Leading Edge
Review

Cell

Next-Generation Machine Learning for Biological Networks

Diogo M. Camacho,[1] Katherine M. Collins,[1,2] Rani K. Powers,[3] James C. Costello,[3,*] and James J. Collins[1,4,5,*]
[1]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA
[2]Department of Brain & Cognitive Sciences and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Computational Bioscience Program, Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[4]Department of Biological Engineering and Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
*Correspondence: james.costello@ucdenver.edu (J.C.C.), jimjc@mit.edu (J.J.C.)
https://doi.org/10.1016/j.cell.2018.05.015

Machine learning, a collection of data-analytical techniques aimed at building predictive models from multi-dimensional datasets, is becoming integral to modern biological research. By enabling one to generate models that learn from large datasets and make predictions on likely outcomes, machine learning can be used to study complex cellular systems such as biological networks. Here, we provide a primer on machine learning for life scientists, including an introduction to deep learning. We discuss opportunities and challenges at the intersection of machine learning and network biology, which could impact disease biology, drug discovery, microbiome research, and synthetic biology.

Introduction

Over the last decade, we have seen a dramatic increase in the number of large, highly complex datasets being generated from biological experiments, quantifying molecular variables such as gene, protein, and metabolite abundance, microbiome composition, and population-wide genetic variation, to name just a few. Community efforts across research disciplines are regularly generating petabytes of data. For example, The Cancer Genome Atlas has sampled multiple -omics measurements from over 30,000 patients across dozens of different cancer types, totaling over 2.5 petabytes of raw data. Projects of similar scope, such as the Human Microbiome Project, the ENCODE Project Consortium, and the 100,000 Genomes Project, are generating overwhelming amounts of data from bacteria to humans.

These datasets present the raw material needed to gain insights into biological systems and complex diseases, but the potential of these data can only be realized through higher-level analysis. The above projects illustrate why it is becoming imperative to focus our data-analytical approaches on tools and techniques specifically tailored to handle large, heterogeneous, complex datasets. Machine learning, an area of long-standing and growing interest in biological research, aims to address this complexity, providing next-level analyses that allow one to take new perspectives and generate novel hypotheses about living systems.

Machine learning is a discipline in computer science wherein machines (i.e., computers) are programmed to learn patterns from data. The learning itself is based on a set of mathematical rules and statistical assumptions. A common goal in machine learning is to develop a predictive model based on statistical associations among features from a given dataset. The learned model can then be used to predict any range of outputs, such as binary responses, categorical labels, or continuous values. Briefly, for a problem of interest—say, the identification and annotation of genes in a newly sequenced genome—a machine-learning algorithm will learn key properties of existing annotated genomes, such as what constitutes a transcriptional start site and specific genomic properties of genes such as GC content and codon usage, and will then use this knowledge to generate a model for finding genes given all of the genomic sequences on which it was trained. For a newly sequenced genome, the algorithm will apply what it has learned from the training data to make predictions about the putative functional organization of the genome.

Applications of machine learning are becoming ubiquitous in biology and encompass not only genome annotation (see, e.g., Leung et al., 2016; Yip et al., 2013), but also predictions of protein binding (see, e.g., Alipanahi et al., 2015; Ballester and Mitchell, 2010), the identification of key transcriptional drivers of cancer (Califano and Alvarez, 2017; Carro et al., 2010), predictions of metabolic functions in complex microbial communities (Langille et al., 2013), and the characterization of transcriptional regulatory networks (Djebali et al., 2012; Marbach et al., 2012), to name just a few. In short, any task where a pattern can be learned and then applied to a new dataset falls under the auspices of machine learning. A key advantage is that machine-learning methods can sift through volumes of data to find patterns that

would be missed otherwise. In the age of big data in biological and biomedical research, machine learning plays a critical role in finding predictive patterns in complex biological systems.

Here, we provide a high-level description of machine learning as it relates to biological research and explore opportunities at the intersection of machine learning and network biology, an area of research that deals with biological networks and large multi-dimensional datasets. Network biology involves the study of the complex interactions of biomolecules that contribute to the structures and functions of living cells. The field plays a central role in the modeling of biological systems, complemented by the highly complex datasets generated across a myriad of multi-omics programs. Network biology involves both the reconstruction and analysis of large-scale endogenous biological networks (in the context of systems biology), as well as the design and construction of small-scale synthetic gene networks (in the context of synthetic biology). The area has benefited from machine learning largely in the identification of network architectures (Butte and Kohane, 2000; Cahan et al., 2014; Faith et al., 2007; Friedman, 2004; Margolin et al., 2006). The diversity of approaches for network inference is extensive, and we direct the reader to important reviews and commentary on the subject (e.g., De Smet and Marchal, 2010; Hill et al., 2016; Marbach et al., 2012). These reverse-engineering approaches have shown a remarkable ability to learn patterns from input data to generate biologically relevant gene regulatory networks, with interesting applications in the identification of drivers of drug response or disease phenotypes (e.g., Akavia et al., 2010; di Bernardo et al., 2005; Costello et al., 2014; Walsh et al., 2017). As we will discuss below, there are many additional opportunities for moving the field forward through the integration of multi-omics datasets and phenotypic measurements with novel machine-learning methods.

This Review is intended for biological researchers who are curious about recent developments and applications in machine learning and its potential for advancing network biology given the vast amounts of data being generated today. We start by leading the reader through a primer on machine learning, where we discuss key concepts needed to understand how machine learning approaches can be applied and utilized in network biology. We include a brief introduction to deep learning, a powerful form of next-generation machine learning. This is followed by a discussion on how next-generation machine learning methods could be used to expand our understanding of biological networks and disease biology, to discover and develop novel therapeutic compounds, to study and characterize host-microbiome interactions, and to identify design rules and functional network architectures for synthetic gene circuits. We highlight these opportunities, as well as an array of challenges that need to be addressed to fully realize the potential of machine learning in network biology.

A Primer on Machine Learning

Below, we introduce and describe the basic concepts, general workflows, and main categories of machine learning. We offer some thoughts on principles that should be considered when designing and implementing a machine-learning method in biological research. We also include a brief discussion on deep learning, a next-generation machine-learning approach that is increasingly being applied in medicine and biology. This primer is intended for readers with little to no knowledge of machine-learning algorithms.

Basics of Machine Learning

Machine-learning methods aim to generate predictive models based on an underlying algorithm and a given dataset. The input data to a machine-learning algorithm typically consist of "features" and "labels" across a set of samples. Features are the measurements across all samples, either raw or mathematically transformed, while labels are what the machine-learning model aims to predict—that is, the output of the model. As we discuss below, machine learning algorithms can also deal with datasets lacking labels. Illustrated in Figure 1, the general machine-learning workflow is to first, process the input data; second, learn or train the underlying model (a set of mathematical formulas and statistical assumptions that define the learning rules); and third, use the machine learning model to make predictions on new data.

