Received: 12 September 2018 | Revised: 30 November 2018 | Accepted: 13 December 2018

DOI: 10.1002/wcms.1408

WILEY WIREs COMPUTATIONAL MOLECULAR SCIENCE

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,

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]

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

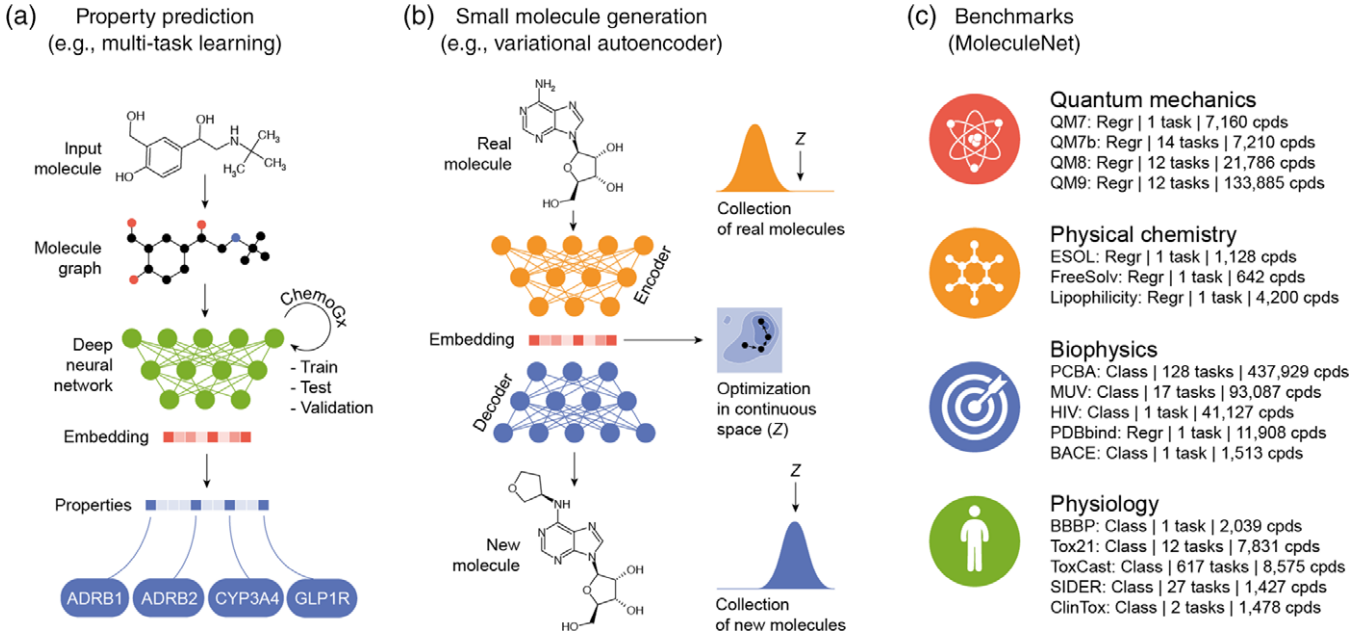

FIGURE 1 Deep learning in chemoinformatics. (a) A classical multitarget prediction exercise based on chemogenomics (ChemoGx) data. Deep neural networks can read a molecule structure as a graph (e.g., convolutional graph networks), and be trained to optimally perform a multitask classification. An inner (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

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

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]

