Information Fusion 50 (2019) 71–91

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities

Marinka Zitnik[a,*], Francis Nguyen[b,c], Bo Wang[d], Jure Leskovec[a,e,*], Anna Goldenberg[f,g,h,*], Michael M. Hoffman[b,c,g,h,*]

[a] Department of Computer Science, Stanford University, Stanford, CA, USA
[b] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[c] Princess Margaret Cancer Centre, Toronto, ON, Canada
[d] Hikvision Research Institute, Santa Clara, CA, USA
[e] Chan Zuckerberg Biohub, San Francisco, CA, USA
[f] Genetics & Genome Biology, SickKids Research Institute, Toronto, ON, Canada
[g] Department of Computer Science, University of Toronto, Toronto, ON, Canada
[h] Vector Institute, Toronto, ON, Canada

ARTICLE INFO

Keywords:
Computational biology
Personalized medicine
Systems biology
Heterogeneous data
Machine learning

ABSTRACT

New technologies have enabled the investigation of biology and human health at an unprecedented scale and in multiple dimensions. These dimensions include a myriad of 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.

1. Introduction

Understanding complex biological systems has been an ongoing quest for many researchers. The rapidly decreasing costs of high-throughput sequencing, development of massively parallel technologies, and new sensor technologies have enabled generation of data that describe biological systems on multiple dimensions. These dimensions include DNA sequence [1], epigenomic states [2], single-cell gene 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 of different types as researchers develop data 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 (Fig. 1). Much of the complexity in biology and medicine thus remains unexplained. If the field relies strictly on single-data-type studies, it never will be explained.

Ideally, one 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 can have numerous meanings, however, it is used here 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.*, [8–10]), a data integration approach can achieve a more thorough and informa-

* Corresponding authors.
E-mail addresses: marinka@cs.stanford.edu (M. Zitnik), jure@cs.stanford.edu (J. Leskovec), anna.goldenberg@utoronto.ca (A. Goldenberg), michael.hoffman@utoronto.ca (M.M. Hoffman).

https://doi.org/10.1016/j.inffus.2018.09.012
Received 28 June 2018; Accepted 20 September 2018
Available online 21 September 2018

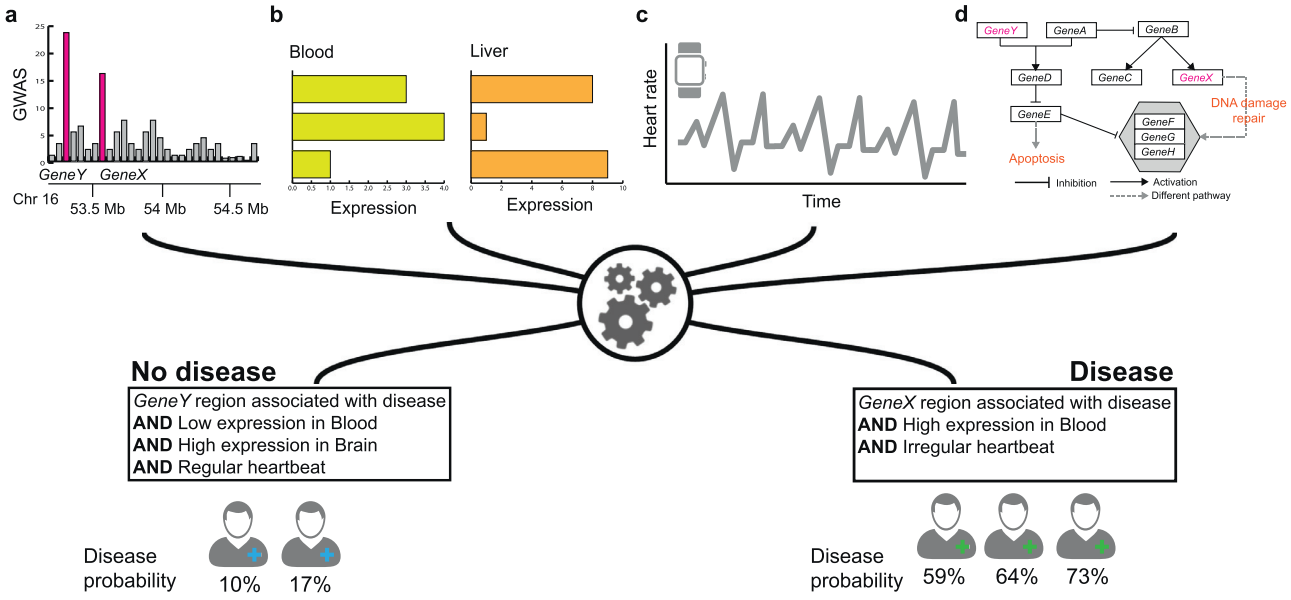

Fig. 1. The importance of data integration in biomedicine. Considering variation in only a single data type can miss many important patterns that can only be observed by considering multiple levels of biomedical data. Shown is a hypothetical example using disease diagnostics as a point of interest. When 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.

tive 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. A complete model of a system like the human body is only likely to be discovered if information from different dimensions is considered, 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 shortcomings 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 [12–14], 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 arises from limitations of measurement technology [15], natural and physical constraints [11,16], and investigative biases [17]. For example, information on what chemical compounds bind to what genes is available for only several thousands of genes, even when considering information from across organisms [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 [21–23] 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. As biomedical 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 chal-

lenging 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 the 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 spatial and temporal scales as *vertical data integration,* which depends on integration of cellular, cell type, tissue, organism, and population models at several temporal scales [23,26,27]. In contrast, horizontal data integration focuses on combining datasets and models at one particular level [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 analysis stage at which integration takes place [8,31–33] (Fig. 2). *Early integration* (Fig. 2b) begins by transforming all datasets into a single, feature-based table or a graph-based representation, which is then used as input to a machine learning method. Theoretically, this approach is powerful because the machine learning method can consider any type of dependence between the features as long as individual datasets are not collapsed before analysis. 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* (Fig. 2d), 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].

In *intermediate integration* (Fig. 2c), a model, such as multiple kernel learning [40,41], collective matrix factorization [33,42,43] or deep neural network [44,45] learns a joint representation of many datasets. Intermediate integration relies on algorithms that explicitly address the multiplicity of datasets and fuse them through inference of a joint model. Importantly, an intermediate data integration method 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 analysis 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 analyze a single dataset (Fig. 3). One area of a particular interest is the prediction of quantitative or categorical properties (*labels, e.g.*, gene functions) for biomedical entities (*e.g.,* genes). For example, many studies integrate a large number of 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 a drug's structural similarity, a 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). Finally, data integration methods exist to identify *complex structures*, such as gene modules or clusters detected in a combined gene interaction network [49] (Section 8.2), and to generate structured outputs, such as gene regulatory networks inferred from mixed data distributions [50].

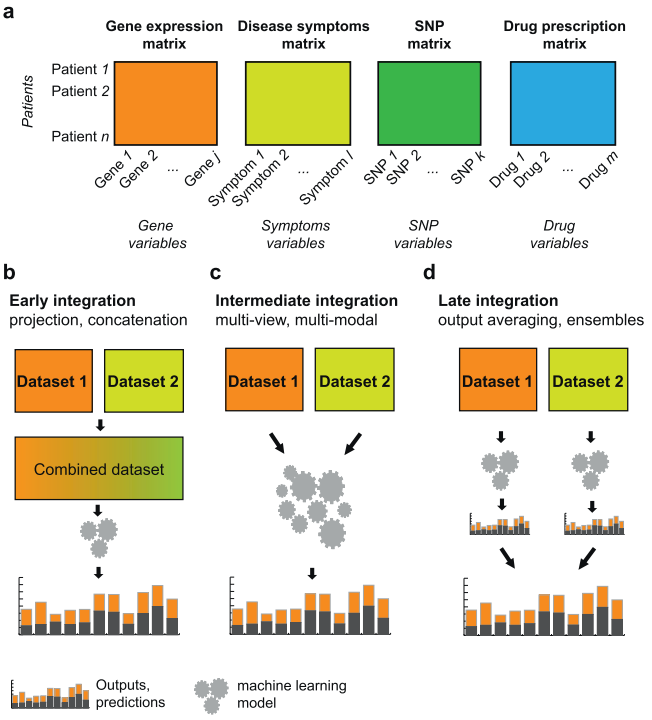

Fig. 2. Categorization of approaches for data integration. (a) Examples of multi-omics data about patients. (b–d) Data integration approaches can be divided into three categories. (b) *Early integration approaches* involve combining datasets from different data types at the raw or processed level before 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.

4. Focus of this Review

This Review is intended for computational researchers who are interested in 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 address various biomedical questions, 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 new machine learning methods 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 question. 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 [8,53–55] survey recent advances in high-throughput technologies and data integration-based

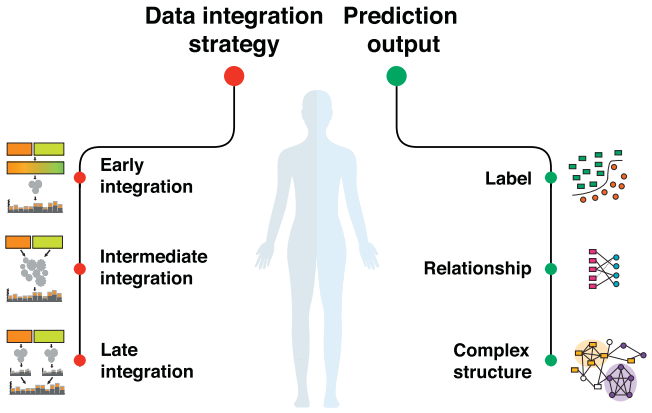

Fig. 3. Data integration. Data integration approaches combine multiples sources of information in a statistically meaningful way to provide a comprehensive analysis of biomedical dara. Broadly, existing approaches use three distinct strategies (*i.e.*, early, intermediate, and late integration; see also Fig. 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).

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 of implementing integrative approaches in clinics and for personalized medicine. Teschendorff et al. [10] survey 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 biomedical levels. 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 (Sections 5 and 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 (Sections 9 and 10). Finally, we discuss the potential for combining diverse 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 [56–58], 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 that DNA wraps around can undergo various chemical changes known as *histone modifications* [59]. The chromatin immunoprecipitation-sequencing (ChIP-seq) [69–72] 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 base pairs (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 genomic 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 on 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 [2,28,81], and one can combine these into a data matrix. One can divide this data matrix into row vectors, one for each assay. 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 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 [82,83,85,86], mouse [87], and fruit fly [88], enabling researchers to quickly assign function to genomic regions.

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.	[69–72]
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]

Methods such as HMMSeg [85], ChromHMM [83], Segway [82], EpiCSeg [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 [90–95] 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 1600 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 edit-

ing the genome in ways that might cause unexpected side effects. Furthermore, these assays all require more biological material than researchers can typically obtain from 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 [101–104]. 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] (see Table 2).

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 epigenomic 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 multi-layer perceptron to integrate these diverse data types and others, learning different hyperparameters 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, 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 Hi-C and ChIP-seq data as input. 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 noncoding 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] finds 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 cause 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 eigendecomposition 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, which can contain a myriad of different types of cells. *Cell type,* such as muscle or nerve, is typically defined based on function of a tissue in which the cells reside and unique morphological properties of that tissue [141]. However, considerable cell-to-cell variation in cells within a single cell type indicates the existence of distinct cell states (*e.g.*, mitotic, migratory) and various cell behaviors, which depend on local activity of each cell in a particular microenviroment. Even within a single tissue, there are diverse populations of cells, representing different manifestations of that tissue.