The learning process itself refers to finding the optimal set of model parameters that translate the features in the input data into accurate predictions of the labels. The parameters are found through a series of back and forth steps, where parameters are estimated, the model performance is evaluated, errors are identified and corrected, and then the process repeats. This process is called training and will proceed until the model performance cannot be improved upon, which is assessed by the minimization of the model error. Once the optimal parameters are identified, the model can be used to make predictions using new data.

In biological applications, features can include one or more types of data, such as gene expression profiles, a genomic sequence, protein-protein interactions, metabolite concentrations, or copy number alterations. Features can be continuous (e.g., gene expression values), categorical (e.g., gene functional annotation), or binary (e.g., genes on or off). Labels, like features, can be continuous (e.g., growth rate), categorical (e.g., stage of disease), or binary (e.g., pathogenic or non-pathogenic). As labels can be continuous or discrete, many machine-learning methods fall under regression or classification tasks, respectively, where a regression task involves the prediction of a continuous output variable and classification tasks involve the prediction of discrete output variables.

As noted above, the goal of training a machine-learning model is to use it to make predictions on new data. If the model is accurate on the training data, as well as on independent datasets (e.g., test data), then the model is said to have been properly learned. However, a given machine-learning model can be trained to predict the training data with high accuracy while failing to make accurate predictions on test data. This is referred to as overfitting and occurs when the parameters for the model are fit so specifically to the training data that they do not provide predictive power outside these data. It is also possible to have an underfit machine learning model, where the model does not accurately predict the training data. Overfitting and underfitting are major causative factors underlying poor performance of machine-learning approaches. The former can arise when the machine-learning model is too complex (too many adjustable parameters) relative to the number of samples in the training

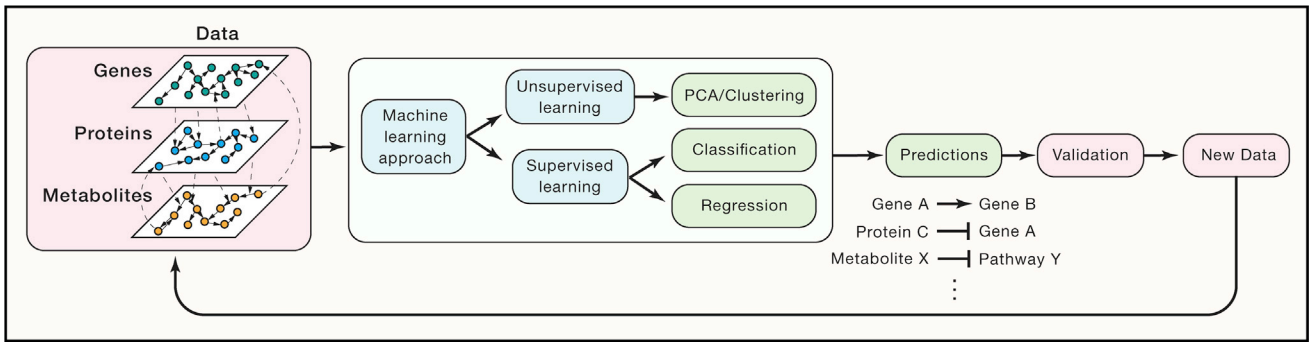

Figure 1. Machine-Learning Applications Build Models to Interpret and Analyze Datasets
Data consist of features measured over many samples, including quantification of genes, proteins, metabolites, and edges within networks. A machine-learning approach is selected based on the prediction task, underlying properties of the data, and if the data are labeled or unlabeled. If the data are unlabeled, then an unsupervised approach is needed, such as PCA or hierarchical clustering. If the data are labeled, then a supervised approach can be applied, which will generate a predictive model for either regression or classification of the data based on input labels. After applying the appropriate machine learning approach, the predictions must be validated. New data can be generated or collected and used to refine the learned model, improve prediction performance, and develop novel biological hypotheses.

dataset, while the latter occurs when the model is too simple. Overfitting can be addressed by increasing the size of the training dataset and/or decreasing the complexity of the learning model, whereas underfitting can be remediated by increasing the model's complexity (Domingos, 2012).

The quality of the input data, in addition to the quantity of the training data, is key to the entire machine-learning process. The old computer-science adage of "garbage in, garbage out" was never truer than it is with machine-learning applications. The performance of any given machine-learning algorithm is dependent on the data used to train the model. Properly formatting, cleaning, and normalizing the input data constitute critical first steps. The input dataset might have many missing values and, thus, is incomplete. The options for dealing with missing data include inferring the missing values directly (e.g., imputation) or simply removing sparse features. Moreover, not every input feature in a given biological dataset will be informative for predicting the output labels. In fact, including irrelevant features can lead to overfitting and therefore hinder the performance of the machine-learning model. A process called feature selection is often used to identify informative features. An example of a feature selection technique is to correlate all input features with the labels and retain only those features that meet a pre-defined threshold. For additional insight into input data and feature selection, we refer the reader to several excellent articles (Chandrashekar and Sahin, 2014; Domingos, 2012; Guyon and Elisseeff, 2003; Little and Rubin, 1987; Saeys et al., 2007).

Categories of Machine-Learning Methods

There are two overarching categories of machine learning methods—namely, unsupervised and supervised learning (see James et al., 2013; Rencher, 2002). Unsupervised approaches are used when the labels on the input data are unknown; these methods learn only from patterns in the features of the input data. Commonly used unsupervised methods include principal components analysis (PCA) and hierarchical clustering. The goal of unsupervised approaches is to group or cluster subsets of the data based on similar features and to identify how many groups or clusters are present in the data. While the machine is used to identify clusters or reduce the dimensions of data directly, an independent predictive model is not produced. In practice, when new data become available, there are two options: (1) the new data can be mapped into the clustered or dimension-reduced space or (2) the clustering or reduction of dimensions can be performed once again with all of the data included. Using either of these approaches, one can determine where the new data fit with respect to the original data (Ghahramani, 2004).

Unsupervised techniques can be advantageous in certain situations. For instance, in a case where the sample labels are missing or incorrect, unsupervised methods can still identify patterns, since the clustering is performed purely on the input data. Additionally, unsupervised methods are well suited for visualization of high-dimensional input data. As an example, by plotting the first two principal components of a PCA, one can judge the relative distance (a metric of similarity) between samples on a simple two-dimensional plot summarizing information from hundreds or thousands of features (Abdi and Williams, 2010; Shlens, 2014).

Supervised methods, on the other hand, are applied when labels are available for the input data. In this case, the labels are used to train the machine-learning model to recognize patterns that are predictive of the data labels. Supervised methods are more typically associated with machine-learning applications because the trained model is a predictive one; thus, when new input data become available, predictions using the trained model can be directly made. Of note, the output of unsupervised approaches can be used as input to supervised approaches. For example, the clusters discovered in hierarchical clustering can be used as input features to supervised methods. Additionally, supervised models can use the output of PCA as input and work directly on the reduced feature space, as opposed to the full set of input features.

