



# HHS Public Access

Author manuscript

*Inf Fusion*. Author manuscript; available in PMC 2020 October 01.

Published in final edited form as:

*Inf Fusion*. 2019 October ; 50: 71–91. doi:10.1016/j.inffus.2018.09.012.

## Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities

Marinka Zitnik<sup>a,i</sup>, Francis Nguyen<sup>b,c</sup>, Bo Wang<sup>d</sup>, Jure Leskovec<sup>a,e,i</sup>, Anna Goldenberg<sup>f,g,h,i</sup>, and Michael M. Hoffman<sup>b,c,g,h,i</sup>

<sup>a</sup>Department of Computer Science, Stanford University, Stanford, CA, USA

<sup>b</sup>Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

<sup>c</sup>Princess Margaret Cancer Centre, Toronto, ON, Canada

<sup>d</sup>Hikvision Research Institute, Santa Clara, CA, USA

<sup>e</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA

<sup>f</sup>Genetics & Genome Biology, SickKids Research Institute, Toronto, ON, Canada

<sup>g</sup>Department of Computer Science, University of Toronto, Toronto, ON, Canada

<sup>h</sup>Vector Institute, Toronto, ON, Canada

### Abstract

New technologies have enabled the investigation of biology and human health at an unprecedented scale and in multiple dimensions. These dimensions include myriad properties describing genome, epigenome, transcriptome, microbiome, phenotype, and lifestyle. No single data type, however, can capture the complexity of all the factors relevant to understanding a phenomenon such as a disease. Integrative methods that combine data from multiple technologies have thus emerged as critical statistical and computational approaches. The key challenge in developing such approaches is the identification of effective models to provide a comprehensive and relevant systems view. An ideal method can answer a biological or medical question, identifying important features and predicting outcomes, by harnessing heterogeneous data across several dimensions of biological variation. In this Review, we describe the principles of data integration and discuss current methods and available implementations. We provide examples of successful data integration in biology and medicine. Finally, we discuss current challenges in biomedical integrative methods and our perspective on the future development of the field.

<sup>i</sup> Corresponding authors marinka@cs.stanford.edu (Marinka Zitnik), jure@cs.stanford.edu (Jure Leskovec), anna.goldenberg@utoronto.ca (Anna Goldenberg), michael.hoffman@utoronto.ca (Michael M. Hoffman).

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

## Keywords

computational biology; personalized medicine; systems biology; heterogeneous data; machine learning

---

## 1. Introduction

Understanding complex biological systems has been an on-going quest for many researchers. The rapidly decreasing costs of high-throughput sequencing, development of massively parallel technologies, and new sensor technologies have enabled us to generate data on multiple dimensions of biological systems. These dimensions include DNA sequence [1], epigenomic state [2], single-cell expression activity [3], proteomics [4], functional and phenotypic measurements [5], and ecological and lifestyle properties [6]. These technological advances in data generation have driven the field of bioinformatics for the past decade, producing ever increasing amounts of data as researchers develop complementary analysis tools. Many of these data types have associated analytical methods designed to examine one data type specifically. Using these methods, we have assembled some of the puzzle of biological architecture. Usually, however, the factors necessary to understand a phenomenon such as a disease, cannot be captured by a single data type (Figure 1). Much of the complexity in biology and medicine thus remains unexplained. If we rely strictly on single-data-type studies, it never will be explained.

Ideally, we can combine different types of data to create a holistic picture of the cell, human health, and disease. Researchers have developed multiple approaches to do this, and therefore address the challenges brought forward by large and heterogeneous biomedical data. For example, one can identify DNA sequence variation through association studies in family- and population-based data, and then integrate it with molecular pathway information to predict the risk of developing a particular disease [7]. Data integration approaches can have numerous meanings. In this Review, however, we use data integration to mean the process by which different types of biomedical data in their broadest sense are combined as predictor variables to allow for more thorough and comprehensive modeling of biomedically relevant outcomes. As reviewed previously (*e.g.*, [8, 9, 10]), a data integration approach can achieve a more thorough and informative analysis of biomedical data than an approach that uses only a single data type. Combining multiple data types can compensate for missing or unreliable information in any single data type, and multiple sources of evidence pointing to the same outcome are less likely to lead to false positives. We are only likely to discover a complete model of a system like the human body if we include information from different dimensions, from the genome and transcriptome to organismal environment.

In this Review, we describe the principles of data integration, and provide a taxonomy of machine learning methods presently in use to integrate biomedical data. We discuss current methods, implementations of these methods, and their successful applications in biology and medicine. Furthermore, we discuss challenges in optimally combining and interpreting data from multiple sources and the advantages of integrating multiple data types. For example, one technology may address short-comings of another to provide a more precise insight into

human disease. In addition, we provide our perspective on how integrative data analysis might develop in the future.

## 2. Challenges in data integration for biology and medicine

When one develops machine learning approaches to integrate biomedical data, several challenges arise. Biological and medical datasets have inherent complexity beyond their large sizes. Biomedical datasets are also high-dimensional, incomplete, biased, heterogeneous, dynamic, and noisy. We briefly describe these challenges below.

Biomedical data is often high-dimensional but sparse. This contrasts with large datasets in other domains, such as social networks, computer vision, and natural language, that typically contain a large number of high-quality examples. A typical genome-wide association study (GWAS) [11] genotypes hundreds of thousands of single-nucleotide polymorphisms for every individual. However, these data can often be collected for only a relatively small number of individuals with a particular phenotype. Furthermore, the sparse nature of these data, *i.e.*, each polymorphism is only present in a small number of all individuals, presents an additional challenge for downstream analytic applications. It remains a major challenge to convert these data into biologically and clinically meaningful insights. Without integrating other types of data, such as pathway or molecular network information [12, 13, 14], GWAS data alone can struggle to identify meaningful patterns associated with the phenotype of interest.

Another important challenge arises from the often incomplete and biased nature of biomedical data. This challenge comes from limitations of measurement technology [15], natural and physical constraints [16, 11], and investigative biases [17]. For example, only several thousands of genes from across organisms have information about which chemical compounds they bind [18]. Furthermore, the number of associated compounds for each gene is highly uneven [19], with many uncharacterized genes playing important roles in drug action [20]. Additionally, biomedical data are hierarchically organized and span molecules, pathways, cells, tissues, organs, patients, and populations [21, 22, 23] and also cover a wide spectrum of timescales and species. Clearly, full understanding of biology requires multiscale modeling, from describing atomic details of molecules to the emergent properties of organismal populations. Furthermore, when biomedical outcomes change over time, machine learning methods integrating the outcomes need to account for these dynamics. For example, cancer cells, bacteria, and viruses evolve rapidly to gain drug resistance [24] and ignoring the dynamics of drug response can lead to poor performance in predicting drug efficacy and toxicity.

A fundamental challenge in biomedical data science lies in discovering new knowledge outside of the existing domain of knowledge, *e.g.*, extrapolating a drug response from an animal model to that in a human patient. Existing approaches typically assume that the dataset on which the algorithm is trained is representative of all the data to which the algorithm can be applied. However, it is challenging to build a model to predict, *e.g.*, efficacy of an anticancer drug in a given patient, as a new patient might be unique and might fall outside of the hypothesis space of the trained model. Asbiomedical datasets are

incomplete and reflect scientific knowledge discovered so far, the models can be trained on only these partially complete datasets and thus can perform poorly when new data become available. For these reasons, it is especially challenging to deploy machine learning systems to support decision making in risk-sensitive discovery and clinical practice [25], e.g., the system might make conflicting predictions about utility of a particular anticancer drug for a given patient depending on the type of input data used for prediction.

In summary, due to the complex and interconnected nature of biomedical systems, any single model trained on any single dataset can touch only a small part of the entire biomedical knowledge. It is thus critical to integrate diverse sources of information to gain a comprehensive understanding of biology and medicine.

### 3. Conceptual organization of methods for data integration

We broadly categorize data integration methods into two types of approaches. We refer to approaches that combine models and datasets across different scales as vertical data integration and we refer to combinations of models and datasets that operate within one scale as horizontal data integration. The vertical modeling typically depends on multiscale integration [26, 27, 23] of molecular, network, tissue, organism, and population models at several spatial and temporal scales. In contrast, horizontal integration methods focus on combining datasets and models that describe a biomedical point of interest at a particular scale [28, 29], for example, at the microbiome [30] or at the epigenome level [2].

More technically, the methods implement one of the following three distinct approaches to data integration depending on the modeling stage at which integration takes place [31, 32, 33, 8] (Figure 2). Early integration (Figure 2a) begins by transforming all datasets into a single, feature-based table or a graph-based representation. One can then use this table or graph as input to a machine learning method. This approach provides great theoretical power because the model can consider any type of dependence between the features as long as individual datasets are not collapsed prior to modeling. Early integration approaches often rely on methods for automatic feature learning, such as dimensionality reduction [34] and representation learning [35, 36], to project raw high-dimensional datasets into a low-dimensional vector space and then combine these low-dimensional representations through concatenation or other simple aggregation techniques. In late integration (Figure 2c), a first-level model is built for each dataset or data type independently. These first-level models are then combined by training a second-level model that uses predictions of the first-level models as features or via a meta-predictor [37] that takes a majority vote or combines prediction weights of the first-level models [38, 39]. We refer to a third category of data integration approaches as intermediate integration (Figure 2b). In intermediate integration, a model, such as multiple kernel learning [40, 41], collective matrix factorization [42, 43, 33] or deep neural network [44, 45] learns a joint representation of many datasets. Intermediate integration relies on algorithms that can explicitly address the multiplicity of datasets and fuse them through inference of a joint model. Importantly, it does not combine input data nor does it develop a separate model for each dataset. Instead, it aims to preserve the structure of data and only merge them during the modeling stage. The intermediate integration approach

can lead to superior performance, however it often requires development of a new algorithm and cannot be used with off-the-shelf software tools.

Finally, methods for data integration can generate diverse types of prediction outputs similar to methods that focus on one dataset or one data type (Figure 3). One area of a particular interest is the prediction of biomedical entities (*e.g.*, genes) with quantitative or categorical characteristics (*labels*). For example, many studies use a large number of genome-scale networks, including protein-protein and genetic interaction networks, which are now available for several organisms, to predict genes that cause a particular phenotype or have a particular function [46, 47] (Section 8.1). Beyond predicting labels of individual entities, many studies aim to predict *relationships*, *i.e.*, molecular interactions, functional associations, or causal relationships between biomedical entities. For example, a multiple kernel learning approach can combine kernels derived from diverse data, such as drug's structural similarity, drug's phenotypic similarity, and target similarity, to predict new *relationships* between a drug and proteins that the drug might target [48], *i.e.*, drug-target interactions (Section 9.1). Some data integration methods can identify *complex structures*, such as gene modules or clusters detected in an combined gene interaction network [49] (Section 8.2), and to generate structured outputs, such as gene regulatory networks inferred from diverse data distributions [50].

#### 4. Focus of this review

This Review is intended for computational researchers who are curious about recent developments and applications of machine learning to biology and medicine and its potential for advancing biomedicine given the vast amounts of heterogeneous data being generated today. In the Review, we focus on statistical approaches and machine learning methods for data integration. We describe the principles of integrative approaches and provide an overview of some of the methods used to predict various biomedical points of interest, the tools available to implement these analyses, and the various strengths and weaknesses of integrative approaches. Additionally, we highlight outstanding challenges and opportunities that are ripe for exploration using next-generation machine learning, and provide our perspective on how integrative approaches might develop in the future.

Several reviews cover related data integration topics from different perspectives, or with a special focus on a particular biomedical problem. For example, Rider *et al.* [51] focus on methods for network inference with a special focus on probabilistic methods. Bebek *et al.* [52] and Cowen *et al.* [49] focus on methods for construction and statistical analysis of biological networks from multiple biological datasets, as well as on visualization tools. Related reviews in [53, 8, 54, 55] survey recent advances in high-throughput technologies and data integration-based methods for translational medicine and list the tools that are available to domain scientists. Karczewski *et al.* [9] describe applications of data integration that combine diverse types of data to understand, diagnose and inform treatment of diseases. They discuss technical challenges to implementation of integrative approaches in clinics and for personalized medicine. Teschendorff *et al.* [10] surveys algorithms for drawing inferences from biological sequence data with a focus on statistical analysis of genome sequencing data.

In this Review, we survey advances in data integration at multiple levels of biomedical systems. We organize our presentation according to the flow of genetic information from the genome level to the transcriptome level and, ultimately, to the phenome level.

Heterogeneous data exist within and between these levels. We start at the DNA sequence level, describing methylation patterns and other epigenetic markers (Section 5 and Section 6), proceed at the single-cell level of gene expression (Section 7), protein variation and cellular phenotypes (Section 8), and reach the patient population levels (Section 9 and Section 10). Finally, we discuss the potential for combining diverse types of data and the central role of integrative approaches in human health and disease (Section 11).

## 5. Epigenomic variation and gene regulation

Individual cells within a multicellular organism usually have nearly identical DNA sequences, but still develop distinct cellular identities. These cellular identities manifest as diverse physical forms and behaviors, but ultimately represent differing programs of gene expression. The different gene expression programs also materialize in site-specific physical and chemical changes to the DNA and the thousands of biomolecules that interact with it. These include chemical modification of DNA bases [56, 57, 58], and of the *histone* proteins that package DNA [59, 60] into *nucleosome* structures. The collection of DNA, its packaging, and associated biomolecules is known as *chromatin*. Biologists often refer to the state of physical and chemical chromatin changes as a cell's *epigenome* [61] (Table 1), and measure its properties base-by-base along the genome.

Researchers use investigative experiments known as *assays* to determine epigenomic properties of each region in the genome (Table 2). For example, the histones DNA wrap around can undergo various chemical changes known as *histone modifications* [59]. The chromatin immunoprecipitation-sequencing (ChIP-seq) [69, 70, 71, 72] assay can map histone modifications, one at a time. As another example, nucleosomes often consistently locate at particular DNA regions in particular cell types. Nucleosome-free regions or *open chromatin* play a critical role in the control of gene regulation. A variety of techniques map nucleosomes and open chromatin, which include deoxyribonuclease-sequencing (DNase-seq) [74] and assay for transposase-accessible chromatin (ATAC-seq) [62].

Epigenomic sequencing assays usually break genomic DNA into fragments around 200 bp in length. This fragmentation enriches for chromatin with some epigenomic property of interest, such as a particular histone modification. These assays end by sequencing the pool of fragments enriched for the sought-after property. In other kinds of epigenomic sequencing experiments we might find the genetic variation in produced sequencing reads interesting. Instead, in an epigenomics sequencing assay, we are usually interested primarily in where these reads map in a reference genome—and how often. For each position in the genome, we can count the number of reads mapped to that position and treat that as a signal of the strength or frequency of the epigenomic property under analysis. Thus, we can treat the result of the experiment as a numerical vector across the genome. Usually we include other normalization steps to account for differences in experimental parameters, such as dividing by the total number of mapped reads. This transforms the initial integer counts into a real-

valued vector. For the human genome at full resolution, this vector would have 3 billion components.

Since epigenomic data might bear only an indirect connection to biological phenomena of interest, machine learning appeals as an aid for interpretation [78]. Researchers have devised numerous ways to draw conclusions about the control of gene expression and its effect of phenotype from epigenomic data [79, 80]. In this section, we survey several problems in the analysis of epigenomic data and some methods designed to solve them.

### 5.1. Semi-automated genome annotation

To get a complete picture of the epigenomic state of each part of the genome, researchers must combine the results of a number of assays. Large consortia have produced datasets that examine many aspects of epigenomic state [81, 2, 28], and one can combine these into a data matrix. One can divide this data matrix into row vectors, one for each assay, as above. Alternatively, one can divide the matrix into column vectors, one for each position in the genome. Either way, the raw signal data proves difficult to interpret and explore on its own.

Semi-automated genomic annotation (SAGA) methods [29] aid in this process by clustering regions of the genome by similarity in terms of epigenomic properties. One might describe the task in terms of identifying clusters of similar column vectors in the data matrix. However, we cannot assume independence between the column vectors. In fact, data in each column vector is highly dependent on its neighbors. Therefore, SAGA methods also simultaneously segment the genome, defining the width of a region dynamically and heterogeneously. This process results in a partition of the genome called a *segmentation*, with every region assigned to a different cluster, usually called a *label* [82] or *chromatin state* [83].

We can almost completely automate the simultaneous segmentation and clustering process of a SAGA method. The “semi-” in “semi-automated genome annotation” refers to the interpretation of the resulting clusters, conducted by a human expert. The expert examines both individual segments and aggregate features of each cluster, and describes the captured pattern in terms of a putative biological role. The identified roles may include the start of a gene, the end of a gene, and an *enhancer*—a kind of genomic element that drives expression of apparently distant genes—as well as many others. All of these have a characteristic epigenomic pattern, and SAGA methods help to characterize new instances of this pattern [84]. Researchers have used these methods to annotate many genomes, including human [85, 82, 83, 86], mouse [87], and fruit fly [88], enabling researchers to quickly assign function to genomic regions.

Methods like HMMSeg [85], ChromHMM [83], Segway [82], EpiCSeq [89] and IDEAS [86] provide an unsupervised learning approach to finding regions with similar characteristics. Most of these methods employ graphical models to find similarities in epigenomic data across genomic regions. These models treat the observed data as being emitted by some theoretical state with defined parameters, reflecting the function of that region. The first SAGA method, HMMSeg [85], takes a collection of input epigenomic assays, smooths the data with wavelets, and uses a hidden Markov model [90, 91, 92, 93, 94,

95] where the hidden state represents cluster membership. ChromHMM [83] uses a hidden Markov model that models input signals as vectors of random Bernoulli variables. The Bernoulli vectorization binarizes input data into discrete “on” or “off” categories for each region, based on whether or not the signal in that region exceeds a significance threshold based on a Poisson background distribution. EpiCSig [89] uses a similar approach, although it takes raw sequencing counts and models them as emissions from negative binomial distributions instead. Segway [82], conversely, uses single- or multiple-component Gaussians to model real-valued signal data [96]. Segway generalizes the hidden Markov model with a dynamic Bayesian network [97] that can impose hard constraints on segment lengths. Segway can also perform semi-supervised learning, and an extension enables using it in a fully-supervised pipeline [98]. IDEAS [86], finally, iteratively segments the genome for multiple input cell types at once, and classifies similar regions from across cell types using an infinite-state hidden Markov model.

## 5.2. Transcription factor binding site prediction

*Transcription factors* form a class of proteins that bind to chromatin and activate or repress gene expression. There are over 1,600 likely transcription factors, each with a characteristic pattern of binding in different cell types [99, 100]. Understanding where transcription factors bind, and why, is crucial to a mechanistic understanding of gene regulation. As transcription factors influence the rate of gene expression, knowing where transcription factors bind can help predict when transcription occurs. The most widely-used method to determine transcription factor binding in living cells is ChIP-seq [69]. These methods sequence protein-bound DNA, determining the positions at which the DNA comes in close proximity to a particular transcription factor. Related methods such ChIP-exo [67], ChIP-nexus [68], and Cleavage Under Targets and Release Using Nuclease (CUT&RUN) [73] improve on the initial approach.

The existing assays for determining transcription factor binding locations fail under many conditions. Most of these methods, require an antibody specific to the target of interest, which sometimes cannot be produced. Other methods, like CETCh-seq [65], require editing the genome in ways that might cause unexpected side effects. Furthermore, these assays all require more biological material than researchers can obtain from some precious patient samples.

Computational approaches, however, can predict binding for many transcription factors at once without requiring specific antibodies or large numbers of cells. These approaches have the goal of predicting a transcription factor’s binding at each genomic region. Several methods tackle prediction by inferring transcription factor occupancy from DNA-binding *motifs*. These motifs consist of short, recurring DNA sequences to which one transcription factor binds [101, 102, 103, 104]. Most often, we represent a motif as a *position weight matrix* [105, 106] which characterizes the expected frequency of each base’s occurrence within a binding sequence. Motifs can come from ChIP-seq data but often come from simple extracellular experiments such as protein-binding microarrays [107] or HT-SELEX (high throughput systematic evolution of ligands by exponential enrichment) [108]. The MEME method for motif elucidation searches for recurring motifs in a given set of genomic regions

using an expectation maximization algorithm [109]. When given transcription factor binding positions from ChIP-seq data, this reveals recurring motifs for that transcription factor. Unfortunately predictions that use sequence motifs alone [110] do not identify experimentally verifiable binding sites with sufficient utility for genome-wide use. A pair of observations state this principle: the *futility conjecture* [106] and the *dual futility conjecture* [111].

To move beyond the futility of predicting transcription factor binding sites with sequence alone, most methods integrate additional data. Sometimes these data include other epigenomic data, such as chromatin accessibility data, that either already exist in public databases or that one can obtain much more easily than a new ChIP-seq assay. CENTIPEDE [112] predicts binding sites using a transcription factor's position weight matrix along with open chromatin or histone modification epigenomic data. It first finds all regions which match a known sequence motif, then uses the shape of signal in other epigenomic assays to cluster each match. CENTIPEDE calculates the posterior probability that a transcription factor binds a genomic region given other information from other epigenomic assays. For instance, a transcription factor bound to DNA will leave an inaccessible region in chromatin accessibility data. Since chromatin accessibility assays mark regions with bound transcription factors as inaccessible, searching for these inaccessible regions can inform whether or not a transcription factor is bound. HINT [113] searches for the same patterns in chromatin accessibility and histone modifications, but delineates regions by detecting sudden changes in epi genomic signal. By modeling ChIP-seq data from histone modifications and an input chromatin accessibility experiment using a hidden Markov model, HINT can find transcription factor binding without motif information. It can also incorporate transcription factor motifs and rank them. Methylphet [114] incorporates *DNA methylation* information, training a random forest on bisulfite sequencing (BS-seq) data and ChIP-seq on one transcription factor. This random forest can then predict transcription factor binding sites using only BS-seq data on another sample.

Other methods use increasing numbers of data types to predict transcription factor binding sites. FactorNet [115] applies a deep neural network to this problem. FactorNet trains on input DNA sequences, chromatin accessibility, gene expression, and the binding status of a given transcription factor. It uses this network to predict the binding status of new input sequences, chromatin accessibility, and expression levels. Keilwagen et al. [116] combine features from both previous genomic annotations, *de novo* motifs from ChIP-seq and DNase-seq, and raw sequence-level data including RNA-seq. They model each of these features in a different manner. Gaussians model numerical features like RNA-seq expression levels, binomial distributions model discrete features like gene annotations, and they use a third order Markov model for genomic sequence. For a new cell type, they then take an average of the prediction scores from these models to obtain a new prediction of transcription factor occupancy. This algorithm tied for best performance in the ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge [117]. Virtual ChIP-seq [111] deemphasizes motifs, relying more on open chromatin data and ChIP-seq data from other cell types [111]. It also uses data from RNA-seq, a method for determining steady-state gene expression. Virtual ChIP-seq uses a multilayer perceptron to integrate

these diverse data types and others, learning different hyper parameters and weights for each transcription factor.