A traditional approach to studying tissues relies on a pooled assay and uses a weighted average of a *bulk sample of cells* from a particular tissue (*i.e.*, a large population of cells), which can obscure variations across cells within the sample. 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 profile individual cells from various perspectives, including genomic [143], epigenomic [144], transcriptomic [145], and proteomic [146] perspectives. 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 of individual cells and to characterize heterogeneity and functional diversity of cell populations [148]. To characterize a cell population one needs to identify what genes are expressed in each cell and how strong that gene expression is. Information on heterogeneity of cells in a given sample can answer questions that cannot be addressed by traditional ensemble-based methods, in which gene expression measurements are averaged over all cells pooled together. Recent studies have demonstrated that *de novo* cell type discovery and identification of functionally distinct cell subpopulations are possible via unbiased 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 distances between cells become more and more similar

as dimension of the space they reside within increases, *e.g.*, mammalian expression profiles are frequently studied as vectors with about 20,000 genes) [147].

To address these challenges, various unsupervised computational algorithms [151–155] 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 dimensionality 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-cell distance by a detected tree structure outlining the hierarchical relationship between single-cell samples to introduce biological prior knowledge. Cleary et al. [153] take another perspective by utilizing compressed sensing together with the assumption that scRNA-seq data are collected in a compressed format, as composite measurements of linear combinations of genes. However, a clear disadvantage of these dimensionality reduction methods is a strong statistical assumption of data following an appropriate distribution. Such assumptions do not always hold and depend on a particular scRNA-seq technology or platform.

Different from dimensionality reduction methods, ensemble methods 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-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.

The scRNA-seq data analysis methods described so far 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 scRNA-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 (Fig. 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 datasets and generates hypotheses by measuring correlations between datasets. For example, several studies [168–171] apply canonical correlation analysis (CCA) [172–174], a method that has been widely used in bulk data analysis, to estimate correlations between single-cell DNA methylation and scRNA-seq data. CCA learns a

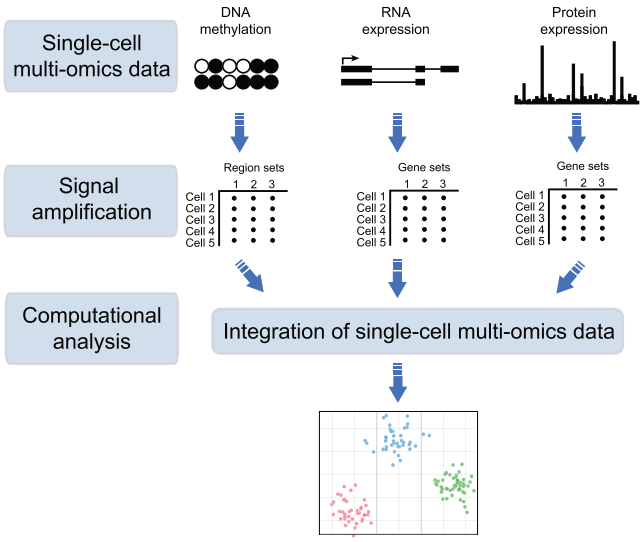

Fig. 4. Single-cell multi-omics data integration. A typical single-cell data analysis consists of three steps. First, the raw data are preprocessed, filtered, and quality-controlled separately for each assayed omics dimension, accounting for analytical challenges, such as technical variation, sparse 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 with similar functions or similar DNA methylation levels across genomic regions bound by the same transcription factors. Finally, data are integrated into one multi-omics map.

low-dimensional representation of omics datasets that captures common information shared across all datasets. 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. Similarly, scImpute [176] pulls information from groups of similar cells to complete a sparse data matrix and obtain a better representation of cell-cell 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 [170,177–179]. 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 omics dataset 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 omics dataset. 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 uti-

lize: (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 multi-layer neural network using memory-efficient mini-batch stochastic gradient descent. For example, Lin et al. [183] apply deep auto-encoders to obtain low-dimensional representations that optimize the reconstruction of original noisy inputs. Similarly, SAUCIE (Sparse Autoencoder 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 of dealing 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] projects, have expanded the breadth of available data that lend themselves to protein function prediction (Fig. 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 (*i.e.*, 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 4225 proteins remain functionally unannotated (*i.e.*, orphan proteins) and a similar proportion applies to *Saccharomyces cerevisiae*.

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 [46,195,197], metabolic interactions [198], genetic interactions [199], evolutionary relationships [200], 3D structural information [201], mining of biomedical text [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 (GO) [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) [355], scan the sequence data banks for homologous proteins of known function or structure and transfer 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, Zitnik and Zupan [15], Cho et al. [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 embed-

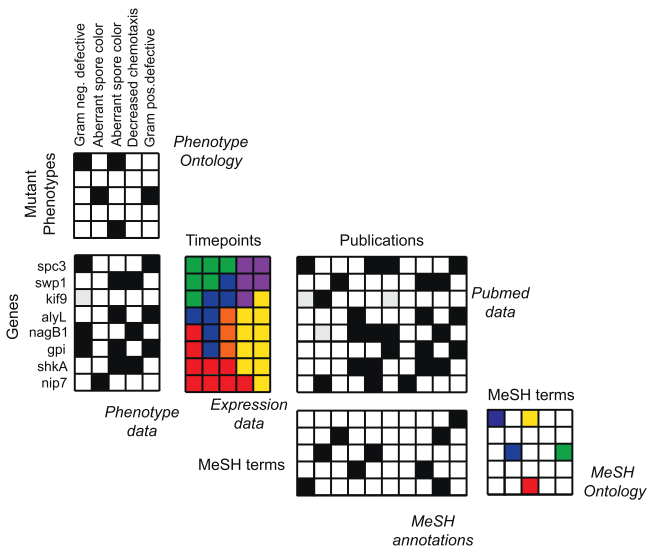

Fig. 5. A matrix-based representation of diverse datasets relevant for gene function prediction. Let us consider a hypothetical gene function prediction task. Here, the function is *response to bacterial infection* [187], meaning that the task is to identify genes in an eukaryotic organism that determine 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 predict task by establishing a rigorous statistical correspondence between different input dimensions of these seemingly disparate data matrices [27,33,43,48,188–193]. For example, genes can be linked to Medical Subject Headings (MeSH concepts) 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].

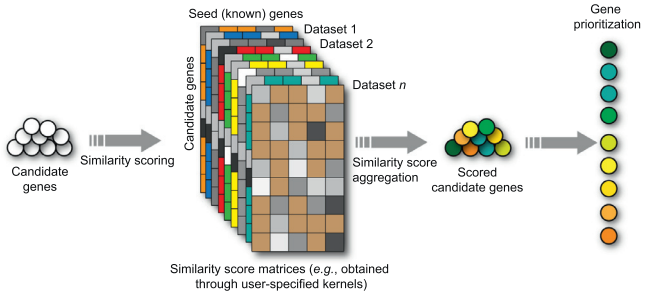

Fig. 6. Gene prioritization. Gene prioritization aims to identify the most promising genes among a list of candidate genes with respect to a biological process of interest. The biological process of interest is most often represented by a small set of seed genes that are known to be involved in the process. 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 [194,222,224,227–232]. Some of these approaches have been implemented as publicly available tools and several of these approaches have been experimentally validated [194,224,225,228,229].

ding (*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 embeddings were then used as input to a clustering algorithm. Many matrix decomposition [34] and tensor factorization [205] methods have proven useful for protein function prediction [206]. For example, Li et al. [207], Ou-Yang et al. [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, [22,211,212] 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 [22,23,213,214] or different species [215–220]. 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 within each layer (*i.e.*, protein-protein interactions) as well as edges across layers (*i.e.*, tissue-tissue similarities).

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* (Fig. 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 [223–225], and (4) methods that construct a separate gene-gene correlation network for each dataset and combine the networks under supervision of seed genes [46,226].

Supervised methods for function prediction use a classification of protein functions in the GO [203] to specify a supervised prediction task. The task presents four interesting challenges for machine learning methods. First, functions of proteins are classified into over 40,000 GO terms, and this large and complex space represents a challenge for any classification method. Second, there are dependencies between GO terms 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 [188,190], multi-label learning [46], and ensemble learning [38,216,233,234]. A number of machine learning methods were also developed to integrate regulatory networks and pathway information to predict functional modules, *i.e.*groups of functionally related proteins [50,234–238], 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 manually-curated GO annotations. Another common approach is to use neural networks to predict protein functions. For example, Zitnik and Leskovec [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 GO hierarchical relationships between protein functions. Along similar lines, Ma et al. [242] use several million genotypes to train a neural network whose architecture is determined by the GO hierarchy. As an example of biological application, Ma et al. [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 (Fig. 7) [49].

With the recent advances in experimental techniques, the number of identified PPIs keeps increasing [243]. However, we are still far from the 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 ChromNet [245], which pre-

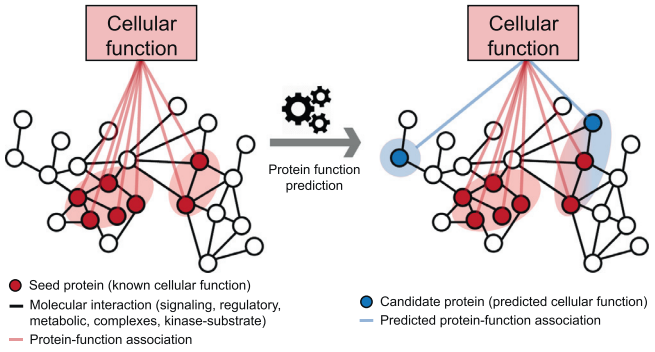

Fig. 7. A network-based approach to cellular function prediction. Biological networks are a powerful representation for the discovery of interactions and emergent properties in biological systems, ranging from cell type identification at a single-cell level to disease treatment at a patient level. Fundamental 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 [23,247–249] 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.

dicts 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 9000 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 [20,252]). 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 target proteins 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 target proteins 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. (Table 3).

Many recent efforts focus on using machine learning for drug-target interaction prediction. These efforts are based on the *guilt-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 target proteins as input features (or kernels) [255–257]. 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 (Fig. 8) and predict target proteins 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 feed 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 is used to infer 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 classic matrix factorization is that it 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 matrix factorizations, which are shallow feature learning algorithms, 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

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 $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, *e.g.*, adverse drug reaction.
medical indication	The use of a drug for treating a disease, *e.g.*, 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, then 5% of the drug is active in the human body and causes pharmacological effects.
target protein	Protein that is addressed and controlled by a chemical compound. Modulation of a target protein can have a therapeutic effect. Target protein is *druggable* when 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.

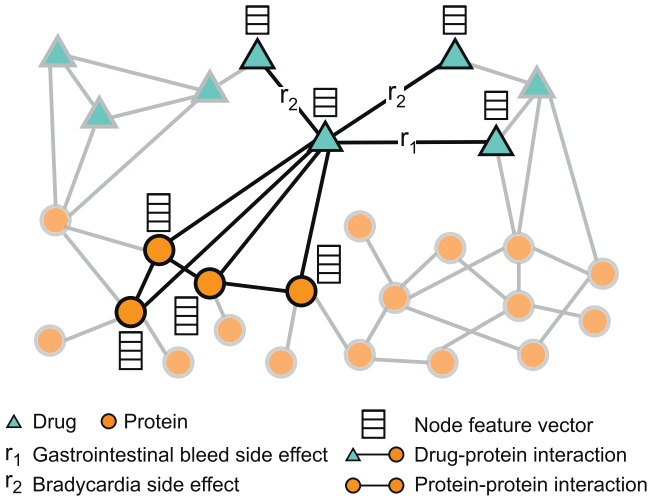

Fig. 8. Drug-target and drug-drug interactions. A heterogeneous network representation of drugs and proteins targeted by the drugs. In addition to interaction information, *e.g.*, drug-drug interactions, drug-protein interactions, and protein-protein interactions (Section 8), each node in the network 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 [19,260,264,265], 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 [270,273–275], 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.

