactions (with no apparent support in any of the metapaths screened) (Fig. 6a). Remarkably, after three updated versions of HuRI, almost half of the interactions can be classified as potentially novel according to the selected metapaths. Moreover, although only 5825 (11%) of the interactions were supported by GEN-ppi-GEN embeddings, mostly coming from previous versions of HuRI[46,47], our analysis suggests that a higher proportion can be recovered. In fact, at 0.05 FDR ("Methods"), the GEN-ppi-GEN embedding recovered 18% of HuRI-III, retrieving 5456 (94%) of previously known interactions while finding 3994 new pairs (Fig. 6b). On the other hand, we observed a substantial number of physical interactions presumably involved in similar pathways (GEN-ass-PWY), cellular components (GEN-has-CMP), or protein domains (GEN-has-DOM). At 0.05 FDR, these metapaths alone recovered 6905 unique interactions of which 4484 (65%) were not obvious from the physical interaction space (Fig. 6c). To delve into the correlation and relative importance of the metapath for explaining PPIs, we used the *P* values as features for a tree-based machine learning model trained to identify HuRI-III edges. We then assessed the importance of each metapath for the prediction using Shapley values[49]. As visually anticipated from the heatmap, the model achieved a reasonable performance (AUROC: 0.69), mostly relying on previously known physical interactions, cellular components, protein domains, and pathways, all of them showing a certain degree of agreement (Supplementary Fig. 4). Interestingly, we also identified successfully predicted cases with little to no evidence from physical PPIs. For instance, our metapath distance-based model predicted the interaction between the neuronal proteins HOMER1 and SHANK2, the tRNA-splicing endonuclease TSEN54 and the polyribonucleotide CLP1, and the adenosine deaminase ADARB1 and the protein kinase PRKRA, none of which had any reported evidence in protein interaction databases but showed strong positive support in the GEN-ass-PWY, GEN-has-CMP, and GEN-has-DOM metapaths, respectively (Fig. 6d). Indeed, some of these associations have been related in other contexts[50–52], but with no indication of physical interactions before HuRI-III.

We have shown how the continuous and interpretable dimensional space of the Bioteque embeddings provides a powerful framework for characterizing individual observations, which can, in turn, be exploited to guide the interpretation of the entire dataset and, to some extent, assess the novelty of the data.

Discovery of drug repurposing opportunities using the multiple scopes offered by the embeddings

Drug repurposing is often regarded as an attractive opportunity to quickly develop new therapies[53]. However, perhaps with the exception of cancer, where abundant models and molecular data are available, it is difficult to generate data-driven predictors to suggest new uses for approved or investigational drugs, mainly due to the lack of disease descriptors and the small number of known drug-disease indications. Indeed, according to the last update of repoDB, half of the drugs (1097) have only one approved indication, and a third of the diseases (458) are treated with only one drug (Supplementary Fig. 5). Thus, training models with all the known drug-disease associations and later transfer of the insights gained to underexplored treatment areas would be highly desirable[54,55].

To explore whether the Bioteque could be useful in this scenario, we set out to predict new compound-disease indication pairs introduced in repoDB in 2020 (v2) training a model on the previous version (v1), launched in 2017 ("Methods"). We mapped all disease terms to the Disease Ontology, removed redundant indications (according to the ontology), and trained a conventional random forest classifier to predict whether a given CPD-DIS corresponds to a true therapeutic indication. We used two sets of metapath embeddings: one in which we used L1 metapaths (*Short*) based on the drug targets (CPD-int-GEN) and gene associations (DIS-ass-GEN), and another in which we used L3 metapath (*Long)* linking the pharmacological class and the treatment of known CPD and DIS to those sharing drug target (CPD-int-GEN-int-CPD-has-PHC) or gene associations (DIS-ass-GEN-ass-DIS-trt-CPD), respectively. We chose to use drug targets and gene associations because we observed that their embeddings broadly recapitulate the pharmacological class and the disease treatment for a sufficient number of nodes (Supplementary Fig. 5). Moreover, to assess the capacity of the gene-based similarities to correctly infer the treatment, we also tested a metapath (*Long-b*) in which we prevented the CPDs and DISs from being linked, thus making the association with PHC or treatment purely based on the gene-driven similarity to other CPD or DIS. To avoid trivial predictions, we removed associations with PHCs or treatments for drugs and disease unique to the repoDB v2 in all *Long* metapaths. As a basal model, we used chemical fingerprints (ECFP4, 2048 bits) for the CPDs and either one-hot identity vectors (*Basal1*) or binary gene annotations (*Basal2*) for the DISs.