### 5.3. Topologically associated domain prediction

While computational biologists usually represent the genome as a simple string of letters, it actually has a complex three-dimensional structure. Beyond the fine-scale structure inherent in nucleosome positioning (~146 bp), each chromosome in a cell's nucleus has higher-order structures that persist in 3D. These structures bring together regions of the genome distant in one dimension, resulting in long-range chromatin interactions between genes and enhancers.

Chromosome conformation capture (3C) assays quantify spatial proximity between specific genomic regions. Some of these assays, such as Hi-C [75] and ChIA-PET [66], interrogate spatial proximity in a whole-genome all-versus-all fashion. Another recent technique, Hi-ChIP [77], combines methods from CHIP-seq to only find large regions nearby a protein of interest. These techniques have found self-interacting regions at various scales that are *conserved* across species [118]. *Topologically associated domains* (TADs) are persistent structures of spatial proximity approximately 1 Mbp in length [118, 119]. Rather than producing a vector like other epigenomic sequencing assays, these techniques produce a triangular matrix of each potential interaction. Unfortunately, as the number of potential interactions grows with the square of the number of regions interrogated, the sequencing necessary to produce it becomes rather expensive.

Many methods predict TAD locations from Hi-C data, such as Chrom3D [120] and TADbit [121]. These tools use 3C-class data to get the proximity of genomic regions to each other, and use this information to infer TAD positioning. Chrom3D [120] uses a Monte Carlo simulation to model histones as beads-on-a-string. Its Monte Carlo simulation minimizes a loss-score function with an input Hi-C and ChIP-seq data. The final output includes both a visualization of the chromatin, and the position of the identified TADs. TADbit [121] uses a breakpoint detection method to segment the genome by finding the optimal balance between the amount of Hi-C interactions upstream, downstream, and within TADs. An optimal segmentation will maximize the total log-likelihood such that all three interaction categories are equal.

Rao et al. [119] have shown that chromatin compartmentalizes itself into either gene dense, highly expressed regions, or lowly expressed regions. They used a Gaussian hidden Markov model on Hi-C interaction data to find large-scale self-interacting regions, and inferred compartmentalization from this. Methods like BACH-MIX [122], and MEGABASE [123], have been developed to determine which compartment each genomic region belongs to. BACH-MIX uses Markov chain Monte Carlo techniques to converge on a 3D model of chromatin that agrees with experimental 3C-class data. Since this experimental data can assay a heterogeneous population, where chromatin can freely move between multiple states, BACH-MIX takes into account multiple spatial rearrangements simultaneously. It models each genomic region as two substructures whose spatial arrangement varies in the sample assayed. By modeling the uncertainty between the possible arrangements with a mixture component model, it reconstructs likely chromatin architectures and their compartmentalization. MEGABASE predicts structure without 3C-class data, instead

determining chromatin compartmentalization from histone modifications. It models DNA as a polymer of self-interacting loci based on ChIP-seq data, and trains a neural network to predict compartmentalization based on this model.

#### 5.4. Histone modification and DNA methylation prediction

Histone modification prediction also benefits from computational alternatives to ChIP-seq. Epigram [124] identifies sequence motifs across cell types that strongly hint at histone modifications. Epigram then employs a random forest classifier to predict histone modification and DNA methylation from these motifs. ChromImpute [125] predicts, from a core set of commonly performed epigenomic assays, signal from other epigenomic assays. To do this, ChromImpute trains regression trees on samples where the data type of interest exists. By averaging the results of the trees from these previous experiments, ChromImpute infers signal from unperformed experiments. PREDICTD [126] imputes missing histone modification and methylation signals with large factor decomposition.

### 6. Noncoding variant effects

Researchers and medical professionals often want to know what effects DNA changes will have on cellular and organismal phenotype. While interpreting the effects of changes to the sequence coding for proteins is relatively easy, interpreting the *noncoding* sequence that makes up most of a complex organism's genome has proven far more challenging. Many non-coding sequence variants are associated with particular phenotypic traits or genetic diseases [127]. Noncoding changes often cause phenotypic effects mediated through epigenomic and gene expression changes [128]. We wish to distinguish benign noncoding variants from those that are *deleterious*. Deleterious noncoding effects often occur in specific regions that control gene regulation, called *regulatory regions* as a class. Regulatory regions include enhancers [129] and regions at the start of a gene [130].

Some methods aim to identify regulatory regions and deleterious noncoding changes based on sequence alone. For example, gkm-SVM [131, 132] find short sequences (k-mers) that are indicative of enhancer activity. It then uses a support vector machine (SVM) to find enriched k-mers in the training set versus a background of random sequences. It also allows these k-mers to have an arbitrary number of breaks, or gaps, in the sequence. The training dataset generally consists of binding sites for a given transcription factor. The kernel for this SVM computes a similarity score between two sequences, which are represented as short sequences including gaps. DeepSEA [133] trains a deep convolutional neural network on genomic sequence to predict epigenomic state. It can predict both transcription factor binding and histone modification status. DeepSEA examines the impact of sequence changes by comparing predictions made for both unmodified and modified sequence. Basset [134] learns chromatin accessibility from sequence alone. It uses a deep convolutional neural network on the sequence to obtain probability predictions of DNase-seq signal.

We can also determine a mutation's deleteriousness by integrating genomic *conservation* data. Conservation measures how little a sequence has changed over the course of evolution. Mutations almost certainly have occurred in conserved regions over evolutionary time, but those that decrease organismal fitness will have greatly diminished prevalence today. We

therefore assume that sequences that remain conserved across species or among populations in the same species indicate that mutations there would be highly deleterious, cause disease or death.

Several methods use conservation to identify deleterious mutations. Combined Annotation Dependent Depletion (CADD) integrates 63 features, including annotations drawn from conservation and epigenomic data, using a linear kernel SVM [135]. To label the SVM's training data, the CADD authors distinguish between common sequence variants that have changed since the human–chimpanzee common ancestor, and depleted simulated variants. Eigen, by contrast, applies an unsupervised method that uses conservation scores, protein function scores, and allele frequencies from a variety of mutation databases [136]. By combining these into a block matrix, and taking the eigende-composition of that matrix, Eigen finds each mutation's predictive accuracy for deleteriousness.

Some methods for predicting deleterious noncoding sequence variants rely on Inference of Natural Selection from Interspersed Genomically coherent elements (INSIGHT) [137] to identify the strength of natural selection on these variants. INSIGHT uses a complex evolutionary model that incorporates knowledge from multiple species and accounts for heterogeneous observations at different parts of the genome. The fitCons method clusters DNase-seq, RNA-seq, and histone modification data not unlike the SAGA methods above [138]. It then estimates the fraction of bases within each cluster that INSIGHT identifies as strongly under natural selection. fitCons labels each genomic region with an importance score based on INSIGHT's natural selection probability. LINSIGHT uses mostly the same procedure as fitCons, but eschews fitCons' clustering step for a generalized linear model relating observed epigenomic features to INSIGHT scores [139]. Like fitCons, it outputs INSIGHT-scored fitness for each genomic region.

## 7. Integrative single-cell analysis

A major question in biology is how to describe and quantify every cell in a multicellular organism [140], such as human, that contains a myriad of different types of cells. *Cell types*, *e.g.*, muscle and nerve, were originally defined by the functions of the tissues in which they reside and their unique morphologies [141]. However, considerable cell-to-cell variation in cells within a single cell type reflects various cell states (*e.g.*, mitotic, migratory, etc.) and various cell behaviors that depend on the local activity of each cell in a particular microenvironment. Even within a single tissue, there are diverse populations of cells, representing different manifestations of that tissue.

A traditional approach to studying tissues rely on a pooled assay and use a weighted average of a *bulk cell sample* from a particular tissue (*i.e.*, a large population of cells), obscuring cell-to-cell variation. Advances in single-cell technologies have enabled measurements at *single-cell resolution* and have opened new avenues to investigate the heterogeneity of cells across tissues and within cell populations [142]. Single-cell technologies can profile individual cells from various perspectives, including genomic [143], epigenomic [144], transcriptomic [145], and proteomic [146] perspective. However, multi-omics single-cell measurements pose a significant challenge for data analysis, integration, and interpretation

[147], one that could benefit from machine learning. Integrative single-cell analyses focus on: (1) identification and characterization of cell types and the study of their organization in space and over time, and (2) inference of gene regulatory networks from multi-omics data and assessment of network robustness across cells.

### 7.1. Cell type discovery and exploration

Single-cell RNA sequencing (scRNA-seq) is a powerful technology to measure gene expression levels of individual cells and reveal heterogeneity and functional diversity of cell populations [148]. Quantifying the variation across gene expression profiles of individual cells is a key to the identification and analysis of complex cell populations for many biological applications. The heterogeneity identified across individual cells can answer questions irresolvable by traditional ensemble-based methods, where gene expression measurements are averaged over a population of cells pooled together. Recent studies have demonstrated that *de novo* cell type discovery and identification of functionally distinct cell subpopulations are possible via un-biased analysis of all transcriptomic information provided by scRNA-seq data [149]. However, compared with bulk RNA-seq data, unique challenges associated with scRNA-seq include high dropout rate [150] (where a large number of genes have zero reads in some cells, but relatively high expression in the remaining cells) and *curse of dimensionality* (where the number of cells is much less than the number of genes, *e.g.*, Mammalian expression profiles are frequently studied as vectors with about 20,000 genes) [147]. To address these challenges, various unsupervised computational algorithms [151, 152, 153, 154, 155] have been proposed since the first study of scRNA-seq [156]. Most of these computational algorithms either rely on dimension-reduction techniques [152, 153, 155] or utilize a consensus from multiple clustering results [151, 154]. For example, Zero Inflated Factor Analysis (ZIFA), one of the very first dimension reduction methods to address the dropout events, assumes that the dropout rate for a gene follows a double exponential distribution with respect to the expected expression level of that gene in the population [152]. CellTree [157] incorporates a Latent Dirichlet Allocation model with latent gene groups to measure cell-to-cell distance by a detected tree structure outlining the hierarchical relationship between single-cell samples to introduce biological prior knowledge. [153] takes another perspective by utilizing compressed sensing together with the underlying assumption that scRNA-seq data might be collected in a compressed format, as composite measurements of linear combinations of genes. However, one clear disadvantage of these dimension-reduction methods is that strong statistical assumptions of data distributions have to be made first to facilitate the computational algorithms. These assumptions may not always hold for all the scRNA-seq technologies or platforms. Different from dimension-reduction methods, ensemble-based methods usually first generate multiple approximate representations or clusterings for cells and then integrate them in a principled way. For instance, SIMLR [151] first generates multiple kernels to represent approximate cell-to-cell variabilities and then uses a non-convex optimization framework to refine and integrate these kernels and output a detailed and fine-grained description of cell-to-cell similarity matrix. This learned similarity matrix can enable efficient clustering and visualization for scRNA-seq data. SC3 [154] takes a similar strategy in that it first generates multiple clustering results with different subsets of genes and then combine these clustering results with majority voting. However, the methods mentioned

above deal with scRNA-seq data generated by a single experiment. When it comes to integrative analysis of scRNA-seq data from multiple patient groups, different samples across tissues, and multiple conditions, the number of available methods is limited. The unique challenge lies in the fact that the accompanying biological and technical variation tends to dominate the signals for clustering the pooled single cells from the multiple populations. A recent effort [158] developed a multi-task clustering method to address the problem. This method introduces a multi-task learning method with embedded feature selection to simultaneously capture the differentially expressed genes among cell clusters and across all cell populations or experiments to achieve better single-cell clustering accuracy.

## 7.2. Single-cell multi-omics analysis

Beyond single-cell RNA-seq data, other single-cell sequencing techniques measure various biological dimensions, such as DNA methylation [159], histone modifications [160], open chromatin (scATAC-seq and scDNase-seq [161, 162]), chromosomal conformation [163], proteome [164], and metabolome [165]. Single-cell multi-omics data are potentially more powerful to provide a comprehensive understanding of the cells than any single-omics data [166], but their analysis poses interesting challenges for machine learning. In particular, one needs to discover not only information shared across various omics data but also complementary signals that are specific to a particular omics data type (Figure 4). Current methods for analysis of single-cell multi-omics data are correlation-based or clustering-based [167]. First, a prevailing approach considers pairs of omics data and generates hypotheses by measuring correlations between different omics. For example, several studies [168, 169, 170, 171] apply canonical correlation analysis (CCA) [172, 173, 174], a method that has been widely used in bulk data analysis to estimate correlations between single-cell DNA methylation and single-cell RNA-seq data. CCA learns a low-dimensional representation of the omics data that captures common information shared across all the data. However, the CCA-based analysis is limited as it does not take into account *dropout events*. Dropout events are a special type of missing values caused by the low number of RNA transcriptomes in the sequencing experiment and the stochastic nature of gene expression at a single cell level. Consequently, these dropout events become zeros in a gene-cell expression matrix and these “false zeros” mix with “true zeros” representing genes not expressed in a cell at all. To conquer this dropout issue, imputation methods use correlations between multi-omics data to impute the missing values. For example, MAGIC [175] imputes the missing values by applying a diffusion model to gene-gene correlation matrix. Another example is scImpute [176], which pulls information from groups of similar cells to fill in sparse data matrices for better representations of the original correlations.

Another direction for integrating single-cell multi-omics data uses a two-stage approach: first, construct a separate clustering for each omics dataset, and then combine these clusterings for comparison and analysis [170, 177, 178, 179]. The advantage of such an approach is its ability to infer importance of each data type and to identify information common to all data types. For example, studies [178, 179] adopt the method that first clusters cells based on each omic and then perform extensive comparisons between clusters using statistical association tests. Along similar lines, MATCHER [180] uses manifold

alignment of single-cell multi-omics data. MATCHER first uses a Gaussian process latent variable model to independently cluster every cell in each omic. It then aligns these clusterings and combines them into a global clustering of cells. These clustering approaches have the advantage of detecting both complementary and common patterns in single-cell multi-omics data. Nevertheless, they can suffer from computational complications caused by extensive generation and statistical comparison of many clusterings.

### 7.3. Large-scale single-cell bioinformatics

As single-cell technologies advance, the number of cells generated per each experiment increases, demanding for efficient and large-scale bioinformatics [150]. Present approaches for large single-cell data utilize: (1) approximate inference [181] and fast software implementations [182], or (2) adopt deep learning methods that take small batches of cells as input [183, 184]. For example, bigScale [181] uses large sample sizes to approximate an accurate numerical model of noise and cluster datasets with millions of cells. SCANPY [182], however, provides an efficient Python-based implementation that is easy to interface with other machine learning packages, such as Tensorflow [185]. Another direction within this vein is to use deep-learning based methods, since they can naturally train a multilayer neural network using memory-efficient mini-batch stochastic gradient descent. For example, [183] apply deep auto-encoders to obtain low-dimensional representations that optimize the reconstruction of original noisy inputs. Similarly, SAUCIE (Sparse Auto encoder for Unsupervised Clustering, Imputation, and Embedding) [184] uses a multi-task deep auto-encoder and performs several key tasks for single-cell data analysis including clustering, batch correction, visualization, denoising, and imputation. SAUCIE is trained to reconstruct its own input after reducing its dimensionality in a 2D embedding layer, which can be used to visualize the data. Different from traditional deep auto-encoders, SAUCIE uses two additional model regularizations: (1) an information dimension regularization to penalize the entropy as computed on the normalized activation values of each neural layer and thereby encourage binary-like encodings amenable to clustering, and (2) a maximal mean discrepancy (MMD) penalty to correct for batch effects. Although these deep learning methods achieve promising results and are capable to deal with large single-cell data, their black-box nature and lack of interpretability limit their wide adoption in practice.

## 8. Cellular phenotype and function

Our ability to generate sequence data has been improving at a rapid rate for the past decade, and this trend is likely to continue for the next decade (Section 5). A vast majority of these sequences are of proteins of unknown function and their worth could be substantially increased by knowing the biological roles that they play. Accurate annotation of protein function is a key to understanding life at the molecular level and has great biomedical and pharmaceutical implications. To this aim, numerous research efforts, such as the Encyclopedia of DNA Elements (ENCODE) [1] (Section 5) and the Genotype-Tissue Expression (GTEx) [186], have expanded the breadth of available data that lend themselves to protein function prediction (Figure 5).

*Protein function* is a concept describing biochemical and cellular aspects of molecular events that involve proteins. Protein functions can be divided into three major categories: (1) molecular functions, *e.g.*, the specific reaction catalyzed by an enzyme, (2) biological processes, *e.g.*, the metabolic pathway the enzyme is involved in, and (3) systems or physiological events, *e.g.*, if the enzyme is involved in respiration, photosynthesis or cell signaling. One can also consider a fourth level, *i.e.*, cellular components, describing cell compartments in which proteins have a role, such as a cell membrane and organelles. Functions of proteins can also vary in space and time as in the case of moonlighting proteins (*e.g.*, multitask proteins). Furthermore, many protein functions are carried out by groups of interacting proteins and these interactions can be predicted.

Most proteins are poorly characterized experimentally and we know little about their functions. Furthermore, vast majority of proteins with known functions are from model organisms, but even for those organisms, a significant part of all proteins coded in their genomes remain to be characterized. For example, in *Escherichia coli*, about one third of the 4,225 proteins remain functionally unannotated (*i.e.*, orphan proteins) and a similar proportion applies to *Saccharomyces cerevisiae*. These observations have put protein function prediction at the fore-front of computational biology.

### 8.1. Protein function prediction

Protein functions can be inferred on the basis of amino acid sequence similarity [195], gene expression [196], protein-protein interactions [46, 197, 195], metabolic interactions [198], genetic interactions [199], evolutionary relationships [200], 3D structural information [201], mining of biological literature [202], and any combination of these data. At the most basic level, protein function prediction methods can be categorized into two categories: (1) unsupervised similarity-based methods using a principle that similar proteins share similar functions, and (2) supervised methods using a classification of protein functions in the Gene Ontology [203].

Similarity-based prediction methods relate a functionally uncharacterized protein with proteins whose functions are already known. The simplest and most often used approach uses sequence similarity search. Given a query protein, similarity search programs, such as Basic Local Alignment Search Tool (BLAST) ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)), scans the sequence data banks for homologous proteins of known function or structure and transfers their functions to the query protein. If the query protein is not homologous to any protein with known function, it is possible to *de novo* predict functions of the query protein. A *de novo* prediction uses diverse information about the query protein to identify biological properties that are shared among all proteins with the same function (*e.g.*, proteins with the same function might act similarly in similar conditions, for example, in a particular human tissue). These properties are then used to select proteins whose functions are transferred to the query protein [47]. For example, [15, 204] developed a low-dimensional matrix decomposition approach that combined genetic interaction networks with other types of gene-gene similarity networks. These approaches used networks to learn an embedding (*i.e.*, a feature vector) for every protein. This was accomplished by optimizing a network reconstruction objective, assuming that each protein's embedding depended on embeddings of protein's

neighbors in the network. The learned em-beddings were then used as input to clustering algorithm. Many matrix decomposition [34] and tensor factorization [205] methods have proven useful for protein function prediction [206]. For example, [207, 208] used tensor computations to combine many weighted co-expression gene similarity networks. The same approach was also used to identify protein complexes, *i.e.*, groups of two or more proteins that form a molecular machinery and together perform a particular function [209, 210]. Along similar lines, [211, 212, 22] used Bayesian latent factor models and combined gene expression, copy number variation (CNV), and methylation data to predict protein functions. As a final example, many approaches aim to understand protein functions by combining data from different tissues [22, 213, 23, 214] or different species [215, 216, 217, 218, 219, 220]. For example, OhmNet [23] organizes 107 human tissues in a multi-layer network, in which each layer represents a tissue-specific protein-protein interaction network. OhmNet models the dependencies between network layers (*i.e.*, tissues) using a tissue hierarchy and develops an unsupervised feature learning method then learns an embedding for every node (*i.e.*, protein) in the multi-layer network by considering edges (*i.e.*, protein-protein interactions) within each layer as well as inter-layer relationships.

If there are examples of proteins with a particular function, they can be used to identify additional proteins with the same function. This is accomplished by *gene prioritization* (Figure 6). Given a set of genes with unknown function, gene prioritization ranks them by their similarity to genes with known function (*i.e.*, seed genes). Genes at the top of the ranked list are most similar to seed genes and thus are likely to have the same function as seed genes. Gene prioritization methods can be categorized into four groups: (1) similarity scoring methods that use filtering techniques to independently analyze each dataset [221], (2) methods that aggregate gene feature vectors from different datasets, *e.g.*, by concatenation, and then use the aggregated vectors as input to a downstream classifier [222], (3) methods that use each dataset separately to estimate the similarity of genes with seed genes and then combine similarity scores via a linear or nonlinear weighting [223, 224, 225], and (4) methods that construct a separate gene-gene correlation network for each dataset and combine the networks under supervision of seed genes [226, 46].

Supervised methods for function prediction use a classification of protein functions in the Gene Ontology (GO) [203] to specify a supervised prediction task. Supervised protein function prediction present four interesting challenges for machine learning methods. First, functions of proteins are classified into over 40,000 classes in the GO, and this large and complex space represents a challenge for any classification method. Second, there are dependencies between classes in the GO that lead to situations, where proteins are assigned to multiple functions in the GO, at different levels of abstraction (*e.g.*, *cellular transport* versus *extracellular amino acid transport*). Furthermore, proteins typically have multiple different functions, making the function prediction inherently a multi-label, multi-class problem. Finally, high-level physiological functions, such as *inter-cellular transport* or *regulation of heart rate*, go beyond simple molecular interactions and require many proteins to participate, and thus such functions usually cannot be predicted by considering a single protein in isolation. To take on these challenges, many approaches use joint latent factor models [190, 188], multi-label learning [46], and ensemble learning [38, 233, 216, 234]. A number of machine learning methods we also developed to integrate regulatory networks

and pathway information to predict functional modules, *i.e.* groups of functionally related proteins [235, 236, 50, 234, 237, 238], which only implicitly invoke the similarity principle described above.

Another consideration is a direct inference of a functional ontology (*i.e.*, a hierarchy of protein functions) from data [239, 240]. For example, [239] use a hierarchical network community detection algorithm together with protein-protein interaction network of *Saccharomyces cerevisiae* to infer an ontology whose coverage is comparable to the manually curated GO. An-other common approach is to use neural networks to predict protein functions. For example, [23] use a neural network to predict tissue-specific protein functions, *i.e.*, functions taking place in a particular cell type, tissue, organ, or organ system. Another example that employs neural networks is [241], who use deep learning to learn protein embeddings using protein sequence data, cross-species protein-protein interaction network, and the hierarchical relationships between protein functions in the GO. Along similar lines, [242] use several million geno-types to train a neural network whose architecture is determined by the hierarchy in the GO. As an example of biological application, [242] demonstrate that neural model can simulate cellular growth almost as accurately as laboratory experiments.