Two notable sub-classes of machine-learning methods that fall under the umbrella of supervised methods are semi-supervised learners and ensemble learners. Semi-supervised methods can be utilized in situations where the labels are

incomplete, e.g., only a small amount of the training data are labeled. This occurs quite often in biological contexts, e.g., for a set of genes of interest, only a small subset may be functionally annotated. With semi-supervised learning, the labeled data are used to infer labels for the unlabeled data, and/or the unlabeled data are utilized to gain insights on the structure of the training dataset. Semi-supervised learning aims to surpass the model performance that can be achieved either by ignoring the labels and conducting unsupervised learning or by ignoring the unlabeled data and conducting supervised learning. Ensemble learners, on the other hand, combine multiple independent machine-learning models into a single predictive model so as to obtain better predictive performance. These methods are based on the fact that all machine-learning approaches are biased to identify method-specific patterns. Thus, combining multiple learners can produce better and more robust predictions compared to an individual learner (Dietterich, 2000; Marbach et al., 2012; Rokach, 2010).

Applying Machine Learning in Biological Contexts

There are several factors to consider when selecting and applying machine-learning algorithms to biological questions, particularly given the variability of biological data and the different experimental platforms and protocols used to collect such data. Due to both technical and biological differences, a machine-learning model trained on one dataset may not generalize well to other datasets. Any new dataset should match the general properties of the data used to train the model. The new data should also be processed using the same pipeline as the training data. Should the new data differ significantly from the training data, the predictions from the machine-learning model will most likely be spurious.

Machine-learning methods, much like molecular biology techniques, are context specific. Both machine learning and molecular biology experiments require careful experimental design to properly test a hypothesis. While the broad goal of machine learning is to develop predictive models, the algorithms that underlie the predictors make different assumptions, and their performance may change under different conditions. All methodological choices have tradeoffs; this concept is widely appreciated in computer science and has been termed the "no free lunch theorem" (Wolpert and Macready, 1997).

The performance of machine-learning methods can be affected by multiple factors, including feature selection, user-defined parameters, and the implementation of the methods themselves. Direct evidence that these factors are major drivers of machine-learning performance in biological applications can be found in the Dialogue for Reverse Engineering Assessment and Methodology (DREAM) challenges, an open-data, crowdsourcing effort to find solutions to big-data research questions in network biology and medicine (Costello and Stolovitzky, 2013). Examples of previous challenges include the inference of genome-scale gene regulatory networks (Marbach et al., 2012) and the prediction of drug sensitivities and synergies using multi-omic datasets (Bansal et al., 2014; Costello et al., 2014). Many network inference approaches can be defined as unsupervised learning, where the input data are used to predict interactions (edges) between biomolecular entities (i.e., features) given the set of experimental observations. A second category of network inference algorithms uses supervised learning approaches, where an underlying inferred network is used to make predictions on a novel sample. Such approaches have been highly successful in the characterization of drug mechanism of action (di Bernardo et al., 2005; Bisikirska et al., 2016; Costello et al., 2014) or drivers of disease states (e.g., Akavia et al., 2010; Mezlini and Goldenberg, 2017).

Each DREAM challenge presents the network biology research community with a specific question and the necessary data to address it. Computational models, commonly machine-learning methods, are needed to address each challenge, but there are no restrictions placed on the types of models that can be applied. A fundamental component to each challenge is a gold standard, an evaluation dataset that is hidden from all participants and used to assess each method's performance, thus providing an independent, unbiased assessment to rank the different methods. With several dozen DREAM challenges completed (Saez-Rodriguez et al., 2016), it is possible to identify consistent patterns that can be distilled into three "rules of thumb" for applying machine learning approaches in network biology:

(1) Simple is often better: Regardless of the challenge, it is almost certain that a straightforward machine learning approach will be among the top performing models. These models often include linear regression-based models (e.g., elastic nets), which perform well across a range of machine learning tasks and thus present an excellent starting point.
(2) Prior knowledge improves performance: The application of domain-specific knowledge almost always helps any predictive model. For example, a challenge was run to reverse engineer signaling networks in breast cancer using phospho-proteomic measurements (Hill et al., 2016). The use of prior knowledge of elements and connections in the signaling network enhanced the ability of machine learning approaches to predict causal signaling interactions.
(3) Ensemble models produce robust results: As discussed above, ensemble models integrate predictions from multiple, independent predictors. If done properly, the strongest signals across predictors will rise to the top. Ensemble predictors consistently performed among the best across challenges and tended to be the most robust to noise in the datasets.

The DREAM challenges present ideal sets of results to analyze and compare the performance of different machine learning methods. Across different challenges, it can be seen that no single machine-learning method or class of methods always performs best. Thus, there is no "magic bullet" method that will optimally solve all machine learning tasks in network biology. For additional insight into machine learning in the context of biological research, we refer the reader to several excellent review articles (Califano et al., 2012; Pe'er and Hacohen, 2011; Zhang et al., 2017).

Deep Learning: Next-Generation Machine Learning

Next-generation sequencing technologies introduced a shift in the throughput, scalability, and speed with which nucleotide

sequences could be analyzed. Here, we use the term "next generation" to describe machine-learning approaches that are being developed and used to deal with the explosion of data in many fields, including biology and medicine. We focus our discussion on deep learning, a next-generation machine-learning approach that is increasingly being applied to cope with the complexity and volume of these data.

Deep-learning methods typically utilize neural networks. Loosely modeled after neurons in the human brain, neural networks transmit information through layers of weighted, interconnected computational units or "neurons" (McCulloch and Pitts, 1943; Parker, 1985; Rumelhart et al., 1986; Werbos, 1974). The simplest neural network architecture has three layers: an input layer, a middle or hidden layer, and an output or prediction layer. The neurons in the input layer take the raw data as input and pass the information to the hidden layer, which uses a mathematical function to transform the raw data into a "representation" that helps the machine learn patterns within the data. The output layer relays back to the problem at hand—classification or regression—based on the transformation performed by the hidden layer (Angermueller et al., 2016). The objective is to train the neural network such that it learns the appropriate representations to accurately predict output values for new sets of input data.

A deep neural network is a neural network that includes multiple hidden layers (Figure 2A); the greater the number of hidden layers, the deeper the neural network. The hidden layers are connected sequentially such that each of the hidden layers learns properties about the structure of the data by taking as input the transformed representation produced from the previous hidden layer. Researchers can define the number and size of the hidden layers depending on the purpose of the learning model. For example, a recurrent neural network (RNN) takes as input one-dimensional sequential data, such as words in a sentence or bases in a DNA sequence (Angermueller et al., 2016; LeCun et al., 2015). RNNs have "thin" hidden layers, often comprised of single neurons connected in a linear architecture. A convolutional neural network (CNN), on the other hand, processes data with two or more dimensions, such as a two-dimensional image or a high-dimensional multi-omics dataset. CNNs often have complex hidden layers consisting of many neurons in each layer (Ching et al., 2018; LeCun et al., 2015).