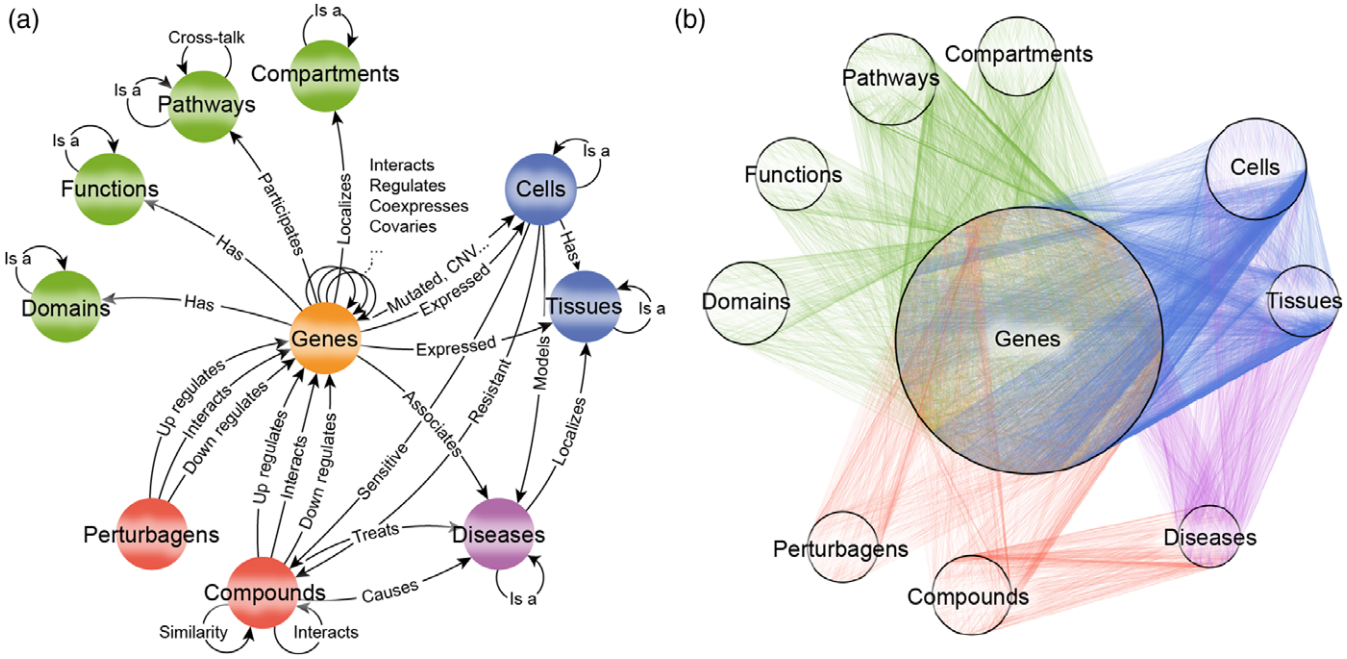

FIGURE 2 Heterogeneous network of biology. (a) A meta-graph of an in-house heterogeneous network, mostly inspired by Hetionet[101] and complemented with the Harmonizome.[72] For simplicity, only the most representative edge types are shown. "Is a" and "has" relationships typically refer to ontologies. (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

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]

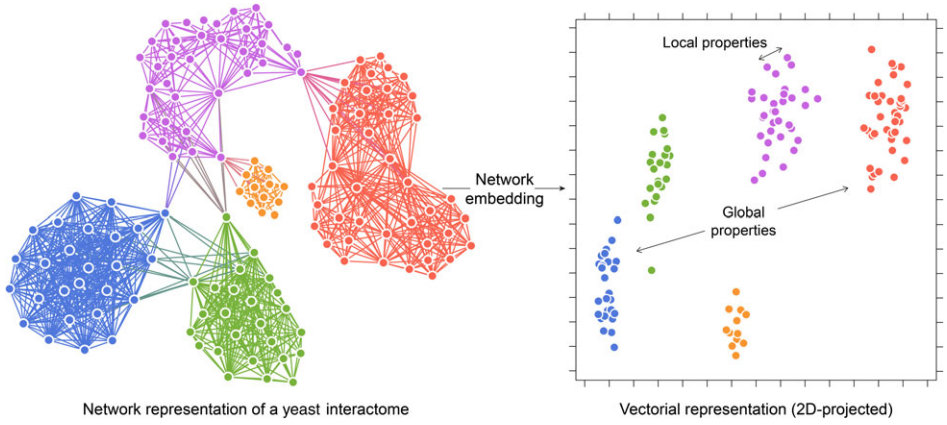

FIGURE 3 Network embedding example. The aim of network embedding is to represent graph entities (typically nodes) as numerical vectors (embeddings) that preserve graph properties, such as local distances, modularity and global organization. Here, we have embedded a fraction (~1%) of the yeast interactome[109] 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)

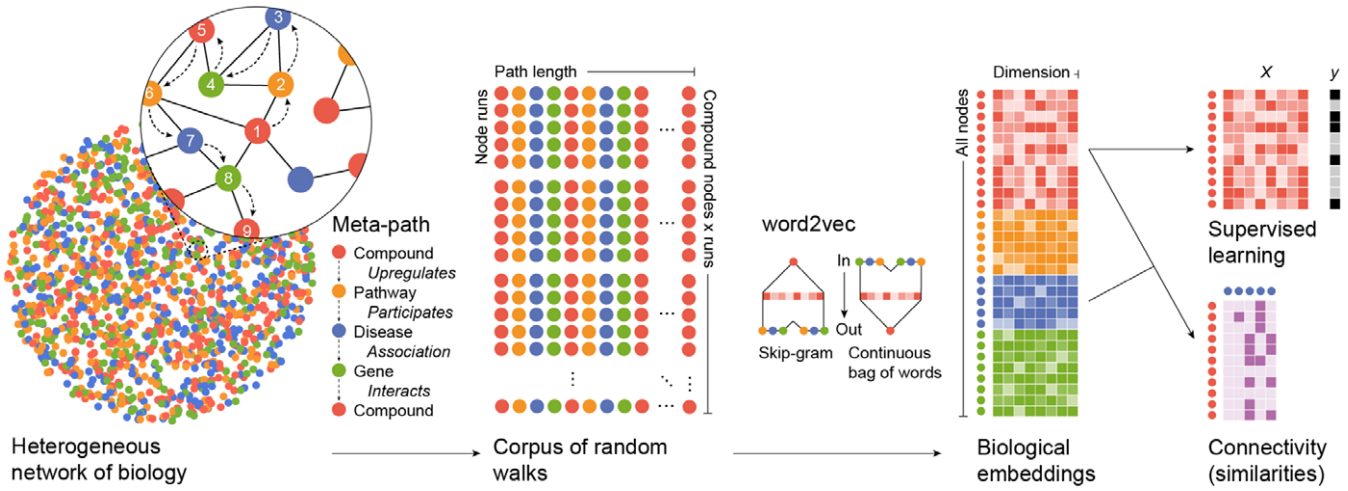

FIGURE 4 Biological embeddings. Given a heterogeneous network, the random walk algorithm can be run under the dictation of a certain meta-path. This will result in a "corpus" (text-like) output that can be apprehended with word2vec (using the skip-gram model or the continuous bag of words model). As 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

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

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.

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.

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.

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.

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*. 2018;e1408. https://doi.org/10.1002/wcms.1408