## 8.2. Protein-protein interaction prediction

One major strategy to study cellular phenotype and function is to analyze networks of physical interactions between proteins. These physical protein-protein interaction (PPI) networks carry out the core functions of cells, since interacting proteins tend to be linked to similar phenotypes and participating in similar functions [17]. Protein-protein interactions also orchestrate complex biological processes including signaling and catalysis (Figure 7) [49].

With the recent advances in experimental techniques, the number of identified PPIs keeps increasing [243]. However, we are still far from complete knowledge of PPIs and their characterization at the network level. Computational methods to predict PPIs have thus recently become popular due to the significant increase in other types of protein data, such as protein sequence and structural information, which is indicative of PPIs.

Proteins can interact with or co-localize with a variety of other biomolecules and can form stable complexes. These complexes can bind to DNA, alter gene expression, and alter cell phenotype. A predictive method by Jansen *et al.* [244] improves analyses based on pull-down assays, which experimentally find proteins interacting with an input protein. However, these assays tend to be noisy and are often incomplete. To address this issue, Jansen *et al.*'s [244] method uses Bayesian inference across pairs of interacting proteins from a variety of datasets, along with transcriptomic and essentiality information to find complete interaction networks. Another example is Chrom-Net [245], which predicts PPIs among chromatin-interacting proteins such as transcription factors using epigenomic data. It does this by identifying conditional dependence structures between proteins present at specific genomic regions. In another example [246], over 9,000 mass spectrometry protein interaction datasets from a variety of human and animal cells and tissues were combined into a comprehensive map of human protein complexes and predict PPIs. Interestingly, the combined map revealed

thousands of PPIs that were not identified by any individual mass spectrometry experiment, thus demonstrating the value of data integration. This analysis was accomplished by a network-based protein complex discovery pipeline. The computational pipeline first generated an integrated protein interaction network using features from all input datasets. To predict PPIs, the approach trained a protein interaction classifier based on support vector machines (SVMs). To predict protein complexes, the approach then employed a Markov clustering algorithm for graphs and optimized the clustering parameters relative to a training set of literature-curated protein complexes.

## 9. Computational pharmacology

The goal of computational pharmacology is to use data to predict and better understand how drugs affect the human body, support decision making in the drug discovery process, improve clinical practice and avoid unwanted side effects (for an excellent review, see [252, 20]). The properties of drugs and their interactions with the human body can be described in a variety of ways and measured at the physicochemical, pharmacological, and phenotypic levels. One can measure the physicochemical properties of a drug, such as chemical structure, melting point, or hydrophobicity. One can also measure interactions between a drug and its protein targets by quantifying binding strength, kinetic activity, and the change in a cellular state or gene expression. Furthermore, one can use phenotypic data, such as information about diseases that a particular drug treats, drug side effects, and interactions of a drug with other drugs. Such data lend themselves to mathematical representations, which are then analyzed to guide drug discovery and *in vivo* experiments in a laboratory.

### 9.1. Drug-target interaction prediction

At the most basic level, drugs have an impact on the human body by binding with protein targets and affecting their downstream activity. Identification of drug-target interactions is thus important for understanding key properties of drugs, including drug side effects, therapeutic mechanisms, and medical indications. Traditional prediction of drug-target interactions uses *molecular docking* [253], an approach that combines 3D modeling and computer simulation to dock a candidate drug into a protein-binding pocket and then score the likelihood of the pair's interaction. This approach provides insights into the structural nature of the interaction, however, the performance of molecular docking is limited when the 3D structures of target proteins are not available. As molecular docking can be computationally very demanding, *ligand-based methods* [254] have emerged as an alternative approach to drug-target interaction prediction. A ligand-based approach specifies an abstract model of chemical properties that are considered important for the interaction with the chosen target protein and then it aligns and scores candidate drugs against this model. However, ligand-based approaches perform poorly when the chosen target protein has only a small number of known binding ligands and the quality of the abstract model is low.

Many recent efforts focus on using machine learning for drug-target interaction prediction. These efforts are based on the *guilty-by-association principle*, a principle that similar drugs tend to share similar target proteins and vice versa. Using this principle, prediction can be

formulated as a binary classification task, which aims to predict whether a drug-target interaction is present or not. This straightforward classification approach considers known drug-target interactions as positive labels and uses chemical structure of drugs and DNA sequence of protein targets as input features (or kernels) [255, 256, 257]. Additionally, many methods integrate side information into the classification model, such as drug side effects [18, 258], gene expression profiles [259], drug-disease associations [260], and genes' functional information [261]. Such data provide a multi-view learning setting for drug-target interaction prediction [262, 263]. For example, [262] use kernelized matrix factorization and combine multiple types of data (*i.e.*, views), each data type is treated as a different kernel, to obtain better prediction performance than single-kernel scenarios. Another common approach is to represent multiple types of data as a heterogeneous network (Figure 8) and predict protein targets using random walks. These methods use diffusion distributions to calculate a score for each node (protein) in the network, such that the score reflects the probability that the protein is targeted by a particular drug [260, 264, 265]. In addition to random walks, one can use meta-paths [266] to extract drug and protein feature vectors from a heterogeneous network and then fed them into a classifier [267].

However, hand-engineered features, such as meta-paths, often require expert knowledge and intensive effort in feature engineering and can thus prevent methods from being scaled to large datasets. For these reasons, matrix factorization algorithms are used to learn an optimal projection of a heterogeneous network into a latent feature space. The learned latent space can be transformed into a drug-target network via a sequence of matrix operations and the resulting drug-target network is used to predict drug-target interactions [268]. A potential limitation of matrix factorization is that a classic factorization algorithm takes as input a homogeneous network and thus one needs to collapse a heterogeneous network into a homogeneous one, discarding potentially useful information. This limitation is overcome by multi-view, collective, and tensor factorization approaches to drug-target interaction prediction [262, 269, 270]. In addition to using shallow matrix factorizations, one can use deep feature learning algorithms, such as deep autoencoders [271] to integrate drug-related information. These algorithms generate a feature vector for every drug and protein in the dataset. Using the learned drug and protein features, the method finds the best projection from the drug space onto the protein space such that the projected feature vectors of drugs are geometrically close to the feature vectors of proteins that are targeted by these drugs [19]. The projection is learned to minimize prediction error on a training dataset of drug-target interactions [272]. After model training, the method predicts target proteins for a particular drug by ranking the proteins based on their geometric proximity to the drug's vector in the projected space.

## 9.2. Drug-drug interaction and drug combination prediction

The use of drug combinations is a common treatment practice. Many patients take multiple drugs at the same time to treat complex diseases or co-existing conditions [276]. A drug combination consists of multiple drugs, each of which has generally been used as a single effective medication in a patient population [277]. Since drugs in a drug combination can modulate the activity of distinct proteins, drug combinations can improve the therapeutic efficacy by overcoming the redundancy in underlying biological processes [278]. While the

use of multiple drugs may be a good practice for the treatment of many diseases, a major consequence of a drug combination for a patient is a much higher risk of side effects which can be due to drug-drug interactions [189, 279]. Such side effects can emerge because the activity of one drug may change if taken with another drug. This means that a combination of drugs leads to an exaggerated response in patients that is over and beyond the response we would expect under no interaction.

Drug-drug interactions are one of the major concerns in drug discovery. They are extremely difficult to identify manually because there are combinatorially many ways in which a given combination of drugs can clinically manifest and each combination is valid in only a certain subset of patients. It is also practically impossible to test all possible pairs of drugs [280], and observe side effects in relatively small clinical testing. Given the large number of drugs, experimental screens of pairwise combinations of drugs pose a formidable challenge in terms of cost and time. For example, given  $n$  drugs, there are  $n(n-1)/2$  pairwise drug combinations and many more higher-order combinations. Furthermore, unwanted side effects are recognized as an increasingly serious problem in the health care system affecting nearly 15% of the U.S. population [281]. To address this combinatorial explosion of candidate drug combinations, computational methods were developed to identify drug pairs that potentially interact [282].

Drug-drug interactions are defined through the concepts of synergy and antagonism [283, 284] and are quantified biologically by measuring the dose-effect curves [285, 286] or cell viability [287, 288, 289, 290, 280, 291, 292]. Computational methods use these measurements to identify combinations of drugs, most often pairs of drugs, that potentially interact. These methods predict drug-drug interactions by estimating the scores representing the overall strength of an interaction for a drug pair. Existing approaches are classification- or similarity-based. *Classification-based approaches* consider drug-drug interaction prediction as a binary classification problem [293, 288, 290, 280, 292, 294]. They use known interacting drug pairs as positive examples and other drug pairs as negative examples. The methods first obtain a feature representation of each drug pair. For example, they use a linear or nonlinear dimensionality reduction algorithm on each data type to derive a feature vector for each drug [290, 295], followed by an aggregation of feature vectors of individual drugs to obtain integrated feature vectors of drug pairs. Finally, the methods train a binary classifier, such as logistic regression, support vector machines, or neural network on feature representations of drug pairs. In contrast, *similarity-based approaches* assume that similar drugs have similar interaction patterns [296, 297, 287, 252, 33, 289, 298]. These methods combine different kinds of drug-drug similarity measures defined on drug chemical substructures, structural interaction fingerprints, drug side effects, off-target side effects, and connectivity of molecular targets. The methods aggregate similarity measures through clustering or label propagation in order to identify potential drug-drug interactions [299, 300, 301].

Moving beyond predicting the chance of drug-drug interaction occurrence, recent methods identify how exactly, if at all, a given drug pair manifests clinically within a patient population [45, 302, 303]. These methods use molecular, drug, and patient data to predict side effects associated with pairs of drugs. For example, Decagon [45] constructs a

multimodal graph of protein-protein interactions, drug-protein interactions, and drug-drug interactions (Figure 8). The approach represents each type of side effects as a different edge type in the multi-modal graph. Decagon uses the graph to develop a graph convolutional neural network, a type of neural network designed for graph data [304], that can predict side effects of drug pairs.

### 9.3. Drug repurposing

Drug repurposing (also called “drug repositioning”, Figure 9) is an area of computational pharmacology that seeks to find new uses for known drugs as well as for novel molecules. Fundamental to drug repurposing are the following two observations. First, many drugs have multiple protein targets [305] and hence a multi-target drug might be used for more than one purpose. Second, different diseases share genetic factors, molecular pathways, and clinical manifestations [306, 17] and hence a drug acting on such overlapping factors might be beneficial to more than one disease.

At a high level, drug repurposing approaches can be categorized into four groups: (1) methods that predict new uses for existing drugs on the basis of protein target interaction networks [307, 308, 309, 310, 272], (2) methods that make predictions by analyzing gene expression activation following various drug treatment regimes [311, 312], (3) methods that make predictions based on drug side effects [313, 314, 315, 316], and (4) methods that consider a variety of disease similarity and drug similarity measures, each capturing a different type of biomedical knowledge [317, 318, 319, 260, 320, 321, 322].

For example, [323, 260, 321, 272] used random walks on a heterogeneous similarity network to rank candidate drugs for a given disease. In another example, [321] designed similarity measures to construct a drug-drug similarity network, a disease-disease similarity network and a drug-disease interaction network, and then used random walks to predict medical indications. The method is based on the observation that similar drugs are used to treat similar diseases. Along similar lines, the work of [317, 318] used multiple types of drug-drug and disease-disease similarity measures and combined them via a large-margin method or logistic regression to solve the drug repurposing task.

## 10. Disease subtyping and biomarker discovery

Majority of diseases are characterized by incredible heterogeneity among patients. This includes many common diseases of which neuropsychiatric and autoimmune disorders (*e.g.*, Autism Spectrum Disorder (ASD), Attention Deficit Hyperactive Disorder (ADHD), Obsessive Compulsive Disorder (OCD), arthritis, lupus, chronic fatigue syndrome (CFS), etc) are among the most diverse. This means that individuals present at the clinic with widely ranging symptoms. ASD patients, for example, range from those with mild behavioral challenges to in-ability to speak; arthritis can affect a very particular type of joint or present itself systemically, affecting multiple organs and tissues. For a lot of common diseases, there exist classifications into *subtypes* that can be distinguished clinically (Figure 10). Consequently, treatment maybe guided by that clinical distinction. On the other hand, diseases such as cancer present themselves, for example, as a solid mass in a given organ (*e.g.*, lung, breast, stomach, etc) and clinically seem similar, however biopsy and the

consequent cellular profiling revealed that these masses may widely differ, conferring different risks and prognoses for the patients. A good example is breast cancer, where at least four different subtypes are currently distinguished in the clinic based on gene expression biomarkers (Luminal A and B, Her2+, Triple Negative/Basal-like). Further research on breast cancer has shown that there maybe closer to ten subtypes [325] or even more. It thus appears that there is both clinical and biological heterogeneity across multiple diseases. The cancer scenario tells us that clinical and biological subcategorizations of disease might not agree, indeed, the symptoms with which breast cancer patients present in the clinic are not indicative of their molecular subtypes.

Determining subtypes computationally presents a challenge. In theory, subtyping a disease means identification of homogeneous subgroups of patients, *i.e.*, clustering, yet we see that in practice, clustering of different types of patient information (clinical vs molecular data) leads to different subgroupings of patients. This inconsistency is not only present between molecular and clinical data, it is also present among molecular sub-types. For example, [326] showed that clustering of gene expression vs methylation of medulloblastoma (brain cancer) patients resulted in inconsistent subgroups which were resolved using *integration* of gene expression and methylation. Another example, is the case of glioblastoma multiform (GBM), a very aggressive adult-onset brain cancer. An earlier analysis combining gene expression and CNVs yielded two subtypes [327], whereas a later analysis, driven primarily by gene expression analysis, yielded 4 subtypes [328]. Interestingly, that while methylation data was available in [328], it was used only to explain the clusters obtained with gene expression and thus it was found to be uninformative. Analysis that used methylation as the driving signal identified a very prominent and now well recognized *IDH1* subtype, a mutation that leads to a hypermethylation across the genome that corresponds to a younger subpopulation of GBM patients with better clinical prognosis. To summarize, analyzing each of the molecular data types independently resulted in inconsistent findings that were difficult to consolidate. These examples illustrate the importance of data integration to identifying subtypes. Indeed, the more completely we can define the patients, the more faithful and hopefully, clinically relevant, will our subtypes be.

Many methods for data integration have been developed with the purpose of identifying disease subtypes. The simplest commonly used method is the concatenation of all the available data types and then clustering patients using the long concatenated vectors. The problem with this approach is that it completely disregards the structure present in each of the datasets, thus diluting the often weak signal even further. Another simple method that avoids this issue is Cluster-Of-Cluster-Assignments (COCA), which was originally developed to define subclasses in the cancer genome atlas (TCGA) breast cancer cohort [329]. COCA first clusters patients according to each of the individual data types and then takes these assignments as binary vector inputs and re-clusters patients according to those vectors thus providing consensus. The problem with this assignment is that it is mostly driven by the common signal across all data types, not making use of the complementary information potentially provided by the different data types. This approach was used by the TCGA to integrate five data types including mRNA, DNA methylation, protein array (RPPA), CNV and miRNA data across 12 cancer types and they successfully re-identified majority of the cancer types [330]. The reality, however, is that one can obtain very similar

accuracy by clustering these samples using mRNA only. The problem was with the borderline cases that multiple types of data disagreed on. COCA was unfortunately not particularly useful for most of those cases.

There are many more sophisticated approaches that try to capture internal structure, latent dimensions and nonlinearity. For example, iCluster is a Gaussian latent variable model with sparsity regularization in Lasso-type optimization framework [331]. The main assumption behind this method is that there exists latent space that captures the true subgrouping of the patients. Each of the different data types are then used jointly to estimate this latent space. This method was applied to identify 10 breast cancer subtypes from the METABRIC cohort [325]. In our experience, iCluster results tend to be dominated by the strongest single data type signal. Another drawback of iCluster is that it cannot naturally handle thousands of variables (genes), thus gene pre-selection has to be applied to the data first. This pre-selection imposes a bias and if the pre-selected features do not contain signal relevant to the true subgroups, it will be hard to impossible to recover them in the post-selection integration. Patient Specific Data Fusion (PSDF) [332] is another latent variable approach. PSDF is a nonparametric Bayesian model for discovering subtypes by combining gene expression and copy number variation. PSDF estimates a latent variable per patient, minimizing samples on which the combined data types contradict each other. While a powerful non-parametric framework, PSDF suffers from high computational costs due to the necessity to infer a large number of parameters and the restriction to combine only two data types.

Another type of methods for integrating data to identify subtypes is network-based. An example of such an approach is Similarity Network Fusion (SNF) [43]. Instead of trying to combine data in the original measurement space that are hard to calibrate and compare across a variety of data types, SNF combines data in the patient similarity space. In short, SNF consists of two steps, first it creates a similarity patient network for each of the available data types and once all the networks are constructed, it combines these networks in an iterative non-linear fashion relying on an idea of extension of random walks across multiple graphs. SNF was shown to outperform the above mentioned methods [43] on five cancers and has subsequently been applied outside of cancer to combine images and clinical data as well as a variety of lab tests across multiple diseases [326, 333, 334, 335, 336]. In its spirit SNF is similar to the Multiple Kernel Learning (MKL), which can also be used to construct and combine similarities [337]. The main difference between SNF and MKL is the linear nature of MKL which hurts its performance during integration as shown in [43]. While there are not as yet many methods that perform subtyping using network fusion, a short review on the topic can be found in [338].

When it comes to biomarker discovery, there is a myriad of papers, however, when it comes to the integrative analysis, the number of approaches identifying truly integrative biological markers is sparse. One of the early and very interesting approaches is called Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) [339]. In a nut-shell, this method models activity levels of each gene, which are represented as latent variables. The method relies on a large public network of genes, including activating and inhibitory interactions. This network is then transformed into a Bayes Net, for which the

following biological assumptions are followed: for each gene, CNA affects expression, which affects protein levels, which affect the latent protein activity. This graph represents the reference (normal) state. Given the data for a particular disease, a joint posterior distribution is computed for all latent activity nodes. By comparing pre- and post-activity levels, PARADIGM obtains a quantitative measure of the alteration induced by the disease. This approach was applied in the pancancer study [330] and biologically relevant dis-regulations were identified.

## 11. Challenges and future directions

There are great opportunities at the intersection of machine learning and biomedical data integration. However, there are equally great challenges that need to be overcome. In particular, the days of studying biomedical datasets in isolation and independently of each other are slowly coming to an end and the reductionist paradigms of looking for ‘low-hanging fruit’ (*i.e.*, the single variables that explain some portion of trait variability) are becoming less prevalent. The realization that performing all analyses within only one data type can limit the discovery of new biomedical insights has led to the development of many new ideas and methods for integrating biomedical data. However, these approaches are only in their beginning and little is known about key principles of their optimal design. In addition, gold standard methods for many biomedical points of interest, such as identifying noncoding DNA variants (Section 6), multi-omics profiling of single cells (Section 7), and stratifying patient populations (Section 10) are only emerging. Furthermore, the analysis of different data types using new machine learning approaches enables one to ask fundamentally new biomedical questions.

There are many directions available to take on these challenges. Below, we highlight outstanding problems and opportunities that need to be addressed to fully realize the potential of machine learning for integrating biomedical data.

### 11.1. Combining mixed-technology data

The structure and distributions of data generated by different technologies (*e.g.*, gene expression data generated by a sequencing-based or an array-based technology [340]) can be very different and it is challenging to combine such data. Data normalization is thus an essential first step when analyzing mixed-technology data. Furthermore, there is a deluge of types of biological assays (*e.g.*, Table 2 and Section 7) and normalizing data derived from different assays prior to downstream integrative analysis remains a major challenge. Normalization is important in the analysis pipeline because it can adjust for unwanted biological and technical noise that can mask the signal of interest. For example, one widely used normalization strategy in single-cell transcriptomics is global scaling [341] that removes cell-specific biases by scaling gene expression measurements within each cell by a constant factor. There are many opportunities for moving data normalization approaches forward by using next-generation machine learning methods. For example, one could use generative adversarial networks (GANs) to generate data with the properties of real data and then use the created data to normalize the real data. Future approaches may include *integrated strategies*, where normalization is intrinsic to a specific type of analysis (*e.g.*,

[342]), and *generic tools*, which normalize the data that can then be used as input to any downstream analysis (*e.g.*, [343, 344, 345]).

## 11.2. Multi-scale and higher-order approaches

A central goal of computational biology is to assemble a predictive model of a cell that would be able to predict a range of cellular phenotypes and answer biological questions. To be able to predict a range of phenotypes, rather than only one type of outcome, we need to understand how phenotypes are interrelated with each other. Here, multi-scale models come into play because the cell is organized in a hierarchical manner, both in 3D structure and in function [21]. Similarly, the higher-order structure and function of the cell might emerge from many molecular measurements and interaction datasets if only one could figure out how to combine these measurements properly. A multi-scale predictive model of a cell is a very general framework, but whether it can capture the full extent of biological complexity remains to be seen. Furthermore, it is not clear how to combine or extrapolate cell models to the scale of an organism (*i.e.*, human patient). This gap between a cell model and a model of a whole organism poses fundamentally new challenges that must be eventually met. Moreover, because the parameters of a machine learning model are typically fixed after the model is trained, such a model is incompatible with biological evolution. First critical steps to address these challenges have already been taken. For example, recent advances in the theory of multi-level graphs and network motifs enabled us to study, for example, higher-order organization of gene regulatory networks [346, 347] and multi-layer nature of ecological systems [26]. Furthermore, these challenges present an excellent opportunity for next-generation machine learning algorithms, such as those based on deep representation learning and topological data analysis, to develop multi-scale [23] and higher-order [348] models of a cell, and eventually of a human patient.

## 11.3. Interpretability and explainability

The black-box nature of many machine learning methods presents an additional challenge for biomedical applications. It is often difficult to interpret the output of a given model from a biomedical perspective, which limits the utility of the model in providing insights into biomedical mechanisms. This is especially the case for advanced machine learning methods, such as deep neural networks that transform the input data in such a way that it can be difficult to determine the relative importance of each feature or whether a feature is positively or negatively correlated with the outcome. Understanding black-box predictions is an open challenge in machine learning, with great attention being given to the interpretation of how a particular model relates the input to its output [349, 350, 351, 352]. There is a critical need to develop means to translate the *black boxes* of deep learning into *white boxes* that can be opened up and interpreted meaningfully from a biomedical perspective. An early application of explainability in biomedicine includes [353], an approach that integrates high fidelity data from a hospital's information management system (*e.g.*, data from patient monitors and anesthesia machines, medications, laboratory results, and electronic medical records) to predict the risk of hypoxemia during surgery and explain the patient- and surgery-specific factors that lead to that risk. In a similar way, [242] used a neural network and integrated into the neural model prior biological knowledge given by the Gene Ontology [203]. A particular genotype-phenotype association could then be explained by a hierarchy

of cellular systems from the Gene Ontology, which was identified as a neural activation map of a particular genotype-phenotype association.