https://doi.org/10.1038/s41467-022-33026-0

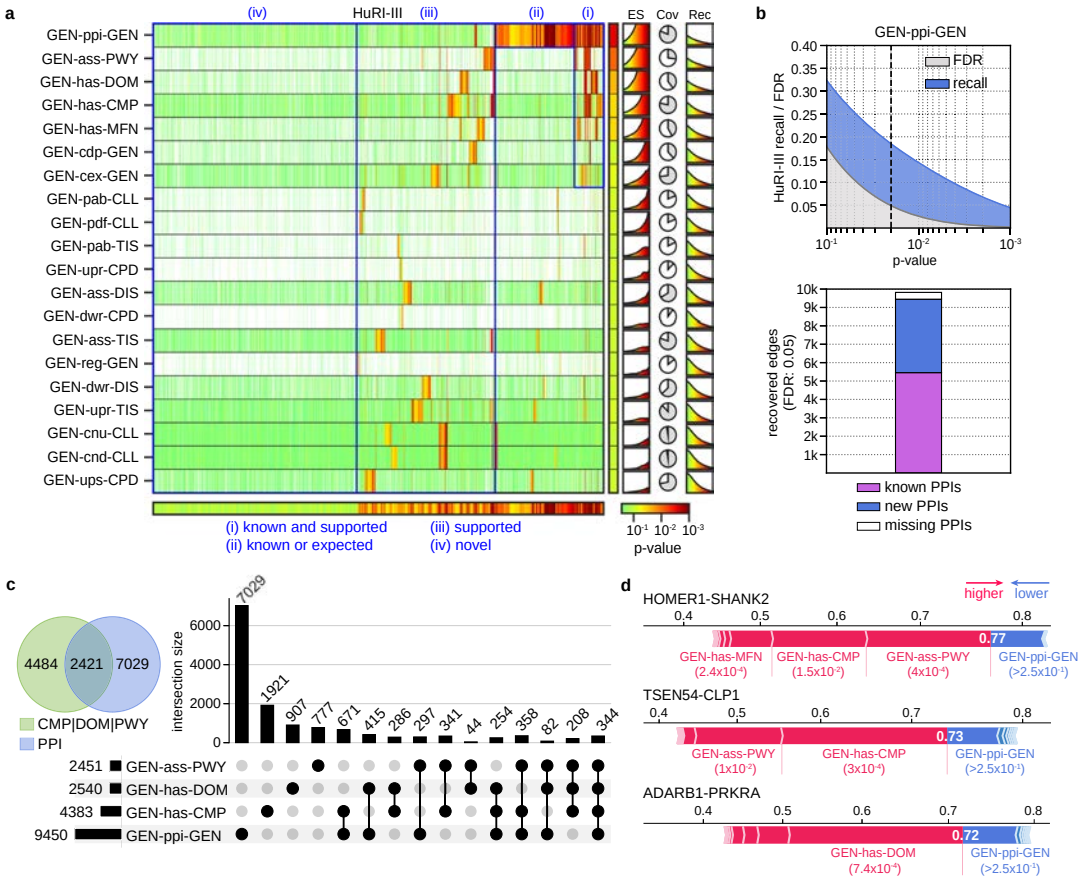


Fig. 6 | Assessing the novelty of the HuRI-III interactome. a Embedding distance *P* values are calculated for each PPI in HuRI-III (x-axis) using the corresponding gene-gene (GEN–GEN) embeddings from a subset of metapaths (y-axis). Please, note that these *P* values do not reflect the significance of any statistical test, but indicate the normalized quantile rank position of a given observation in a background distance distribution ("Methods"). Red tones (lower *P* values) indicate similarity according to a given embedding space. The column and row next to the heatmap show the 10th percentile of the *P* value distribution for each metapath and the lowest *P* value for each edge, respectively. In blue, we grouped edges according to four levels of support. On the right, it is shown the enrichment scores (ES) (capped between 1 and 5 on the y-axis) across *P* values, the coverage (Cov), and the cumulative recall (Rec) across *P* values. **b** (Top) Recovery of HuRI-III edges (recall) and randomly permuted edges (FDR) by 'protein interacts protein' (GEN-ppi-GEN) embeddings across the *P* values (x-axis). The dashed line is placed at the 0.05 FDR (corresponding to a *P* value of 0.02). (Bottom) Number of HuRI-III interactions recovered by the GEN-ppi-GEN embedding at 0.05 FDR stratified by those covered in the original network (known PPIs), those not available in the network, hence, predicted by the embeddings (new PPIs), and those present in the original network but not covered at the given *P* value (missing PPIs). **c** Number of unique HuRI-III edges recovered at 0.05 FDR by the GEN-ppi-GEN and/or the three most supportive metapaths, including 'gene has cellular components' (GEN-has-CMP), 'protein has domain' (GEN-has-DOM), and 'gene associates with pathway' (GEN-ass-PWY). **d** Shapley force plots corresponding to the prediction of three PPIs with no direct evidence of physical interaction before HuRI-III was released. Red segments are metapath-specific *P* values that pushed predictions toward a high probability of interactions, while blue segments pulled predictions towards a low probability. The length of the segments is proportional to their impact on the prediction. The final output probability given by the model is found where both forces equalize (shown in white).

We considered two use cases: a drug repurposing exercise, in which we ranked all the diseases predicted to be potentially treated with a given compound, and a prescription exercise, in which we ranked all compounds that might be useful to treat a given disease. In both scenarios, the three metapath embeddings showed remarkable predictive power compared to the basal models, with the model built from *Long* embeddings being the one with superior performance (Fig. 7a). Specifically, for half the tested compounds, the *Long* embeddings model found a new validated therapeutic purpose within the top 2% of disease predictions (corresponding to the top 10 ranked diseases). Analogously, for roughly 50% of the diseases, the model found a correct treatment within the top 1% of compound predictions (corresponding to the top 8 ranked compounds). Furthermore, although with poorer performance, our biological embeddings were able to yield correct predictions for compounds and diseases with minimal evidence available (i.e., with only one known indication or treatment in repoDB v1) (Fig. 7a, dotted lines). In contrast, the best performing basal model (*Basal2*) found correct predictions for 32% of the compounds and 41% of the diseases within the same ranking range. Moreover, the Bioteque-based models were better at consistently up-ranking indications (or treatments) of compounds (or diseases) with multiple new annotations in repoDB v2 (Fig. 7b). In fact, among our top

https://doi.org/10.1038/s41467-022-33026-0

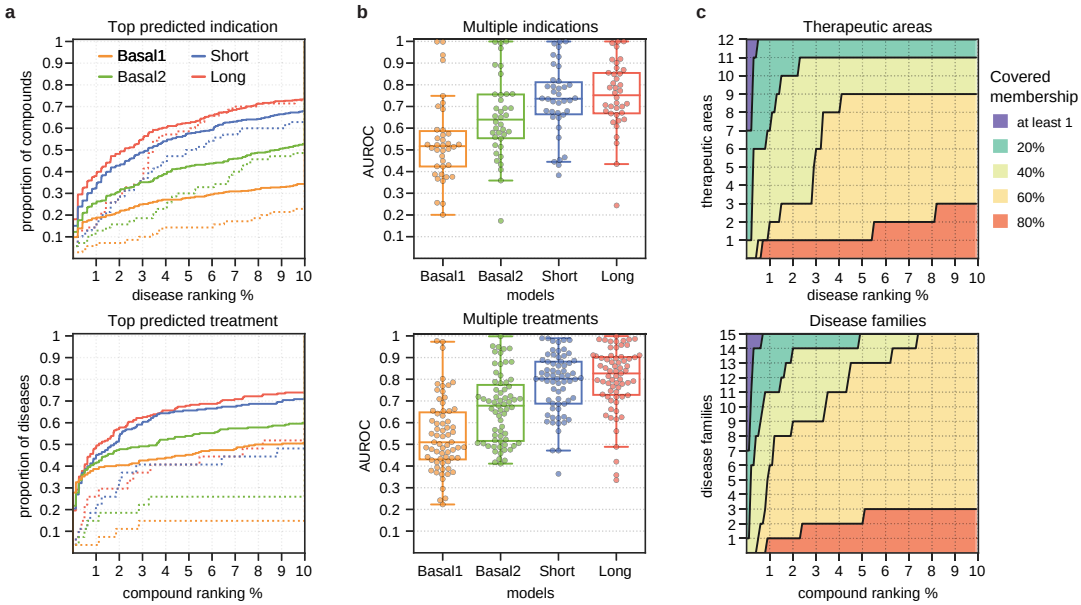


Fig. 7 | Prediction of drug indications and disease treatments from repoDB. **a** Cumulative distribution (y-axis) of compounds (top) and diseases (bottom) according to the ranked position (x-axis) of the top predicted disease indication (top) or compound treatment (bottom) for the four tested models. The rankings are shown in percentages and only for the first 10% of compound/disease predictions (corresponding to the top 50 and 80 diseases and compounds, respectively). Dotted lines show the distribution for those compounds or diseases with only one positive indication in repoDB v1. **b** Classification performance obtained for each compound ($n = 38$, top plot) and disease ($n = 67$, bottom plot) with multiple (≥5) new indications reported in repoDB v2. Box plots indicate median (middle line), 25th, 75th percentile (box), and max and min value within the 1.5*25th and 1.5*75th percentile range (whiskers). **c** Number of different therapeutic areas (top) and disease families (bottom) covered by the predictions of the *Long* model. We considered a given therapeutic area or disease family to be covered when the model predicted one true indication or treatment (as in panel (**a**)) for at least 1%, 20%, 40%, 60%, or 80% of its instances.

predictions, we found repurposing cases that reached clinical trials (Supplementary Fig. 6a). For instance, while both Verapamil and Ranolazine drugs have been approved for the treatment of angina pectoris, our model correctly predicted the repurposing effect of Verapamil in the treatment of ischaemic stroke (clinical trial: NCT02823106) and Ranolazine in the treatment of atrial fibrillation (clinical trial: NCT03162120) in the top 1 and 2 positions, respectively (Supplementary Fig. 6b). Interestingly, our model highlights hyperinsulinemia as the top repurposing for Ranolazine. While this link is not included in repoDB, we have found diverse studies supporting the correlation of Ranolazine with insulin levels[56–58]. Finally, we verified that these predictions covered a broad range of therapeutic areas and disease families. Indeed, we found that within the top 1% of predictions, the *Long* model successfully predicted one indication or treatment for 20% of all the compounds and diseases in each therapeutic area or disease family (Fig. 7c and Supplementary Fig. 6e). These results were reproduced with the *Long-b* model, showing that, as expected, the genes associated with drugs or diseases of known treatment can indeed be used to better infer the activity of drugs and diseases with unknown indication (Supplementary Fig. 6c, d).