A crucial aspect of deep learning is that the behavior of these layers—that is, how they transform the data—can be learned by the machine rather than defined by the researcher (Angermueller et al., 2016; LeCun et al., 2015). Deep neural networks accomplish this by iteratively tuning their internal parameters to minimize prediction error, typically via a process known as backpropagation. With backpropagation, an error signal based on the difference between the model's output and the target output is computed and sent back through the system (Mitchell, 1997). The parameters (or weights) in each layer of the neural network are then adjusted so that the error for each neuron and the error for the network as a whole are minimized. This process is repeated many times until the difference between the model's output (prediction) and the target output are reduced to an acceptable level. Because deep learning methods attempt to construct hidden layers that learn features that best predict successful outcomes for a given task, they can recognize novel patterns in complex datasets that would have been missed by other techniques (Angermueller et al., 2016; Krizhevsky et al., 2012; LeCun et al., 2015). This is an especially powerful tool for biological applications and enables one to extract the most predictive features from complex datasets.

A key drawback of the deep learning paradigm is that training a deep neural network requires massive datasets of a size often not be attainable in many biological studies. This is due to the need to train the many hidden layers in a deep neural network. Moreover, the complex architecture and training process involved in deep learning largely prevent one from understanding *how* a deep neural network calculates a prediction, as one can only control the input data and some parameters in the model (e.g., number and size of hidden layers). This can limit the interpretability of the model's predictions, thereby constraining its utility for yielding insights on underlying biological mechanisms (Ching et al., 2018). In the next section, we discuss these and other challenges and offer some thoughts on how to address them in the context of network biology.

Intersection of Machine Learning and Network Biology

As we gather increasingly large and diverse data on the many layers of biological systems, one can devise machine-learning approaches that take advantage of these datasets to build more complex and biologically realistic network models across multiple levels, from gene regulation to interspecies interactions (Karr et al., 2012, 2014). Additionally, next-generation machine-learning methods provide tools that can enhance the utilization of these network models for a variety of biomedical applications. Below, we highlight outstanding problems and opportunities in network biology that span disease biology, drug discovery, microbiome research, and synthetic biology and that are ripe for exploration under a next-generation machine learning lens. We also discuss key challenges that need to be overcome to fully realize the potential of machine-learning methods in network biology.

Disease Biology

Network biology can help us gain a better understanding of the intricacies of disease biology. While traditional approaches rely on the identification and characterization of particular aspects of a disease, such as the discovery of disease-associated genes, network biology takes a more holistic approach and, as such, is poised to provide us with a more comprehensive view of the factors that drive disease phenotypes. Rather than simply identifying potential biomarkers, network biology allows us to characterize networks and sub-networks of biomolecular interactions critical for the emergence of a disease state (Barabási and Oltvai, 2004; Bordbar and Palsson, 2012; Chuang et al., 2007; Goh et al., 2007; Greene et al., 2015; Margolin et al., 2013; Schadt and Lum, 2006).

In defining network-specific characteristics of a disease, one can rationalize the use of machine learning algorithms to help understand and define the underlying disease mechanisms. As an example application, one could use existing network knowledge from sources such as BioGRID (Chatr-Aryamontri et al., 2017; Stark et al., 2006)—a database of gene interactions, protein-protein interactions, chemical interactions and post-translational

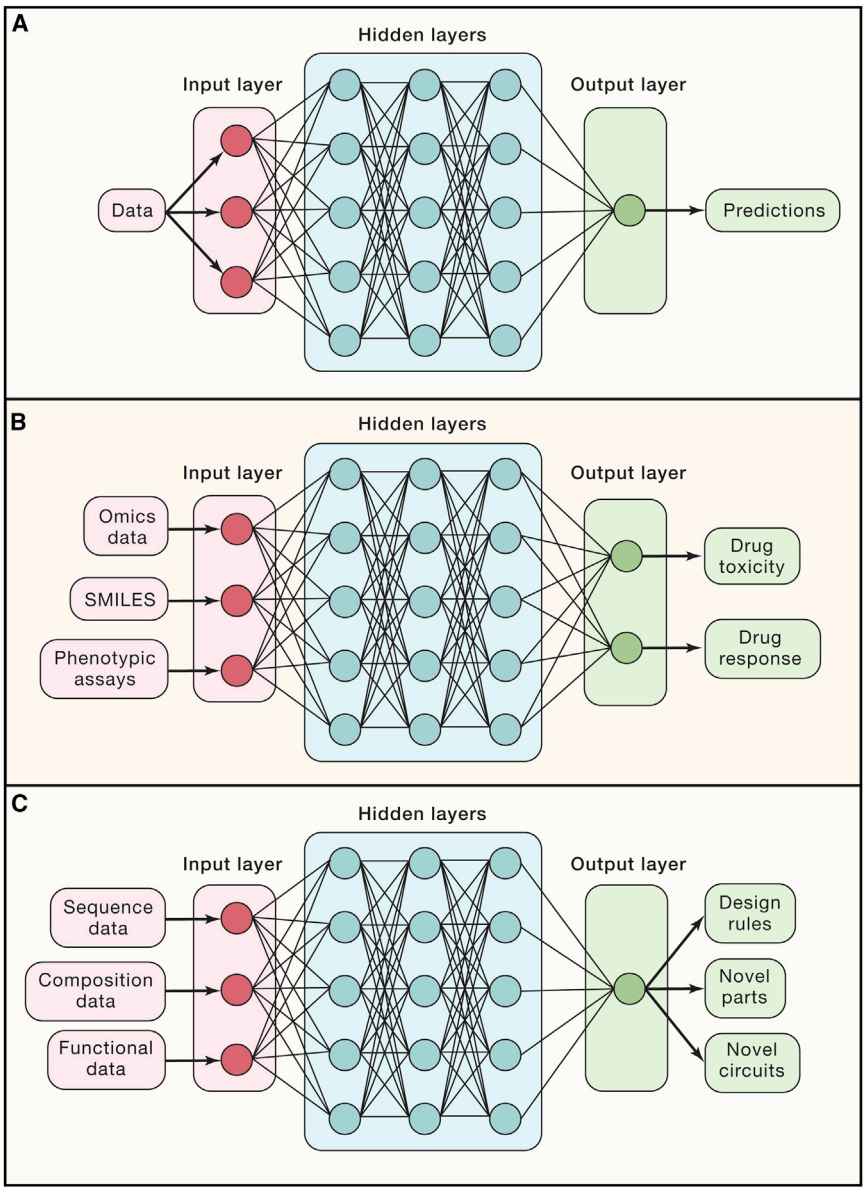

Figure 2. Next-Generation Machine-Learning Approaches and Applications

(A) Deep learning approaches consist of neural network models in which the depth of the network structure itself is defined as the number of hidden layers being considered. These algorithms generate predictive models based on an input layer, the hidden (deep) layers, and an output layer. The data are processed and fed into the input layer. Next, the hidden layer transforms the data into a representation that can be learned and fed forward to the next layer, which again transforms the data into a new representation. Errors made based on the training data labels are back-propagated through the network and the model is tuned for higher performance. The output layer generates a prediction (classification or regression) based on the tuned hidden layers.