#### 11.4. Integration of self-reported, lifestyle, and ecological data

While the cost and speed of generating genomic data have come down dramatically in recent years, advances in the collection of phenome data (*i.e.*, the set of all phenotypic information for a single organism or individual, see Section 10) have not kept pace. To begin to address the phenomics challenge, new research models are needed that facilitate both broad and deep phenotyping and maximize the utility of gathered data while minimizing the burden on individuals. Although studies have traditionally used medical records as the gold standard information about medical conditions, emerging research models consider internet and mobile technologies as a viable method for broad phenotyping in large populations.

Relative to medical record review, *internet-based phenotyping* can be fast (*e.g.*, [354] assessed more than 20,000 people for 50 phenotypes, such as Crohn's disease, inflammatory bowel disease and diabetes, in approximately 12 months using only a small team of people). Emerging research has demonstrated the value of combining these self-reported data with genomic information about individuals. For example, [11] conducted a genome-wide association analysis of self-reported morningness (*i.e.*, a morning person prefers to rise and rest early) and then analyzed the newly identified genetic variants using biological pathways. Along similar lines, [16] recently used self-reported data from more than 300,000 individuals and combined them with a genome-wide association study to identify genetic variants associated with depression. Furthermore, integration of other types of lifestyle and ecological data together with molecular information has a large potential to reveal new biological mechanisms. For example, [30] is an early work in this area that combined human gut microbiome data with lifestyle information. The combined data revealed striking differences in gut microbial communities between seasons that depended on seasonal availability of different types of food.

## 12. Conclusions

Machine learning is becoming integral to modern biomedical research. Importantly, machine learning approaches have emerged that can integrate data from many different sources. These approaches aim to bridge the gap between our ability to generate vast amounts of data and our understanding of biomedical systems and thus reflect the intricate complexity of biology. Ongoing methodological developments and emerging applications of machine learning promise an exciting future for biomedical data integration, although it is likely that no single method will perform best for all biomedical points of interest. Approaches thus need to be selected according to different types of domain-specific models, specific types of data, and different types of biomedical outcomes. In this Review, we described various machine learning approaches that can currently be implemented to perform powerful integrative analyses. As integrative approaches becomes more readily available, systems biology and systems medicine are likely to become a central computational strategy to generate new knowledge in biology and medicine.

## Acknowledgements

M.Z. and J.L. were supported in part by NSF IIS-1149837, NIH BD2K U54EB020405, DARPA SIMPLEX, Stanford Data Science Initiative and Chan Zuckerberg Biohub. F.N. and M.M.H. were supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H.).

## References

- [1]. Consortium EP, et al., An integrated encyclopedia of DNA elements in the human genome, *Nature* 489 (7414) (2012) 57. [PubMed: 22955616]
- [2]. Kundaje A, et al., Integrative analysis of 111 reference human epigenomes, *Nature* 518 (7539) (2015) 317–330. [PubMed: 25693563]
- [3]. Quake SR, Wyss-Coray T, Darmanis S, Consortium TM, et al., Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris, *bioRxiv* (2018) 237446.
- [4]. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al., Mass-spectrometry-based draft of the human proteome, *Nature* 509 (7502) (2014) 582. [PubMed: 24870543]
- [5]. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al., A global genetic interaction network maps a wiring diagram of cellular function, *Science* 353 (6306) (2016) aaf1420.
- [6]. Li X, Dunn J, Salins D, Zhou G, Zhou W, Rose SMS-F, Perelman D, Colbert E, Runge R, Rego S, et al., Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information, *PLoS Biology* 15 (1) (2017) e2001402. [PubMed: 28081144]
- [7]. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H, Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies, *Nature Genetics* 45 (4) (2013) 400. [PubMed: 23455638]
- [8]. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D, Methods of integrating data to uncover genotype-phenotype interactions, *Nature Reviews Genetics* 16 (2) (2015) 85–97.
- [9]. Karczewski KJ, Snyder MP, Integrative omics for health and disease, *Nature Reviews Genetics*
- [10]. Teschendorff AE, Relton CL, Statistical and integrative system-level analysis of DNA methylation data, *Nature Reviews Genetics* 19 (3) (2018) 129.
- [11]. Hu Y, Shmygelska A, Tran D, Eriksson N, Tung JY, Hinds DA, GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person, *Nature Communications* 7 (2016) 10448.
- [12]. Linghu B, Snitkin ES, Hu Z, Xia Y, DeLisi C, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, *Genome Biology* 10 (9) (2009) R91. [PubMed: 19728866]
- [13]. Hofree M, Shen JP, Carter H, Gross A, Ideker T, Network-based stratification of tumor mutations, *Nature Methods* 10 (11) (2013) 1108–1115. [PubMed: 24037242]
- [14]. Lundby A, Rossin EJ, Steffensen AB, Acha MR, Newton-Cheh C, Pfeufer A, Lynch SN, Olesen S-P, Brunak S, Ellinor PT, et al., Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics, *Nature Methods* 11 (8) (2014) 868–874. [PubMed: 24952909]
- [15]. Zitnik M, Zupan B, Data imputation in epistatic maps by network-guided matrix completion, *Journal of Computational Biology* 22 (6) (2015) 595–608. [PubMed: 25658751]
- [16]. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, Tung JY, Hinds DA, Perlis RH, Winslow AR, Identification of 15 genetic loci associated with risk of major depression in individuals of European descent, *Nature Genetics* 48 (9) (2016) 1031–1036. [PubMed: 27479909]
- [17]. Menche J, et al., Uncovering disease-disease relationships through the incomplete interactome, *Science* 347 (6224) (2015) 1257601.
- [18]. Campillos M, et al., Drug target identification using side-effect similarity, *Science* 321 (5886) (2008) 263–266. [PubMed: 18621671]

- [19]. Zong N, Kim H, Ngo V, Harismendy O, Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations, *Bioinformatics* (2017) btx160.
- [20]. Hodos RA, et al., In silico methods for drug repurposing and pharmacology, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 8 (3) (2016) 186–210. [PubMed: 27080087]
- [21]. Carvunis A-R, Ideker T, Siri of the cell: what biology could learn from the iPhone, *Cell* 157 (3) (2014) 534–538. [PubMed: 24766803]
- [22]. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Seal- fon SC, et al., Understanding multicellular function and disease with human tissue-specific networks, *Nature Genetics* 47 (6) (2015) 569. [PubMed: 25915600]
- [23]. Zitnik M, Leskovec J, Predicting multicellular function through multi- layer tissue networks, *Bioinformatics* 33 (14) (2017) i190–i198. [PubMed: 28881986]
- [24]. Bicker J, et al., Elucidation of the impact of p-glycoprotein and breast cancer resistance protein on the brain distribution of catechol-*o*- methyltransferase inhibitors, *Drug Metabolism and Disposition* 45 (12) (2017) 1282–1291. [PubMed: 28916530]
- [25]. Mullainathan S, Obermeyer Z, Does machine learning automate moral hazard and error?, *American Economic Review* 107 (5) (2017) 476–80. [PubMed: 28781376]
- [26]. Pilosof S, Porter MA, Pascual M, K’fi S, The multilayer nature of ecological networks, *Nature Ecology & Evolution* 1 (2017) 0101.
- [27]. Zitnik M, Zupan B, Jumping across biomedical contexts using compressive data fusion, *Bioinformatics* 32 (12) (2016) i90–i100. [PubMed: 27307649]
- [28]. Bujold D, Morais D.A.d.L., Gauthier C, Co’te’ C Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T, Caron B, Veilleux A, Jacques P–, Bourque G, The International Human Epigenome Consortium Data Portal, *Cell Systems* 3 (5) (2016) 496–499. [PubMed: 27863956]
- [29]. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS, Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression., *Genome Research* 25 (4) (2015) 544– 57. [PubMed: 25677182]
- [30]. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, et al., Seasonal cycling in the gut microbiome of the hadza hunter-gatherers of tanzania, *Science* 357 (6353) (2017) 802–806. [PubMed: 28839072]
- [31]. Pavlidis P, Weston J, Cai J, Noble WS, Learning gene functional classifications from multiple data types, *Journal of Computational Biology* 9 (2) (2002) 401–411. [PubMed: 12015889]
- [32]. Maragos P, Gros P, Katsamanis A, Papandreou G, Cross-modal integration for performance improving in multimedia: a review, in: *Multi- modal Processing and Interaction*, Springer, 2008, pp. 1–46.
- [33]. Zitnik M, Zupan B, Data fusion by matrix factorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (1) (2015) 41–53. [PubMed: 26353207]
- [34]. Zitnik M, Zupan B, Nimfa: A python library for nonnegative matrix factorization, *Journal of Machine Learning Research* 13 (2012) 849–853.
- [35]. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *Journal of Machine Learning Research* 11 (2010) 3371–3408.
- [36]. Sarajlic’ A, Malod-Dognin N, Yaverog’lu O’N, Prz’ulj N, Graphlet-based characterization of directed networks, *Scientific Reports* 6 (2016) 35098. [PubMed: 27734973]
- [37]. Yang P, Hwa Yang Y, B Zhou B, Y Zomaya A, A review of ensemble methods in bioinformatics, *Current Bioinformatics* 5 (4) (2010) 296–308.
- [38]. Wu C-C, Asgharzadeh S, Triche TJ, D’argenio DZ, Prediction of human functional genetic networks from heterogeneous data using rvm- based ensemble learning, *Bioinformatics* 26 (6) (2010) 807–813. [PubMed: 20134029]
- [39]. Iam-On N, Boongoen T, Garrett S, LCE: a link-based cluster ensemble method for improved gene expression data analysis, *Bioinformatics* 26 (12) (2010) 1513–1519. [PubMed: 20444838]

- [40]. Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, Tahi F, Towards a piRNA prediction using multiple kernel fusion and support vector machine, *Bioinformatics* 30 (17) (2014) i364–i370. [PubMed: 25161221]
- [41]. Mariette J, Villa-Vialaneix N, Unsupervised multiple kernel learning for heterogeneous data integration, *Bioinformatics*
- [42]. Zitnik M, Zupan B, Survival regression by data fusion, *Systems Biomedicine* 2 (3) (2014) 47–53.
- [43]. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A, Similarity network fusion for aggregating data types on a genomic scale, *Nature Methods* 11 (3) (2014) 333–337. [PubMed: 24464287]
- [44]. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene CS, Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks, *Cell Systems* 5 (1) (2017) 63–71. [PubMed: 28711280]
- [45]. Zitnik M, Agrawal M, Leskovec J, Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics*
- [46]. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q, Gene- MANIA: a real-time multiple association network integration algorithm for predicting gene function, *Genome Biology* 9 (1) (2008) S4.
- [47]. Carreras-Puigvert J, Zitnik M, Jemth A-S, Carter M, Unterlass JE, Hallström B, Loseva O, Karem Z, Caldero'n-Montan˜o JM, Lindskog C, et al., A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family, *Nature Communications* 8 (1) (2017) 1541.
- [48]. Goˆnen M, Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization, *Bioinformatics* 28 (18) (2012) 2304–2310. [PubMed: 22730431]
- [49]. Cowen L, Ideker T, Raphael BJ, Sharan R, Network propagation: a universal amplifier of genetic associations, *Nature Reviews Genetics* 18 (2017) 551562.
- [50]. Zitnik M, Zupan B, Gene network inference by fusing data from diverse distributions, *Bioinformatics* 31 (12) (2015) i230–i239. [PubMed: 26072487]
- [51]. Rider AK, Chawla NV, Emrich SJ, A survey of current integrative network algorithms for systems biology, in: *Systems Biology*, Springer, 2013, pp. 479–495.
- [52]. Bebek G, Koyuturk M, Price ND, Chance MR, Network biology methods integrating biological data for translational science, *Briefings in Bioinformatics* 13 (4) (2012) 446–459. [PubMed: 22390873]
- [53]. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L, Principles and methods of integrative genomic analyses in cancer, *Nature Reviews Cancer* 14 (5) (2014) 299– 313. [PubMed: 24759209]
- [54]. Gligorijevic V, Malod-Dognin N, Pržulj N, Integrative methods for analyzing big data in precision medicine, *Proteomics* 16 (5) (2016) 741– 58. [PubMed: 26677817]
- [55]. Malod-Dognin N, Petschnigg J, Pržulj N, Precision medicine promising, yet challenging road lies ahead, *Current Opinion in Systems Biology*
- [56]. Klose RJ, Bird AP, Genomic DNA methylation: the mark and its mediators, *Trends in Biochemical Sciences* 31 (2) (2006) 89–97. [PubMed: 16403636]
- [57]. Severin PMD, Zou X, Schulten K, Gaub HE, Effects of cytosine hydroxymethylation on DNA strand separation., *Biophysical Journal* 104 (1) (2013) 208–15. [PubMed: 23332073]
- [58]. Spruijt CG, Vermeulen M, DNA methylation: old dog, new tricks?, *Nature Structural & Molecular Biology* 21 (11) (2014) 949–954.
- [59]. Rothbart SB, Strahl BD, Interpreting the language of histone and DNA modifications., *Biochimica et biophysica acta* 1839 (8) (2014) 627–43. [PubMed: 24631868]
- [60]. Stirzaker C, Taberlay PC, Statham AL, Clark SJ, Mining cancer methylomes: prospects and challenges, *Trends in Genetics* 30 (2) (2014) 75–84. [PubMed: 24368016]
- [61]. Lappalainen T, Grealley JM, Associating cellular epigenetic models with human phenotypes, *Nature Reviews Genetics* 18 (7) (2017) 441–451.

- [62]. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position., *Nature Methods* 10 (12) (2013) 1213–8. [PubMed: 24097267]
- [63]. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, *Nature* 452 (7184) (2008) 215–219. [PubMed: 18278030]
- [64]. Arnold P, Schöler A, Pachkov M, Balwiercz PJ, Jørgensen H, Stadler MB, van Nimwegen E, Schübeler D, Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting., *Genome Research* 23 (1) (2013) 60–73. [PubMed: 22964890]
- [65]. Savic D, Partridge EC, Newberry KM, Smith SB, Meadows SK, Roberts BS, Mackiewicz M, Mendenhall EM, Myers RM, CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins., *Genome Research* 25 (10) (2015) 1581–9. [PubMed: 26355004]
- [66]. Fullwood MJ, Liu MH, Pan YF, et al., An oestrogen-receptor- $\alpha$ -bound human chromatin interactome, *Nature* 462 (7269) (2009) 58–64. [PubMed: 19890323]
- [67]. Rhee HS, Pugh BF, ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy., *Current protocols in molecular biology* Chapter 21 (2012) Unit 21.24.
- [68]. He Q, Johnston J, Zeitlinger J, ChIP-nexus enables improved detection of in vivo transcription factor binding footprints, *Nature Biotechnology* 33 (4) (2015) 395–401.
- [69]. Johnson DS, Mortazavi A, Myers RM, Wold B, Genome-Wide Mapping of in Vivo Protein-DNA Interactions, *Science* 316 (5830) (2007) 1497–1502. [PubMed: 17540862]
- [70]. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nature Methods* 4 (8) (2007) 651–657. [PubMed: 17558387]
- [71]. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K, High-Resolution Profiling of Histone Methylations in the Human Genome, *Cell* 129 (4) (2007) 823–837. [PubMed: 17512414]
- [72]. Mikkelsen TS, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature* 448 (7153) (2007) 553–560. [PubMed: 17603471]
- [73]. Skene PJ, Henikoff S, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites., *eLife* 6.
- [74]. Song L, Crawford GE, DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells., *Cold Spring Harbor protocols* 2010 (2) (2010) pdb.prot5384.
- [75]. Lieberman-Aiden E, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome., *Science (New York, N.Y.)* 326 (5950) (2009) 289–93.
- [76]. de Wit E, de Laat W, A decade of 3C technologies: insights into nuclear organization., *Genes & development* 26 (1) (2012) 11–24. [PubMed: 22215806]
- [77]. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY, HiChIP: efficient and sensitive analysis of protein-directed genome architecture, *Nature Methods* 13 (11) (2016) 919–922. [PubMed: 27643841]
- [78]. Holder LB, Haque MM, Skinner MK, Machine learning for epigenetics and future medical applications, *Epigenetics* 12 (7) (2017) 505–514. [PubMed: 28524769]
- [79]. Widschwendter M, et al., Epigenome-based cancer risk prediction: rationale, opportunities and challenges, *Nature Reviews Clinical Oncology* 15 (5) (2018) 292–309.
- [80]. Stricker SH, Köferle A, Beck S, From profiles to function in epigenomics, *Nature Reviews Genetics* 18 (1) (2017) 51–66.
- [81]. Consortium TEP, The ENCODE (ENCyclopedia Of DNA Elements) Project., *Science (New York, N.Y.)* 306 (5696) (2004) 636–40.
- [82]. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS, Unsupervised pattern discovery in human chromatin structure through genomic segmentation, *Nature Methods* 9 (5) (2012) 473–476. [PubMed: 22426492]

- [83]. Ernst J, Kellis M, ChromHMM: automating chromatin-state discovery and characterization, *Nature Methods* 9 (3) (2012) 215–216. [PubMed: 22373907]
- [84]. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS, Integrative annotation of chromatin elements from ENCODE data, *Nucleic Acids Research* 41 (2) (2013) 827–841. [PubMed: 23221638]
- [85]. Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS, Unsupervised segmentation of continuous genomic data, *Bioinformatics* 23 (11) (2007) 1424–1426. [PubMed: 17384021]
- [86]. Zhang Y, An L, Yue F, Hardison RC, Jointly characterizing epigenetic dynamics across multiple human cell types., *Nucleic acids research* 44 (14) (2016) 6721–31. [PubMed: 27095202]
- [87]. Yue F, et al., A comparative encyclopedia of DNA elements in the mouse genome, *Nature* 515 (7527) (2014) 355–364. [PubMed: 25409824]
- [88]. Kharchenko PV, et al., Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*., *Nature* 471 (7339) (2011) 480–5. [PubMed: 21179089]
- [89]. Mammanna A, Chung H-R, Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome, *Genome Biology* 16 (1) (2015) 151. [PubMed: 26206277]
- [90]. Rabiner L, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* 77 (2) (1989) 257–286.
- [91]. Baum LE, Petrie T, Soules G, Weiss N, A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *The Annals of Mathematical Statistics* 41 (1) (1970) 164–171.
- [92]. Baum LE, Petrie T, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, *The Annals of Mathematical Statistics* 37 (6) (1966) 1554–1563.
- [93]. Baum L, An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process, *Inequalities* 3 (1972) 1–8.
- [94]. Baum LE, Sell G, Growth transformations for functions on manifolds, *Pacific Journal of Mathematics* 27 (2) (1968) 211–227.
- [95]. Blakley GR, Homogeneous nonnegative symmetric quadratic transformations, *Bulletin of the American Mathematical Society* 70 (5) (1964) 712–716.
- [96]. Chan RCW, Libbrecht MW, Roberts EG, Bilmes JA, Noble WS, Hoffman MM, Birol I, Segway 2.0: Gaussian mixture models and minibatch training, *Bioinformatics* 34 (4) (2018) 669–671. [PubMed: 29028889]
- [97]. Dagum P, Galper A, Horvitz E, Seiver A, Uncertain reasoning and forecasting, *International Journal of Forecasting* 11 (1) (1995) 73–87.
- [98]. Libbrecht MW, Rodriguez O, Weng Z, Hoffman M, Bilmes JA, Noble WS, A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types, *bioRxiv* (2018) 086025.
- [99]. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM, A census of human transcription factors: function, expression and evolution, *Nature Reviews Genetics* 10 (4) (2009) 252–263.
- [100]. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT, The Human Transcription Factors., *Cell* 172 (4) (2018) 650–665. [PubMed: 29425488]
- [101]. D’haeseleer P, What are DNA sequence motifs?, *Nature Biotechnology* 24 (4) (2006) 423–425.
- [102]. Bailey TL, Gribskov M, Combining evidence using p-values: application to sequence homology searches, *Bioinformatics* 14 (1) (1998) 48–54. [PubMed: 9520501]
- [103]. Grant CE, Bailey TL, Noble WS, FIMO: scanning for occurrences of a given motif, *Bioinformatics* 27 (7) (2011) 1017–1018. [PubMed: 21330290]
- [104]. Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohee S, van Helden J, RSAT: regulatory sequence analysis tools, *Nucleic Acids Research* 36 (Web Server) (2008) W119–W127. [PubMed: 18495751]

- [105]. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A, Use of the Perceptron algorithm to distinguish translational initiation sites in *E. coli*, *Nucleic Acids Research* 10 (9) (1982) 2997–3011. [PubMed: 7048259]
- [106]. Wasserman WW, Sandelin A, Applied bioinformatics for the identification of regulatory elements, *Nature Reviews Genetics* 5 (4) (2004) 276–287.
- [107]. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang C-F, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML, Diversity and complexity in DNA recognition by transcription factors., *Science (New York, N.Y.)* 324 (5935) (2009) 1720–3.
- [108]. Ogawa N, Biggin MD, High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro, in: *Gene Regulatory Networks*, Springer, 2012, pp. 51–63.
- [109]. Bailey TL, Elkan C, Unsupervised learning of multiple motifs in biopolymers using expectation maximization, *Machine Learning* 21 (1– 2) (1995) 51–80.
- [110]. Jayaram N, Usvyat D, Martin ACR, Evaluating tools for transcription factor binding site prediction, *BMC Bioinformatics* (2016) 1. [PubMed: 26817711]
- [111]. Karimzadeh M, Hoffman MM, Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome, *bioRxiv* (2018) 168419.
- [112]. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data., *Genome research* 21 (3) (2011) 447–55. [PubMed: 21106904]
- [113]. Gusmao EG, Dieterich C, Zenke M, Costa IG, Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications, *Bioinformatics* 30 (22) (2014) 3143–3151. [PubMed: 25086003]
- [114]. Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, Yao B, Zhang F, Jin P, Wu H, Qin ZS, Base-resolution methylation patterns accurately predict transcription factor bindings in vivo, *Nucleic Acids Research* 43 (5) (2015) 2757–2766. [PubMed: 25722376]
- [115]. Quang D, Xie X, FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data, *bioRxiv* (2017) 151274.
- [116]. Keilwagen J, Posch S, Grau J, Learning from mistakes: Accurate prediction of cell type-specific transcription factor binding, *bioRxiv* (2017) 230011.
- [117]. ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge - syn6131484 (2017).
- [118]. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B, Topological domains in mammalian genomes identified by analysis of chromatin interactions, *Nature* 485 (7398) (2012) 376–380. [PubMed: 22495300]
- [119]. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping., *Cell* 159 (7) (2014) 1665–80. [PubMed: 25497547]
- [120]. Paulsen J, Sekelja M, Oldenburg AR, Barateau A, Briand N, Del-barre E, Shah A, Sørensen AL, Vigouroux C, Buendia B, Collas P, Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamina-genome contacts, *Genome Biology* 18 (1) (2017) 21. [PubMed: 28137286]
- [121]. Serra F, Bau` D, Goodstadt M, Castillo D, Filion GJ, MartiRenom MA, Automatic analysis and 3D-modelling of Hi-C data using TAD-bit reveals structural features of the fly chromatin colors, *PLOS Computational Biology* 13 (7) (2017) e1005665. [PubMed: 28723903]
- [122]. Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS, Bayesian inference of spatial organizations of chromosomes., *PLoS computational biology* 9 (1) (2013) e1002893. [PubMed: 23382666]
- [123]. Di Pierro M, Cheng RR, Lieberman Aiden E, Wolynes PG, Onuchic JN, De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture., *Proceedings of the National Academy of Sciences of the United States of America* 114 (46) (2017) 12126–12131. [PubMed: 29087948]
- [124]. Whitaker JW, Chen Z, Wang W, Predicting the human epigenome from DNA motifs, *Nature Methods* 12 (3) (2015) 265–272. [PubMed: 25240437]