Overall, we showed how Bioteque embeddings can be directly plugged into machine learning models, and how, by combining different context associations into larger metapaths, they can increase the performance of drug-disease prediction models. Indeed, we used a preliminary version of Bioteque embeddings to successfully identify potential targets for a set of kinase inhibitors from perturbational profiles, including drug-induced transcriptional changes and cell sensitivity data, in several cell lines[59].

The Bioteque resource

We built an online resource to facilitate access to all the pre-calculated Bioteque embeddings (https://bioteque.irbbarcelona.org). The Bioteque web offers a visual way to explore over one thousand metapaths by selecting the nodes to connect, as well as the type of relationship between them. For a selected metapath, we provide an analytical card displaying a 2D representation of the embedding, a ROC curve assessing the preservation of the original network, distance distributions of the embedding space, and biological associations that are best recapitulated by the metapath of interest.

Furthermore, the web page also offers a section were metapath embeddings and other metadata can be downloaded. The generated file contains the embeddings for each node, the nearest neighbours of each node in the space, and the analytical card displayed on the web. Additionally, we make available executable notebooks showing how to download our embedding resource programmatically as well as how to perform most of the downstream analyses presented throughout this manuscript. More specifically, we illustrate how to (i) generate 2D (interactive) visualizations that can be coloured and annotated according to side information (e.g., colour cell lines by tissue of origin), (ii) identify similar nodes (close neighbours) for a given entity of interest, (iii) cluster the embedding space and (iv) build a predictor model trained on our embeddings.

The Bioteque web also provides information on the specific sources used to construct each metapath, and some general statistics on the contents of the current version of this web resource. We also provide a link to our GitLab repository, which contain the full code necessary to pre-process the data to generate and analyse biological embeddings (http://gitlabsbnb.irbbarcelona.org/bioteque). The entire

resource, including the underlying data and biological embeddings, will be updated once per year, or as soon as a major dataset is released.

Discussion

With the accumulation of large-scale molecular and cell biology datasets, coming from ever-growing literature, omics experiments and high-throughput screenings, new frameworks for integrative data analysis are necessary. For a given biological entity (e.g., a gene), we are now able to stack multiple layers of its biological complexity (e.g., its structure, function, regulation, or interactions), which offers an opportunity for a more complete, systemic view of biological phenomena, but brings along several challenges, including the handling of different data structures, nomenclatures, signal strengths, and variable dimensionalities.

To tackle these challenges, we have developed the Bioteque, a resource of pre-calculated, fixed-format vector embeddings built from a comprehensive biomedical knowledge graph (KG). The KG contains physical entities like genes, cell lines, and compounds, as well as concepts like pathways, molecular functions, and pharmacological classes. Embeddings capture the connections between nodes in the KG according to a certain metapath, i.e., a sequence of semantic and/or mechanistic relationships between entities. We have shown how this approach is useful to (i) produce compact descriptors that broadly preserve the original data, (ii) systematically characterize biological datasets such as cancer cell line transcriptional signatures, (iii) assess the novelty of a given omics experiment and (iv) mine for drug repurposing opportunities based on multiple associations between drugs and diseases.

In the Bioteque, we have incorporated datasets from over 150 distinct sources, keeping the integrity of the original data to a feasible extent and applying standard transformations when required. Note that the accuracy of the Bioteque is determined by the quality of the source data. As experimental technologies continue to evolve, new information will populate these databases and novel standards will emerge, opening the door for more comprehensive and higher quality embeddings. In addition, as a first attempt, we used a network embedding technique that purely relies on the graph topology built from the biomedical data, in contrast to other techniques that also leverage node and edge attributes (e.g., Graph Neural Networks, GNN). While these methods may contribute to improving the embedding space, their quality depends on the availability of enough data and meaningful node features, while requiring a thorough fine-tuning of the hyperparameters[60,61]. Taken together, the proper implementation of these methods becomes unfeasible for the systematic embedding of thousands of networks. Additionally, the incorporation of external node features in the network could compromise the controlled identity of the metapaths. Nevertheless, Bioteque descriptors can be easily recycled as node features for new task-specific networks, thus transferring the learning encoded from orthogonal biomedical datasets to more complex, attribute-aware models. Finally, we would like to point out that there are parts of the current biomedical knowledge that have not yet been included in the resource, such as antibody-target interactions and metabolomics. As a molecular/cell-centric resource, the Bioteque also lacks patient-derived data[13], including interactions with the microbiome[62]. Updated versions of the Bioteque will have to be complemented with the incorporation of other fields of biological knowledge, the re-accommodation of the datasets in the resource (based on updated standards), and the improvement of embedding strategies to account for side-features of the nodes or incorporate unseen (external) nodes in the embedding space. Moreover, future developments will explore the adoption of biological descriptors as features for a variety of downstream-specific tasks, including a systematic screening of the biological support of wet lab experiments or the modelling of complex diseases to guide the generation of new chemical entities to tackle them[20].

Methods

Building the metagraph

All gathered data was stored in a graph database (KG) in which nodes represent biological or chemical entities and edges represent associations between them.

Nodes (entities). The nodes in the graph can belong to one of 12 types (aka metanodes). For each entity type, we predefined a universe of nodes and chose a reference vocabulary based on standard terminologies. These 12 entity types are (in alphabetical order):

Chemical entities (CHE). Chemistry terminologies extracted from the Chemical Entities of Biological Interest (ChEBI) ontology[63].

Cells (CLL). Cell lines used in biomedical research and extracted from the Cellosaurus resource[64].

Cellular Components (CMP). Biomolecular structures and complexes as defined by the Gene Ontology[65] (extracted from the basic filtered ontology).

Compounds (CPD). Small molecules codified with the standard InChIKey. As we do not use any predefined library of compounds, the universe will be determined by the union of compounds included in other datasets (e.g., drug–target interactions).

Diseases (DIS). Abnormal conditions, drug side effects and symptoms. We used the Disease Ontology[66] as a reference vocabulary.

Domains (DOM). Functional and structural protein domains extracted from InterPro[67].

Genes and proteins (GEN). Genes and proteins were unified and stored by Uniprot[68] accession code (UniProtAC). We worked on the reviewed Human proteome.

Molecular functions (MFN). Biological function of the proteins defined by the basic Gene Ontology[65].

Perturbagens (PGN). CRISPR, overexpression, and shRNA perturbations. Note that PGNs are always mapped to the corresponding perturbed gene when constructing the metapath. Therefore, instead of providing PGN labels, we provide the UniProtAC of the perturbed genes.