(B) Deep learning architectures present great opportunities in drug discovery. Taking in multiple types of data, such as multi-omics data, the SMILES representation of a given compound, or the output of many different phenotypic assays, deep learning networks could be designed to perform a myriad of predictive tasks. Here, we exemplify and simplify a multi-task learning application, in which the drug toxicity and drug response are predicted based on the input data.

(C) Deep learning applications for synthetic biology include the prediction of novel design rules, molecular components, and gene circuitries, based on input data such as genomic sequences, composition data, and functional data from existing components and gene circuits.

modifications—to explore how the relationships between different biomolecules change in disease states compared to healthy states. Starting with data from a healthy cohort, one could train a deep learning algorithm (e.g., a deep neural network) to learn the fundamental characteristics that define healthy states. After training, the algorithm could be provided data from a patient cohort and used to predict differences between the healthy and disease states, identifying differentiating sets of regulatory interactions and biomolecules that could be validated and explored further. Similar approaches have been utilized in the context of network inference, where topological features are identified that can be attributable to differences in phenotypic observations at the expression level (de la Fuente, 2010; Mall et al., 2017).

As noted above, there is a need to better understand the complex, hierarchical structure of biological networks underlying disease and how the dysregulation of these networks may lead to a disease state. Here is where capsule networks (Hinton et al., 2011; Sabour et al., 2017), a next-generation machine learning method, could be of high value. Capsule networks involve a new type of neural network architecture, where CNNs are encapsulated in interconnected modules. As described earlier, CNNs are a special kind of deep neural network that processes multi-dimensional data, such as the -omics datasets found in network biology. A capsule network, on the other hand, is a representation of a deep neural network as a set of modules (capsules), which allows for the learning of data structures in a manner that preserves hierarchical aspects of the data itself. This representation has been particularly useful in the analyses of image data, as it allows for the algorithms to learn features of images independent of viewing angle of the image, a common problem with CNN applications.

Capsule networks are ripe for application in network biology and disease biology given that biological networks are highly modular in nature, with specified layers for the many biomolecules, while allowing each of these layers to interact with other layers. In the context of capsule networks, each biological layer could be treated as a capsule; with data generated across the different biological layers (e.g., transcriptomics, proteomics,

metabolomics), CNNs associated with each capsule could be trained to learn the specific properties of each of these layers independently. Applying the premises of dynamic routing (i.e., the act of relaying information) between capsules would allow for the different capsules to take as inputs the output of any other capsule, thereby enabling the model to learn how each layer interacts and depends on the others. This approach would allow one to study highly modular systems such as biological networks comprised of genes, proteins, metabolites, etc., and analyze how the functional organization and interplay of such networks and their sub-networks are disrupted in disease states.

We are not aware of any biological applications of capsule networks, but their unique features could enable us to disentangle and tackle the complexities of human disease. As we describe below, the successful implementation of capsule networks and other deep learning methods will depend critically upon the availability of suitably large, high-quality, well-annotated datasets.

Drug Discovery

In drug discovery, there is a critical need to characterize the mode of action of compounds, identify off-target effects of drugs, and develop effective drug combinations to treat complex diseases (Chen and Butte, 2016). Network biology approaches, along with machine-learning algorithms, have been successfully applied in these areas; for example, inferred network models and transcriptomics have been used to predict the likely targets of compounds of interest (e.g., di Bernardo et al., 2005; Woo et al., 2015). However, significant challenges remain, particularly in closing the gaps between the biological and chemical aspects of drug discovery and development. Below, we highlight how next-generation machine-learning algorithms, in the context of network biology, could bring added capabilities to address these challenges and accelerate efforts in drug discovery.

Extensive multi-omics data from drug treatments (Barretina et al., 2012; Basu et al., 2013; Garnett et al., 2012; Goodspeed et al., 2016; Musa et al., 2017; Rees et al., 2016; Seashore-Ludlow et al., 2015; Shoemaker, 2006; Yang et al., 2013), together with large amounts of genotypic data collected and stored in repositories such as dbGAP (Mailman et al., 2007) and the GTEx Portal (Lonsdale et al., 2013), bring the raw biological material needed to generate comprehensive network models for machine-learning applications. It is exciting to consider, from a machine-learning perspective, how one might integrate these network models and biological datasets with the wealth of information available on chemical matter via outlets such PubChem (Kim et al., 2016), a database of chemical molecules and their biological activities; DrugBank (Wishart et al., 2006, 2008), which contains data on drugs and drug targets; and the ZINC database (Sterling and Irwin, 2015), which includes structural information on over 100 million drug-like compounds.

Multi-task-learning neural networks are well suited for these types of applications, where a given system may include many labels (e.g., response to drug, disease state) across a multitude of data types (e.g., expression profiles, chemical structures) comprised of many independent features (Figure 2B). Typical machine-learning applications define a single task, where a model is trained to predict a single label. If a new label is to be learned using the same input data, then a new model is trained; that is, the learning tasks are treated as independent events. However, in some cases, there is important information that can be learned from one task that can inform the learning of another task. The idea underlying multi-task learning is to co-learn a set of tasks simultaneously (Caruana, 1998). Single-task learners aim to optimize the performance for the single task, while the goal of a multi-task learner is to optimize the performance for all tasks together. Multi-task learners take multiple representations to learn the system as a whole, thereby learning multiple tasks at once.

In multi-task learning, multiple related tasks are learned *at the same time*, leveraging differences and similarities across the tasks. This approach is based on the premise that learning related concepts imposes a generalization on the learning model, which results in improved performance over a single-task-learning approach while avoiding model overfitting (Caruana, 1998). Importantly, multi-task-learning neural networks can integrate or synthesize data from many distinct sources and assays. Thus, a multi-task learner could be trained to predict the physiological response to a given drug as well as its toxicity simultaneously by taking into account regulatory network relationships as well as data from multi-omics experiments, high-throughput drug screens, biological activity assays, and phenotypic observations from drug treatments. Recent successes, such as the prediction of drug toxicity in cancer cell lines and drug sensitivity in breast cancer cell lines (Ammad-Ud-Din et al., 2017; Costello et al., 2014; Tan, 2016), highlight the power of multi-task learning for drug discovery.

It is exciting to consider how multi-task learners could be used to bridge the gap between the biological and chemical aspects of drug discovery by incorporating structural data on chemical entities. One could, for example, use simplified molecular-input line-entry system (SMILES) representations of drugs (Anderson et al., 1987) as input data to the learner. The SMILES representation translates the structure of chemical species into a linear text string, which can be readily incorporated into machine-learning applications. Providing a multi-task-learning algorithm with the SMILES representations and identified targets of different compounds, along with their transcriptional and toxicity profiles, could enable the algorithm to be trained to predict potential side effects or likely targets of new compounds under consideration. Additionally, one could use natural language-processing techniques such as word embeddings (Mikolov et al., 2013a, 2013b) to learn specific properties of drugs based on their SMILES representations, complementing the multi-task learner while allowing for the identification of key properties and/or structural features of compounds that could be incorporated or removed in subsequent drug design efforts.