- [125]. Ernst J, Kellis M, Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues, *Nature Biotechnology* 33 (4) (2015) 364–376.
- [126]. Durham TJ, Libbrecht MW, Howbert JJ, Bilmes J, Noble WS, PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition, *Nature Communications* 9 (1) (2018) 1402.
- [127]. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits., *Proceedings of the National Academy of Sciences of the United States of America* 106 (23) (2009) 9362–7. [PubMed: 19474294]
- [128]. Prensner JR, Chinnaiyan AM, The emergence of lncRNAs in cancer biology., *Cancer discovery* 1 (5) (2011) 391–407. [PubMed: 22096659]
- [129]. Shlyueva D, Stampfel G, Stark A, Transcriptional enhancers: from properties to genome-wide predictions, *Nature Reviews Genetics* 15 (4) (2014) 272–286.
- [130]. Riethoven J-JM, Regulatory regions in DNA: promoters, enhancers, silencers, and insulators, in: *Computational Biology of Transcription Factor Binding*, Springer, 2010, pp. 33–42.
- [131]. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA, gkmSVM: an R package for gapped-kmer SVM., *Bioinformatics (Oxford, England)* 32 (14) (2016) 2205–7.
- [132]. Ghandi M, Lee D, Mohammad-Noori M, Beer MA, Enhanced regulatory sequence prediction using gapped k-mer features., *PLoS computational biology* 10 (7) (2014) e1003711. [PubMed: 25033408]
- [133]. Zhou J, Troyanskaya OG, Predicting effects of noncoding variants with deep learning based sequence model, *Nature Methods* 12 (10) (2015) 931–934. [PubMed: 26301843]
- [134]. Kelley DR, Snoek J, Rinn JL, Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks., *Genome research* 26 (7) (2016) 990–9. [PubMed: 27197224]
- [135]. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J, A general framework for estimating the relative pathogenicity of human genetic variants, *Nature Genetics* 46 (3) (2014) 310–315. [PubMed: 24487276]
- [136]. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD, A spectral approach integrating functional genomic annotations for coding and non-coding variants, *Nature Genetics* 48 (2) (2016) 214–220. [PubMed: 26727659]
- [137]. Gronau I, Arbiza L, Mohammed J, Siepel A, Inference of natural selection from interspersed genomic elements based on polymorphism and divergence., *Molecular biology and evolution* 30 (5) (2013) 1159–71. [PubMed: 23386628]
- [138]. Gulko B, Hubisz MJ, Gronau I, Siepel A, A method for calculating probabilities of fitness consequences for point mutations across the human genome, *Nature Genetics* 47 (3) (2015) 276–283. [PubMed: 25599402]
- [139]. Huang Y-F, Gulko B, Siepel A, Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data, *Nature Genetics* 49 (4) (2017) 618–624. [PubMed: 28288115]
- [140]. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al., Science forum: the human cell atlas, *Elife* 6 (2017) e27041. [PubMed: 29206104]
- [141]. Clevers H, Rafelski S, Elowitz M, Klein A, Shendure J, Trapnell C, Lein E, Lundberg E, Uhlen M, Martinez-Arias A, et al., What is your conceptual definition of cell type in the context of a mature organism?, *Cell Systems* 4 (3) (2017) 255–259. [PubMed: 28334573]
- [142]. Kelsey G, Stegle O, Reik W, Single-cell epigenomics: Recording the past and predicting the future, *Science* 358 (6359) (2017) 69–75. [PubMed: 28983045]
- [143]. Gawad C, Koh W, Quake SR, Single-cell genome sequencing: current state of the science, *Nature Reviews Genetics* 17 (3) (2016) 175.
- [144]. Schwartzman O, Tanay A, Single-cell epigenomics: techniques and emerging applications, *Nature Reviews Genetics* 16 (12) (2015) 716.
- [145]. Stegle O, Teichmann SA, Marioni JC, Computational and analytical challenges in single-cell transcriptomics, *Nature Reviews Genetics* 16 (3) (2015) 133.

- [146]. Wu M, Singh AK, Single-cell protein analysis, *Current opinion in biotechnology* 23 (1) (2012) 83–88. [PubMed: 22189001]
- [147]. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, Guo G, Irizarry R, Kharchenko P, Kim J, Orkin S, et al., Challenges and emerging directions in single-cell analysis, *Genome biology* 18 (1) (2017) 84. [PubMed: 28482897]
- [148]. Shapiro E, Biezuner T, Linnarsson S, Single-cell sequencing-based technologies will revolutionize whole-organism science, *Nature Reviews Genetics* 14 (9) (2013) 618.
- [149]. Poirion OB, Zhu X, Ching T, Garmire L, Single-cell transcriptomics bioinformatics and computational challenges, *Frontiers in genetics* 7 (2016) 163. [PubMed: 27708664]
- [150]. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al., Massively parallel digital transcriptional profiling of single cells, *Nature communications* 8 (2017) 14049.
- [151]. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglu S, Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning, *Nature methods* 14 (4) (2017) 414. [PubMed: 28263960]
- [152]. Pierson E, Yau C, Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis, *Genome biology* 16 (1) (2015) 241. [PubMed: 26527291]
- [153]. Cleary B, Cong L, Cheung A, Lander ES, Regev A, Efficient generation of transcriptomic profiles by random composite measurements, *Cell* 171 (6) (2017) 1424–1436. [PubMed: 29153835]
- [154]. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al., Sc3: consensus clustering of single-cell rna-seq data, *Nature methods* 14 (5) (2017) 483. [PubMed: 28346451]
- [155]. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R, Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nature biotechnology* 36 (5) (2018) 411.
- [156]. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al., mrna-seq whole-transcriptome analysis of a single cell, *Nature methods* 6 (5) (2009) 377. [PubMed: 19349980]
- [157]. Yotsukura S, Nomura S, Aburatani H, Tsuda K, et al., Celltree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data, *BMC bioinformatics* 17 (1) (2016) 363. [PubMed: 27620863]
- [158]. Zhang H, Lee CA, Li Z, Garbe JR, Eide CR, Petegrosso R, Kuang R, Tolar J, A multitask clustering approach for single-cell rna-seq analysis in recessive dystrophic epidermolysis bullosa, *PLoS computational biology* 14 (4) (2018) e1006053. [PubMed: 29630593]
- [159]. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity, *Nature methods* 11 (8) (2014) 817. [PubMed: 25042786]
- [160]. Rotem A, Ram O, Shosh N, Sperling RA, Goren A, Weitz DA, Bernstein BE, Single-cell chip-seq reveals cell subpopulations defined by chromatin state, *Nature biotechnology* 33 (11) (2015) 1165.
- [161]. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ, Single-cell chromatin accessibility reveals principles of regulatory variation, *Nature* 523 (7561) (2015) 486. [PubMed: 26083756]
- [162]. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing, *Science* 348 (6237) (2015) 910–914. [PubMed: 25953818]
- [163]. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P, Single-cell hi-c reveals cell-to-cell variability in chromosome structure, *Nature* 502 (7469) (2013) 59. [PubMed: 24067610]
- [164]. Frei AP, Bava F-A, Zunder ER, Hsieh EW, Chen S-Y, Nolan GP, Gherardini PF, Highly multiplexed simultaneous detection of rnas and proteins in single cells, *Nature methods* 13 (3) (2016) 269. [PubMed: 26808670]
- [165]. Fessenden M, *Metabolomics: Small molecules, single cells* (2016).

- [166]. Macaulay IC, Ponting CP, Voet T, Single-cell multiomics: multiple measurements from single cells, *Trends in Genetics* 33 (2) (2017) 155–168. [PubMed: 28089370]
- [167]. Bock C, Farlik M, Sheffield NC, Multi-omics of single cells: strategies and applications, *Trends in biotechnology* 34 (8) (2016) 605–608. [PubMed: 27212022]
- [168]. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al., Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity, *Nature methods* 13 (3) (2016) 229. [PubMed: 26752769]
- [169]. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, et al., Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas, *Cell research* 26 (3) (2016) 304. [PubMed: 26902283]
- [170]. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al., G&t-seq: parallel sequencing of single-cell genomes and transcriptomes, *Nature methods* 12 (6) (2015) 519. [PubMed: 25915121]
- [171]. Han KY, Kim K-T, Joung J-G, Son D-S, Kim YJ, Jo A, Jeon H-J, Moon H-S, Yoo CE, Chung W, et al., Sidr: simultaneous isolation and parallel sequencing of genomic dna and total rna from single cells, *Genome research* 28 (1) (2018) 75–87. [PubMed: 29208629]
- [172]. Witten DM, Tibshirani R, Hastie T, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics* 10 (3) (2009) 515–534. [PubMed: 19377034]
- [173]. Waaijenborg S, de Witt Hamer PCV, Zwinderman AH, Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis, *Statistical applications in genetics and molecular biology* 7 (1).
- [174]. Le Cao K-A, Martin PG, Robert-Granic C, Besse P, Sparse canonical methods for biological data integration: application to a cross-platform study, *BMC bioinformatics* 10 (1) (2009) 34. [PubMed: 19171069]
- [175]. van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D, Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data, *BioRxiv* (2017) 111591.
- [176]. Li WV, Li JJ, An accurate and robust imputation method scimpute for single-cell rna-seq data, *Nature Communications* 9 (1) (2018) 997.
- [177]. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, Tan DS, Robson P, Loh Y-H, Quake SR, et al., Single-cell multi-modal profiling reveals cellular epigenetic heterogeneity, *nature methods* 13 (10) (2016) 833. [PubMed: 27525975]
- [178]. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA, Multiplexed quantification of proteins and transcripts in single cells, *Nature biotechnology* 35 (10) (2017) 936.
- [179]. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P, Simultaneous epitope and transcriptome measurement in single cells, *nature methods* 14 (9) (2017) 865. [PubMed: 28759029]
- [180]. Welch JD, Hartemink AJ, Prins JF, Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics, *Genome biology* 18 (1) (2017) 138. [PubMed: 28738873]
- [181]. Iacono G, Mereu E, Guillaumet-Adkins A, Corominas R, Cusco I, Rodríguez-Esteban G, Gut M, Pérez-Jurado LA, Gut I, Heyn H, bigscale: an analytical framework for big-scale single-cell data, *Genome research* 28 (6) (2018) 878–890. [PubMed: 29724792]
- [182]. Wolf FA, Angerer P, Theis FJ, Scanpy: large-scale single-cell gene expression data analysis, *Genome biology* 19 (1) (2018) 15. [PubMed: 29409532]
- [183]. Lin C, Jain S, Kim H, Bar-Joseph Z, Using neural networks for reducing the dimensions of single-cell rna-seq data, *Nucleic acids research* 45 (17) (2017) e156–e156. [PubMed: 28973464]
- [184]. Amodio M, Srinivasan K, van Dijk D, Mohsen H, Yim K, Muhle R, Moon KR, Kaech S, Sowell R, Montgomery R, et al., Exploring single-cell data with multitasking deep neural networks, *bioRxiv* (2017) 237065.

- [185]. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467.
- [186]. Consortium G, et al., The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans, *Science* 348 (6235) (2015) 648–660. [PubMed: 25954001]
- [187]. Typas A, Sourjik V, Bacterial protein networks: properties and functions, *Nature Reviews Microbiology* 13 (9) (2015) 559. [PubMed: 26256789]
- [188]. Gligorijevic V, Janjić V, Przulj N, Integration of molecular network data reconstructs gene ontology, *Bioinformatics* 30 (17) (2014) i594–i600. [PubMed: 25161252]
- [189]. Zitnik M, Zupan B, Matrix factorization-based data fusion for drug-induced liver injury prediction, *Systems Biomedicine* 2 (1) (2014) 16–22.
- [190]. Zitnik M, Zupan B, Matrix factorization-based data fusion for gene function prediction in bakers yeast and slime mold, in: *Pacific Symposium on Biocomputing*, NIH Public Access, 2014, p. 400.
- [191]. Gligorijevic V, Malod-Dognin N, Przulj N, Fuse: multiple network alignment via data fusion, *Bioinformatics* 32 (8) (2015) 1195–1203. [PubMed: 26668003]
- [192]. Strazar M, Zitnik M, Zupan B, Ule J, Curk T, Orthogonal matrix factorization enables integrative analysis of multiple rna binding proteins, *Bioinformatics* 32 (10) (2016) 1527–1535. [PubMed: 26787667]
- [193]. Gligorijevic V, Malod-Dognin N, Przulj N, Patient-specific data fusion for cancer stratification and personalised treatment, in: *Pacific Symposium on Biocomputing*, World Scientific, 2016, pp. 321–332.
- [194]. Zitnik M, Nam EA, Dinh C, Kuspa A, Shaulsky G, Zupan B, Gene prioritization by compressive data fusion and chaining, *PLoS Computational Biology* 11 (10) (2015) e1004552.
- [195]. Zhang C, Freddolino PL, Zhang Y, Cofactor: improved protein function prediction by combining structure, sequence and protein–protein interaction information, *Nucleic Acids Research* (2017) gkx366.
- [196]. Wan C, Lees JG, Minneci F, Orengo CA, Jones DT, Analysis of temporal transcription expression profiles reveal links between protein function and developmental stages of drosophila melanogaster, *PLoS Computational Biology* 13 (10) (2017) e1005791. [PubMed: 29045400]
- [197]. Amar D, Shamir R, Constructing module maps for integrated analysis of heterogeneous biological networks, *Nucleic Acids Research* 42 (7) (2014) 4208–4219. [PubMed: 24497192]
- [198]. Manichaikul A, Ghamisari L, Hom EF, Lin C, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, et al., Metabolic network analysis integrated with transcript verification for sequenced genomes, *Nature Methods* 6 (8) (2009) 589–592. [PubMed: 19597503]
- [199]. Kuzmin E, VanderSluis B, Wang W, Tan G, Deshpande R, Chen Y, Usaj M, Balint A, Usaj MM, van Leeuwen J, et al., Systematic analysis of complex genetic interactions, *Science* 360 (6386) (2018) eaao1729.
- [200]. Gaudet P, Livstone MS, Lewis SE, Thomas PD, Phylogenetic-based propagation of functional annotations within the gene ontology consortium, *Briefings in Bioinformatics* 12 (5) (2011) 449–462. [PubMed: 21873635]
- [201]. Konc J, Janežič D, Binding site comparison for function prediction and pharmaceutical discovery, *Current Opinion in Structural Biology* 25 (2014) 34–39. [PubMed: 24878342]
- [202]. You R, Zhu S, Deeptext2go: Improving large-scale protein function prediction with deep semantic text representation, in: *IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, 2017, pp. 42–49.
- [203]. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al., Gene ontology: tool for the unification of biology, *Nature Genetics* 25 (1) (2000) 25. [PubMed: 10802651]
- [204]. Cho H, Berger B, Peng J, Compact integration of multi-network topology for functional analysis of genes, *Cell Systems* 3 (6) (2016) 540–548. [PubMed: 27889536]
- [205]. Nickel M, Tresp V, Kriegel H-P, A three-way model for collective learning on multi-relational data, in: *ICML*, Vol. 11, 2011, pp. 809–816.

- [206]. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al., A large-scale evaluation of computational protein function prediction, *Nature Methods* 10 (3) (2013) 221. [PubMed: 23353650]
- [207]. Li W, Liu C-C, Zhang T, Li H, Waterman MS, Zhou XJ, Integrative analysis of many weighted co-expression networks using tensor computation, *PLoS Computational Biology* 7 (6) (2011) e1001106. [PubMed: 21698123]
- [208]. Ou-Yang L, Wu M, Zhang X-F, Dai D-Q, Li X-L, Yan H, A two-layer integration framework for protein complex detection, *BMC Bioinformatics* 17 (1) (2016) 100.
- [209]. Bugge K, Papaleo E, Haxholm GW, Hopper JT, Robinson CV, Olsen JG, Lindorff-Larsen K, Kragelund BB, A combined computational and structural model of the full-length human prolactin receptor, *Nature Communications* 7 (2016) 11578.
- [210]. Shi Y, Pellarin R, Fridy PC, Fernandez-Martinez J, Thompson MK, Li Y, Wang QJ, Sali A, Rout MP, Chait BT, A strategy for dissecting the architectures of native macromolecular assemblies, *Nature Methods* 12 (12) (2015) 1135–1138. [PubMed: 26436480]
- [211]. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG, Discovery of biological networks from diverse functional genomic data, *Genome Biology* 6 (13) (2005) R114. [PubMed: 16420673]
- [212]. Ray P, Zheng L, Lucas J, Carin L, Bayesian joint analysis of heterogeneous genomics data, *Bioinformatics* 30 (10) (2014) 1370–1376. [PubMed: 24489367]
- [213]. Ori A, Toyama BH, Harris MS, Bock T, Iskar M, Bork P, Ingolia NT, Hetzer MW, Beck M, Integrated transcriptome and proteome analyses reveal organ-specific proteome deterioration in old rats, *Cell Systems* 1 (3) (2015) 224–237. [PubMed: 27135913]
- [214]. Andrews SV, Ellis SE, Bakulski KM, Sheppard B, Croen LA, Hertz-Picciotto I, Newschaffer CJ, Feinberg AP, Arking DE, Ladd-Acosta C, et al., Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder, *Nature Communications* 8 (1) (2017) 1011.
- [215]. Deng J, Deng L, Su S, Zhang M, Lin X, Wei L, Minai AA, Hassett DJ, Lu LJ, Investigating the predictability of essential genes across distantly related organisms using an integrative approach, *Nucleic Acids Research* 39 (3) (2010) 795–807. [PubMed: 20870748]
- [216]. Hooghe B, Broos S, Van Roy F, De Bleser P, A flexible integrative approach based on random forest improves prediction of transcription factor binding sites, *Nucleic Acids Research* 40 (14) (2012) e106–e106. [PubMed: 22492513]
- [217]. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, Agius P, Huse JT, Holland EC, Leslie CS, Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma, *Molecular Systems Biology* 8 (1) (2012) 605. [PubMed: 22929615]
- [218]. Penfold CA, Millar JB, Wild DL, Inferring orthologous gene regulatory networks using interspecies data fusion, *Bioinformatics* 31 (12) (2015) i97–i105. [PubMed: 26072515]
- [219]. Imam S, Noguera DR, Donohue TJ, An integrated approach to reconstructing genome-scale transcriptional regulatory networks, *PLoS Computational Biology* 11 (2) (2015) e1004103. [PubMed: 25723545]
- [220]. Ihekwaba AE, Mura I, Walshaw J, Peck MW, Barker GC, An integrative approach to computational modelling of the gene regulatory network controlling *Clostridium botulinum* type A1 toxin production, *PLoS Computational Biology* 12 (11) (2016) e1005205. [PubMed: 27855161]
- [221]. Franke L, Van Bakel H, Diosdado B, Van Belzen M, Wapenaar M, Wijmenga C, TEAM: a tool for the integration of expression, and linkage and association maps, *European Journal of Human Genetics* 12 (8) (2004) 633. [PubMed: 15114375]
- [222]. Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y, eXtasy: variant prioritization by genomic data fusion, *Nature Methods* 10 (11) (2013) 1083–1084. [PubMed: 24076761]
- [223]. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS, A statistical framework for genomic data fusion, *Bioinformatics* 20 (16) (2004) 2626–2635. [PubMed: 15130933]

- [224]. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, et al., Gene prioritization through genomic data fusion, *Nature Biotechnology* 24 (5) (2006) 537–544.
- [225]. Tranchevent L-C, Ardeshirdavani A, ElShal S, Alcaide D, Aerts J, Auboeuf D, Moreau Y, Candidate gene prioritization with endeavour, *Nucleic Acids Research* 44 (W1) (2016) W117–W121. [PubMed: 27131783]
- [226]. Koehler S, Bauer S, Horn D, Robinson PN, Walking the interactome for prioritization of candidate disease genes, *The American Journal of Human Genetics* 82 (4) (2008) 949–958. [PubMed: 18371930]
- [227]. De Bie T, Tranchevent L-C, Van Oeffelen LM, Moreau Y, Kernel-based data fusion for gene prioritization, *Bioinformatics* 23 (13) (2007) i125–i132. [PubMed: 17646288]
- [228]. Chen J, Bardes EE, Aronow BJ, Jegga AG, ToppGene suite for gene list enrichment analysis and candidate gene prioritization, *Nucleic Acids Research* 37 (suppl 2) (2009) W305–W311. [PubMed: 19465376]
- [229]. Robinson PN, Koehler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, et al., Improved exome prioritization of disease genes through cross-species phenotype comparison, *Genome Research* 24 (2) (2014) 340–348. [PubMed: 24162188]
- [230]. Simões SN, Martins DC, Pereira CA, Hashimoto RF, Brentani H, NERI: network-medicine based integrative approach for disease gene prioritization by relative importance, *BMC Bioinformatics* 16 (19) (2015) S9.
- [231]. Himmelstein DS, Baranzini SE, Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes, *PLoS Computational Biology* 11 (7) (2015) e1004259. [PubMed: 26158728]
- [232]. Kumar AA, Van Laer L, Alaerts M, Ardeshirdavani A, Moreau Y, Laukens K, Loeys B, Vandeweyer G, pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion, *Bioinformatics* 1 (2018) 9.
- [233]. Pandey G, Zhang B, Chang AN, Myers CL, Zhu J, Kumar V, Schadt EE, An integrative multi-network and multi-classifier approach to predict genetic interactions, *PLoS Computational Biology* 6 (9) (2010) e1000928. [PubMed: 20838583]
- [234]. Bonnet E, Calzone L, Michoel T, Integrative multi-omics module network inference with lemon-tree, *PLoS Computational Biology* 11 (2) (2015) e1003983. [PubMed: 25679508]
- [235]. Heiser LM, Wang NJ, Talcott CL, Laderoute KR, Knapp M, Guan Y, Hu Z, Ziyad S, Weber BL, Laquerre S, et al., Integrated analysis of breast cancer cell lines reveals unique signaling pathways, *Genome Biology* 10 (3) (2009) R31. [PubMed: 19317917]
- [236]. Nibbe RK, Koyuturk M, Chance MR, An integrative-omics approach to identify functional sub-networks in human colorectal cancer, *PLoS Computational Biology* 6 (1) (2010) e1000639. [PubMed: 20090827]
- [237]. Rudolph JD, de Graauw M, van de Water B, Geiger T, Sharan R, Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks, *Cell Systems* 3 (6) (2016) 585–593. [PubMed: 28009266]
- [238]. Piccolo SR, Hoffman LM, Conner T, Shrestha G, Cohen AL, Marks JR, Neumayer LA, Agarwal CA, Beckerle MC, Andrulis IL, et al., Integrative analyses reveal signaling pathways underlying familial breast cancer susceptibility, *Molecular Systems Biology* 12 (3) (2016) 860. [PubMed: 26969729]
- [239]. Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T, A gene ontology inferred from molecular networks, *Nature Biotechnology* 31 (1) (2013) 38–45.
- [240]. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA, Uberon, an integrative multi-species anatomy ontology, *Genome Biology* 13 (1) (2012) R5.
- [241]. Kulmanov M, Khan MA, Hoehndorf R, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, *Bioinformatics* 34 (2017) 660–668.
- [242]. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T, Using deep learning to model the hierarchical structure and function of a cell, *Nature Methods* 15 (4) (2018) 290. [PubMed: 29505029]