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. Furthermore, it is 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 [280,287–292]. 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 [280,288,290,292–294]. 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 classifier, support vector machine, or neural network on feature representations of drug pairs. In contrast, *similarity-based approaches* assume that similar drugs have similar interaction patterns [33,252,287,289,296–298]. These methods combine different kinds of drug-drug similarity

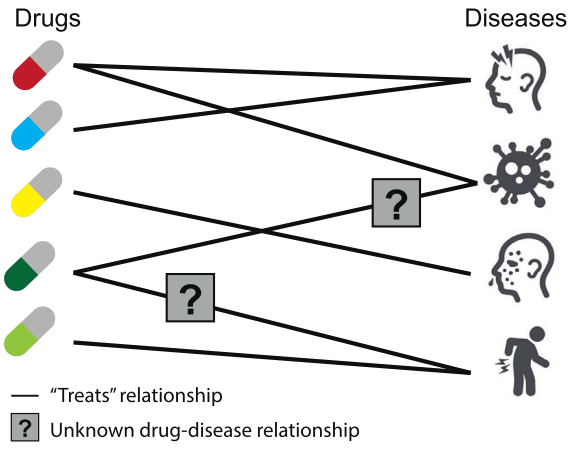

Fig. 9. Drug repurposing. Drug repurposing uses computational methods to find new uses for existing drugs [20,252]. Given a disease, the task is to predict drugs (*e.g.*, among all drugs approved for use by the U.S. Food and Drug Administration) that might treat that disease. Integrative methods for drug repurposing comprise similarity-based methods [317], network approaches [260,272,322], and matrix factorization [324].

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 to predict new drug-drug interactions [299–301].

Moving beyond predicting the chance of drug-drug interaction occurrence, recent methods identify how 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 (Fig. 8). The approach represents each type of side effect as a different edge type in the multimodal graph. Decagon uses the graph to develop a graph convolutional neural network, a type of neural network designed for graph data [304], to predict side effects of drug pairs.

9.3. *Drug repurposing*

Drug repurposing (also called "drug repositioning", Fig. 9) 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 target proteins [305] and hence a multi-target drug might be used for more than one purpose. Second, different diseases share genetic factors, molecular pathways, and symptoms [17,306] 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 [272,307–310], (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 [313–316], and (4) methods that consider a variety of disease similarity and drug similarity measures, each capturing a different type of biomedical knowledge [260,317–322].

For example, [260,272,321,323] used random walks on a heterogeneous similarity network to rank candidate drugs for a given disease. In another example, Luo et al. [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.

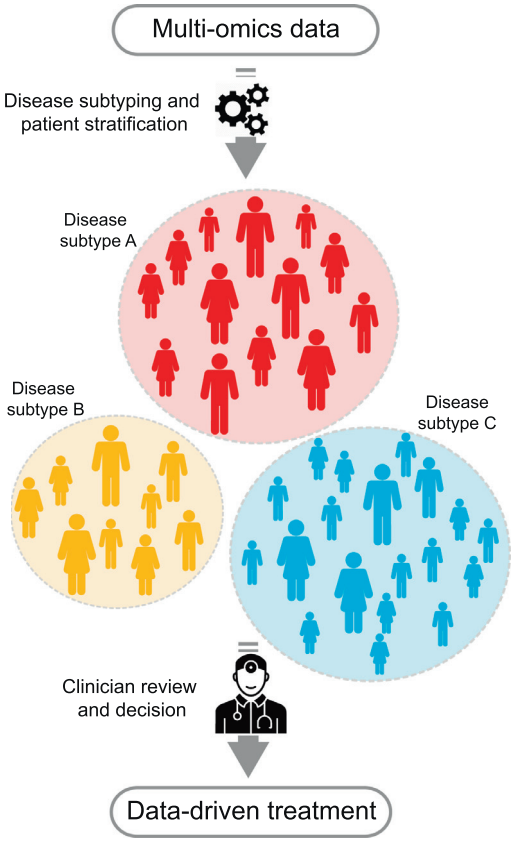

Fig. 10. Disease subtyping. Many diseases are heterogeneous. *Disease subtyping* stratifies a heterogeneous group of patients with a particular disease into homogeneous subgroups, *i.e.*, subtypes, based on clinical, molecular, and other types of patient features. Accurate clustering of patients into subtypes is an important step towards personalized medicine and can inform clinical decision making and treatment matching.

10. Disease subtyping and biomarker discovery

Many 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)) 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 inability 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 (Fig. 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 heterogene-

ity 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 subtypes. For example, Cavalli et al. [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 multiforme (GBM), a very aggressive adult-onset brain cancer. An earlier analysis combining gene expression and copy number variations (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 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 patient 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, reverse phase 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 non-linearity. For example, iCluster is a Gaussian latent variable model with sparsity regularization in a Lasso-type optimization framework [331]. The main assumption behind this method is that there exists a 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 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 [326,333–336]. In spirit, SNF is similar to 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 nutshell, 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 network, for which the following biological assumptions are followed: for each gene, copy number alteration (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.*, a single variable that would fully explain a trait) are becoming less prevalent. The realization that performing all analyses within only one data type can limit the potential for 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 questions, 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, combinations of heterogeneous data and new machine learning methods has enabled us 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 distribution of data generated by different technologies (*e.g.*, gene expression data generated by a sequencing-based versus 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 different assays (*e.g.*, Table 2 and Section 7) and appropriate normalization of data derived from these assays prior to downstream analysis remains a major challenge. Normalization is important 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.*, [343–345]).

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 phenotypes and answer biological questions. To be able to predict many 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, 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 the 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 cell models and organismal models poses fundamentally new challenges that must eventually be met. Moreover, because parameters of most current machine learning models are fixed after the model is trained, such models are 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 difficult to interpret the output of such methods from a biomedical perspective, a challenge that limits methods' utility for providing insights. This is especially the case for advanced 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 [349–352]. There is a critical need to transform *black-box* methods into *white-box* methods that can be opened up and interpreted meaningfully. 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, Ma et al. [242] used a neural network and integrated prior biological knowledge from the GO into the neural model [203]. A particular genotype-phenotype association could then be explained by a hierarchy of cellular systems, which was identified as a neural activation map.

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 is needed to 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 gold standard information about medical conditions, emerging research considers internet and mobile technologies as a viable method for broad phenotyping of large populations.

Relative to medical record review, *internet-based phenotyping* can be fast. For example, Tung et al. [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, Hu et al. [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, Hyde et al. [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, Smits et al. [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, approaches have emerged that can integrate data from many different biomedical datasets. 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 problems. 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 approaches that can currently be implemented to perform powerful integrative analyses. As integrative ap-

proaches become 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 National Science Foundation IIS-1149837, NIH BD2K U54EB020405, DARPA SIMPLEX, Stanford Data Science Initiative and the Chan Zuckerberg Biohub. F.N. and M.M.H. were supported by the Natural Sciences and Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H.).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2018.09.012.