These machine-learning approaches could also be utilized to study and exploit the "dirtiness" of drug compounds. Most, if not all, compounds hit more than their primary target, and these effects vary in a dose- and network-dependent fashion. Multi-task-learning neural networks are well suited to learn aspects from diverse data types (e.g., pharmacokinetic and pharmacodynamic properties of different drugs, multi-omics data from cellular screens of such drugs, etc.) so as to better understand and predict input-output relationships (e.g., between the biophysical and structural properties of various chemical entities, their molecular targets, and the biological responses they

induce). Many drug developers view off-target effects as detrimental artifacts; however, one can envision using machine-learning approaches to turn such effects into advantageous properties that could be exploited and/or accounted for in drug combinations. For example, it is conceivable that a capsule-network model might be used to study a complex disease and, in doing so, predict that multiple targets need to be inhibited in order to treat the disease; these predictions could be utilized by a multi-task learner to identify dirty compounds or combinations of such compounds that hit all of the needed targets. Accordingly, we foresee multi-task learning, in conjunction with other deep learning approaches, as being instrumental to tackling the problem of biological and chemical data integration in drug discovery and creating multi-layered predictive network models that help advance rational drug design.

Microbiome Research

The human microbiome consists of the microorganisms—bacteria, archaea, viruses, fungi, protozoa—that live on or inside the human body. The diversity of microbes at each body site is staggering (Human Microbiome Project Consortium, 2014), and it is now accepted that these microbiota, which exist in dynamic interconnected ecosystems, play a central role in health, disease, and development. There is a deluge of metagenomic data on the human microbiome, but converting these data into biologically and clinically meaningful mechanistic insights remains a major challenge. This presents an excellent opportunity for network biology approaches that harness the power of next-generation machine-learning algorithms.

Microbes and host cells at different body sites interact by producing, exchanging, and utilizing small biomolecules, primarily metabolites. These interactions lead to metabolic networks within cells, across cells, across species, and across kingdoms. This creates an opportunity to generate meta-metabolic network models, based on shared metabolites, for any given microbiota-host system. Such models could be used to map, dissect, and understand polymicrobial-host interactions, as well as predict and gain insights into the synergistic and dysbiotic relationships that can arise between hosts and their microbial passengers.

Metabolic network models have been constructed for a number of microbial model organisms (e.g., *Escherichia coli*), as well as human cells. These models, which provide a global picture of how metabolites interact via biochemical reactions in a given cell, could be leveraged, modified, and integrated to create meta-networks that span multiple organisms or cell types. Unfortunately, our understanding of the metabolic networks or functions in many microbes is limited or non-existent due to sparse data and measurements on such microbes. This poses a significant challenge for the generation of meta-metabolic network models, one that could benefit from an area of machine learning known as transfer learning. In contrast to multi-task learning, transfer learning aims at learning a specific task from knowledge acquired while learning a different but related task (Pan and Yang, 2010). Biological systems share many characteristics, suggesting that data generated in one system can help inform another. The challenge becomes how to best apply the knowledge learned in a given system to a novel system for which limited data exist.

Transfer learning provides paradigms that allow one to make inferences and predictions on new systems based on observations made in other systems. Specifically, transfer learning enables one to repurpose a model used to learn a particular task as the starting point to learn a different but related task. The concepts behind transfer learning readily apply to problems in biology. Consider the case of metabolism and metabolic networks as an example—the immutable nature of biochemical compounds (i.e., ''glucose'' in *E. coli* is the same organic compound as ''glucose'' in *B. anthracis*) provides a basis for inductive transfer of knowledge between organisms, which implies that machine-learning models optimized in model organisms can be reused or repurposed to learn features on a different organism for which data are scarce.

This opens an exciting avenue for studying the metabolic intricacies of microbial communities, where one can ''transfer'' or use learned information on the metabolic network from a well-studied species such as *E. coli* to inform the network models of under-studied species and thereby accelerate our understanding of multiple species in the microbiome. Similar to learning features on different microbes in a complex microbial community based on transfer learning, we can conceptualize machine-learning models in which we leverage the knowledge gained on simpler systems to understand more complex systems. In this way, one may be able to build comprehensive models of the metabolic interactions and relationships between the microbiota and host. Such models could be trained on appropriate datasets spanning healthy and disease states and used to predict how the disappearance, introduction, or outgrowth of a particular species might disrupt or enhance the metabolic balance of the ecosystem, producing, for example, beneficial metabolites that promote health or toxic metabolic by-products that damage host tissue. Of note, these advanced machine-learning techniques and network biology approaches need not be limited to human health applications—they could be readily extended to microbiota found in agricultural, environmental, and industrial settings.

Synthetic Biology

Synthetic biology is focused, in part, on creating synthetic gene networks out of molecular components, and using these circuits to reprogram living cells, endowing them with novel capabilities (Cameron et al., 2014; Elowitz and Leibler, 2000; Gardner et al., 2000; Mukherji and van Oudenaarden, 2009; Purnick and Weiss, 2009). The design and construction of synthetic gene circuits, however, is far from straightforward—early versions of circuits rarely function as intended and typically require many weeks or months of post hoc tweaking. These development efforts are hindered by a limited understanding of core design principles for gene circuits and a lack of diverse, well-characterized components for network construction. As synthetic biology extends its reach into broad application areas (e.g., health, agriculture, energy, environment) (Khalil and Collins, 2010), there is a growing need to take on these challenges so as to make biological design more predictable, straightforward, and time efficient; this creates marvelous opportunities for deep learning approaches, as we highlight below.

Multiple levels of organization exist in synthetic gene circuits, and these could be explicitly accounted for in deep learning

algorithms, as noted above. At the base level in a synthetic circuit, there are individual molecular components, such as genes, promoters, operators, terminators, and ribosome binding sites (RBSs). At the intermediate level, there are regulatory units made up of multiple components, such as gene-promoter pairs. At the highest level, regulatory units interact to create a particular gene circuit, e.g., two gene-promoter pairs can be arranged in a mutually inhibitory network to create a genetic toggle switch. At each of these levels, one can identify sequence representations that define certain aspects of regulation and control, as well as compositional relationships (e.g., spatial arrangement and orientation) and interactions between biomolecules, molecular components, and/or sub-components that impact functional outputs and behaviors.

To create appropriate training datasets for deep learning approaches, one could generate, sequence, and functionally characterize large, diverse sets of molecular components, regulatory units, and synthetic gene circuits (Figure 2C). The functional characterization could include quantifying the strength of RBSs, the Hill coefficients for promoter-gene pairs, and the response times of gene circuits, among many other variables. Since deep learning approaches rely heavily on large amounts of data, it would be useful to develop and implement fast "component-to-readout" experimental workflows, and similar workflows for regulatory units and gene circuits, that integrate robotics with plate-based assays and machine-interpretable functional readouts.