- [243]. Rolland T, Tasan M, Charloreaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, et al., A proteome-scale map of the human interactome network, *Cell* 159 (5) (2014) 1212–1226. [PubMed: 25416956]
- [244]. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M, A Bayesian networks approach for predicting protein-protein interactions from genomic data., *Science (New York, N.Y.)* 302 (5644) (2003) 449–53.
- [245]. Lundberg SM, Tu WB, Raught B, Penn LZ, Hoffman MM, Lee S-I, ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data, *Genome Biology* 17 (1) (2016) 82. [PubMed: 27139377]
- [246]. Drew K, Lee C, Huizar RL, Tu F, Borgeson B, McWhite CD, Ma Y, Wallingford JB, Marcotte EM, Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes, *Molecular Systems Biology* 13 (6) (2017) 932. [PubMed: 28596423]
- [247]. Li Y, Patra JC, Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network, *Bioinformatics* 26 (9) (2010) 1219–1224. [PubMed: 20215462]
- [248]. Blatti C, Sinha S, Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks, *Bioinformatics* 32 (14) (2016) 2167–2175. [PubMed: 27153592]
- [249]. Liu Y, Zeng X, He Z, Zou Q, Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources, *IEEE Transactions on Computational Biology and Bioinformatics* 14 (4) (2017) 905–915. [PubMed: 27076459]
- [250]. Scannell JW, Blanckley A, Boldon H, Warrington B, Diagnosing the decline in pharmaceutical r&d efficiency, *Nature Reviews Drug Discovery* 11 (3) (2012) 191. [PubMed: 22378269]
- [251]. Yeh PJ, Hegreness MJ, Aiden AP, Kishony R, Drug interactions and the evolution of antibiotic resistance, *Nature Reviews Microbiology* 7 (6) (2009) 460. [PubMed: 19444248]
- [252]. Li J, et al., A survey of current trends in computational drug repositioning, *Briefings in Bioinformatics* 17 (1) (2015) 2–12. [PubMed: 25832646]
- [253]. Donald BR, Algorithms in structural molecular biology, MIT Press, 2011.
- [254]. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK, Relating protein pharmacology by ligand chemistry, *Nature Biotechnology* 25 (2) (2007) 197.
- [255]. Bleakley K, Yamanishi Y, Supervised prediction of drug–target interactions using bipartite local models, *Bioinformatics* 25 (18) (2009) 2397–2403. [PubMed: 19605421]
- [256]. van Laarhoven T, Nabuurs SB, Marchiori E, Gaussian interaction pro- file kernels for predicting drug–target interaction, *Bioinformatics* 27 (21) (2011) 3036–3043. [PubMed: 21893517]
- [257]. Wang S, Peng J, Network-assisted target identification for haplo insufficiency and homozygous profiling screens, *PLoS Computational Biology* 13 (6) (2017) e1005553. [PubMed: 28574983]
- [258]. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y, Relating drug–protein interaction network with drug side effects, *Bioinformatics* 28 (18) (2012) i522–i528. [PubMed: 22962476]
- [259]. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, et al., Discov- ery of drug mode of action and drug repositioning from transcriptional responses, *Proceedings of the National Academy of Sciences* 107 (33) (2010) 14621–14626.
- [260]. Wang W, Yang S, Zhang X, Li J, Drug repositioning by integrating target information through a heterogeneous network model, *Bioinformatics* 30 (20) (2014) 2923–2930. [PubMed: 24974205]
- [261]. Yang F, Xu J, Zeng J, Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data, in: *Pacific Symposium on Biocomputing*, World Scientific, 2014, pp. 148–159.
- [262]. Gonen M, Khan S, Kaski S, Kernelized bayesian matrix factorization, in: *ICML*, 2013, pp. 864–872.
- [263]. Zhang X, Li L, Ng MK, Zhang S, Drug–target interaction prediction by integrating multiview network data, *Computational Biology and Chemistry* 69 (2017) 185–193. [PubMed: 28648470]
- [264]. Breinig M, Klein FA, Huber W, Boutros M, A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells, *Molecular Systems Biology* 11 (12) (2015) 846. [PubMed: 26700849]

- [265]. Lee S, Zhang C, Liu Z, Klevstig M, Mukhopadhyay B, Bergentall M, Cinar R, Sta<sup>o</sup>hlman M, Sikanic N, Park JK, et al., Network analyses identify liver-specific targets for treating liver diseases, *Molecular Systems Biology* 13 (8) (2017) 938.
- [266]. Sun Y, Han J, Yan X, Yu PS, Wu T, Pathsim: Meta path-based top- k similarity search in heterogeneous information networks, *Proceedings of the VLDB* 4 (11) (2011) 992–1003.
- [267]. Fu G, Ding Y, Seal A, Chen B, Sun Y, Bolton E, Predicting drug target interactions using meta-path-based semantic network analysis, *BMC Bioinformatics* 17 (1) (2016) 160. [PubMed: 27071755]
- [268]. Zheng X, Ding H, Mamitsuka H, Zhu S, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, in: *Proceedings of the KDD, ACM, 2013*, pp. 1025–1033.
- [269]. Narita A, Hayashi K, Tomioka R, Kashima H, Tensor factorization using auxiliary information, *Data Mining and Knowledge Discovery* 25 (2) (2012) 298–324.
- [270]. Zitnik M, Zupan B, Collective pairwise classification for multi-way analysis of disease and drug data, in: *Pacific Symposium on Biocomputing, Vol. 21, NIH Public Access, 2016*, p. 81.
- [271]. Hinton GE, Salakhutdinov RR, Reducing the dimensionality of data with neural networks, *Science* 313 (5786) (2006) 504–507. [PubMed: 16873662]
- [272]. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, *Nature Communications* 8 (2017) 573.
- [273]. Vilar S, Uriarte E, Santana L, Lorberbaum T, Hripcsak G, Friedman C, Tatonetti NP, Similarity-based modeling in large-scale prediction of drug-drug interactions, *Nature Protocols* 9 (9) (2014) 2147. [PubMed: 25122524]
- [274]. Cheng F, Zhao Z, Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, *Journal of the American Medical Informatics Association* 21 (e2) (2014) e278–e286. [PubMed: 24644270]
- [275]. Sridhar D, Fakhraei S, Getoor L, A probabilistic approach for collective similarity-based drug–drug interaction prediction, *Bioinformatics* 32 (20) (2016) 3175–3182. [PubMed: 27354693]
- [276]. Han K, et al., Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions, *Nature Biotechnology*
- [277]. Jia J, et al., Mechanisms of drug combinations: interaction and network perspectives, *Nature Reviews Drug Discovery* 8 (2) (2009) 111–128. [PubMed: 19180105]
- [278]. Sun Y, et al., Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer, *Nature Communications* 6 (2015) 8481.
- [279]. Woo HG, Choi J-H, Yoon S, Jee BA, Cho EJ, Lee J-H, Yu SJ, Yoon J-H, Yi N-J, Lee K-W, et al., Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer, *Nature Communications* 8 (1) (2017) 839.
- [280]. Chen X, et al., NLLSS: predicting synergistic drug combinations based on semi-supervised learning, *PLoS Computational Biology* 12 (7) (2016) e1004975. [PubMed: 27415801]
- [281]. Kantor ED, et al., Trends in prescription drug use among adults in the United States from 1999–2012, *JAMA* 314 (17) (2015) 1818–1830. [PubMed: 26529160]
- [282]. Ryall KA, Tan AC, Systems biology approaches for advancing the discovery of effective drug combinations, *Journal of Cheminformatics* 7 (1) (2015) 7. [PubMed: 25741385]
- [283]. Loewe S, The problem of synergism and antagonism of combined drugs, *Arzneimittel-Forschung* 3 (1953) 285–290. [PubMed: 13081480]
- [284]. Lewis R, et al., Synergy maps: exploring compound combinations using network-based visualization, *Journal of Cheminformatics* 7 (1) (2015) 36. [PubMed: 26236402]
- [285]. Bansal M, et al., A community computational challenge to predict the activity of pairs of compounds, *Nature Biotechnology* 32 (12) (2014) 1213–1222.
- [286]. Takeda T, et al., Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge, *Journal of Cheminformatics* 9 (1) (2017) 16. [PubMed: 28316654]

- [287]. Huang L, et al., Drugcomboranker: drug combination discovery based on target network analysis, *Bioinformatics* 30 (12) (2014) i228–i236. [PubMed: 24931988]
- [288]. Huang H, et al., Systematic prediction of drug combinations based on clinical side-effects, *Scientific Reports* 4.
- [289]. Sun Y, et al., Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer, *Nature Commun* 6 (2015) 8481. [PubMed: 26412466]
- [290]. Zitnik M, Zupan B, Collective pairwise classification for multi-way analysis of disease and drug data, in: *Pacific Symposium on Biocomputing*, Vol. 21, 2016, p. 81.
- [291]. Chen D, et al., Synergy evaluation by a pathway–pathway interaction network: a new way to predict drug combination, *Molecular BioSystems* 12 (2) (2016) 614–623. [PubMed: 26687590]
- [292]. Shi J-Y, et al., Predicting combinative drug pairs towards realistic screening via integrating heterogeneous features, *BMC Bioinformatics* 18 (12) (2017) 409. [PubMed: 29072137]
- [293]. Cheng F, Zhao Z, Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, *JAMIA* 21 (e2) (2014) e278–e286. [PubMed: 24644270]
- [294]. Zheng W, Lin H, Luo L, Zhao Z, Li Z, Zhang Y, Yang Z, Wang J, An attention-based effective neural model for drug-drug interactions extraction, *BMC Bioinformatics* 18 (1) (2017) 445. [PubMed: 29017459]
- [295]. Zhao Z, Yang Z, Luo L, Lin H, Wang J, Drug drug interaction extraction from biomedical literature using syntax convolutional neural network, *Bioinformatics* 32 (22) (2016) 3444–3453. [PubMed: 27466626]
- [296]. Gottlieb A, et al., INDI: a computational framework for inferring drug interactions and their associated recommendations, *Mol. Syst. Biol* 8 (1) (2012) 592. [PubMed: 22806140]
- [297]. Vilar S, et al., Drug-drug interaction through molecular structure similarity analysis, *JAMIA* 19 (6) (2012) 1066–1074. [PubMed: 22647690]
- [298]. Li X, et al., Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles, *Artificial Intelligence in Medicine*
- [299]. Zhang P, Wang F, Hu J, Sorrentino R, Label propagation prediction of drug-drug interactions based on clinical side effects, *Scientific Reports* 5.
- [300]. Ferdousi R, et al., Computational prediction of drug-drug interactions based on drugs functional similarities, *Journal of Biomedical Informatics* 70 (2017) 54–64. [PubMed: 28465082]
- [301]. Zhang W, et al., Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, *BMC Bioinform* 18 (1) (2017) 18.
- [302]. Ma T, Xiao C, Zhou J, Wang F, Drug similarity integration through attentive multi-view graph auto-encoders, in: *Proceedings of IJCAI*, 2018, pp. 1–7.
- [303]. Ryu JY, Kim HU, Lee SY, Deep learning improves prediction of drug–drug and drug–food interactions, *Proceedings of the National Academy of Sciences* 115 (18) (2018) E4304–E4311.
- [304]. Hamilton WL, Ying R, Leskovec J, Representation learning on graphs: Methods and applications, *IEEE Data Engineering Bulletin*
- [305]. Guney E, Menche J, Vidal M, Barabasi A-L, Network-based in silico drug efficacy screening, *Nature Communications* 7 (2016) 10331.
- [306]. Zitnik M, Janjic V, Larminie C, Zupan B, Natasa P, Discovering disease-disease associations by fusing systems-level molecular data, *Scientific Reports* 3.
- [307]. Li J, Zhu X, Chen JY, Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts, *PLoS Computational Biology* 5 (7) (2009) e1000450. [PubMed: 19649302]
- [308]. Wu Z, Wang Y, Chen L, Network-based drug repositioning, *Molecular Bio Systems* 9 (6) (2013) 1268–1281.
- [309]. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y, Prediction of drug-target interactions and drug repositioning via network-based inference, *PLoS Computational Biology* 8 (5) (2012) e1002503. [PubMed: 22589709]
- [310]. Zhao S, Li S, A co-module approach for elucidating drug–disease associations and revealing their molecular basis, *Bioinformatics* 28 (7) (2012) 955–961. [PubMed: 22285830]

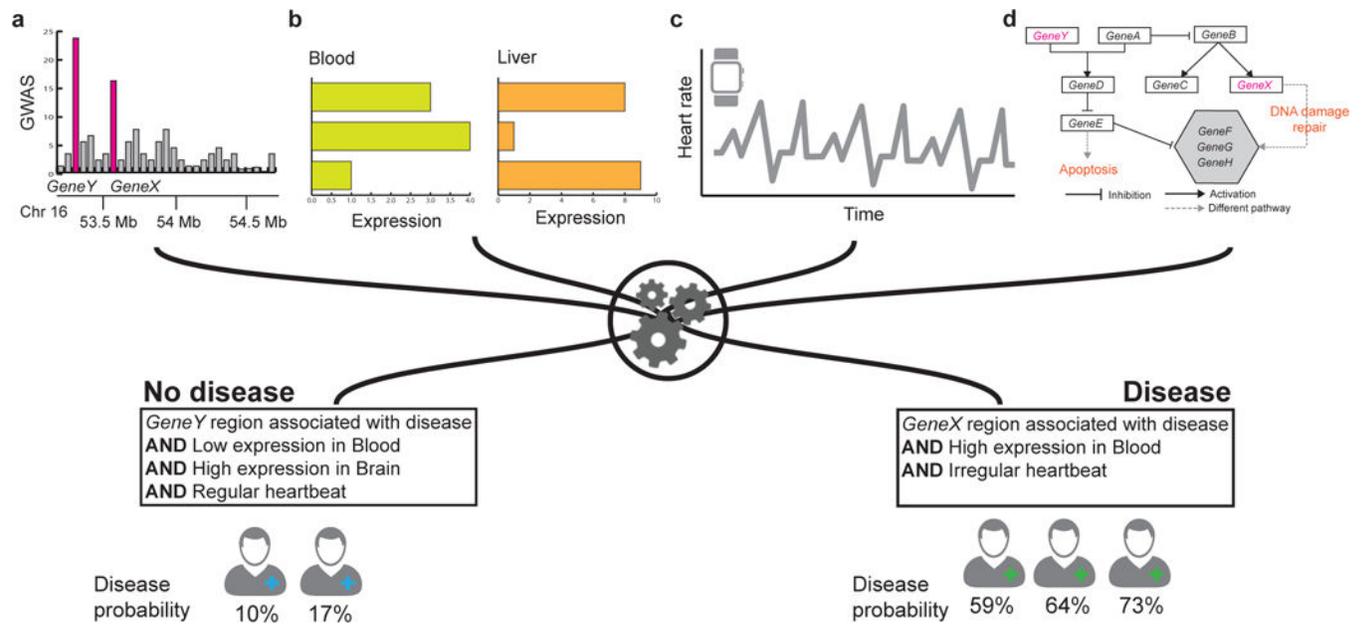
- [311]. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet- Cordero A, Sage J, Butte AJ, Discovery and preclinical validation of drug indications using compendia of public gene expression data, *Science Translational Medicine* 3 (96) (2011) 96ra77–96ra77.
- [312]. Stanfield Z, Coskun M, Koyuturk M, Drug response prediction as a link prediction problem, *Scientific Reports* 7 (2017) 40321. [PubMed: 28067293]
- [313]. Fung KW, Jao CS, Demner-Fushman D, Extracting drug indication information from structured product labels using natural language processing, *Journal of the American Medical Informatics Association* 20 (3) (2013) 482–488. [PubMed: 23475786]
- [314]. Zhang P, Wang F, Hu J, Sorrentino R, Exploring the relationship between drug side-effects and therapeutic indications, in: *Proceedings of the AMIA Annual Symposium*, Vol. 2013, American Medical Informatics Association, 2013, p. 1568.
- [315]. Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin A-C, Bork P, Systematic identification of proteins that elicit drug side effects, *Molecular Systems Biology* 9 (1) (2013) 663. [PubMed: 23632385]
- [316]. Wang F, Zhang P, Cao N, Hu J, Sorrentino R, Exploring the associations between drug side-effects and therapeutic indications, *Journal of Biomedical Informatics* 51 (2014) 15–23. [PubMed: 24727480]
- [317]. Gottlieb A, Stein GY, Ruppin E, Sharan R, PREDICT: a method for inferring novel drug indications with application to personalized medicine, *Molecular Systems Biology* 7 (1) (2011) 496. [PubMed: 21654673]
- [318]. Zhang P, Agarwal P, Obradovic Z, Computational drug repositioning by ranking and integrating multiple data sources, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, pp. 579–594.
- [319]. Li J, Lu Z, Pathway-based drug repositioning using causal inference, *BMC Bioinformatics* 14 (16) (2013) S3.
- [320]. Yu L, Su R, Wang B, Zhang L, Zou Y, Zhang J, Gao L, Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk, *IEEE Transactions on Computational Biology and Bioinformatics* 14 (4) (2017) 966–977. [PubMed: 27076463]
- [321]. Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, Pan Y, Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm, *Bioinformatics* 32 (17) (2016) 2664–2671. [PubMed: 27153662]
- [322]. Himmelstein DS, Lizée A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, *bioRxiv* (2017) 087619.
- [323]. Wang W, Yang S, Li J, Drug target predictions based on heterogeneous graph inference, in: *Pacific Symposium on Biocomputing*, World Scientific, 2013, pp. 53–64.
- [324]. Zhang P, Wang F, Hu J, Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity, in: *Proceedings of the AMIA Annual Symposium*, Vol. 2014, 2014, p. 1258.
- [325]. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, *Nature* 486 (7403) (2012) 346–352. [PubMed: 22522925]
- [326]. Cavalli FM, Remke M, Rampasek L, et al., Intertumoral heterogeneity within medulloblastoma subgroups, *Cancer Cell* 31 (6) (2017) 737 – 754.e6. [PubMed: 28609654]
- [327]. Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, Griffin C, Ozburn N, Chen M, Pan E, Koul D, et al., Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma, *Cancer Research* 65 (5) (2005) 1678–1686.
- [328]. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O’Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN, Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized

- by abnormalities in PDGFRA, IDH1, EGFR, and NF1, *Cancer Cell* 17 (1) (2010) 98 – 110. [PubMed: 20129251]
- [329]. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki- Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, et al., Comprehensive molecular portraits of human breast tumours, *Nature* 490 (7418) (2012) 61–70. [PubMed: 23000897]
- [330]. Hoadley KA, Yau C, Wolf DM, et al., Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin, *Cell* 158 (4) (2014) 929 – 944. [PubMed: 25109877]
- [331]. Shen R, Olshen AB, Ladanyi M, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, *Bioinformatics* 25 (22) (2009) 2906–2912. [PubMed: 19759197]
- [332]. Yuan Y, Savage RS, Markowitz F, Patient-specific data fusion defines prognostic cancer subtypes, *PLoS Computational Biology* 7 (10) (2011) e1002227. [PubMed: 22028636]
- [333]. de Vega WC, Erdman L, Vernon SD, Goldenberg A, Mc- Gowan PO, Integration of dna methylation and health scores identifies sub- types in myalgic encephalomyelitis/chronic fatigue syndrome, *Epigenomics* 10 (5) (2018) 539–557. [PubMed: 29692205]
- [334]. Zizzo AN, Erdman L, Feldman BM, Goldenberg A, Similarity network fusion: A novel application to making clinical diagnoses, *Rheumatic Disease Clinics of North America* 44 (2) (2018) 285 – 293, advanced Epidemiologic Methods for the Study of Rheumatic Diseases. [PubMed: 29622295]
- [335]. Stefanik L, Erdman L, Ameis SH, Foussias G, Mulsant BH, Behdinan T, Goldenberg A, O'Donnell LJ, Voineskos AN, Brain- behavior participant similarity networks among youth and emerging adults with schizophrenia spectrum, autism spectrum, or bipolar disorder and matched controls, *Neuro psychopharmacology* 43 (5) (2017) 1180–1188.
- [336]. Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, Robertson AG, Cherniack AD, Gupta M, Getz G, et al., Integrated genomic characterization of pancreatic ductal adenocarcinoma, *Cancer Cell* 32 (2) (2017) 185 – 203.e13. [PubMed: 28810144]
- [337]. Huang H-C, Chuang Y-Y, Chen C-S, Affinity aggregation for spectral clustering, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012, pp. 773–780.
- [338]. Pai S, Bader GD, Patient similarity networks for precision medicine, *Journal of Molecular Biology*
- [339]. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM, Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, *Bioinformatics* 26 (12) (2010) i237–i245. [PubMed: 20529912]
- [340]. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, et al., The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance, *Nature Biotechnology* 32 (9) (2014) 926.
- [341]. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC, Normalizing single-cell RNA sequencing data: challenges and opportunities, *Nature Methods*
- [342]. Hiranuma N, Lundberg SM, Lee S-I, AIControl: Replacing matched control experiments with machine learning improves ChIP-seq peak identification, *bioRxiv* (2018) 278762.
- [343]. Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C, SCnorm: robust normalization of single-cell RNA-seq data, *Nature Methods* 14 (6) (2017) 584–586. [PubMed: 28418000]
- [344]. Taroni JN, Greene CS, Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously, *bioRxiv* (2017) 118349.
- [345]. Wang B, Pourshafeie A, Zitnik M, Zhu J, Bustamante CD, Bat-zoglou S, Leskovec J, Network enhancement: a general method to de- noise weighted biological networks, *arXiv* (2018) 1805.03327.
- [346]. Milenkovic T, Przulj N, Uncovering biological network function via graphlet degree signatures, *Cancer Informatics* 6 (2008) CIN–S680.
- [347]. Benson AR, Gleich DF, Leskovec J, Higher-order organization of complex networks, *Science* 353 (6295) (2016) 163–166. [PubMed: 27387949]