References

[1] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57.
[2] A. Kundaje, et al., Integrative analysis of 111 reference human epigenomes, Nature 518 (7539) (2015) 317–330.
[3] S.R. Quake, T. Wyss-Coray, S. Darmanis, Tabula Muris Consortium, et al., Single–cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris, bioRxiv (2018) 237446.
[4] M. Wilhelm, J. Schlegl, H. Hahne, A.M. Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al., Mass-spectrometry-based draft of the human proteome, Nature 509 (7502) (2014) 582.
[5] M. Costanzo, B. VanderSluis, E.N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S.D. Lee, et al., A global genetic interaction network maps a wiring diagram of cellular function, Science 353 (6306) (2016). aaf1420
[6] X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S.M.S.-F. Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, et al., Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information, PLoS Biol. 15 (1) (2017). e2001402
[7] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S.J. Chanock, J.-H. Park, Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies, Nat. Genet. 45 (4) (2013) 400.
[8] M.D. Ritchie, E.R. Holzinger, R. Li, S.A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype-phenotype interactions, Nat. Rev. Genet. 16 (2) (2015) 85–97.
[9] K.J. Karczewski, M.P. Snyder, Integrative omics for health and disease, Nat. Rev. Genet. 19 (5) (2018) 299.
[10] A.E. Teschendorff, C.L. Relton, Statistical and integrative system-level analysis of DNA methylation data, Nat. Rev. Genet. 19 (3) (2018) 129.
[11] Y. Hu, A. Shmygelska, D. Tran, N. Eriksson, J.Y. Tung, D.A. Hinds, GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person, Nat. Commun. 7 (2016) 10448.
[12] B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. DeLisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biol. 10 (9) (2009) R91.
[13] M. Hofree, J.P. Shen, H. Carter, A. Gross, T. Ideker, Network-based stratification of tumor mutations, Nat. Methods 10 (11) (2013) 1108–1115.
[14] A. Lundby, E.J. Rossin, A.B. Steffensen, M.R. Acha, C. Newton-Cheh, A. Pfeufer, S.N. Lynch, S.-P. Olesen, S. Brunak, P.T. Ellinor, et al., Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics, Nat. Methods 11 (8) (2014) 868–874.
[15] M. Zitnik, B. Zupan, Data imputation in epistatic MAPs by network-guided matrix completion, J. Comput. Biol. 22 (6) (2015) 595–608.
[16] C.L. Hyde, M.W. Nagle, C. Tian, X. Chen, S.A. Paciga, J.R. Wendland, J.Y. Tung, D.A. Hinds, R.H. Perlis, A.R. Winslow, Identification of 15 genetic loci associated with risk of major depression in individuals of European descent, Nat. Genet. 48 (9) (2016) 1031–1036.
[17] J. Menche, et al., Uncovering disease-disease relationships through the incomplete interactome, Science 347 (6224) (2015) 1257601.
[18] M. Campillos, et al., Drug target identification using side-effect similarity, Science 321 (5886) (2008) 263–266.
[19] N. Zong, H. Kim, V. Ngo, O. Harismendy, Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations, Bioinformatics (2017). btx160
[20] R.A. Hodos, et al., In silico methods for drug repurposing and pharmacology, Wiley Interdiscip. Rev. Syst. Biol. Med. 8 (3) (2016) 186–210.
[21] A.-R. Carvunis, T. Ideker, Siri of the cell: what biology could learn from the iPhone, Cell 157 (3) (2014) 534–538.
[22] C.S. Greene, A. Krishnan, A.K. Wong, E. Ricciotti, R.A. Zelaya, D.S. Himmelstein, R. Zhang, B.M. Hartmann, E. Zaslavsky, S.C. Sealfon, et al., Understanding multicellular function and disease with human tissue-specific networks, Nat. Genet. 47 (6) (2015) 569.
[23] M. Zitnik, J. Leskovec, Predicting multicellular function through multi-layer tissue networks, Bioinformatics 33 (14) (2017) i190–i198.
[24] J. Bicker, et al., Elucidation of the impact of P-glycoprotein and breast cancer resistance protein on the brain distribution of catechol-O-methyltransferase inhibitors, Drug Metab. Dispos. 45 (12) (2017) 1282–1291.
[25] S. Mullainathan, Z. Obermeyer, Does machine learning automate moral hazard and error? Am. Econ. Rev. 107 (5) (2017) 476–480.
[26] S. Pilosof, M.A. Porter, M. Pascual, S. Kéfi, The multilayer nature of ecological networks, Nature Ecology & Evolution 1 (2017) 0101.
[27] M. Zitnik, B. Zupan, Jumping across biomedical contexts using compressive data fusion, Bioinformatics 32 (12) (2016) i90–i100.
[28] D. Bujold, D.A.d. L. Morais, C. Gauthier, C. Côté, M. Caron, T. Kwan, K.C. Chen, J. Laperle, A.N. Markovits, T. Pastinen, B. Caron, A. Veilleux, P. Jacques, G. Bourque, The International Human Epigenome Consortium Data Portal, Cell Syst. 3 (5) (2016) 496–499.
[29] M.W. Libbrecht, F. Ay, M.M. Hoffman, D.M. Gilbert, J.A. Bilmes, W.S. Noble, Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression, Genome Res. 25 (4) (2015) 544–557.
[30] S.A. Smits, J. Leach, E.D. Sonnenburg, C.G. Gonzalez, J.S. Lichtman, G. Reid, R. Knight, A. Manjurano, J. Changalucha, J.E. Elias, et al., Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania, Science 357 (6353) (2017) 802–806.
[31] P. Pavlidis, J. Weston, J. Cai, W.S. Noble, Learning gene functional classifications from multiple data types, J. Comput. Biol. 9 (2) (2002) 401–411.
[32] P. Maragos, P. Gros, A. Katsamanis, G. Papandreou, Cross-modal integration for performance improving in multimedia: a review, in: Multimodal Processing and Interaction, Springer, 2008, pp. 1–46.
[33] M. Zitnik, B. Zupan, Data fusion by matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 41–53.
[34] M. Zitnik, B. Zupan, Nimfa: A Python library for nonnegative matrix factorization, J. Mach. Learn. Res. 13 (2012) 849–853.
[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (12) (2010) 3371–3408.
[36] A. Sarajlić, N. Malod-Dognin, Ö. N. Yaveroğlu, N. Pržulj, Graphlet-based characterization of directed networks, Sci. Rep. 6 (2016) 35098.
[37] P. Yang, Y. Hwa Yang, B. B Zhou, A. Y Zomaya, A review of ensemble methods in bioinformatics, Curr. Bioinform. 5 (4) (2010) 296–308.
[38] C.-C. Wu, S. Asgharzadeh, T.J. Triche, D.Z. D'argenio, Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning, Bioinformatics 26 (6) (2010) 807–813.
[39] N. Iam-On, T. Boongoen, S. Garrett, LCE: a link-based cluster ensemble method for improved gene expression data analysis, Bioinformatics 26 (12) (2010) 1513–1519.
[40] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, F. Tahi, Towards a piRNA prediction using multiple kernel fusion and support vector machine, Bioinformatics 30 (17) (2014) i364–i370.
[41] J. Mariette, N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, Bioinformatics (2017).
[42] M. Zitnik, B. Zupan, Survival regression by data fusion, Systems Biomedicine 2 (3) (2014) 47–53.
[43] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (3) (2014) 333–337.
[44] J. Tan, G. Doing, K.A. Lewis, C.E. Price, K.M. Chen, K.C. Cady, B. Perchuk, M.T. Laub, D.A. Hogan, C.S. Greene, Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks, Cell Syst. 5 (1) (2017) 63–71.
[45] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks., Bioinformatics 34 (13) (2018) 457466.
[46] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, Q. Morris, GeneMANIA: a real–time multiple association network integration algorithm for predicting gene function, Genome Biol. 9 (1) (2008) S4.
[47] J. Carreras-Puigvert, M. Zitnik, A.-S. Jemth, M. Carter, J.E. Unterlass, B. Hallström, O. Loseva, Z. Karem, J.M. Calderón-Montaño, C. Lindskog, et al., A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family, Nat. Commun. 8 (1) (2017) 1541.
[48] M. Gönen, Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization, Bioinformatics 28 (18) (2012) 2304–2310.
[49] L. Cowen, T. Ideker, B.J. Raphael, R. Sharan, Network propagation: a universal amplifier of genetic associations, Nat. Rev. Genet. 18 (2017) 551–562.
[50] M. Zitnik, B. Zupan, Gene network inference by fusing data from diverse distributions, Bioinformatics 31 (12) (2015) i230–i239.
[51] A.K. Rider, N.V. Chawla, S.J. Emrich, A survey of current integrative network algorithms for systems biology, in: Systems Biology, Springer, 2013, pp. 479–495.
[52] G. Bebek, M. Koyutürk, N.D. Price, M.R. Chance, Network biology methods integrating biological data for translational science, Brief. Bioinform. 13 (4) (2012) 446–459.
[53] V.N. Kristensen, O.C. Lingjærde, H.G. Russnes, H.K.M. Vollan, A. Frigessi, A.-L. Børresen-Dale, Principles and methods of integrative genomic analyses in cancer, Nat. Rev. Cancer 14 (5) (2014) 299–313.
[54] V. Gligorijević, N. Malod-Dognin, N. Pržulj, Integrative methods for analyzing big data in precision medicine, Proteomics 16 (5) (2016) 741–758.
[55] N. Malod-Dognin, J. Petschnigg, N. Pržulj, Precision medicine–a promising, yet challenging road lies ahead, Current Opinion in Systems Biology (2017).
[56] R.J. Klose, A.P. Bird, Genomic DNA methylation: the mark and its mediators, Trends Biochem. Sci. 31 (2) (2006) 89–97.

[57] P.M.D. Severin, X. Zou, K. Schulten, H.E. Gaub, Effects of cytosine hydroxymethylation on DNA strand separation, Nat. Struct. Mol. Biol. 21 (11) (2014) 949–954.
[58] C.G. Spruijt, M. Vermeulen, DNA methylation: old dog, new tricks? Nature Structural & Molecular Biology 21 (11) (2014) 949–954.
[59] S.B. Rothbart, B.D. Strahl, Interpreting the language of histone and DNA modifications, Biochimica et Biophysica Acta 1839 (8) (2014) 627–643.
[60] C. Stirzaker, P.C. Taberlay, A.L. Statham, S.J. Clark, Mining cancer methylomes: prospects and challenges, Trends Genet. 30 (2) (2014) 75–84.
[61] T. Lappalainen, J.M. Greally, Associating cellular epigenetic models with human phenotypes, Nat. Rev. Genet. 18 (7) (2017) 441–451.
[62] J.D. Buenrostro, P.G. Giresi, L.C. Zaba, H.Y. Chang, W.J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, Nat. Methods 10 (12) (2013) 1213–1218.
[63] S.J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, S.E. Jacobsen, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, Nature 452 (7184) (2008) 215–219.
[64] P. Arnold, A. Schöler, M. Pachkov, P.J. Balwierz, H. Jørgensen, M.B. Stadler, E. van Nimwegen, D. Schübeler, Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting, Genome Res. 23 (1) (2013) 60–73.
[65] D. Savic, E.C. Partridge, K.M. Newberry, S.B. Smith, S.K. Meadows, B.S. Roberts, M. Mackiewicz, E.M. Mendenhall, R.M. Myers, CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins, Genome Res. 25 (10) (2015) 1581–1589.
[66] M.J. Fullwood, M.H. Liu, Y.F. Pan, et al., An oestrogen-receptor-α-bound human chromatin interactome, Nature 462 (7269) (2009) 58–64.
[67] H.S. Rhee, B.F. Pugh, 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] Q. He, J. Johnston, J. Zeitlinger, ChIP-nexus enables improved detection of in vivo transcription factor binding footprints, Nat. Biotechnol. 33 (4) (2015) 395–401.
[69] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-Wide Mapping of in Vivo Protein-DNA Interactions, Science 316 (5830) (2007) 1497–1502.
[70] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, S. Jones, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, Nat. Methods 4 (8) (2007) 651–657.
[71] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-Resolution Profiling of Histone Methylations in the Human Genome, Cell 129 (4) (2007) 823–837.
[72] T.S. Mikkelsen, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (7153) (2007) 553–560.
[73] P.J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites, eLife 6 (2017).
[74] L. Song, G.E. Crawford, DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells, Cold Spring Harb. Protoc. 2010 (2) (2010). pdb.prot5384
[75] E. Lieberman-Aiden, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326 (5950) (2009) 289–293.
[76] E. de Wit, W. de Laat, A decade of 3C technologies: insights into nuclear organization, Genes & development 26 (1) (2012) 11–24.
[77] M.R. Mumbach, A.J. Rubin, R.A. Flynn, C. Dai, P.A. Khavari, W.J. Greenleaf, H.Y. Chang, HiChIP: efficient and sensitive analysis of protein-directed genome architecture, Nat. Methods 13 (11) (2016) 919–922.
[78] L.B. Holder, M.M. Haque, M.K. Skinner, Machine learning for epigenetics and future medical applications, Epigenetics 12 (7) (2017) 505–514.
[79] M. Widschwendter, et al., Epigenome-based cancer risk prediction: rationale, opportunities and challenges, Nat. Rev. Clin. Oncol. 15 (5) (2018) 292–309.
[80] S.H. Stricker, A. Köferle, S. Beck, From profiles to function in epigenomics, Nat. Rev. Genet. 18 (1) (2017) 51–66.
[81] The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project, Science 306 (5696) (2004) 636–640.
[82] M.M. Hoffman, O.J. Buske, J. Wang, Z. Weng, J.A. Bilmes, W.S. Noble, Unsupervised pattern discovery in human chromatin structure through genomic segmentation, Nat. Methods 9 (5) (2012) 473–476.
[83] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, Nat. Methods 9 (3) (2012) 215–216.
[84] M.M. Hoffman, J. Ernst, S.P. Wilder, A. Kundaje, R.S. Harris, M. Libbrecht, B. Giardine, P.M. Ellenbogen, J.A. Bilmes, E. Birney, R.C. Hardison, I. Dunham, M. Kellis, W.S. Noble, Integrative annotation of chromatin elements from ENCODE data, Nucleic Acids Res. 41 (2) (2013) 827–841.
[85] N. Day, A. Hemmaplardh, R.E. Thurman, J.A. Stamatoyannopoulos, W.S. Noble, Unsupervised segmentation of continuous genomic data, Bioinformatics 23 (11) (2007) 1424–1426.
[86] Y. Zhang, L. An, F. Yue, R.C. Hardison, Jointly characterizing epigenetic dynamics across multiple human cell types, Nucleic Acids Res. 44 (14) (2016) 6721–6731.
[87] F. Yue, et al., A comparative encyclopedia of DNA elements in the mouse genome, Nature 515 (7527) (2014) 355–364.
[88] P.V. Kharchenko, et al., Comprehensive analysis of the chromatin landscape in Drosophila melanogaster, Nature 471 (7339) (2011) 480–485.
[89] A. Mammana, H.-R. Chung, Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome, Genome Biol. 16 (1) (2015) 151.
[90] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[91] L.E. Baum, T. Petrie, G. Soules, N. Weiss, 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] L.E. Baum, T. Petrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, The Annals of Mathematical Statistics 37 (6) (1966) 1554–1563.
[93] L. Baum, An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process, Inequalities 3 (1972) 1–8.
[94] L.E. Baum, G. Sell, Growth transformations for functions on manifolds, Pac. J. Math. 27 (2) (1968) 211–227.
[95] G.R. Blakley, Homogeneous nonnegative symmetric quadratic transformations, Bulletin of the American Mathematical Society 70 (5) (1964) 712–716.
[96] R.C.W. Chan, M.W. Libbrecht, E.G. Roberts, J.A. Bilmes, W.S. Noble, M.M. Hoffman, I. Birol, Segway 2.0: Gaussian mixture models and minibatch training, Bioinformatics 34 (4) (2018) 669–671.
[97] P. Dagum, A. Galper, E. Horvitz, A. Seiver, Uncertain reasoning and forecasting, Int. J. Forecast. 11 (1) (1995) 73–87.
[98] M.W. Libbrecht, O. Rodriguez, Z. Weng, M. Hoffman, J.A. Bilmes, W.S. Noble, A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types, bioRxiv (2018) 086025.
[99] J.M. Vaquerizas, S.K. Kummerfeld, S.A. Teichmann, N.M. Luscombe, A census of human transcription factors: function, expression and evolution, Nat. Rev. Genet. 10 (4) (2009) 252–263.
[100] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T.R. Hughes, M.T. Weirauch, The Human Transcription Factors., Cell 172 (4) (2018) 650–665.
[101] P. D'haeseleer, What are DNA sequence motifs? Nat. Biotechnol. 24 (4) (2006) 423–425.
[102] T.L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, Bioinformatics 14 (1) (1998) 48–54.
[103] C.E. Grant, T.L. Bailey, W.S. Noble, FIMO: scanning for occurrences of a given motif, Bioinformatics 27 (7) (2011) 1017–1018.
[104] M. Thomas-Chollier, O. Sand, J.-V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohee, J. van Helden, RSAT: regulatory sequence analysis tools, Nucleic Acids Res. 36 (Web Server) (2008) W119–W127.
[105] G.D. Stormo, T.D. Schneider, L. Gold, A. Ehrenfeucht, Use of the Perceptron algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Res. 10 (9) (1982) 2997–3011.
[106] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat. Rev. Genet. 5 (4) (2004) 276–287.
[107] G. Badis, M.F. Berger, A.A. Philippakis, S. Talukder, A.R. Gehrke, S.A. Jaeger, E.T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D.E. Newburger, Q. Morris, T.R. Hughes, M.L. Bulyk, Diversity and complexity in DNA recognition by transcription factors, Science 324 (5935) (2009) 1720–1723.
[108] N. Ogawa, M.D. Biggin, High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro, in: Gene Regulatory Networks, Springer, 2012, pp. 51–63.
[109] T.L. Bailey, C. Elkan, Unsupervised learning of multiple motifs in biopolymers using expectation maximization, Mach. Learn. 21 (1-2) (1995) 51–80.
[110] N. Jayaram, D. Usvyat, A.C. R. Martin, Evaluating tools for transcription factor binding site prediction, BMC Bioinformatics (2016) 1.
[111] M. Karimzadeh, M.M. Hoffman, Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome, bioRxiv (2018) 168419.
[112] R. Pique-Regi, J.F. Degner, A.A. Pai, D.J. Gaffney, Y. Gilad, J.K. Pritchard, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data., Genome Res. 21 (3) (2011) 447–455.
[113] E.G. Gusmao, C. Dieterich, M. Zenke, I.G. Costa, Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications, Bioinformatics 30 (22) (2014) 3143–3151.
[114] T. Xu, B. Li, M. Zhao, K.E. Szulwach, R.C. Street, L. Lin, B. Yao, F. Zhang, P. Jin, H. Wu, Z.S. Qin, Base-resolution methylation patterns accurately predict transcription factor bindings in vivo, Nucleic Acids Res. 43 (5) (2015) 2757–2766.
[115] D. Quang, X. Xie, FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data, bioRxiv (2017) 151274.
[116] J. Keilwagen, S. Posch, J. Grau, 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] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature 485 (7398) (2012) 376–380.
[119] S.S.P. Rao, M.H. Huntley, N.C. Durand, E.K. Stamenova, I.D. Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, A.D. Omer, E.S. Lander, E.L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.
[120] J. Paulsen, M. Sekelja, A.R. Oldenburg, A. Barateau, N. Briand, E. Delbarre, A. Shah, A.L. Sørensen, C. Vigouroux, B. Buendia, P. Collas, Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts, Genome Biol. 18 (1) (2017) 21.
[121] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G.J. Filion, M.A. Marti-Renom, Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors, PLOS Comput. Biol. 13 (7) (2017) e1005665.