Pharmacologic classes (PHC). Pharmacologic classes defined by the Anatomical Therapeutic Chemical (ATC) code (http://www.whocc.no).

Pathways (PWY). Biological pathways and processes. We used Reactome[69] as a reference vocabulary.

Tissues (TIS). Anatomical tissues and cell types defined by the BRENDA Tissue Ontology[70].

Please note that in the datasets containing ontological terms (CMP, DIS, MFN and PWY), we removed the least informative terms (i.e., those that are higher up in the ontology). These terms were identified by calculating the information content[71]. The node universe for each entity and the list of removed terms are available in Supplementary Data 1.

Vocabulary mapping. To integrate terminologies, we extracted curated cross-references from the official terminology sources and associated ontologies. As the nomenclatures used to identify diseases and pathways were particularly diverse and rarely cross-referenced, we further increased the mapping of these terms by inferring similarities within concepts as detailed below.

Diseases were mapped by calculating disease term similarities through shared cross-references to the Unified Medical Language System (UMLS), obtained from the DisGeNET mapping resources (https://www.disgenet.org/downloads). Specifically, we encoded each disease term into a binary vector spanning the universe of UMLS terms of all nomenclatures. We then transformed the binary vectors with the corresponding term frequency-inverse document frequency (TF-IDF) values and computed pairwise cosine distances between the Disease Ontology and the rest of the vocabularies. Using the similarities obtained from curated cross-references as reference, we found a cosine similarity cutoff of 0.5 to correspond to an empirical P value of 5×10^{-4}.

Pathway cross-references were extracted from the ComPath resource[72] and extended following the PathCards[73] approach. This approach first clusters the pathways into SuperPaths based on overlapping genes and then uses Jaccard similarities between the SuperPaths genes to define pathway similarity. We used the same parameters described in the PathCards paper (0.9 for the overlap cutoff, 20 minimum genes in the pathways, and a Jaccard similarity of at least 0.7).

Edges (associations). Edges in the graph are used to link biological and/or chemical entities. Since two entities may be connected by multiple edge types (i.e., 'compound treats disease' or 'compound causes disease'), we define the associations as triplets (metapaths) of entity-relationship-entity (CPD-trt-DIS, CPD-cau-DIS).

Homogeneous associations are those concerning entities (metanodes) of the same type (e.g., 'gene is co-expressed with gene', GEN-cex-GEN), while heterogeneous associations are related to entities of different types (e.g., 'tissue has cell', TIS-has-CLL). Note that we annotated only one direction of the heterogeneous associations (in fact, we kept CLL-has-TIS instead of TIS-has-CLL), although both directions are valid when defining metapaths. On the other hand, edges were treated as directional whenever a homogeneous association had only one valid directionality, like in the case of kinase-substrate interactions ('gene phosphorylates gene', GEN-pho-GEN) or transcription factor regulations ('gene regulates gene', GEN-reg-GEN). Finally, edges corresponding to similarity measures required a predefined set of nodes for pairwise comparison, and they were computed only after the rest of the graph was populated.

Populating the knowledge graph with data

For each type of association or metaedge, we can have one or more datasets (Supplementary Data 2). Datasets are not merged but kept as individual sources so that they can be embedded individually or in combination within a given metapath. The dataset processing pipeline consisted of two steps. In the first step, nomenclatures were standardized and cutoffs were applied. In the second, applied only to ontological data, terminologies were mapped and the network was pruned.

Dataset standardization. We processed each dataset individually in order to handle the diversity of formats and data types. The guiding principles of data processing were those defined by the Harmonizome[7].

Datasets that already provided binary data were integrated naturally by converting them into the network format of the KG. If the database provided a measure of confidence (e.g., edge weights or P values), we applied default cutoffs (if given) and/or followed author recommendations in order to remove spurious interactions. To build the network, we did not use any edge weight coming from the original source during the embedding process. This was motivated by the observation that most of these weights are based on a measure of support or confidence, which does not necessarily reflect biological significance/strength. Instead, these scores usually capture biases on the knowledge annotation (e.g., associations for under-studied diseases will be less covered among the different sources and, therefore, are prone to have lower confidence scores) or detectability limitations of the experimental screening (e.g., the abundance level of some proteins are more difficult to detect than others). While weighted edges could provide valuable information for the embedding, we could not find a general way to treat them across the diverse and heterogeneous associations in our resource.

Occasionally, the same dataset can be further divided into different subsets on the basis of a given categorical variable (e.g., curated/inferred). We kept these subsets as independent datasets when applicable. For instance, there is a curated version of DisGeNET and an inferred version of it.

Continuous data requires the application of a cutoff before its integration in the KG. Below, we detail how these cutoffs were chosen depending on the nature of the data.

Transcriptomics and proteomics data. We adapted the strategy followed by Harmonizome, which is based on traditional statistical treatment of gene expression profiles. More specifically, we first mapped the samples and genes to our reference vocabulary and collapsed the duplicates by their mean value. A log2 transformation was then applied followed by a quantile normalization of the genes (unless the dataset was already transformed by the data providers). Next, we subtracted the median and scaled the data according to the quantile range of each gene. Finally, the top 250 most positive and negative genes were selected for each sample and kept in the corresponding metaedges (e.g., CLL-upr-GEN and CLL-dwr-GEN).

Drug sensitivity. To binarize drug sensitivity data, we used the waterfall method first described by Barretina et al.[74], and used since then in different subsequent works (e.g.,[75–77]). This method ranks cell lines on the basis of a drug response measure, for instance, the area under the growth inhibition curve (AUC), and uses the shape of the plot to define a sensitivity threshold. The waterfall method was applied for each compound in the dataset, keeping at least 1% but no more than 20% of sensitive cell lines and requiring an AUC sensitivity value lower than 0.9.

Perturbation experiments. Gene perturbation data required a preliminary step to differentiate the type of perturbation (e.g., 'CRISPR modification silences gene A') from its outcome (e.g., 'silencing gene A results in overexpression of gene B'). First, for each perturbation in the dataset, we created a perturbagen (PGN) node with a unique identifier. We then simplified the two-step relationship (e.g., 'perturbagen that silences gene A upregulates gene B') into a 'perturbagen upregulates gene B' association (PGN-upr-GEN).

Other datasets. For some datasets containing continuous data, we had to apply customized approaches to convert them into a network format. Details about the pre-processing of each particular dataset are provided in Supplementary Data 2, while the corresponding Python scripts can be found on https://bioteque.irbbarcelona.org/sources.

Terminologies and pruning. Six terminologies (namely, CMP, DOM, MFN, PHC and PWY) had semantic relationships between them. In these cases, we propagated all the reported relationships with other terms (e.g., GEN) through the parents of their corresponding

ontologies. To maximize coverage, propagation was done before cross-referencing.

Selection of metapaths

We chose a controlled set of metapaths for which we pre-computed embeddings. These are the embeddings that are deposited in the Bioteque resource. The metapaths were selected as follows.

Length 1 (L1). All possible metapaths of length 1 are embedded except for those capturing cross-references (DIS-xrf-DIS), ontologies (PWY-hsp-PWY), compound-compound similarities (CPD-sim-CPD), and PGN associations. Note that PGN nodes are mapped to the corresponding perturbed genes through the PGN-pdw-GEN or PGN-pup-GEN metapaths (thus, >L1 metapaths).