One can then envision using the aforementioned sequencing and functional characterization data to generate a predictive model across the multiple levels of biological organization in synthetic gene circuits, from molecular components to regulatory units to the circuits themselves. To do so, one could develop a multi-staged deep learning model that captures the essence of each of these organizational levels from a learning model that embeds biological sequences to ones that embed regulatory motifs and circuit structures. For example, recurrent neural networks could be utilized to encode sequences for different components, where a sequence could be treated as a specific "sentence" that allows one to learn specific "sentence properties"—style, syntax, and topic—that equate to specific sequence properties—promoters, binding regions, and terminators. Additionally, convolution neural networks could be used to encode features on the topological organization of regulatory units and synthetic gene circuits. The algorithms could be trained to learn the sequence-function relationships for different components, as well as the composition-function relationships for regulatory units and synthetic gene circuits. In this manner, the model could learn key aspects of synthetic gene circuits both from a regulatory (network control) perspective and a topological (network architecture) perspective.

The generated deep learning model could be used to identify fundamental design principles for synthetic biology. Correspondingly, the platform could be utilized to create components (e.g., inducible promoters, operator sites, etc.) with enhanced or novel functions and thereby expand the number and diversity of molecular parts available for synthetic biology development efforts. The deep learning model could also be purposed to design and identify novel regulatory units and synthetic gene networks, each with desired performance specifications. For example, for a given desired function, the model could be used to generate a set of gene circuitries that are predicted to produce said function. Combining such an approach with mathematical modeling to characterize the regions of stability and operability of each predicted circuit would allow one to iterate, very rapidly, through thousands of potential circuitries. The most promising candidates could be synthesized, tested, and validated. These developments could serve to fast-track design efforts in synthetic biology, facilitating the creation of complex synthetic gene networks for a wide range of biomedical applications.

Challenges and Future Outlook

It is clear from the above discussion that there are great opportunities at the intersection of network biology and next-generation machine learning. However, there are equally great challenges that need to be overcome. The most critical and central to these efforts is the need for massively large datasets. Deep learning methods and other next-generation machine-learning approaches are exceptionally data hungry. We live in the age of big data in biology and medicine, where data are collected on many different layers of biological organization. Data captured from biological systems can be incredibly complex, with many thousands of variables capturing many different facets of the biological system. However, the majority of these datasets are orders of magnitude too small for deep learning algorithms to be applied appropriately.

There are many options available to take on this challenge. The first and foremost is to invest in the collection of suitably large, well-annotated datasets for state-of-the-art studies in network biology. Multi-omics datasets can be prohibitively expensive, and thus, we need to consider alternatives to supplement or complement such data. Increased utilization of imaging data (including video) to characterize morphological or phenotypic changes of cells (e.g., in response to drug treatments) is one attractive possibility, as many deep learning algorithms have been successfully applied to imaging data in the context of diagnostics. Such efforts could be enhanced by creating cell lines with fluorescent or colorimetric readouts that report on cellular responses to various treatments or environmental perturbations. The small, sparse nature of many biological datasets also presents an interesting challenge to machine-learning researchers—namely, producing a new generation of deep learning algorithms specifically designed to handle such datasets.

Another possibility is to generate *in silico* data with properties of real data. For image analysis in the context of deep learning, this is often accomplished using generative adversarial networks (GANs), which learn to create datasets that are similar to the training data. GANs are deep neural network architectures comprised of two neural networks that are pitted against each other—one is a generative model that produces new data that mimic the distributions of the training dataset, while the other is a discriminative model (the adversary) that evaluates the new data and determines whether or not it belongs to the actual training dataset. Competition between the two neural networks serves to improve their methods until the generated datasets are indistinguishable from the training dataset. This machine-learning approach could be readily extended to the

multi-omics datasets that drive network biology. A simple example would be using GANs to generate dramatically larger expression datasets that can be used in the context of network inference to generate predictive models of transcriptional regulation.

The ''black box'' nature of most next-generation machine learning models presents an additional challenge for biological applications. It is often difficult to interpret the output of a given model from a biological perspective, thereby limiting the utility of the model in providing insights into underlying biological mechanisms and functional network architectures. This is not always the case, particularly for simpler machine learning methods. Sparse linear regression models (e.g., elastic net, lasso and ridge regression [Hastie et al., 2009]), for example, learn optimal coefficients that represent a relative weight for each feature. In such instances, model coefficients can inform researchers of the relative ''importance'' of each feature in the model. However, for more advanced machine-learning methods, such as deep neural networks, the training procedure transforms the input data in such a way that it can be difficult to determine the relative importance of features or whether a feature is positively or negatively correlated with the outcome of interest. Interpretation of model features is an open challenge in machine learning, with great attention being given to the interpretation of how particular models relate to input features (Lakkaraju et al., 2016; Letham et al., 2015; Lipton, 2016). There is a critical need to develop means to transform the ''black boxes'' of deep learning into ''white boxes'' that can be opened up and interpreted meaningfully from a biological perspective.

We have a long way to go to uncover and harness the networked intricacies and complexities of living systems, and machine learning itself is far from fulfilling its potential in biological research. Nonetheless, ongoing and emerging developments in the application of machine-learning approaches to better understand complex biological networks enable us to predict an exciting and deep future for network biology.

ACKNOWLEDGEMENTS

This work is supported by the Paul G. Allen Frontiers Group, the Wyss Institute for Biologically inspired Engineering, and the Boettcher Foundation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Abdi, H., and Williams, L.J. (2010). Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. *2*, 433–459.

Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. Cell *143*, 1005–1017.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Ammad-Ud-Din, M., Khan, S.A., Wennerberg, K., and Aittokallio, T. (2017). Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics *33*, i359–i368.

Anderson, E., Veith, G.D., and Weininger, D. (1987). SMILES: a line notation and computerized interpreter for chemical structures (U.S. Environmental Protection Agency, Environmental Research Laboratory).

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. *12*, 878.

Ballester, P.J., and Mitchell, J.B. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics *26*, 1169–1175.

Bansal, M., Yang, J., Karan, C., Menden, M.P., Costello, J.C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R., et al.; NCI-DREAM Community; NCI-DREAM Community (2014). A community computational challenge to predict the activity of pairs of compounds. Nat. Biotechnol. *32*, 1213–1222.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Basu, A., Bodycombe, N.E., Cheah, J.H., Price, E.V., Liu, K., Schaefer, G.I., Ebright, R.Y., Stewart, M.L., Ito, D., Wang, S., et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell *154*, 1151–1161.

Bisikirska, B., Bansal, M., Shen, Y., Teruya-Feldstein, J., Chaganti, R., and Califano, A. (2016). Elucidation and Pharmacological Targeting of Novel Molecular Drivers of Follicular Lymphoma Progression. Cancer Res. *76*, 664–674.

Bordbar, A., and Palsson, B.O. (2012). Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J. Intern. Med. *271*, 131–141.

Butte, A.J., and Kohane, I.S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac. Symp. Biocomput., 418–429.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903–915.

Califano, A., and Alvarez, M.J. (2017). The recurrent architecture of tumour initiation, progression and drug sensitivity. Nat. Rev. Cancer *17*, 116–130.

Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat. Genet. *44*, 841–847.