[122] M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, J.S. Liu, Bayesian inference of spatial organizations of chromosomes, PLoS Comput. biol. 9 (1) (2013) e1002893.
[123] M. Di Pierro, R.R. Cheng, E. Lieberman Aiden, P.G. Wolynes, J.N. Onuchic, De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture, Proc. Natl. Acad. Sci. 114 (46) (2017) 12126–12131.
[124] J.W. Whitaker, Z. Chen, W. Wang, Predicting the human epigenome from DNA motifs, Nat. Methods 12 (3) (2015) 265–272.
[125] J. Ernst, M. Kellis, Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues, Nat. Biotechnol. 33 (4) (2015) 364–376.
[126] T.J. Durham, M.W. Libbrecht, J.J. Howbert, J. Bilmes, W.S. Noble, PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition, Nat. Commun. 9 (1) (2018) 1402.
[127] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad. Sci. 106 (23) (2009) 9362–9367.
[128] J.R. Prensner, A.M. Chinnaiyan, The emergence of lncRNAs in cancer biology, Cancer Discov. 1 (5) (2011) 391–407.
[129] D. Shlyueva, G. Stampfel, A. Stark, Transcriptional enhancers: from properties to genome-wide predictions, Nat. Rev. Genet. 15 (4) (2014) 272–286.
[130] J.-J. M. Riethoven, Regulatory regions in DNA: promoters, enhancers, silencers, and insulators, in: Computational Biology of Transcription Factor Binding, Springer, 2010, pp. 33–42.
[131] M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, M.A. Beer, gkmSVM: an R package for gapped-kmer SVM, Bioinformatics 32 (14) (2016) 2205–2207.
[132] M. Ghandi, D. Lee, M. Mohammad-Noori, M.A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features., PLoS computational biology 10 (7) (2014) e1003711.
[133] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learningbased sequence model, Nat. Methods 12 (10) (2015) 931–934.
[134] D.R. Kelley, J. Snoek, J.L. Rinn, Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks, Genome Res. 26 (7) (2016) 990–999.
[135] M. Kircher, D.M. Witten, P. Jain, B.J. O'Roak, G.M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants, Nat. Genet. 46 (3) (2014) 310–315.
[136] I. Ionita-Laza, K. McCallum, B. Xu, J.D. Buxbaum, A spectral approach integrating functional genomic annotations for coding and noncoding variants, Nat. Genet. 48 (2) (2016) 214–220.
[137] I. Gronau, L. Arbiza, J. Mohammed, A. Siepel, Inference of natural selection from interspersed genomic elements based on polymorphism and divergence, Mol. Biol. Evol. 30 (5) (2013) 1159–1171.
[138] B. Gulko, M.J. Hubisz, I. Gronau, A. Siepel, A method for calculating probabilities of fitness consequences for point mutations across the human genome, Nat. Genet. 47 (3) (2015) 276–283.
[139] Y.-F. Huang, B. Gulko, A. Siepel, Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data, Nat. Genet. 49 (4) (2017) 618–624.
[140] A. Regev, S.A. Teichmann, E.S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, et al., Science forum: the human cell atlas, wlife 6 (2017) e27041.
[141] H. Clevers, S. Rafelski, M. Elowitz, A. Klein, J. Shendure, C. Trapnell, E. Lein, E. Lundberg, M. Uhlen, A. Martinez-Arias, et al., What is your conceptual definition of "cell type" in the context of a mature organism? Cell Syst. 4 (3) (2017) 255–259.
[142] G. Kelsey, O. Stegle, W. Reik, Single-cell epigenomics: Recording the past and predicting the future, Science 358 (6359) (2017) 69–75.
[143] C. Gawad, W. Koh, S.R. Quake, Single-cell genome sequencing: current state of the science, Nat. Rev. Genet. 17 (3) (2016) 175.
[144] O. Schwartzman, A. Tanay, Single-cell epigenomics: techniques and emerging applications, Nat. Rev. Genet. 16 (12) (2015) 716.
[145] O. Stegle, S.A. Teichmann, J.C. Marioni, Computational and analytical challenges in single-cell transcriptomics, Nat. Rev. Genet. 16 (3) (2015) 133.
[146] M. Wu, A.K. Singh, Single-cell protein analysis, Curr. Opin. Biotech. 23 (1) (2012) 83–88.
[147] G.-C. Yuan, L. Cai, M. Elowitz, T. Enver, G. Fan, G. Guo, R. Irizarry, P. Kharchenko, J. Kim, S. Orkin, et al., Challenges and emerging directions in single-cell analysis, Genome Biol. 18 (1) (2017) 84.
[148] E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science, Nat. Rev. Genet. 14 (9) (2013) 618.
[149] O.B. Poirion, X. Zhu, T. Ching, L. Garmire, Single-cell transcriptomics bioinformatics and computational challenges, Frontiers in Genetics 7 (2016) 163.
[150] G.X. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, et al., Massively parallel digital transcriptional profiling of single cells, Nat. Commun. 8 (2017) 14049.
[151] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning, Nat. Methods 14 (4) (2017) 414.
[152] E. Pierson, C. Yau, ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis, Genome Biol. 16 (1) (2015) 241.
[153] B. Cleary, L. Cong, A. Cheung, E.S. Lander, A. Regev, Efficient generation of transcriptomic profiles by random composite measurements, Cell 171 (6) (2017) 1424–1436.
[154] V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W. Reik, M. Barahona, A.R. Green, et al., Sc3: consensus clustering of single-cell RNA-seq data, Nat. Methods 14 (5) (2017) 483.
[155] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. 36 (5) (2018) 411.
[156] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B.B. Tuch, A. Siddiqui, et al., mRNA-Seq whole-transcriptome analysis of a single cell, Nat. Methods 6 (5) (2009) 377.
[157] S. Yotsukura, S. Nomura, H. Aburatani, K. Tsuda, 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.
[158] H. Zhang, C.A. Lee, Z. Li, J.R. Garbe, C.R. Eide, R. Petegrosso, R. Kuang, J. Tolar, A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa, PLoS Comput. Biol. 14 (4) (2018) e1006053.
[159] S.A. Smallwood, H.J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S.R. Andrews, O. Stegle, W. Reik, G. Kelsey, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity, Nat. Methods 11 (8) (2014) 817.
[160] A. Rotem, O. Ram, N. Shoresh, R.A. Sperling, A. Goren, D.A. Weitz, B.E. Bernstein, Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state, Nat. Biotechnol. 33 (11) (2015) 1165.
[161] J.D. Buenrostro, B. Wu, U.M. Litzenburger, D. Ruff, M.L. Gonzales, M.P. Snyder, H.Y. Chang, W.J. Greenleaf, Single-cell chromatin accessibility reveals principles of regulatory variation, Nature 523 (7561) (2015) 486.
[162] D.A. Cusanovich, R. Daza, A. Adey, H.A. Pliner, L. Christiansen, K.L. Gunderson, F.J. Steemers, C. Trapnell, J. Shendure, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing, Science 348 (6237) (2015) 910–914.
[163] T. Nagano, Y. Lubling, T.J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E.D. Laue, A. Tanay, P. Fraser, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure, Nature 502 (7469) (2013) 59.
[164] A.P. Frei, F.-A. Bava, E.R. Zunder, E.W. Hsieh, S.-Y. Chen, G.P. Nolan, P.F. Gherardini, Highly multiplexed simultaneous detection of RNAs and proteins in single cells, Nat. Methods 13 (3) (2016) 269.
[165] M. Fessenden, Metabolomics: Small molecules, single cells, 2016.
[166] I.C. Macaulay, C.P. Ponting, T. Voet, Single-cell multiomics: multiple measurements from single cells, Trends Genet. 33 (2) (2017) 155–168.
[167] C. Bock, M. Farlik, N.C. Sheffield, Multi-omics of single cells: strategies and applications, Trends Biotechnol. 34 (8) (2016) 605–608.
[168] C. Angermueller, S.J. Clark, H.J. Lee, I.C. Macaulay, M.J. Teng, T.X. Hu, F. Krueger, S.A. Smallwood, C.P. Ponting, T. Voet, et al., Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity, Nat. Methods 13 (3) (2016) 229.
[169] Y. Hou, H. Guo, C. Cao, X. Li, B. Hu, P. Zhu, X. Wu, L. Wen, F. Tang, Y. Huang, et al., Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas, Cell Res. 26 (3) (2016) 304.
[170] I.C. Macaulay, W. Haerty, P. Kumar, Y.I. Li, T.X. Hu, M.J. Teng, M. Goolam, N. Saurat, P. Coupland, L.M. Shirley, et al., G&t-seq: parallel sequencing of single-cell genomes and transcriptomes, Nat. Methods 12 (6) (2015) 519.
[171] K.Y. Han, K.-T. Kim, J.-G. Joung, D.-S. Son, Y.J. Kim, A. Jo, H.-J. Jeon, H.-S. Moon, C.E. Yoo, W. Chung, et al., SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells, Genome Res. 28 (1) (2018) 75–87.
[172] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[173] S. Waaijenborg, P.C.V. de Witt Hamer, A.H. Zwinderman, Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis, Statistical Applications in Genetics and Molecular Biology 7(1).
[174] K.-A. Lê Cao, P.G. Martin, C. Robert-Granié, P. Besse, Sparse canonical methods for biological data integration: application to a cross-platform study, BMC Bioinformatics 10 (1) (2009) 34.
[175] D. van Dijk, J. Nainys, R. Sharma, P. Kathail, A.J. Carr, K.R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data, BioRxiv (2017) 111591.
[176] W.V. Li, J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nat. Commun. 9 (1) (2018) 997.
[177] L.F. Cheow, E.T. Courtois, Y. Tan, R. Viswanathan, Q. Xing, R.Z. Tan, D.S. Tan, P. Robson, Y.-H. Loh, S.R. Quake, et al., Single-cell multimodal profiling reveals cellular epigenetic heterogeneity, Nat. Methods 13 (10) (2016) 833.
[178] V.M. Peterson, K.X. Zhang, N. Kumar, J. Wong, L. Li, D.C. Wilson, R. Moore, T.K. McClanahan, S. Sadekova, J.A. Klappenbach, Multiplexed quantification of proteins and transcripts in single cells, Nat. Biotechnol. 35 (10) (2017) 936.
[179] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P.K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells, Nat. Methods 14 (9) (2017) 865.
[180] J.D. Welch, A.J. Hartemink, J.F. Prins, MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics, Genome Biol. 18 (1) (2017) 138.
[181] G. Iacono, E. Mereu, A. Guillaumet-Adkins, R. Corominas, I. Cuscó, G. Rodríguez-Esteban, M. Gut, L.A. Pérez-Jurado, I. Gut, H. Heyn, bigscale: an analytical framework for big-scale single-cell data, Genome Res. 28 (6) (2018) 878–890.
[182] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (1) (2018) 15.