- [348]. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R, Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development, *Nature Biotechnology* 35 (6) (2017) 551–560.
- [349]. Ribeiro MT, Singh S, Guestrin C, Why should i trust you?: Explaining the predictions of any classifier, in: *KDD, ACM*, 2016, pp. 1135– 1144.
- [350]. Lundberg SM, Lee S-I, A unified approach to interpreting model predictions, in: *NIPS*, 2017, pp. 4768–4777.
- [351]. Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Bengio Y, et al., A closer look at memorization in deep networks, in: *ICML*, 2017, pp. 1–10.
- [352]. Koh PW, Liang P, Understanding black-box predictions via influence functions, in: *ICML*, 2017, pp. 1–11.
- [353]. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, et al., Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery, *bioRxiv* (2017) 206540.
- [354]. Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB, Francke U, Naughton BT, Mountain JL, Wojcicki A, et al., Efficient replication of over 180 genetic associations with self-reported medical data, *PloS One* 6 (8) (2011) e23473. [PubMed: 21858135]

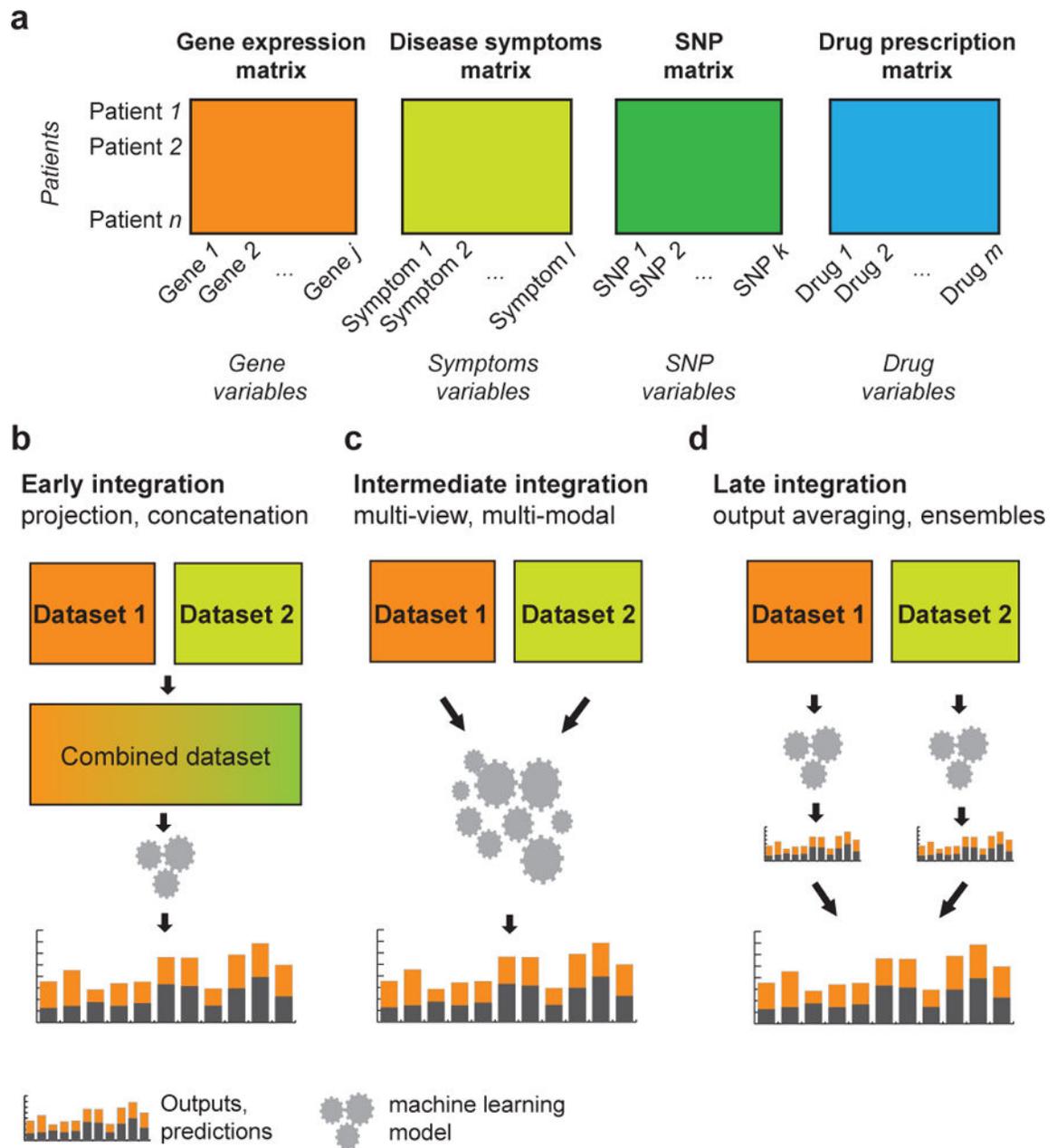
### Highlights

- New technologies measure biology and human health at scale and in multiple dimensions
- Large and diverse biomedical data present interesting challenges to machine learning
- Approaches combine different types of data to provide a comprehensive systems view
- Data integration creates a holistic picture of the cell, human health, and disease
- Advances in machine learning bring exciting future for biomedical data integration



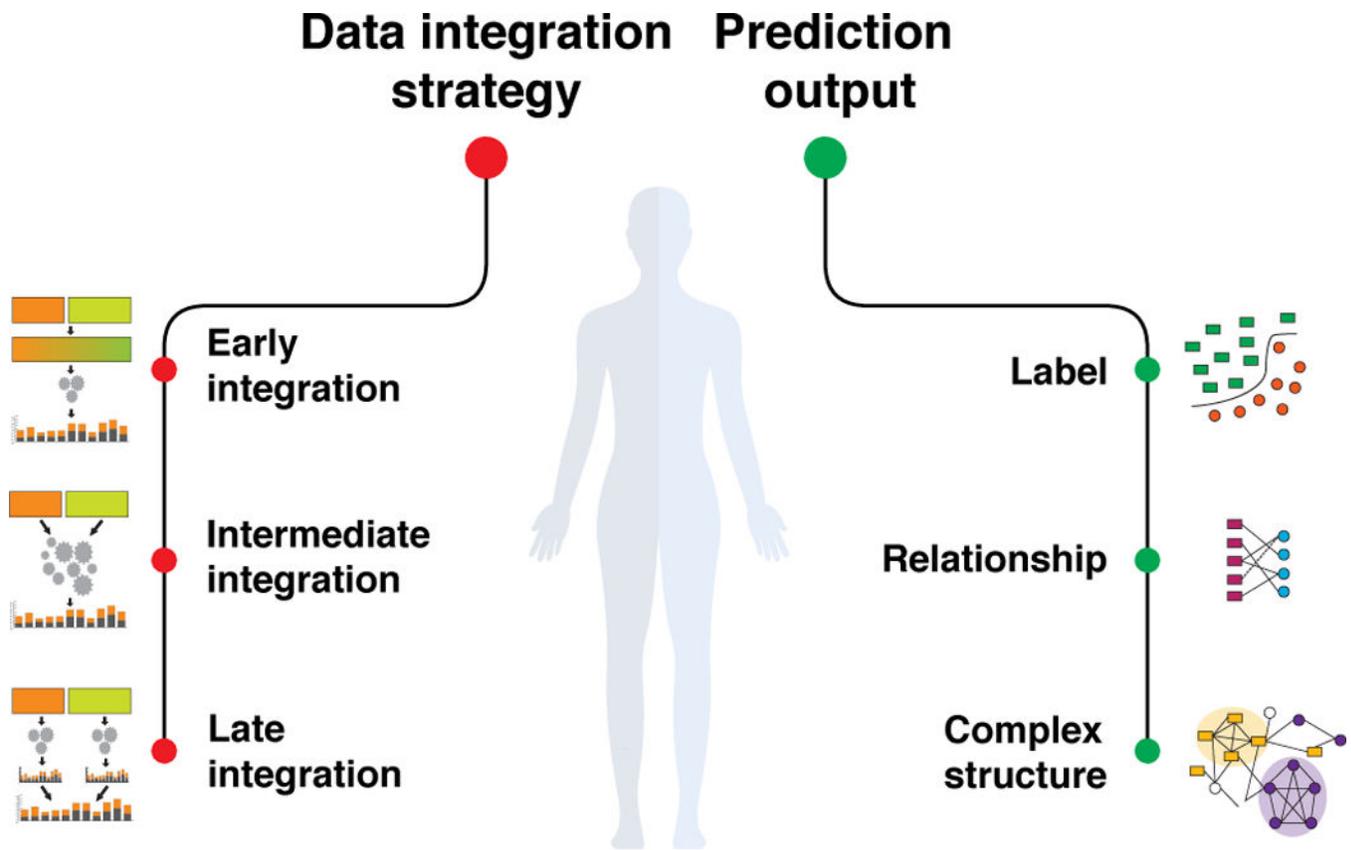
**Figure 1: The importance of data integration in biomedicine.**

Considering variation in only a single data type can miss many important patterns that can only be observed by considering multiple levels of biomedical data. Shown is a hypothetical example using disease diagnostics as a point of interest. When a new patient arrives to the clinic, (a) domain experts sequence the patient's genome and compare it with a database to identify mutations and disease-causing genes, (b) perform laboratory tests using tissue samples, and (c) process information about the patient's behavior and lifestyle. (d) The patient's genomic, transcriptomic, and lifestyle information is combined with curated databases of biomedical knowledge (e.g., disease and metabolic pathways). Finally, a machine learning algorithm predicts probability that the patient will develop a particular disease in near future. To make accurate prediction, the machine learning model needs to use many different types of data. This example illustrates that accurate prediction can only be made by analyzing multiple types of patient's data.



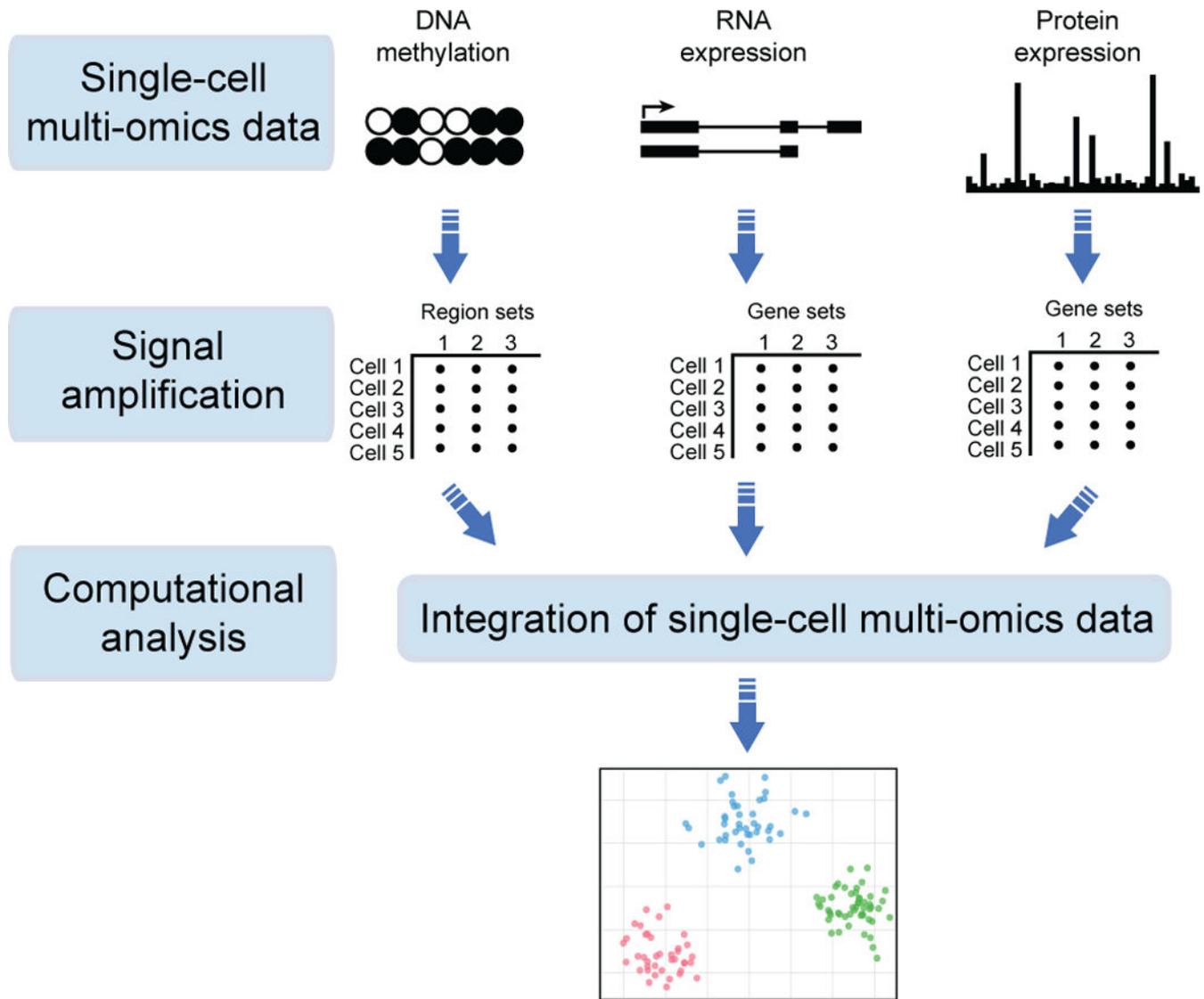
**Figure 2: Categorization of approaches for data integration**

(a) Examples of multiomics data about patients. (b-d) Data integration approaches can be divided into three categories. (b) *Early integration approaches* involve combining datasets from different data types at the raw or processed level before analysis and prediction. (c) *Intermediate integration approaches* transform or map the underlying datasets at the same time as they estimate model parameters. (d) *Late integration approaches* perform analysis on each dataset independently, which is followed by integration of the resulting models to generate predictions, e.g., prognosis for a particular patient. SNP, single-nucleotide polymorphism



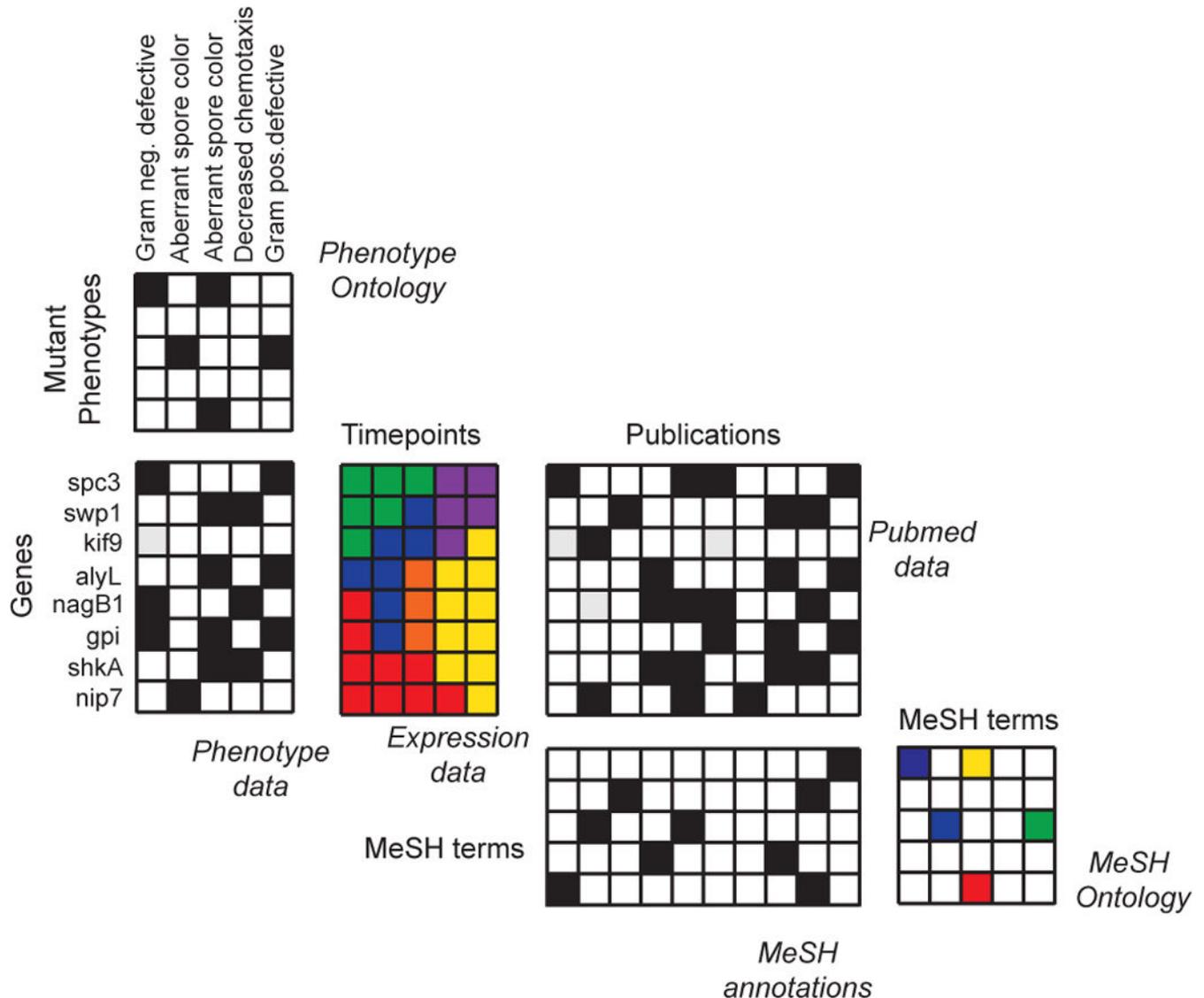
**Figure 3: Data integration.**

Data integration approaches combine multiple sources of information in a statistically meaningful way to provide a comprehensive analysis of a biomedical point of interest. Broadly, existing approaches employ three distinct modeling strategies (*i.e.*, early, intermediate, and late integration; see also Figure 2) and produce three types of prediction outputs (*i.e.*, a label representing probability of an entity belonging to a given class; a relationship representing probability of an association between two entities; and a complex structure, such as an inferred network or a partitioning of entities into groups).



**Figure 4: Scheme of single-cell multi-omics data integration.**

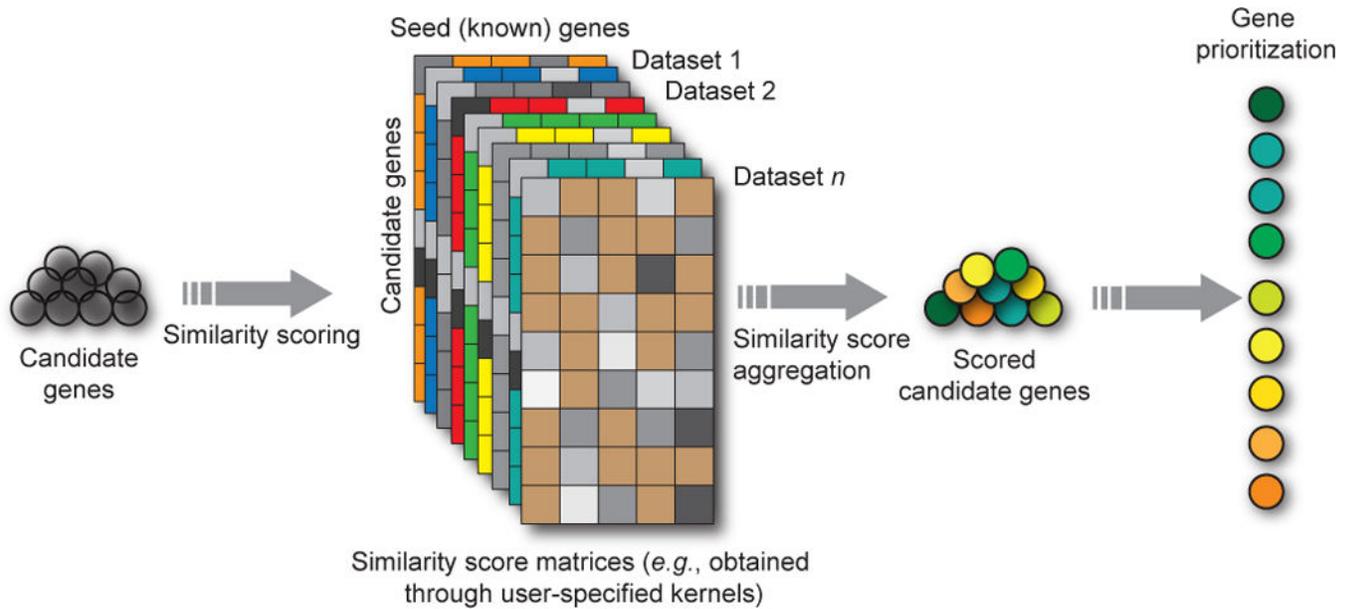
A generic bioinformatic analysis workflow usually includes three steps: first, the raw data are preprocessed, filtered, and quality-controlled separately for each assayed omics dimension, accounting for the analytical challenges of single-cell data, such as technical variation, sparse signal, and amplified artifacts. Second, as single-cell data are intrinsically of low coverage, it is a good practice to increase the signal to noise ratio by aggregating data; for example, by combining expression levels of genes of similar function or similar DNA methylation levels across genomic regions bound by the same transcription factors. Finally, data are integrated into one multi-omics map, representing a data-driven single-cell model.