Cameron, D.E., Bashor, C.J., and Collins, J.J. (2014). A brief history of synthetic biology. Nat. Rev. Microbiol. *12*, 381–390.

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature *463*, 318–325.

Caruana, R. (1998). Multitask Learning. In Learning to Learn (Springer), pp. 95–133.

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. Comput. Electr. Eng. *40*, 16–28.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. *45* (D1), D369–D379.

Chen, B., and Butte, A.J. (2016). Leveraging big data to transform target selection and drug discovery. Clin. Pharmacol. Ther. *99*, 285–297.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface *15*.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Mol. Syst. Biol. *3*, 140.

Costello, J.C., and Stolovitzky, G. (2013). Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clin. Pharmacol. Ther. *93*, 396–398.

Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al.; NCI DREAM Community (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. *26*, 326–333.

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. *8*, 717–729.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., and Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat. Biotechnol. *23*, 377–383.

Dietterich, T.G. (2000). Ensemble methods in machine learning. G. Goos, J. Hartmanis, and J.P. van Leeuwen, eds. International Workshop on Multiple Classifier Systems, 1–15.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Domingos, P. (2012). A few useful things to know about machine learning. Commun. ACM *55*, 78–87.

Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature *403*, 335–338.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. *5*, e8.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. Science *303*, 799–805.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature *403*, 339–342.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Ghahramani, Z. (2004). Unsupervised Learning. In Advanced Lectures on Machine Learning (Berlin, Heidelberg: Springer), pp. 72–112.

Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA *104*, 8685–8690.

Goodspeed, A., Heiser, L.M., Gray, J.W., and Costello, J.C. (2016). Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol. Cancer Res. *14*, 3–13.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res. *3*, 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics).

Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al.; HPN-DREAM Consortium (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. Nat. Methods *13*, 310–318.

Hinton, G.E., Krizhevsky, A., and Wang, S.D. (2011). Transforming Auto-encoders. In Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I. (Berlin, Heidelberg: Springer-Verlag), pp. 44–51.

Human Microbiome Project Consortium (2014). The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. Cell Host Microbe *16*, 276–289.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning : with applications in R (New York: Springer).

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Karr, J.R., Phillips, N.C., and Covert, M.W. (2014). WholeCellSimDB: a hybrid relational/HDF database for whole-cell model predictions. Database J. Biol. Databases Curation *2014*.

Khalil, A.S., and Collins, J.J. (2010). Synthetic biology: applications come of age. Nat. Rev. Genet. *11*, 367–379.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. (2016). PubChem Substance and Compound databases. Nucleic Acids Res. *44* (D1), D1202–D1213.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems 25, F, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds. (Pereira: Curran Associates, Inc.), pp. 1097–1105.

Lakkaraju, H., Bach, S.H., and Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. KDD *2016*, 1675–1684.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Letham, B., Rudin, C., McCormick, T.H., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Ann. Appl. Stat. *9*, 1350–1371.

Leung, M.K.K., Delong, A., Alipanahi, B., and Frey, B.J. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Proceedings of the IEEE *104*, 176–197.

Lipton, Z.C. (2016). The Mythos of Model Interpretability. arXiv, arXiv:1606.03490, http://arxiv.org/abs/1606.03490.

Little, R.J.A., and Rubin, D.B. (1987). Statistical analysis with missing data (New York: Wiley).

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. *39*, 1181–1186.

Mall, R., Cerulo, L., Bensmail, H., Iavarone, A., and Ceccarelli, M. (2017). Detection of statistically significant network changes in complex biological networks. BMC Syst. Biol. *11*, 32.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G.; DREAM5 Consortium (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7* (*Suppl 1*), S7.

Margolin, A.A., Bilal, E., Huang, E., Norman, T.C., Ottestad, L., Mecham, B.H., Sauerwine, B., Kellen, M.R., Mangravite, L.M., Furia, M.D., et al. (2013).

Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci. Transl. Med. *5*, 181re1.

McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. *5*, 115–133.

Mezlini, A.M., and Goldenberg, A. (2017). Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. PLoS Comput. Biol. *13*, e1005580.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. Adv. Neur. Inf. Proc. Sys. *26*, 3111–3119.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. arXiv, arXiv:1301.3781, http://arxiv.org/abs/1301.3781.

Mitchell, T.M. (1997). Machine Learning (New York: McGraw-Hill Education).

Mukherji, S., and van Oudenaarden, A. (2009). Synthetic biology: understanding biological design from synthetic circuits. Nat. Rev. Genet. *10*, 859–871.

Musa, A., Ghoraie, L.S., Zhang, S.-D., Glazko, G., Yli-Harja, O., Dehmer, M., Haibe-Kains, B., and Emmert-Streib, F. (2017). A review of connectivity map and computational approaches in pharmacogenomics. Brief. Bioinform. *18*, 903.

Pan, S.J., and Yang, Q. (2010). A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. *22*, 1345–1359.

Parker, D.B. (1985). Learning-Logic: Casting the Cortex of the Human Brain in Silicon. In Technical Report Tr-47 (MIT Press).

Pe'er, D., and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. Cell *144*, 864–873.

Purnick, P.E.M., and Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. Nat. Rev. Mol. Cell Biol. *10*, 410–422.

Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E.V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E., et al. (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat. Chem. Biol. *12*, 109–116.

Rencher, A.C. (2002). Methods of multivariate analysis (New York: J. Wiley).

Rokach, L. (2010). Ensemble-based classifiers. Artif. Intell. Rev. *33*, 1–39.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533.

Sabour, S., Frosst, N., and Hinton, G.E. (2017). Dynamic Routing Between Capsules. In Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 3859–3869.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics *23*, 2507–2517.

Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., and Stolovitzky, G. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. *17*, 470–486.

Schadt, E.E., and Lum, P.Y. (2006). Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. J. Lipid Res. *47*, 2601–2613.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. *5*, 1210–1223.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. arXiv, arXiv:1404.1100, http://arxiv.org/abs/1404.1100.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.

Sterling, T., and Irwin, J.J. (2015). ZINC 15–Ligand Discovery for Everyone. J. Chem. Inf. Model. *55*, 2324–2337.

Tan, M. (2016). Prediction of anti-cancer drug response by kernelized multi-task learning. Artif. Intell. Med. *73*, 70–77.

Walsh, L.A., Alvarez, M.J., Sabio, E.Y., Reyngold, M., Makarov, V., Mukherjee, S., Lee, K.-W., Desrichard, A., Turcan, Ş., Dalin, M.G., et al. (2017). An Integrated Systems Biology Approach Identifies TRIM25 as a Key Determinant of Breast Cancer Metastasis. Cell Rep. *20*, 1623–1640.

Werbos, P. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences (Harvard University).

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668–D672.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. *36*, D901–D906.

Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. IEEE Trans. Evol. Comput. *1*, 67–82.

Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R., et al. (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441–451.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.

Yip, K.Y., Cheng, C., and Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? Genome Biol. *14*, 205.

Zhang, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov. Today *22*, 1680–1685.