[183] C. Lin, S. Jain, H. Kim, Z. Bar-Joseph, Using neural networks for reducing the dimensions of single-cell RNA-Seq data, Nucleic Acids Res. 45 (17) (2017). e156–e156
[184] M. Amodio, K. Srinivasan, D. van Dijk, H. Mohsen, K. Yim, R. Muhle, K.R. Moon, S. Kaech, R. Sowell, R. Montgomery, et al., Exploring single-cell data with multi-tasking deep neural networks, bioRxiv (2017) 237065.
[185] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, OSDI (2016).
[186] GTEx Consortium, The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans, Science 348 (6235) (2015) 648–660.
[187] A. Typas, V. Sourjik, Bacterial protein networks: properties and functions, Nat. Rev. Microbiol. 13 (9) (2015) 559.
[188] V. Gligorijevic, V. Janjić, N. Przulj, Integration of molecular network data reconstructs Gene Ontology, Bioinformatics 30 (17) (2014) i594–i600.
[189] M. Zitnik, B. Zupan, Matrix factorization-based data fusion for drug-induced liver injury prediction, Systems Biomedicine 2 (1) (2014a) 16–22.
[190] M. Zitnik, B. Zupan, Matrix factorization-based data fusion for gene function prediction in bakers yeast and slime mold, in: Pacific Symposium on Biocomputing, NIH Public Access, 2014b, p. 400.
[191] V. Gligorijević, N. Malod-Dognin, N. Pržulj, Fuse: multiple network alignment via data fusion, Bioinformatics 32 (8) (2015) 1195–1203.
[192] M. Stražar, M. Zitnik, B. Zupan, J. Ule, T. Curk, Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins, Bioinformatics 32 (10) (2016) 1527–1535.
[193] V. Gligorijevic, N. Malod-Dognin, N. Przulj, Patient-specific data fusion for cancer stratification and personalised treatment, in: Pacific Symposium on Biocomputing, World Scientific, 2016, pp. 321–332.
[194] M. Zitnik, E.A. Nam, C. Dinh, A. Kuspa, G. Shaulsky, B. Zupan, Gene prioritization by compressive data fusion and chaining, PLoS Comput. Biol. 11 (10) (2015) e1004552.
[195] C. Zhang, P.L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information, Nucleic Acids Res. (2017) gkx366.
[196] C. Wan, J.G. Lees, F. Minneci, C.A. Orengo, D.T. Jones, Analysis of temporal transcription expression profiles reveal links between protein function and developmental stages of drosophila melanogaster, PLoS Comput. Biol. 13 (10) (2017) e1005791.
[197] D. Amar, R. Shamir, Constructing module maps for integrated analysis of heterogeneous biological networks, Nucleic Acids Res. 42 (7) (2014) 4208–4219.
[198] A. Manichaikul, L. Ghamsari, E.F. Hom, C. Lin, R.R. Murray, R.L. Chang, S. Balaji, T. Hao, Y. Shen, A.K. Chavali, et al., Metabolic network analysis integrated with transcript verification for sequenced genomes, Nat. Methods 6 (8) (2009) 589–592.
[199] E. Kuzmin, B. VanderSluis, W. Wang, G. Tan, R. Deshpande, Y. Chen, M. Usaj, A. Balint, M.M. Usaj, J. van Leeuwen, et al., Systematic analysis of complex genetic interactions, Science 360 (6386) (2018) eaao1729.
[200] P. Gaudet, M.S. Livstone, S.E. Lewis, P.D. Thomas, Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium, Brief. Bioinform. 12 (5) (2011) 449–462.
[201] J. Konc, D. Janežič, Binding site comparison for function prediction and pharmaceutical discovery, Current Opinion in Structural Biology 25 (2014) 34–39.
[202] R. You, S. Zhu, DeepText2Go: improving large-scale protein function prediction with deep semantic text representation, in: IEEE ICBB, IEEE, 2017, pp. 42–49.
[203] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25.
[204] H. Cho, B. Berger, J. Peng, Compact integration of multi-network topology for functional analysis of genes, Cell Syst. 3 (6) (2016) 540–548.
[205] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data., in: ICML, 11, 2011, pp. 809–816.
[206] P. Radivojac, W.T. Clark, T.R. Oron, A.M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, et al., A large-scale evaluation of computational protein function prediction, Nat. Methods 10 (3) (2013) 221.
[207] W. Li, C.-C. Liu, T. Zhang, H. Li, M.S. Waterman, X.J. Zhou, Integrative analysis of many weighted co-expression networks using tensor computation, PLoS Comput. Biol. 7 (6) (2011) e1001106.
[208] L. Ou-Yang, M. Wu, X.-F. Zhang, D.-Q. Dai, X.-L. Li, H. Yan, A two-layer integration framework for protein complex detection, BMC Bioinformatics 17 (1) (2016) 100.
[209] K. Bugge, E. Papaleo, G.W. Haxholm, J.T. Hopper, C.V. Robinson, J.G. Olsen, K. Lindorff-Larsen, B.B. Kragelund, A combined computational and structural model of the full-length human prolactin receptor, Nat. Commun. 7 (2016) 11578.
[210] Y. Shi, R. Pellarin, P.C. Fridy, J. Fernandez-Martinez, M.K. Thompson, Y. Li, Q.J. Wang, A. Sali, M.P. Rout, B.T. Chait, A strategy for dissecting the architectures of native macromolecular assemblies, Nat. Methods 12 (12) (2015) 1135–1138.
[211] C.L. Myers, D. Robson, A. Wible, M.A. Hibbs, C. Chiriac, C.L. Theesfeld, K. Dolinski, O.G. Troyanskaya, Discovery of biological networks from diverse functional genomic data, Genome Biol. 6 (13) (2005) R114.
[212] P. Ray, L. Zheng, J. Lucas, L. Carin, Bayesian joint analysis of heterogeneous genomics data, Bioinformatics 30 (10) (2014) 1370–1376.
[213] A. Ori, B.H. Toyama, M.S. Harris, T. Bock, M. Iskar, P. Bork, N.T. Ingolia, M.W. Hetzer, M. Beck, Integrated transcriptome and proteome analyses reveal organ-specific proteome deterioration in old rats, Cell Syst. 1 (3) (2015) 224–237.
[214] S.V. Andrews, S.E. Ellis, K.M. Bakulski, B. Sheppard, L.A. Croen, I. Hertz-Picciotto, C.J. Newschaffer, A.P. Feinberg, D.E. Arking, C. Ladd-Acosta, et al., Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder, Nat. Commun. 8 (1) (2017) 1011.
[215] J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, A.A. Minai, D.J. Hassett, L.J. Lu, Investigating the predictability of essential genes across distantly related organisms using an integrative approach, Nucleic Acids Res. 39 (3) (2010) 795–807.
[216] B. Hooghe, S. Broos, F. Van Roy, P. De Bleser, A flexible integrative approach based on random forest improves prediction of transcription factor binding sites, Nucleic Acids Res. 40 (14) (2012). e106–e106
[217] M. Setty, K. Helmy, A.A. Khan, J. Silber, A. Arvey, F. Neezen, P. Agius, J.T. Huse, E.C. Holland, C.S. Leslie, Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma, Mol. Syst. Biol. 8 (1) (2012) 605.
[218] C.A. Penfold, J.B. Millar, D.L. Wild, Inferring orthologous gene regulatory networks using interspecies data fusion, Bioinformatics 31 (12) (2015) i97–i105.
[219] S. Imam, D.R. Noguera, T.J. Donohue, An integrated approach to reconstructing genome-scale transcriptional regulatory networks, PLoS Comput. Biol. 11 (2) (2015) e1004103.
[220] A.E. Ihekwaba, I. Mura, J. Walshaw, M.W. Peck, G.C. Barker, An integrative approach to computational modelling of the gene regulatory network controlling Clostridium botulinum type A1 toxin production, PLoS Comput. Biol. 12 (11) (2016) e1005205.
[221] L. Franke, H. Van Bakel, B. Diosdado, M. Van Belzen, M. Wapenaar, C. Wijmenga, TEAM: a tool for the integration of expression, and linkage and association maps, Eur. J. Hum. Genet. 12 (8) (2004) 633.
[222] A. Sifrim, D. Popovic, L.-C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, J.R. Vermeesch, J. Aerts, B. De Moor, Y. Moreau, eXtasy: variant prioritization by genomic data fusion, Nat. Methods 10 (11) (2013) 1083–1084.
[223] G.R. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan, W.S. Noble, A statistical framework for genomic data fusion, Bioinformatics 20 (16) (2004) 2626–2635.
[224] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, et al., Gene prioritization through genomic data fusion, Nat. Biotechnol. 24 (5) (2006) 537–544.
[225] L.-C. Tranchevent, A. Ardeshirdavani, S. ElShal, D. Alcaide, J. Aerts, D. Auboeuf, Y. Moreau, Candidate gene prioritization with endeavour, Nucleic Acids Res. 44 (W1) (2016) W117–W121.
[226] S. Köhler, S. Bauer, D. Horn, P.N. Robinson, Walking the interactome for prioritization of candidate disease genes, The American Journal of Human Genetics 82 (4) (2008) 949–958.
[227] T. De Bie, L.-C. Tranchevent, L.M. Van Oeffelen, Y. Moreau, Kernel-based data fusion for gene prioritization, Bioinformatics 23 (13) (2007) i125–i132.
[228] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res. 37 (suppl_2) (2009) W305–W311.
[229] P.N. Robinson, S. Köhler, A. Oellrich, K. Wang, C.J. Mungall, S.E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, et al., Improved exome prioritization of disease genes through cross-species phenotype comparison, Genome Res. 24 (2) (2014) 340–348.
[230] S.N. Simões, D.C. Martins, C.A. Pereira, R.F. Hashimoto, H. Brentani, NERI: network-medicine based integrative approach for disease gene prioritization by relative importance, BMC Bioinformatics 16 (19) (2015) S9.
[231] D.S. Himmelstein, S.E. Baranzini, Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes, PLoS Comput. Biol. 11 (7) (2015) e1004259.
[232] A.A. Kumar, L. Van Laer, M. Alaerts, A. Ardeshirdavani, Y. Moreau, K. Laukens, B. Loeys, G. Vandeweyer, pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion, Bioinformatics 1 (2018) 9.
[233] G. Pandey, B. Zhang, A.N. Chang, C.L. Myers, J. Zhu, V. Kumar, E.E. Schadt, An integrative multi-network and multi-classifier approach to predict genetic interactions, PLoS Comput. Biol. 6 (9) (2010) e1000928.
[234] E. Bonnet, L. Calzone, T. Michoel, Integrative multi-omics module network inference with lemon-tree, PLoS Comput. Biol. 11 (2) (2015) e1003983.
[235] L.M. Heiser, N.J. Wang, C.L. Talcott, K.R. Laderoute, M. Knapp, Y. Guan, Z. Hu, S. Ziyad, B.L. Weber, S. Laquerre, et al., Integrated analysis of breast cancer cell lines reveals unique signaling pathways, Genome Biol. 10 (3) (2009) R31.
[236] R.K. Nibbe, M. Koyutürk, M.R. Chance, An integrative-omics approach to identify functional sub-networks in human colorectal cancer, PLoS Comput. Biol. 6 (1) (2010) e1000639.
[237] J.D. Rudolph, M. de Graauw, B. van de Water, T. Geiger, R. Sharan, Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks, Cell Syst. 3 (6) (2016) 585–593.
[238] S.R. Piccolo, L.M. Hoffman, T. Conner, G. Shrestha, A.L. Cohen, J.R. Marks, L.A. Neumayer, C.A. Agarwal, M.C. Beckerle, I.L. Andrulis, et al., Integrative analyses reveal signaling pathways underlying familial breast cancer susceptibility, Mol. Syst. Biol. 12 (3) (2016) 860.
[239] J. Dutkowski, M. Kramer, M.A. Surma, R. Balakrishnan, J.M. Cherry, N.J. Krogan, T. Ideker, A gene ontology inferred from molecular networks, Nat. Biotechnol. 31 (1) (2013) 38–45.
[240] C.J. Mungall, C. Torniai, G.V. Gkoutos, S.E. Lewis, M.A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome Biol. 13 (1) (2012) R5.
[241] M. Kulmanov, M.A. Khan, R. Hoehndorf, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, Bioinformatics 34 (2017) 660–668.
[242] J. Ma, M.K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, T. Ideker, Using deep learning to model the hierarchical structure and function of a cell, Nat. Methods 15 (4) (2018) 290.

