Opportunities and obstacles for deep learning in biology and medicine

Travers Ching[1,*], Daniel S. Himmelstein[2], Brett K. Beaulieu-Jones[3], Alexandr A. Kalinin[4], Brian T. Do[5], Gregory P. Way[2], Enrico Ferrero[6], Paul-Michael Agapow[7], Wei Xie[8], Gail L. Rosen[9], Benjamin J. Lengerich[10], Johnny Israeli[11], Jack Lanchantin[12], Stephen Woloszynek[9], Anne E. Carpenter[13], Avanti Shrikumar[14], Jinbo Xu[15], Evan M. Cofer[16], David J. Harris[17], Dave DeCaprio[18], Yanjun Qi[12], Anshul Kundaje[19], Yifan Peng[20], Laura K. Wiley[21], Marwin H.S. Segler[22], Anthony Gitter[23,†], Casey S. Greene[2,†]

[*] Author order was determined with a randomized algorithm

[†] To whom correspondence should be addressed: gitter@biostat.wisc.edu (A.G.) and csgreene@upenn.edu (C.S.G.)

1. Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI
2. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
3. Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
4. Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI
5. Harvard Medical School, Boston, MA
6. Computational Biology and Stats, Target Sciences, GlaxoSmithKline, Stevenage, United Kingdom
7. Data Science Institute, Imperial College London, London, United Kingdom
8. Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN
9. Ecological and Evolutionary Signal-processing and Informatics Laboratory, Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA
10. Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
11. Biophysics Program, Stanford University, Stanford, CA
12. Department of Computer Science, University of Virginia, Charlottesville,

VA
13. Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, MA
14. Department of Computer Science, Stanford University, Stanford, CA
15. Toyota Technological Institute at Chicago, Chicago, IL
16. Department of Computer Science, Trinity University, San Antonio, TX
17. Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL
18. ClosedLoop.ai, Austin, TX
19. Department of Genetics and Department of Computer Science, Stanford University, Stanford, CA
20. National Center for Biotechnology Information and National Library of Medicine, National Institutes of Health, Bethesda, MD
21. Division of Biomedical Informatics and Personalized Medicine, University of Colorado School of Medicine, Aurora, CO
22. Institute of Organic Chemistry, Westfälische Wilhelms-Universität Münster, Münster, Germany
23. Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison and Morgridge Institute for Research, Madison, WI

Abstract

Deep learning, which describes a class of machine learning algorithms, has recently showed impressive results across a variety of domains. Biology and medicine are data rich, but the data are complex and often ill-understood. Problems of this nature may be particularly well-suited to deep learning techniques. We examine applications of deep learning to a variety of biomedical problems -- patient classification, fundamental biological processes, and treatment of patients -- to predict whether deep learning will transform these tasks or if the biomedical sphere poses unique challenges. We find that deep learning has yet to revolutionize or definitively resolve any of these problems, but promising advances have been made on the prior state of the art. Even when improvement over a previous baseline has been modest, we have seen signs that deep learning methods may speed or aid human investigation. More work is needed to address concerns related to interpretability and how to best model each problem. Furthermore, the limited amount of labeled data for training presents problems in some domains, as can legal and privacy constraints on work with sensitive health records. Nonetheless, we foresee deep learning powering changes at the bench and bedside with the potential to transform several areas of biology and medicine.

Introduction to deep learning

Biology and medicine are rapidly becoming data-intensive. A recent

comparison of genomics with social media, online videos, and other data-intensive disciplines suggests that genomics alone will equal or surpass other fields in data generation and analysis within the next decade [1]. The volume and complexity of these data present new opportunities, but also pose new challenges. Automated algorithms that extract meaningful patterns could lead to actionable knowledge and change how we develop treatments, categorize patients, or study diseases, all within privacy-critical environments.

The term deep learning has come to refer to a collection of new techniques that, together, have demonstrated breakthrough gains over existing best-in-class machine learning algorithms across several fields. For example, over the past five years these methods have revolutionized image classification and speech recognition due to their flexibility and high accuracy [2]. More recently, deep learning algorithms have shown promise in fields as diverse as high-energy physics [3], dermatology [4], and translation among written languages [5]. Across fields, "off-the-shelf" implementations of these algorithms have produced comparable or higher accuracy than previous best-in-class methods that required years of extensive customization, and specialized implementations are now being used at industrial scales.

Neural networks were first proposed in 1943 [6] as a model for how our brains process information. The history of neural networks is interesting in its own right [7]. In neural networks, inputs are fed into a hidden layer, which feeds into one or more hidden layers, which eventually produce an output layer. The neural networks used for deep learning have multiple hidden layers. Each layer essentially performs feature construction for the layers before it. The training process used often allows layers deeper in the network to contribute to the refinement of earlier layers. For this reason, these algorithms can automatically engineer features that are suitable for many tasks and customize those features for one or more specific tasks. Deep learning does many of the same things as more familiar approaches. Like a clustering algorithm, it can build features that describe recurrent patterns in data. Like a regression approach, deep learning methods can predict some output. However, deep learning methods combine both of these steps. When sufficient data are available, these methods construct features tuned to a specific problem and combine those features into a predictor. Recently, hardware improvements and very large training datasets have allowed these deep learning techniques to surpass other machine learning algorithms for many problems.

