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

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].

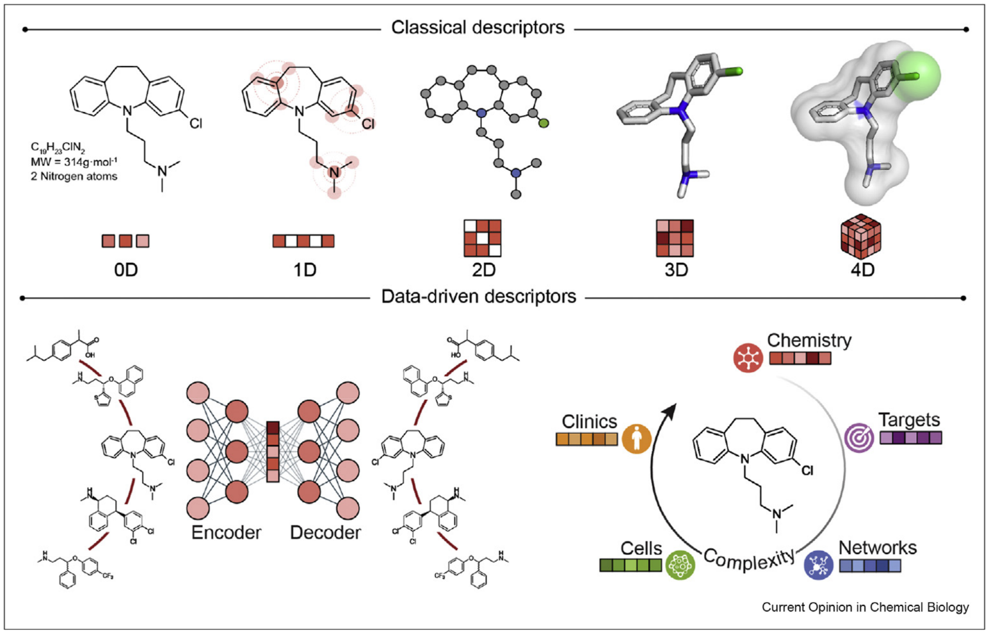
A natural extension of this first generation of data-driven descriptors is to include the wealth of bioactivity information available in the databases, to encapsulate, in the form of 'bioactivity descriptors', the experimental evidence gathered over years of research. Here, we review some recent attempts 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



Current Opinion in Chemical Biology

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 co-ordinates) 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*].

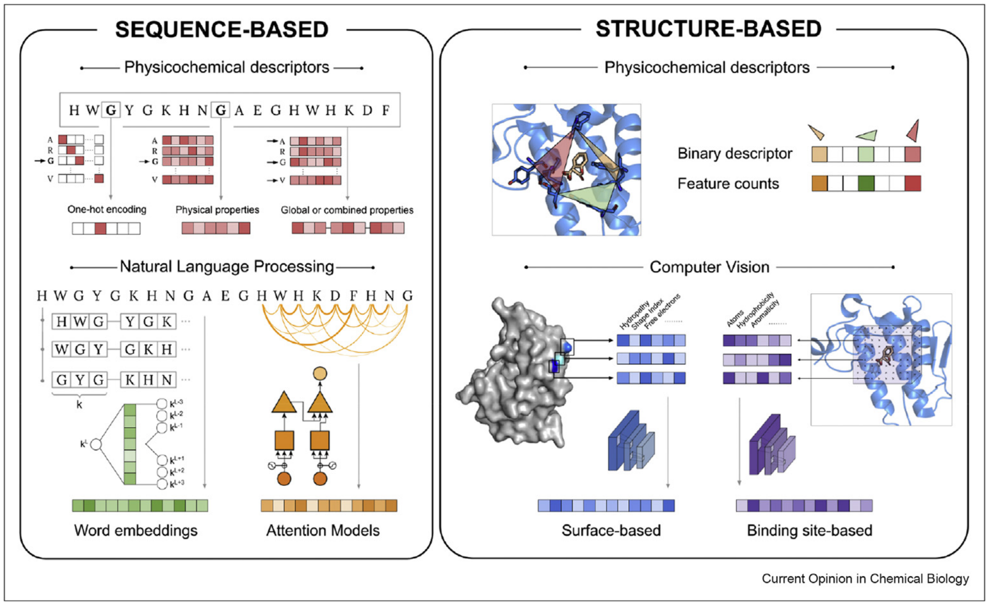
Overall, bioactivity signatures provide a complementary means to describe small molecules, focusing on the integration of multiple types of experimental data [31].

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



Current Opinion in Chemical Biology

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.