Length 2 (L2). Only the mimicking (e.g., CLL-dwr+upr-GEN-dwr+upr-CLL) or reversion (CLL-upr+dwr-GEN-dwr+upr-CLL) of both directions (up/down) are used for metapaths connecting entities through transcriptomic, proteomic or transcription factor signatures. CLL and TIS are always connected through the CLL-has-TIS association. Finally, only the following associations are allowed when linking cells and genes within a metapath: CLL-upr-GEN, CLL-dwr-GEN, CLL-mut-GEN.

Length 3 (L3). L3 metapaths are constructed by linking L1 metapaths with any of the following L2 metapaths: CLL-dwr+upr-GEN-dwr+upr-CLL; CLL-has-TIS-has-CLL; CMP-has-GEN-has-CMP; CPD-has-PHC-has-CPD; CPD-int-GEN-int-CPD; DIS-ass-GEN-ass-DIS; DOM-has-GEN-has-DOM; MFN-has-GEN-has-MFN; TIS-dwr+upr-GEN-dwr+upr-TIS; or PWY-ass-GEN-ass-PWY. GENs from the PGN-pup-GEN or PGN-pdw-GEN are linked through heterogeneous or directed homogeneous associations but not through undirected homogeneous associations.

Length > 3 (>L3). Generated when mapping the source or target PGN to the perturbed genes in L3 metapaths.

In the case of directed homogeneous associations, we used the '_' mark next to the entity that acted as the source of the association. For instance, GEN-_pho-GEN-ass-PWY links the kinases to the pathways associated with their substrates while GEN-pho_-GEN-ass-PWY links the substrates with the pathways associated with their kinases.

Finally, metapaths whose embedding did not preserve the original network or that failed to keep most of the nodes in a single connected component were removed as described in the following section. The entire list of the embedded metapaths is provided in Supplementary Data 3.

Obtaining Bioteque embeddings

To obtain the embeddings we used the node2vec algorithm[31], a well-accepted approach based on random walk trajectories[78], in which metapaths are used as single networks and fed to the node2vec algorithm. We acknowledge that there are embedding methods that allow a direct embedding of the network from metapath walks (e.g., metapath2vec[79]). However, we decided to first pre-compute the source-target networks using the DWPC method, since the resulting network already weighs those source-target associations that are more strongly connected according to the metapath, thus requiring fewer random walker steps to learn the relationship between the source and target nodes. Moreover, this pre-computed network encourages the embedding model to only focus on source-target relations, giving us more control about what information we are encoding in the embedding space while allowing an easier generalization of the model's hyperparameters across different metapaths lengths (i.e., the source and target nodes are always one-hop apart regardless of the metapath length). Notice that, since all our metapath networks are either homogeneous or bipartite, the default skip-gram implementation of metapath2vec is equivalent to node2vec.

Homogeneous and bipartite networks. L1 metapaths already correspond to homogenous or bipartite networks. For >L1 metapaths, the source and target nodes were connected by computing degree-weighted path counts (DWPC)[11] through the corresponding datasets and associations in the metapath. To this end, we sorted the datasets according to the associations of the metapath, represented them as adjacency matrices and kept the same source (rows) and target (columns) node universe as the target and source nodes of the previous and following datasets, respectively. Following the DWPC method, we first downweighted the degree of the nodes in each of the datasets by raising the degrees to the −0.5 power. We then calculated the DWPC values by concatenating the matrix multiplication from the source to the target dataset. As a result, we obtained a new $n \times m$ matrix where n are the source nodes of the first dataset and m are the target nodes of the last dataset. The values of the matrix are the DWPC between the source and target nodes, which are used as weights during the random walker exploration. Finally, we limited the number of edges for each node to 5% of the total possible neighbours (with a minimum of 3 and maximum of 250 edges per node).

Occasionally, we used more than one dataset within the same association or we combined two metapaths into one. This is a common case for >L1 metapaths with transcriptomic signatures where the two directions (CLL-upr-GEN and CLL-dwr-GEN) are often combined (CLL-dwr+upr-GEN-dwr+upr-CPD). To handle these cases, we first obtained an individual network for each metapath or dataset following the approach detailed above. We then merged all the networks by taking the union of the edges (L1 metapaths) or adding the DWPC values (>L1 metapaths).

At the end of the process, we removed network components that cover less than 5% of the entities from the network. And we also removed from the source metapaths that fail to retain 50% of the total nodes within their network components.

Node2vec parameters. The node2vec algorithm consists of a random walk-driven exploration of the network followed by a feature vector learning through a skip-gram neural network architecture.

We implemented a custom random walker (with the node2vec parameters p and q set to 1) and ran 100 walks of length 100 for each node of the network. For >L1 metapaths, we scaled the DWPC values for each node to sum 1 and used them as probabilities to bias the random walker. We used the C++ skip-gram implementation provided by Dong et al.[79] with default parameters to obtain a 128-dimensional vector for each node.

Accounting for node degree biases

The uneven distribution of information across the different knowledge domains and data sources incorporated in our KG inevitably leads to an uneven number of associations across entities, introducing a bias towards nodes with higher degrees. We implemented several measures to mitigate these biases, not only during the generation of the embeddings, but also in the way distances are calculated.

Before generating the embedding. To control the degree of the metapath networks, we implemented the DWPC method (as described in the previous section), which was specifically developed to account for degree biases. Furthermore, we also limited the number of connections a given node can have at the end of the metapath to 5% of the total possible neighbours (with a minimum of 3 and maximum of 250 edges per node). This was implemented since we observed that nodes in longer metapaths often find at least one spurious path to connect to every other node in the network. Although most of them end up having very low weights, the resulting network is very dense, requiring a much larger number of random-walks for the skip-gram model to learn the weight distribution of the network. All these cutoffs were chosen based on the thought exploration made by Himmelstein et al. and after

optimizing for different metapaths in our resource. Importantly, the effect of controlling the degree of the network was fundamental for having embedding spaces of good quality, especially for longer metapaths where these biases get exacerbated due to the combination of high-degree nodes from different datasets (Supplementary Fig. 7).

Additionally, we removed from the KG those nodes whose meaning was too general according to the information content provided in the ontology. This prevented those nodes to attract many connections in the network at the cost of providing very little information (e.g., disease terms such as 'cancer', 'syndrome' or 'genetic disease'; or cell compartments terms such as 'cell', 'membrane' or 'cell periphery'). All the pruned terms are provided in Supplementary Data 1.

After generating the embedding. Most downstream analyses rely on distances between the embeddings. However, even if we have implemented measures to control the degree of the network when producing the embedding, it is expected that nodes having more general implications will be generally closer to the rest than others that are more specific (e.g.'Brain disease' (https://disease-ontology.org/term/DOID:936) will be closer to a much broad set of genes than 'Migraine' (https://disease-ontology.org/term/DOID:6364) which is a specific condition comprised within the family of Brain diseases). Therefore, some terms may be biased to have a closer distance distribution than others just because their edges define broader associations. Although encoding this can be useful in some downstream analysis (e.g., identifying drugs that target proteins specifically associated with particular brain diseases) it also may introduce biases when comparing distance distributions between terms (Supplementary Fig. 7).