[243] T. Rolland, M. Taşan, B. Charloteaux, S.J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, et al., A proteome-scale map of the human interactome network, Cell 159 (5) (2014) 1212–1226.
[244] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, A Bayesian networks approach for predicting protein-protein interactions from genomic data, Science 302 (5644) (2003) 449–453.
[245] S.M. Lundberg, W.B. Tu, B. Raught, L.Z. Penn, M.M. Hoffman, S.-I. Lee, ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data, Genome Biol. 17 (1) (2016) 82.
[246] K. Drew, C. Lee, R.L. Huizar, F. Tu, B. Borgeson, C.D. McWhite, Y. Ma, J.B. Wallingford, E.M. Marcotte, Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes, Mol. Syst. Biol. 13 (6) (2017) 932.
[247] Y. Li, J.C. Patra, Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (9) (2010) 1219–1224.
[248] C. Blatti, S. Sinha, Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks, Bioinformatics 32 (14) (2016) 2167–2175.
[249] Y. Liu, X. Zeng, Z. He, Q. Zou, 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.
[250] J.W. Scannell, A. Blanckley, H. Boldon, B. Warrington, Diagnosing the decline in pharmaceutical r&d efficiency, Nat. Rev. Drug Discov. 11 (3) (2012) 191.
[251] P.J. Yeh, M.J. Hegreness, A.P. Aiden, R. Kishony, Drug interactions and the evolution of antibiotic resistance, Nat. Rev. Microbiol. 7 (6) (2009) 460.
[252] J. Li, et al., A survey of current trends in computational drug repositioning, Brief. Bioinform. 17 (1) (2015) 2–12.
[253] B.R. Donald, Algorithms in structural molecular biology, MIT Press, 2011.
[254] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry, Nat. Biotechnol. 25 (2) (2007) 197.
[255] K. Bleakley, Y. Yamanishi, Supervised prediction of drug–target interactions using bipartite local models, Bioinformatics 25 (18) (2009) 2397–2403.
[256] T. van Laarhoven, S.B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics 27 (21) (2011) 3036–3043.
[257] S. Wang, J. Peng, Network-assisted target identification for haploinsufficiency and homozygous profiling screens, PLoS Comput. Biol. 13 (6) (2017) e1005553.
[258] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, Y. Yamanishi, Relating drug–protein interaction network with drug side effects, Bioinformatics 28 (18) (2012) i522–i528.
[259] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, et al., Discovery of drug mode of action and drug repositioning from transcriptional responses, Proc. Natl. Acad. Sci. 107 (33) (2010) 14621–14626.
[260] W. Wang, S. Yang, X. Zhang, J. Li, Drug repositioning by integrating target information through a heterogeneous network model, Bioinformatics 30 (20) (2014) 2923–2930.
[261] F. Yang, J. Xu, J. Zeng, Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data, in: Pacific Symposium on Biocomputing, World Scientific, 2014, pp. 148–159.
[262] M. Gönen, S. Khan, S. Kaski, Kernelized bayesian matrix factorization, in: ICML, 2013, pp. 864–872.
[263] X. Zhang, L. Li, M.K. Ng, S. Zhang, Drug–target interaction prediction by integrating multiview network data, Comput. Biol. and Chem. 69 (2017) 185–193.
[264] M. Breinig, F.A. Klein, W. Huber, M. Boutros, A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells, Mol. Syst. Biol. 11 (12) (2015) 846.
[265] S. Lee, C. Zhang, Z. Liu, M. Klevstig, B. Mukhopadhyay, M. Bergentall, R. Cinar, M. Ståhlman, N. Sikanic, J.K. Park, et al., Network analyses identify liver-specific targets for treating liver diseases, Mol. Syst. Biol. 13 (8) (2017) 938.
[266] Y. Sun, J. Han, X. Yan, P.S. Yu, T. Wu, Pathsim: Meta path-based top-k similarity search in heterogeneous information networks, VLDB 4 (11) (2011) 992–1003.
[267] G. Fu, Y. Ding, A. Seal, B. Chen, Y. Sun, E. Bolton, Predicting drug target interactions using meta-path-based semantic network analysis, BMC Bioinformatics 17 (1) (2016) 160.
[268] X. Zheng, H. Ding, H. Mamitsuka, S. Zhu, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, in: KDD, ACM, 2013, pp. 1025–1033.
[269] A. Narita, K. Hayashi, R. Tomioka, H. Kashima, Tensor factorization using auxiliary information, Data Mining and Knowledge Discovery 25 (2) (2012) 298–324.
[270] M. Zitnik, B. Zupan, Collective pairwise classification for multi-way analysis of disease and drug data, in: Pacific Symposium on Biocomputing, 21, NIH Public Access, 2016, p. 81.
[271] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[272] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nat. Commun. 8 (2017) 573.
[273] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, N.P. Tatonetti, Similarity-based modeling in large-scale prediction of drug-drug interactions, Nature Protocols 9 (9) (2014) 2147.
[274] F. Cheng, Z. Zhao, 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.
[275] D. Sridhar, S. Fakhraei, L. Getoor, A probabilistic approach for collective similarity-based drug–drug interaction prediction, Bioinformatics 32 (20) (2016) 3175–3182.
[276] K. Han, et al., Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions, Nat. Biotechnol. (2017).
[277] J. Jia, et al., Mechanisms of drug combinations: interaction and network perspectives, Nat. Rev. Drug Discov. 8 (2) (2009) 111–128.
[278] Y. Sun, et al., Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer, Nat. Commun. 6 (2015) 8481.
[279] H.G. Woo, J.-H. Choi, S. Yoon, B.A. Jee, E.J. Cho, J.-H. Lee, S.J. Yu, J.-H. Yoon, N.-J. Yi, K.-W. Lee, et al., Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer, Nat. Commun. 8 (1) (2017) 839.
[280] X. Chen, et al., NLLSS: predicting synergistic drug combinations based on semi-supervised learning, PLoS Comput. Biol. 12 (7) (2016) e1004975.
[281] E.D. Kantor, et al., Trends in prescription drug use among adults in the United States from 1999-2012, Journal of the American Medical Association 314 (17) (2015) 1818–1830.
[282] K.A. Ryall, A.C. Tan, Systems biology approaches for advancing the discovery of effective drug combinations, J. Cheminformatics 7 (1) (2015) 7.
[283] S. Loewe, The problem of synergism and antagonism of combined drugs, Arzneimittel-Forschung 3 (1953) 285–290.
[284] R. Lewis, et al., Synergy Maps: exploring compound combinations using network-based visualization, J. Cheminformatics 7 (1) (2015) 36.
[285] M. Bansal, et al., A community computational challenge to predict the activity of pairs of compounds, Nat. Biotechnol. 32 (12) (2014) 1213–1222.
[286] T. Takeda, et al., Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge, J. Cheminformatics 9 (1) (2017) 16.
[287] L. Huang, et al., DrugComboRanker: drug combination discovery based on target network analysis, Bioinformatics 30 (12) (2014a) i228–i236.
[288] H. Huang, et al., Systematic prediction of drug combinations based on clinical side–effects, Sci. Rep. 4 (2014b).
[289] Y. Sun, et al., Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer, Nat. Commun. 6 (2015) 8481.
[290] M. Zitnik, B. Zupan, Collective pairwise classification for multi-way analysis of disease and drug data, in: Pacific Symposium on Biocomputing, 21, 2016, p. 81.
[291] D. Chen, et al., Synergy evaluation by a pathway–pathway interaction network: a new way to predict drug combination, Mol. BioSyst. 12 (2) (2016) 614–623.
[292] J.-Y. Shi, et al., Predicting combinative drug pairs towards realistic screening via integrating heterogeneous features, BMC Bioinformatics 18 (12) (2017) 409.
[293] F. Cheng, Z. Zhao, 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.
[294] W. Zheng, H. Lin, L. Luo, Z. Zhao, Z. Li, Y. Zhang, Z. Yang, J. Wang, An attention-based effective neural model for drug–drug interactions extraction, BMC Bioinformatics 18 (1) (2017) 445.
[295] Z. Zhao, Z. Yang, L. Luo, H. Lin, J. Wang, Drug–drug interaction extraction from biomedical literature using syntax convolutional neural network, Bioinformatics 32 (22) (2016) 3444–3453.
[296] A. Gottlieb, et al., INDI: a computational framework for inferring drug interactions and their associated recommendations, Mol. Syst. Biol. 8 (1) (2012) 592.
[297] S. Vilar, et al., Drug-drug interaction through molecular structure similarity analysis, Journal of the American Medical Informatics Association 19 (6) (2012) 1066–1074.
[298] X. Li, et al., Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles, Artificial Intelligence in Medicine (2017).
[299] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Label propagation prediction of drug-drug interactions based on clinical side effects, Sci. Rep. 5 (2015).
[300] R. Ferdousi, et al., Computational prediction of drug-drug interactions based on drugs functional similarities, J. Biomed. Inform. 70 (2017) 54–64.
[301] W. Zhang, et al., Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, BMC Bioinformatics 18 (1) (2017) 18.
[302] T. Ma, C. Xiao, J. Zhou, F. Wang, Drug similarity integration through attentive multi-view graph auto-encoders, in: IJCAI, 2018, pp. 1–7.
[303] J.Y. Ryu, H.U. Kim, S.Y. Lee, Deep learning improves prediction of drug–drug and drug–food interactions, Proc. Natl. Acad. Sci. 115 (18) (2018) E4304–E4311.
[304] W.L. Hamilton, R. Ying, J. Leskovec, Representation learning on graphs: Methods and applications, IEEE Data Eng. Bull. (2017).
[305] E. Guney, J. Menche, M. Vidal, A.-L. Barábasi, Network-based in silico drug efficacy screening, Nat. Commun. 7 (2016) 10331.
[306] M. Zitnik, V. Janjic, C. Larminie, B. Zupan, P. Natasa, Discovering disease-disease associations by fusing systems-level molecular data, Sci. Rep. 3 (2013).
[307] J. Li, X. Zhu, J.Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts, PLoS Comput. Biol. 5 (7) (2009) e1000450.
[308] Z. Wu, Y. Wang, L. Chen, Network-based drug repositioning, Mol. BioSyst. 9 (6) (2013) 1268–1281.
[309] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of drug-target interactions and drug repositioning via network-based inference, PLoS Comput. Biol. 8 (5) (2012) e1002503.
[310] S. Zhao, S. Li, A co-module approach for elucidating drug–disease associations and revealing their molecular basis, Bioinformatics 28 (7) (2012) 955–961.
[311] M. Sirota, J.T. Dudley, J. Kim, A.P. Chiang, A.A. Morgan, A. Sweet-Cordero, J. Sage, A.J. Butte, Discovery and preclinical validation of drug indications using compen-