**Figure 5: A matrix-based representation of diverse datasets relevant for gene function prediction.**

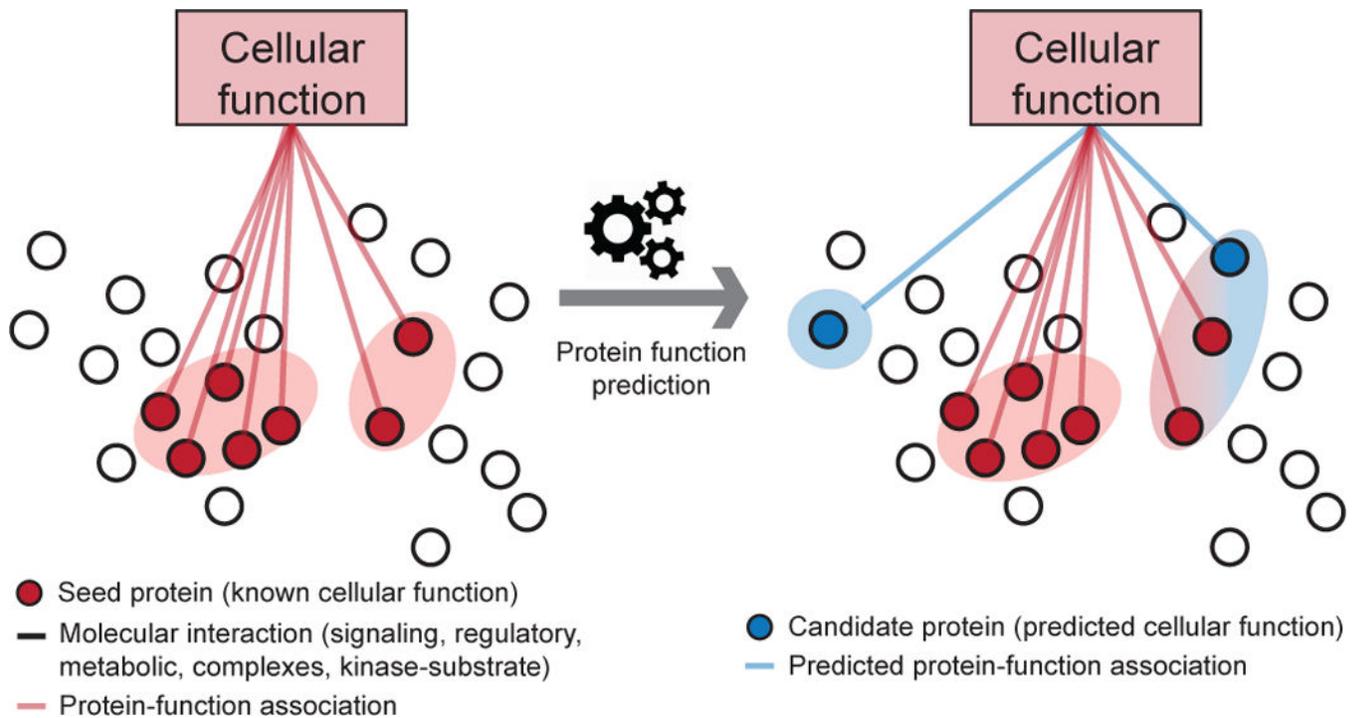
Let us consider a hypothetical gene function prediction task. Here, the function is *response to bacterial infection* [187], meaning that the task is to identify genes in a eukaryotic organism that determine how the organism will respond to a bacterial infection. There is a variety of diverse datasets potentially relevant for this task and each dataset is typically represented with a separate data matrix. Shown is an example with six data matrices, including gene-phenotype associations, gene expression profiles, biomedical literature, and annotations of research papers. Integrative approaches solve the gene function prediction task by establishing a rigorous statistical correspondence between different input dimensions of these seemingly disparate data matrices [48, 43, 188, 189, 190, 191, 33, 27, 192, 193]. For example, genes can be linked to the MeSH concepts in the Medical Subject Headings database via gene-publication relationships (*i.e.*, lists of genes discussed in a given research paper), followed by publication-MeSH relationships (*i.e.*, lists of the MeSH concepts

assigned to a given research paper). For example, a collective matrix factorization approach in [33] can fuse such complex systems of data matrices. The approach has been used to predict gene functions in various species [33, 190] and has subsequently been applied to prioritization of genes mediating bacterial infections [194].



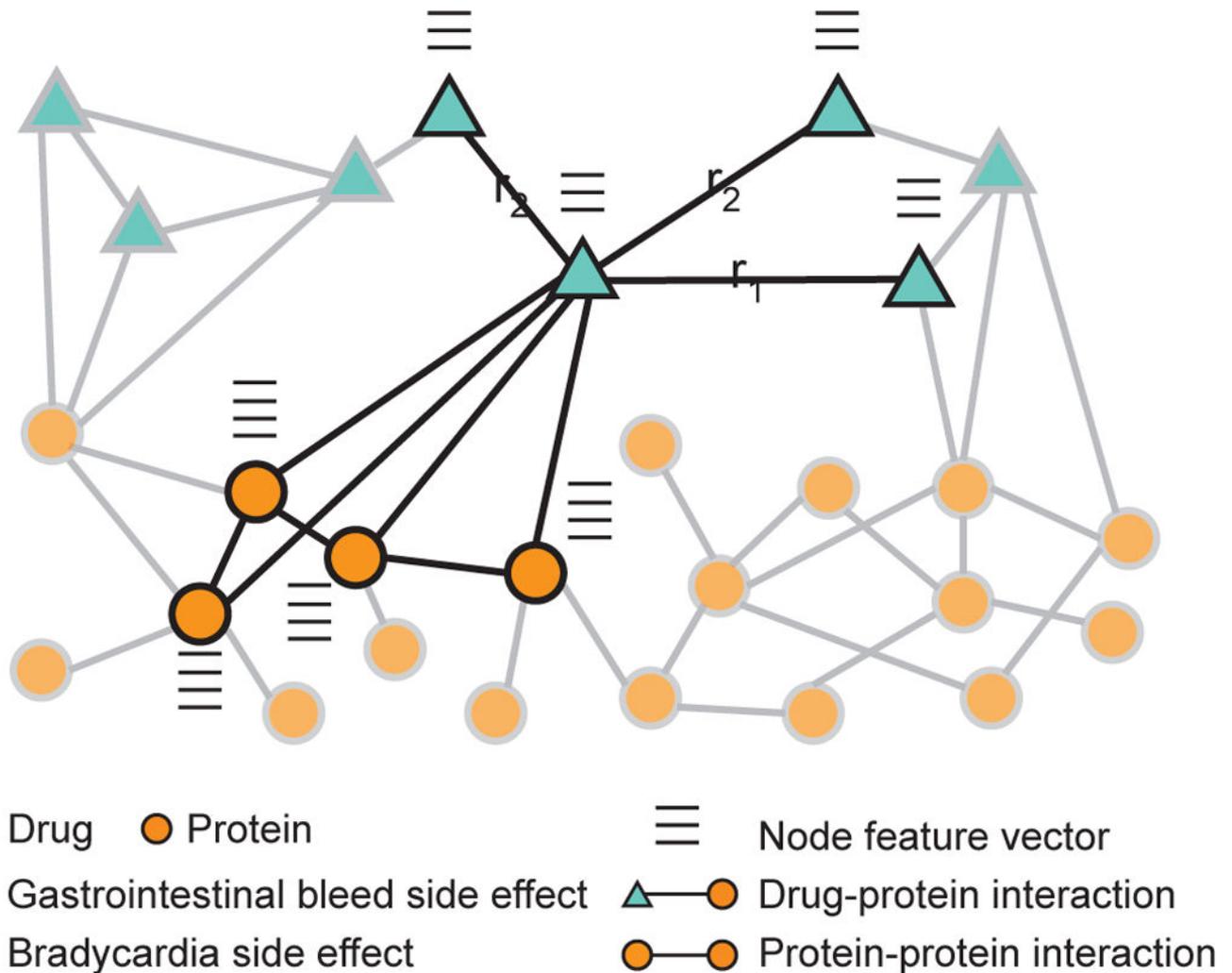
**Figure 6: Gene prioritization.**

Gene prioritization aims to identify the most promising genes among a list of candidate genes with respect to a biological process of interest. The biological process of interest is most often represented by a small set of seed genes that are known to be involved in the process. Typically, gene lists generated by traditional disease gene hunting techniques generate dozens or hundreds of genes among which only one or a few are of primary interest. The overall goal is to identify these genes and, in a second step, experimentally validate these genes only. Many different computational methods that use different algorithms, datasets, and strategies have been developed [224, 227, 228, 222, 229, 194, 230, 231, 232]. Some of these approaches have been implemented as publicly available tools and several of these approaches have been experimentally validated [224, 228, 229, 194, 225].



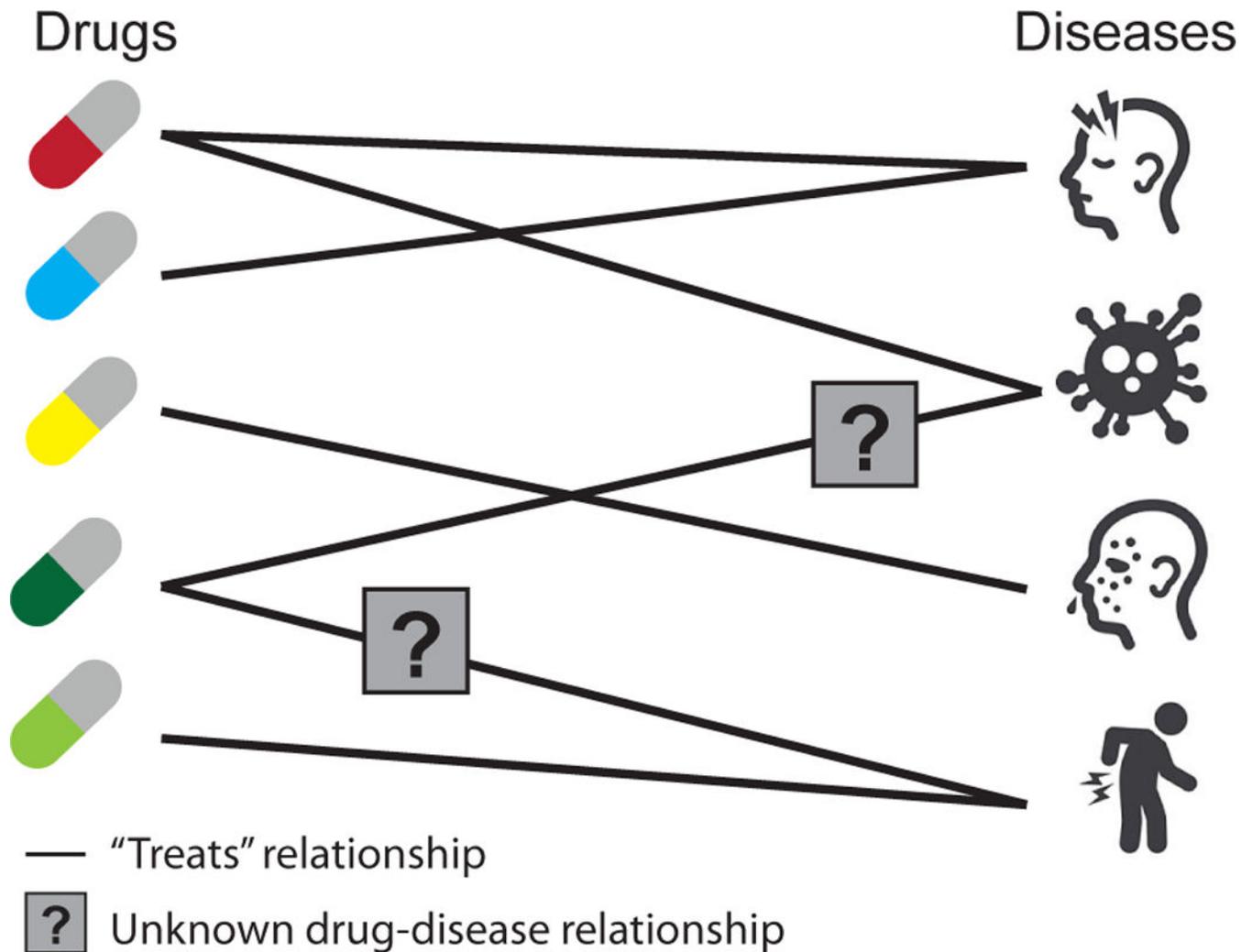
**Figure 7: A network-based approach to cellular function prediction.**

Bio-logical networks are a powerful representation for the discovery of interactions and emergent properties in biological systems, ranging from cell type identification at a single-cell level to disease treatment at a patient level. Fundamental to biological networks is the principle that genes involved in the same cellular function or underlying the same phenotype tend to interact [49]. This principle has been used many times to combine and to amplify signals from individual genes, and has led to remarkable discoveries in biology. For example, network-based methods for protein function prediction [247, 248, 23, 249] often use a heterogeneous protein-protein interaction network and conduct a large number of random walks on the network that are biased towards visiting known proteins associated with a specific function. These methods then calculate a score for each protein representing the probability that a protein is involved in a given cellular function based on how often the protein's node in the network is visited by random walkers.



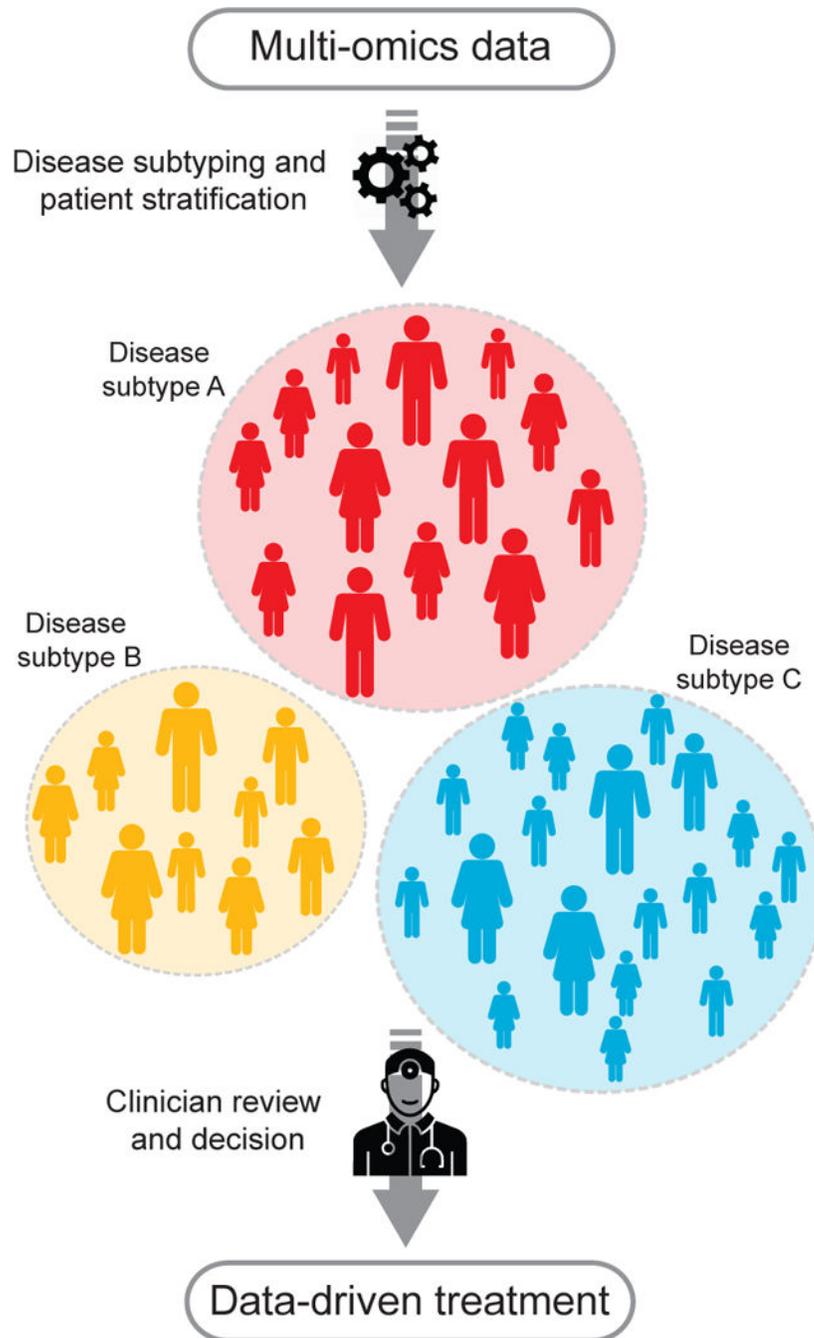
**Figure 8: Drug-target and drug-drug interactions.**

A heterogeneous network representation of drugs and proteins targeted by the drugs. In addition to interaction information, *e.g.*, drug-drug interactions, drug-protein interactions, and protein-protein interactions (Section 8), each node in the network has a feature vector describing important biological characteristics of the node, *e.g.*, drug's chemical structure, and protein's activity in tissues. Such networks are used to address two important tasks in computational pharmacology. The first is the prediction of drug-target interactions [260, 264, 19, 265], which are fundamental to the way that drugs work and often provide an important foundation for other tasks in the computational pharmacology. The second is the prediction of drug-drug interactions [273, 274, 270, 275], which are fundamental to modeling drug combinations and identifying drug pairs whose combination gives an exaggerated response beyond the response expected under no interaction. Zitnik *et al.* [45] use heterogeneous networks, such as the one shown in the figure, and develop a graph convolutional deep network approach to predict which side effects a patient might develop when taking multiple drugs at the same time.



**Figure 9: Drug repurposing.**

One exciting application of computational pharmacology is drug repurposing [252, 20]. Drug repurposing uses computational methods to find new uses for existing drugs. Given a disease, the task is to predict drugs (*e.g.*, among all drugs approved for use by the U.S. Food and Drug Administration) that might treat that disease. Integrative methods for drug repurposing comprise similarity-based methods [317], network modeling [260, 322, 272], and matrix factorization [324].



**Figure 10: Disease subtyping.**

Many diseases are heterogeneous. *Disease sub-typing* stratifies a heterogeneous group of patients with a particular disease into homogeneous subgroups, *i.e.*, subtypes, based on clinical, molecular, and other types of patient features. Accurate clustering of patients into subtypes is an important step towards personalized medicine and can inform clinical decision making and treatment matching.

**Table 1:****Epigenomics glossary.**

Terms referenced in this section.

Term	Description
assay	Laboratory experiment used to measure some physical or chemical aspect of a sample.
chromatin	DNA, its structure for packaging, and the attached biomolecules.
chromatin accessibility	Measurement of the openness of chromatin.
chromatin state	Label summarizing multiple properties of a region of chromatin, which often include histone modifications, chromatin accessibility, and transcription factor binding.
conservation	Measurement of how little a particular sequence changes throughout evolution.
deleterious	Hindering an organism's survival.
DNA methylation	Chemical modification to DNA that alters protein binding affinity without changing sequence.
dual futility conjecture	Conjecture that many transcription factor binding sites cannot be predicted from sequence alone.
epigenome	The collection of site-specific chemical and physical properties of the genome other than its sequence.
enhancer	Genomic region that influences transcription of a gene distant along the one-dimensional chromosome.
futility conjecture	Conjecture that many transcription factor binding sites predicted from sequence alone will have no functional role.
histone	Class of protein that packages DNA into nucleosomes.
histone modifications	Chemical alterations to histones that can alter gene expression.
label	Identifier for a pattern or cluster describing multiple regions of the genome, such as a chromatin state.
motif	Short, recurring sequences recognized by proteins such as transcription factors. Often defined probabilistically as a position weight matrix.
noncoding	Occurring outside the protein-coding sequence of any gene.
nucleosome	Eight histones and the DNA wrapped around them.
open chromatin	Region of chromatin not packaged into a nucleosome. Available for binding by other proteins.
position weight matrix	Probabilistic model that scores how well a motif describes a sequence. The matrix has a column for each position in the sequence and a row for each symbol in the sequence's alphabet.
regulatory region	Region of DNA with a known effect on gene expression.
segmentation	Partition of the genome with a label assigned to every segment.
topologically associated domain	Region of the genome enriched for three-dimensional interactions within.
transcription factor	Class of protein that binds to chromatin and regulates gene expression.

**Table 2:****Epigenomic assays.**

Glossary of assays referenced in this section.

Assay	Property measured	Method	References
ATAC-seq	chromatin accessibility	Uses the Tn5 transposase to insert a short sequence at open chromatin, followed by sequencing.	[62]
BS-seq	DNA methylation	Converts unmethylated cytosines into uracil, followed by sequencing.	[63]
CETCh-seq	associated protein	ChIP-seq against a special protein region added by clustered regularly interspaced short palindromic repeats (CRISPR) genome editing.	[64, 65]
ChIA-PET	long-range interactions	Ligates DNA regions close in three dimensions together with a known linker sequence, followed by sequencing	[66]
ChIP-exo	associated protein	Like ChIP-seq, with a step that cuts DNA fragments closer to a bound protein.	[67]
ChIP-nexus	associated protein	Like ChIP-exo, with an additional self-circularization step that increases library generation efficiency.	[68]
ChIP-seq	associated protein	Pulls down regions associated with a protein using an antibody against that protein, followed by sequencing.	[69, 70, 71, 72]
CUT&RUN	associated protein	Like ChIP-seq, with antibodies that diffuse into the cell to avoid breaking the cell apart prematurely.	[73]
DNase-seq	chromatin accessibility	Uses a deoxyribonuclease (DNase) protein to cut DNA at open chromatin, followed by sequencing.	[74]
Hi-C	long-range interactions	Ligates DNA regions close in three dimensions together, followed by sequencing.	[75, 76]
Hi-ChIP	long-range interactions and associated protein	Combines Hi-C and ChIP-seq. Ligates DNA regions inside of the nucleus with biotin, and applies ChIP-seq on ligated reads.	[77]

**Table 3:****Glossary for computational pharmacology.**

Terms referenced in this section.

Term	Description
drug discovery	The process through which potential new drugs are identified. It currently takes 13–15 years and between US\$2 billion and \$3 billion on average to get a new drug on the market [250].
side effect	Secondary, typically undesirable effect of a drug, <i>i.e.</i> adverse drug reaction.
medical indication	The use of a drug for treating a disease, <i>e.g.</i> , insulin is indicated for the treatment of diabetes.
drug-protein binding	The formation of a drug-protein complex. It describes the ability of a protein to form bonds with a drug in the human body. For example, if a drug is 95% bound to a protein and 5% free, that means that 5% is active in the human body and causing pharmacological effects.
protein target	Protein that has a critical role in a disease but is less significantly involved in other important processes to limit potential side effects. Modulation of the target protein is likely to have a therapeutic effect. Protein target can be <i>druggable</i> , <i>i.e.</i> , it binds with high affinity to a drug.
drug-target interaction	A drug interacting with a target protein in the human body and affecting the protein's activity.
drug-drug interaction	Phenomenon, in which the activity of one drug changes, favorably or unfavorably, if taken with another drug. It is often defined through the concepts of synergy and antagonism [251].
protein-ligand binding	The process by which a ligand (usually a molecule) produces a signal by binding to a site on a target protein.
drug combination	Combinatorial therapy that involves a concurrent use of multiple medications.
structural interaction fingerprint	Binary vector representation of 3D structural information about protein-ligand binding.
on-target side effect	Side effect that results from affecting the desired target protein of treatment.
off-target side effect	Side effect that results from an unwanted interaction of a drug with other proteins.