To address these biases, we first assessed how different distances differentiate between these terms, finding that cosine distances provided more comparable distributions between terms while still preserving the (expected) enrichment of small distance associations of broader terms. Moreover, in order to add a measure of specificity in the distance, we also opted to compute co-ranks quantiles, which requires both nodes to be close to each other in order to consider they are sharing a close relationship (this was used in the HuRI-III exercise and the procedure is detailed in the corresponding section). By doing that, we can normalize the distance values of all entities, making them comparable (e.g., having a 0.1 co-rank quantile means the same regardless of the disease node).

Additionally, network permutations can be used in downstream analysis to control spurious observations made in networks that are being analysed with our embeddings. In fact, in the HuRI-III analysis, we randomly permuted the HuRI-III network (as detailed in the corresponding section) and used the permuted network as a reference to derive statistical significance cutoffs for the embedding distances we calculated.

Embedding evaluation

We used opt-SNE to generate the 2D representation of the embeddings[80]. To assess the quality of the embeddings, we reassembled the network obtained from the metapath using the embedding vectors. To this end, we first computed the cosine distance of each edge in the network using the embedding vectors of the nodes. Next, we generated 100 random permutations for each edge in the network and calculated the cosine distances between them. Finally, we sorted all the distances and computed the area under the ROC curve (AUROC) using the network edges and the random permutations as the positive and negative sets, respectively. When assessing >L1 metapaths, we repeated the same exercise using 3 extra network subsets obtained by keeping, for each node, the top 1%, 25% and 50% closest neighbours according to the DWPC weights of their edges. Embeddings with an AUROC below 0.8 were removed from the resource.

Embedding characterization

To characterize the embeddings, we first preselected a collection of reference networks representing commonly used biological associations. Then, given a set of embeddings corresponding to a certain metapath, we tested their capacity to recapitulate edges from other (orthogonal) datasets (i.e., the reference networks). Two measures were kept, the coverage (i.e., the number of overlapping nodes) and the AUROC, following the approach described above.

Aiming to extend this characterization, for each metapath we sought to characterize nodes separately, based on their entity type. We first calculated the term frequency-inverse document frequency (TF-IDF) values of the nodes from each reference network in our collection. Next, within the same entity type and network, we used the TF-IDF-transformed vectors to compute pairwise cosine similarities between nodes. Finally, we built the entity similarity network by keeping the top 5 closest neighbours for each node. Note that from one heterogeneous (bipartite) network this process yields two homogeneous networks, one for each entity type.

Some of the networks in our collection required customized preprocessing. To represent perturbation associations, we directly linked the perturbed genes (PGN-pup-GEN or PGN-pdw-GEN) and the outcome of such perturbation (e.g., PGN-bfn-CLL or PGN-upr-GEN) through the corresponding associations and datasets. We computed the CHE-has-CPD similarity networks by directly linking each node with the top 3 partners that shared more neighbours. Additionally, some entity similarity networks were gathered from other sources, like the CPD-CPD mechanism of action similarity obtained from our Chemical Checker resource[81].

Embedding-based gene expression analysis of cancer cell lines

We downloaded the RMA-normalized gene expression (GEx) and the drug sensitivity data from the GDSC1000[40] web resource (https://www.cancerrxgene.org). We mapped the cell lines and genes to our reference vocabularies and took the mean value whenever duplicates occurred. We used the tissue of origin annotations from the CLUE cell app (https://clue.io/cell-app), which were already part of our graph (CLL-has-TIS, cl_tissue_clueio). Regarding CCLE data, we used the next-generation data[43] from the Broad Institute Portal (https://portals.broadinstitute.org/ccle/about). We processed the RNAseq data and produced three embeddings (CLL-upr-GEN, CLL-dwr-GEN and CLL-dwr+upr-GEN-dwr+upr-CLL) following the pipeline detailed in the "Dataset standardization" and "Obtaining the embeddings" sections.

In the drug sensitivity prediction exercise, we trained a random forest (RF) classifier for each drug and each GEx input data (i.e., the raw GEx or any of the GEx-derived embeddings). After removing drugs with less than 10 sensitive or resistant cell lines, we modelled 262 drugs. We used the SciKit Learn implementation of the RF algorithm, with a 10-fold stratified cross-validation scheme, and optimized RF hyperparameters over 20 iterations of Hyperopt[82].

Analysis of the HuRI-III protein-protein interaction network

We downloaded HuRI-III from the Human interactome atlas (http://www.interactome-atlas.org/). Next, we considered all L1 metapaths containing a GEN metanode, keeping the dataset with higher coverage for each metapath and discarding those covering less than 10% of the HuRI-III network. As a representative of PPI interactions (GEN-ppi-GEN), we used a version of IntAct dated December 2019 (before publication of the HuRI-III network) from which we removed all entries belonging to the HuRI-III screening (IMEX: IM-25472). Next, we calculated the cosine distance between each PPI in each of the metapath embedding spaces and ranked the distances according to the distance distribution of each of the proteins. Distances and rankings were obtained with FAISS[83]. To derive empirical *P* values, we transformed the rankings into percentiles by normalizing them by the total number of

covered genes in each metapath and kept the geometric mean of the normalized co-ranked pairs.

In parallel, we generated 1000 random permutations of HuRI-III by randomly swapping each of the HuRI-III edges 10 times using the BiRewire bioconductor package (https://doi.org/10.18129/B9.bioc.BiRewire) and, likewise, calculated *P* values for each metapath. For each permuted network, we calculated the recovery of the edges with a sliding *P* value cutoff (between 1 and 0.001) and averaged the counts at each cutoff. After repeating this process with the HuRI-III network, we were able to derive, for each metapath, the expected fold change (FC) across different *P* value cutoffs (i.e., the number of covered HuRI-III edges at a given *P* value cutoff divided by the average number of covered edges in the permuted networks). Moreover, the permuted networks were also used to estimate an empirical FDR for a given *P* value. For instance, for each metapath, we found the *P* value cutoff associated with a 0.05 FDR by calculating the minimum *P* value needed to cover no more than 5% of the permuted network edges. Finally, to build the matrix shown in Fig. 6a, we selected the top 20 metapaths with the highest FC (i.e., FC average in the *P* value range between 0.1 and 0.001), and used their *P* values to cluster the PPIs with the fastcluster package[84] and the ward distance update formula.

To obtain the Shapley values, we trained a XGBoost model to classify GEN–GEN edges as positive (i.e., present in HuRI-III) or negative (i.e., not present in HuRI-III) using the *P* values across metapaths as features. To sample negative pairs, we used the instance of the permuted networks hitting fewer HuRI-III edges (~3%) in order to avoid having the same edge as positive and negative instance at the same time. Furthermore, since the objective of this exercise was to study the interplay between the metapaths, we removed edges that were covered by less than 10 (50%) metapaths, resulting in a dataset of 60k positive and negative pairs. A simple mean imputation was applied to the missing *P* values. At training time, we implemented a 20-fold stratified cross-validation split scheme and fine-tuned the hyperparameters using 20 iterations of Hyperopt[82]. Finally, we obtained the Shapley values from the test splits by implementing the TreeExplainer method[49]. All subsequent analyses and figures were obtained using the SHAP package (https://github.com/slundberg/shap).