dia of public gene expression data, Science Translational Medicine 3 (96) (2011). 96ra77–96ra77
[312] Z. Stanfield, M. Coşkun, M. Koyutürk, Drug response prediction as a link prediction problem, Sci. Rep. 7 (2017) 40321.
[313] K.W. Fung, C.S. Jao, D. Demner-Fushman, Extracting drug indication information from structured product labels using natural language processing, Journal of the American Medical Informatics Association 20 (3) (2013) 482–488.
[314] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Exploring the relationship between drug side-effects and therapeutic indications, in: Proceedings of the AMIA Annual Symposium, 2013, American Medical Informatics Association, 2013, p. 1568.
[315] M. Kuhn, M. Al Banchaabouchi, M. Campillos, L.J. Jensen, C. Gross, A.-C. Gavin, P. Bork, Systematic identification of proteins that elicit drug side effects, Mol. Syst. Biol. 9 (1) (2013) 663.
[316] F. Wang, P. Zhang, N. Cao, J. Hu, R. Sorrentino, Exploring the associations between drug side-effects and therapeutic indications, J. Biomed. Inform. 51 (2014) 15–23.
[317] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, Mol. Syst. Biol. 7 (1) (2011) 496.
[318] P. Zhang, P. Agarwal, Z. Obradovic, 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] J. Li, Z. Lu, Pathway-based drug repositioning using causal inference, BMC Bioinformatics 14 (16) (2013) S3.
[320] L. Yu, R. Su, B. Wang, L. Zhang, Y. Zou, J. Zhang, L. Gao, 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.
[321] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, Y. Pan, Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm, Bioinformatics 32 (17) (2016) 2664–2671.
[322] D.S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S.L. Chen, D. Hadley, A. Green, P. Khankhanian, S.E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Elife 6 (2017) e26726.
[323] W. Wang, S. Yang, J. Li, Drug target predictions based on heterogeneous graph inference, in: Pacific Symposium on Biocomputing, World Scientific, 2013, pp. 53–64.
[324] P. Zhang, F. Wang, J. Hu, Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity, in: Proceedings of the AMIA Annual Symposium, 2014, 2014, p. 1258.
[325] C. Curtis, S.P. Shah, S.-F. Chin, G. Turashvili, O.M. Rueda, M.J. Dunning, D. Speed, A.G. Lynch, S. Samarajiwa, et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature 486 (7403) (2012) 346–352.
[326] F.M. Cavalli, M. Remke, L. Rampasek, et al., Intertumoral heterogeneity within medulloblastoma subgroups, Cancer Cell 31 (6) (2017). 737–754.e6
[327] J.M. Nigro, A. Misra, L. Zhang, I. Smirnov, H. Colman, C. Griffin, N. Ozburn, M. Chen, E. Pan, D. Koul, et al., Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma, Cancer Res. 65 (5) (2005) 1678–1686.
[328] R.G. Verhaak, K.A. Hoadley, E. Purdom, V. Wang, Y. Qi, M.D. Wilkerson, C.R. Miller, L. Ding, T. Golub, J.P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B.A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H.S. Feiler, J.G. Hodgson, C.D. James, J.N. Sarkaria, C. Brennan, A. Kahn, P.T. Spellman, R.K. Wilson, T.P. Speed, J.W. Gray, M. Meyerson, G. Getz, C.M. Perou, D.N. Hayes, 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.
[329] D.C. Koboldt, R.S. Fulton, M.D. McLellan, H. Schmidt, J. Kalicki-Veizer, J.F. McMichael, L.L. Fulton, D.J. Dooling, L. Ding, E.R. Mardis, et al., Comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70.
[330] K.A. Hoadley, C. Yau, D.M. Wolf, et al., Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin, Cell 158 (4) (2014) 929–944.
[331] R. Shen, A.B. Olshen, M. Ladanyi, 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.
[332] Y. Yuan, R.S. Savage, F. Markowetz, Patient-specific data fusion defines prognostic cancer subtypes, PLoS Comput. Biol. 7 (10) (2011) e1002227.
[333] W.C. de Vega, L. Erdman, S.D. Vernon, A. Goldenberg, P.O. McGowan, Integration of dna methylation and health scores identifies subtypes in myalgic encephalomyelitis/chronic fatigue syndrome, Epigenomics 10 (5) (2018) 539–557.
[334] A.N. Zizzo, L. Erdman, B.M. Feldman, A. Goldenberg, 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
[335] L. Stefanik, L. Erdman, S.H. Ameis, G. Foussias, B.H. Mulsant, T. Behdinan, A. Goldenberg, L.J. O'Donnell, A.N. Voineskos, Brain-behavior participant similarity networks among youth and emerging adults with schizophrenia spectrum, autism spectrum, or bipolar disorder and matched controls, Neuropsychopharmacology 43 (5) (2017) 1180–1188.
[336] B.J. Raphael, R.H. Hruban, A.J. Aguirre, R.A. Moffitt, J.J. Yeh, C. Stewart, A.G. Robertson, A.D. Cherniack, M. Gupta, G. Getz, et al., Integrated genomic characterization of pancreatic ductal adenocarcinoma, Cancer Cell 32 (2) (2017). 185–203.e13
[337] H.-C. Huang, Y.-Y. Chuang, C.-S. Chen, Affinity aggregation for spectral clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 773–780.
[338] S. Pai, G.D. Bader, Patient similarity networks for precision medicine, J. Mol. Biol. (2018).
[339] C.J. Vaske, S.C. Benz, J.Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, J.M. Stuart, Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, Bioinformatics 26 (12) (2010) i237–i245.
[340] C. Wang, B. Gong, P.R. Bushel, J. Thierry-Mieg, D. Thierry-Mieg, J. Xu, H. Fang, H. Hong, J. Shen, Z. Su, et al., The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance, Nat. Biotechnol. 32 (9) (2014) 926.
[341] C.A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, J.C. Marioni, Normalizing single-cell RNA sequencing data: challenges and opportunities, Nat. Methods (2017).
[342] N. Hiranuma, S.M. Lundberg, S.-I. Lee, AIControl: Replacing matched control experiments with machine learning improves ChIP-seq peak identification, bioRxiv (2018) 278762.
[343] R. Bacher, L.-F. Chu, N. Leng, A.P. Gasch, J.A. Thomson, R.M. Stewart, M. Newton, C. Kendziorski, SCnorm: robust normalization of single-cell RNA-seq data, Nat. Methods 14 (6) (2017) 584–586.
[344] J.N. Taroni, C.S. Greene, Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously, bioRxiv (2017) 118349.
[345] B. Wang, A. Pourshafeie, M. Zitnik, J. Zhu, C.D. Bustamante, S. Batzoglou, J. Leskovec, Network Enhancement: a general method to denoise weighted biological networks, Nat. Commun. 9 (2018) 3108.
[346] T. Milenkovic, N. Przulj, Uncovering biological network function via graphlet degree signatures, Cancer Inform. 6 (2008) CIN–S680.
[347] A.R. Benson, D.F. Gleich, J. Leskovec, Higher-order organization of complex networks, Science 353 (6295) (2016) 163–166.
[348] A.H. Rizvi, P.G. Camara, E.K. Kandror, T.J. Roberts, I. Schieren, T. Maniatis, R. Rabadan, Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development, Nat. Biotechnol. 35 (6) (2017) 551–560.
[349] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: KDD, ACM, 2016, pp. 1135–1144.
[350] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: NIPS, 2017, pp. 4768–4777.
[351] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., A closer look at memorization in deep networks, in: ICML, 2017, pp. 1–10.
[352] P.W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: ICML, 2017, pp. 1–11.
[353] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery, bioRxiv (2017) 206540.
[354] J.Y. Tung, C.B. Do, D.A. Hinds, A.K. Kiefer, J.M. Macpherson, A.B. Chowdry, U. Francke, B.T. Naughton, J.L. Mountain, A. Wojcicki, et al., Efficient replication of over 180 genetic associations with self-reported medical data, PLoS One 6 (8) (2011) e23473.
[355] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.