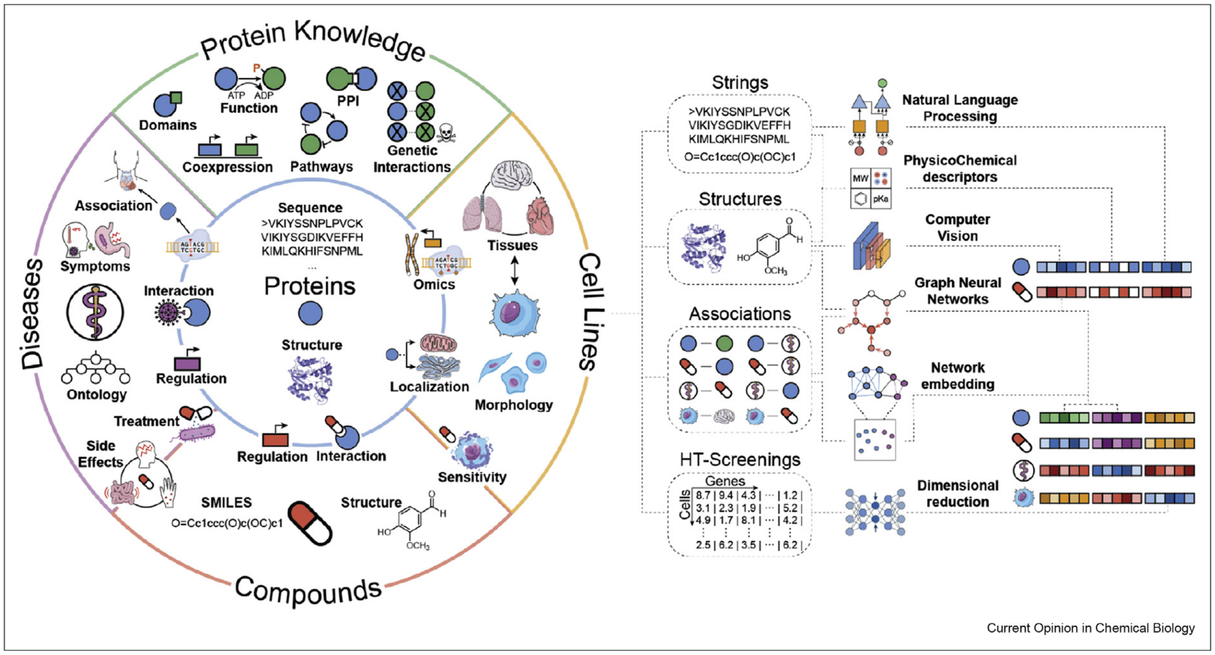
Capturing biological complexity in biomolecular descriptors

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

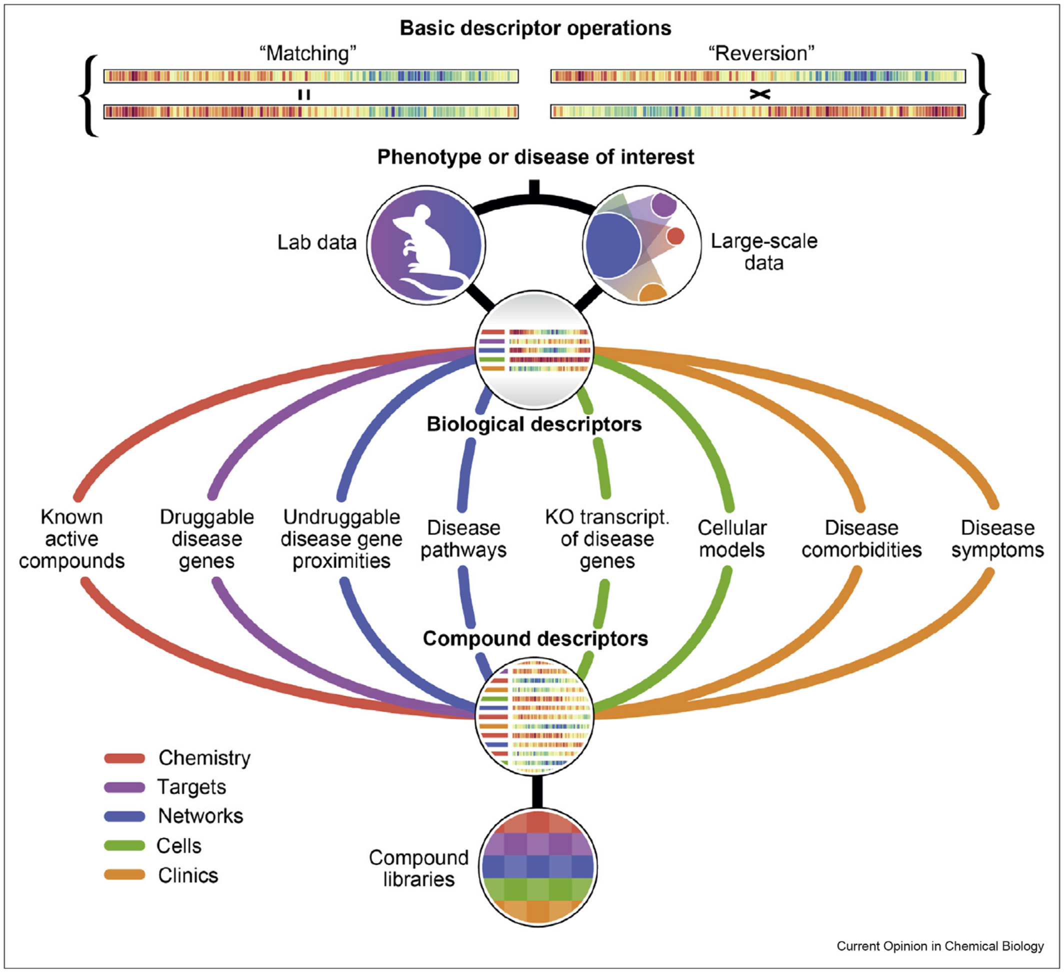
Figure 3



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, liu 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.