Drug repurposing based on drug and disease embeddings

The first release of the repoDB (v1) data was downloaded from http://apps.chiragjpgroup.org/repoDB while the updated release (v2) was obtained from https://unmtid-shinyapps.net/shiny/repodb. Compounds were mapped to InChIKeys and diseases to the Disease Ontology (DO) forcing a 1:1 mapping. As features, we used the following metapaths (datasets) from the Bioteque resource: CPD-int-GEN (curated_targets); DIS-ass-GEN (disgenet_curated+disgenet_inferred); CPD-int-GEN-int-CPD-has-PHC (curated_targets-curated_targets-atc_-drugs); and DIS-ass-GEN-ass-DIS-trt-CPD (disgenet_curated+disgenet_inferred-disgenet_curated+disgenet_inferred-repodb).

Additionally, we obtained the 2048-bit Morgan fingerprints (ECDF4) of the compounds using RDKIT (http://rdkit.org) and used the adjacency matrix of the disease-gene network from DisGeNET as binary descriptors of diseases. Having defined the features of the model, we filtered out those drugs and diseases from repoDB that fell outside the embedding universe and removed redundant pairs by de-propagating the associations to the most specific drug-disease terms according to the Disease Ontology. As a result, the train (repoDB v1) and test (repoDB v2) splits consisted of 2522 and 1187 unique drug-disease associations, respectively (Supplementary Fig. 5). Additionally, to prevent the model from focusing on the most frequently annotated drug and disease entities, we further processed the train data to balance the number of associations (degree of the nodes). More specifically, we capped the number of drug or disease associations to 5% of all possible associations (44 diseases and 26 drugs, respectively). Therefore, the associations of those drugs or diseases exceeding this limit were subsampled by performing a K-means clustering (where K was set to the capping limit) using the CPD-int-GEN or DIS-ass-GEN embeddings as features, and by randomly selecting a representative association from each of the clusters (Supplementary Fig. 5). This step slightly decreased the number of training data to 2326 drug-disease associations.

Next, we produced train negative pairs by aggregating 20 negative networks obtained by randomly swapping the edges of the training data (thus, forcing a ratio of 1:20 between the positive and negative instances), while preventing inconsistencies in the Disease Ontology (i.e., having a negative association that would be obtained by propagating a positive drug-disease association through the ontology). Note that, to comply with the time-split scenario, we did not remove any negative drug-disease pair reported to be positive in the repoDB v2 release.

Once the training data was ready, we ran an RF classifier for each of the explored models using 20 iterations of Hyperopt[82] to fine-tune the hyperparameters. At prediction time, drug-disease associations in repoDB v2 were considered positive test pairs, whereas all the remaining drug-disease pairwise combinations were considered negative pairs. To avoid inconsistencies, we removed those negative pairs that were semantically related to positive pairs according to the Disease Ontology. As a result, we obtained between [460–500] diseases and [750–800] drug predictions for each drug and disease, respectively. As most of the drugs and diseases only had one or two positive instances, we assessed the performance of the models by ranking all the predictions individually for each entity (ranks were used as percentages). Additionally, we calculated ROC curves for those drugs and diseases that had at least 5 positive instances. Finally, we obtained the pharmacological action of the drugs by mapping them to the uppermost level of the Anatomical Therapeutic Chemical (ATC) classification, when available. Likewise, disease families were derived by propagating the disease terms to the first and second levels of the Disease Ontology.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All the embeddings generated in this study are available for direct download from https://bioteque.irbbarcelona.org/downloads. The raw networks that were embedded are provided in the same downloadable file for metapaths of length ≥ 2. To comply with the wide variety of licences associated with the data owners, raw networks for L1 metapaths are not provided. Instead, instructions and code to download and pre-process the data are made available at https://gitlabsbnb.irbbarcelona.org/bioteque/. Accessible links to all the datasets embedded in the Bioteque resource are listed on https://bioteque.irbbarcelona.org/sources. RMA-normalized expression data of the GDSC cell lines was downloaded from https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html. CCLE RNAseq data was downloaded from https://sites.broadinstitute.org/ccle/datasets. Cell line tissue of origin annotations were obtained from clue.io (https://clue.io/cell-app). The HuRI-III network was downloaded from http://www.interactome-atlas.org/download. The first release (v1) of repoDB indications was downloaded from http://apps.chiragjpgroup.org/repoDB/. The second release (v2) of repoDB indications was downloaded from https://unmtid-shinyapps.net/shiny/repodb/. ATC codes were obtained from Drugbank (https://go.drugbank.com/

releases/latest#full), Drugcentral (https://drugcentral.org/download) and KEGG (https://www.genome.jp/kegg-bin/get_htext?br08303+D00731). Curated gene-disease associations were downloaded from DisGeNET (https://www.disgenet.org/downloads).

Code availability

The code used to generate the embedding resource is available at https://gitlabsbnb.irbbarcelona.org/bioteque/. Individual scripts used to download, pre-process and integrate the embedded datasets into the knowledge graph can be obtained from https://bioteque.irbbarcelona.org/sources. Jupyter notebooks exemplifying how to programmatically download embeddings from the Bioteque resource and how to run the downstream tasks illustrated in this manuscript can be downloaded from https://bioteque.irbbarcelona.org/downloads/demo.

References

1. Baker, M. Big biology: the 'omes puzzle. *Nature* **494**, 416–419 (2013).
2. Cantelli, G. et al. The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Res.* **50**, D11–D19 (2022).
3. Rouillard, A. D., Wang, Z. & Ma'ayan, A. Reprint of "Abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction". *Comput. Biol. Chem.* **59**, 123–138 (2015).
4. Rigden, D. J. & Fernandez, X. M. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **49**, D1–D9 (2021).
5. Ma'ayan, A. et al. Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.* **35**, 450–460 (2014).
6. Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* **16**, 1069–1080 (2015).
7. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, https://doi.org/10.1093/database/baw100 (2016).
8. Kawata, K. et al. Trans-omic analysis reveals selective responses to induced and basal insulin across signaling, transcriptional, and metabolic networks. *iScience* **7**, 212–229 (2018).
9. Vitrinel, B. et al. Exploiting interdata relationships in next-generation proteomics analysis. *Mol. Cell Proteom.* **18**, S5–S14 (2019).
10. Argelaguet, R. et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
11. Himmelstein, D. S. & Baranzini, S. E. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS Comput. Biol.* **11**, e1004259 (2015).
12. Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**, https://doi.org/10.7554/eLife.26726 (2017).
13. Santos, A. et al. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01145-6 (2022).
14. Cai, H., Zheng, V. W. & Chang, K. C. -C. A comprehensive survey of graph embedding: problems, techniques and applications. Preprint at https://arxiv.org/abs/1709.07604 (2017).
15. Li, M., Huang, K. & Zitnik, M. Representation learning for networks in biology and medicine: advancements, challenges, and opportunities. Preprint at https://www.arxiv-vanity.com/papers/2104.04883/ (2021).
16. Zitnik, M. & Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* **33**, i190–i198 (2017).
17. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.* **12**, 1796 (2021).
18. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
19. Duran-Frigola, M., Fernández-Torras, A., Bertoni, M. & Aloy, P. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9**, https://doi.org/10.1002/wcms.1408 (2019).
20. Fernandez-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr. Opin. Chem. Biol.* **66**, 102090 (2021).
21. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).
22. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e1417 (2017).
23. Paliwal, S., de Giorgio, A., Neil, D., Michel, J. B. & Lacoste, A. M. Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Sci. Rep.* **10**, 18250 (2020).
24. Geleta, D. et al. Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development. Preprint at *bioRxiv*, https://doi.org/10.1101/2021.10.28.466262 (2021).
25. Sosa, D. N. et al. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pac. Symp. Biocomput.* **25**, 463–474 (2020).
26. Bonner, S. et al. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. Preprint at https://doi.org/10.48550/arXiv.2102.10062 (2021).
27. Orchard, S. et al. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
28. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
29. Carvalho-Silva, D. et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
30. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
31. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. *KDD* **2016**, 855–864 (2016).
32. Vlietstra, W. J., Vos, R., Sijbers, A. M., van Mulligen, E. M. & Kors, J. A. Using predicate and provenance information from a knowledge graph for drug efficacy screening. *J. Biomed. Semant.* **9**, 23 (2018).
33. Iskar, M. et al. Drug-induced regulation of target expression. *PLoS Comput. Biol.* **6**, e1000925 (2010).
34. Wu, G., Liu, J. & Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinform.* **20**, 134 (2019).
35. Kose, F., Kocer, N. E., Sumbul, A. T., Sezer, A. & Yilkan, O. Kaposi's sarcoma following chronic lymphocytic leukemia: a rare entity. *Case Rep. Oncol.* **5**, 271–274 (2012).
36. Belur, A. A., Raajasekar, A. K. A., Nannapaneni, S. & Chelliah, T. A case of Kaposi's sarcoma in a HIV negative patient with CLL treated with rituximab. *Blood* **124**, 4970–4970 (2014).
37. Vučinić, D. et al. Kaposi's sarcoma in an HIV-negative chronic lymphocytic leukemia patient without immunosuppressive therapy: a case report. *SAGE Open Med. Case Rep.* **6**, 2050313X18799239 (2018).
38. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
39. Taskesen, E. & Reinders, M. J. T. 2D representation of transcriptomes by t-SNE exposes relatedness between human tissues. *PLoS ONE* **11**, e0149853 (2016).

40. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
41. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
42. Fernandez-Torras, A., Duran-Frigola, M. & Aloy, P. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med.* **11**, 17 (2019).
43. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
44. Costanzo, M. et al. Environmental robustness of the global yeast genetic interaction network. *Science* **372**, https://doi.org/10.1126/science.abf8424 (2021).
45. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, https://doi.org/10.1126/science.aaz8528 (2020).
46. Rual, J.-F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
47. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
48. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
49. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
50. Yoon, S. et al. Homer1 promotes dendritic spine growth through ankyrin-G and its loss reshapes the synaptic proteome. *Mol. Psychiatry* **26**, 1775–1789 (2021).
51. Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a human endonuclease complex reveals a link between tRNA splicing and Pre-mRNA 3′ end formation. *Cell* **117**, 311–321 (2004).
52. Chung, H. et al. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. *Cell* **172**, 811–824.e814 (2018).
53. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Disco.* **18**, 41–58 (2019).
54. Cai, C. et al. Transfer learning for drug discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).
55. Ma, J. et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
56. Fu, Z. et al. Ranolazine recruits muscle microvasculature and enhances insulin action in rats. *J. Physiol.* **591**, 5235–5249 (2013).
57. Eckel, R. H. et al. Effect of ranolazine monotherapy on glycemic control in subjects with type 2 diabetes. *Diabetes Care* **38**, 1189–1196 (2015).
58. Shah, N. R. et al. Ranolazine in symptomatic diabetic patients without obstructive coronary artery disease: impact on microvascular and diastolic function. *J. Am. Heart Assoc.* **6**, https://doi.org/10.1161/JAHA.116.005027 (2017).
59. Douglass, E. F. A community challenge for a pancancer drug mechanisms of action inference from perturbational profile data. *Cell Rep. Med.* **3**, 100492 (2022).
60. Schumacher, T. et al. The Effects of randomness on the stability of node embeddings. Preprint at https://doi.org/10.48550/arXiv.2005.10039 (2020).
61. Khosla, M., Setty, V. & Anand, A. A Comparative study for unsupervised network representation learning. Preprint at https://doi.org/10.48550/arXiv.1903.07902 (2019).
62. Forslund, S. K. et al. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* **600**, 500–505 (2021).
63. Hastings, J. et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
64. Bairoch, A. The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.* **29**, 25–38 (2018).
65. The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
66. Bello, S. M. et al. Disease Ontology: improving and unifying disease annotations across species. *Dis. Model. Mech.* **11**, https://doi.org/10.1242/dmm.032839 (2018).
67. Blum, M. et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa977 (2020).
68. The UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
69. Rigden, D. J. & Fernández, X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res.* **48**, D1–D8 (2020).
70. Gremse, M. et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507–D513 (2011).
71. Seco, N., Veale, T. & Hayes, J. In *Proceedings of the 16th European Conference on Artificial Intelligence* 1089–1090 (IOS Press, 2004).
72. Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marín-Llaó, J. & Hofmann-Apitius, M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst. Biol. Appl* **5**, 3 (2019).
73. Belinky, F. et al. PathCards: multi-source consolidation of human biological pathways. *Database* **2015**, https://doi.org/10.1093/database/bav006 (2015).
74. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
75. Haibe-Kains, B. et al. Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
76. Cancer Cell Line Encyclopedia, C. & Genomics of Drug Sensitivity in Cancer, C. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
77. Smirnov, P. et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).
78. Qin, Y. et al. A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).
79. Dong, Y., Chawla, N. V. & Swami, A. In *Proceedings of the* 23rd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 135–144 (Association for Computing Machinery, 2017).
80. Belkina, A. C. et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.* **10**, 5415 (2019).
81. Duran-Frigola, M. et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.* **38**, 1087–1096 (2020).
82. Bergstra, J., Yamins, D. & Cox, D. Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. *Proc 12th Python in Science Conference*. https://doi.org/10.25080/majora-8b375195-003 (2013).
83. Johnson, J., Douze, M. & Jegou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 1–1, (2019).
84. Müllner, D. fastcluster: fast hierarchical, agglomerative clustering routines forRandPython. *J. Stat. Softw.* **53**, https://doi.org/10.18637/jss.v053.i09 (2013).
85. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2017).
86. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
87. Smirnov, P. et al. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res.* **46**, D994–D1002 (2017).

Acknowledgements
P.A. acknowledges the support of the Generalitat de Catalunya (RIS3-CAT Emergents CECH: 001-P-001682 and VEIS: 001-P-001647), the Spanish Ministerio de Ciencia, Innovación y Universidades (PID2020-119535RB-I00), the Instituto de Salud Carlos III (IMPaCT-Data), and the European Commission (RiPCoN: 101003633). A.F-.T. is a recipient of an FPI fellowship (BES-2017-083053). We also acknowledge institutional funding from the Spanish Ministry of Science and Innovation through the Centres of Excellence Severo Ochoa Award, and from the CERCA Programme/Generalitat de Catalunya.

Author contributions
A.F.-T., M.D.-F. and P.A. designed the study and wrote the manuscript. A.F.-T. implemented the entire computational strategy and analysis. A.F.-T., M.B. and M.L. implemented the web resource. All authors analyzed the results and read and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33026-0.

Correspondence and requests for materials should be addressed to Patrick Aloy.

Peer review information *Nature Communications* thanks Alberto Santos and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022