Neural networks are most widely associated with supervised machine learning, where the goal is to accurately predict one or more labels associated with each data point. However, deep learning algorithms can also be used in an exploratory, "unsupervised" mode, where the goal is to summarize, explain, or identify interesting patterns in a data set. In a famous and early example,

scientists from Google demonstrated that a neural network "discovered" that cats, faces, and pedestrians were important components of online videos [8] without being told to look for them. What if, more generally, deep learning could solve the challenges presented by the growth of data in biomedicine? Could these algorithms identify the "cats" hidden in our data - the patterns unknown to the researcher - and suggest ways to act on them? In this review, we examine deep learning's application to biomedical science and discuss the unique challenges that biomedical data pose for deep learning methods.

Several important advances make the current surge of work done in this area possible. Easy-to-use software packages have brought the techniques of the field out of the specialist's toolkit to a broad community of computational scientists. Additionally, new techniques for fast training have enabled their application to larger datasets [9]. Dropout of nodes, edges, and layers makes networks more robust, even when the number of parameters is very large. New neural network approaches are also well-suited for addressing distinct challenges. For example, neural networks structured as autoencoders or as adversarial networks require no labels and are now regularly used for unsupervised tasks. In this review, we do not exhaustively discuss the different types of deep neural network architectures. A recent book from Goodfellow et al. [10] covers these in detail. Finally, the larger datasets now available are also sufficient for fitting the many parameters that exist for deep neural networks. The convergence of these factors currently makes deep learning extremely adaptable and capable of addressing the nuanced differences of each domain to which it is applied.

Will deep learning transform the study of human disease?

With this review, we ask the question: what is needed for deep learning to transform how we categorize, study, and treat individuals to maintain or restore health? We choose a high bar for "transform." Andrew Grove, the former CEO of Intel, coined the term Strategic Inflection Point to refer to a change in technologies or environment that requires a business to be fundamentally reshaped [11]. Here, we seek to identify whether deep learning is an innovation that can induce a Strategic Inflection Point in the practice of biology or medicine.

There are already a number of reviews focused on applications of deep learning in biology [12–16], healthcare [17], and drug discovery [18–21]. Under our guiding question, we sought to highlight cases where deep learning enabled researchers to solve challenges that were previously considered infeasible or makes difficult, tedious analyses routine. We also identified approaches that researchers are using to sidestep challenges posed by biomedical data. We find that domain-specific considerations have greatly

influenced how to best harness the power and flexibility of deep learning. Model interpretability is often critical. Understanding the patterns in data may be just as important as fitting the data. In addition, there are important and pressing questions about how to build networks that efficiently represent the underlying structure and logic of the data. Domain experts can play important roles in designing networks to represent data appropriately, encoding the most salient prior knowledge and assessing success or failure. There is also great potential to create deep learning systems that augment biologists and clinicians by prioritizing experiments or streamlining tasks that do not require expert judgment. We have divided the large range of topics into three broad classes: Disease and Patient Categorization, Fundamental Biological Study, and Treatment of Patients. Below, we briefly introduce the types of questions, approaches and data that are typical for each class in the application of deep learning.

Disease and patient categorization

A key challenge in biomedicine is the accurate classification of diseases and disease subtypes. In oncology, current "gold standard" approaches include histology, which requires interpretation by experts, or assessment of molecular markers such as cell surface receptors or gene expression. One example is the PAM50 approach to classifying breast cancer where the expression of 50 marker genes divides breast cancer patients into four subtypes. Substantial heterogeneity still remains within these four subtypes [22,23]. Given the increasing wealth of molecular data available, a more comprehensive subtyping seems possible. Several studies have used deep learning methods to better categorize breast cancer patients: denoising autoencoders, an unsupervised approach, can be used to cluster breast cancer patients [24], and convolutional neural networks (CNNs) can help count mitotic divisions, a feature that is highly correlated with disease outcome in histological images [25]. Despite these recent advances, a number of challenges exist in this area of research, most notably the integration of molecular and imaging data with other disparate types of data such as electronic health records (EHRs).

Fundamental biological study

Deep learning can be applied to answer more fundamental biological questions; it is especially suited to leveraging large amounts of data from high-throughput "omics" studies. One classic biological problem where machine learning, and now deep learning, has been extensively applied is molecular target prediction. For example, deep recurrent neural networks (RNNs) have been used to predict gene targets of microRNAs [26], and CNNs have been applied to predict protein residue-residue contacts and secondary structure [27–29]. Other recent exciting applications of deep learning include recognition

of functional genomic elements such as enhancers and promoters [30–32] and prediction of the deleterious effects of nucleotide polymorphisms [33].

Treatment of patients

Although the application of deep learning to patient treatment is just beginning, we expect new methods to recommend patient treatments, predict treatment outcomes, and guide the development of new therapies. One type of effort in this area aims to identify drug targets and interactions or predict drug response. Another uses deep learning on protein structures to predict drug interactions and drug bioactivity [34]. Drug repositioning using deep learning on transcriptomic data is another exciting area of research [35]. Restricted Boltzmann machines (RBMs) can be combined into deep belief networks (DBNs) to predict novel drug-target interactions and formulate drug repositioning hypotheses [36,37]. Finally, deep learning is also prioritizing chemicals in the early stages of drug discovery for new targets [21].

Deep learning and patient categorization

In a healthcare setting, individuals are diagnosed with a disease or condition based on symptoms, the results of certain diagnostic tests, or other factors. Once diagnosed with a disease, an individual might be assigned a stage based on another set of human-defined rules. While these rules are refined over time, the process is evolutionary rather than revolutionary.

We might imagine that deep learning or artificial intelligence methods could reinvent how individuals are categorized for healthcare. A deep neural network might identify entirely new categories of health or disease that are only present when data from multiple lab tests are integrated. As a potential example, consider the condition Latent Autoimmune Diabetes in Adults (LADA). The history of this disease classification is briefly reviewed in Stenström et al. [38].

Imagine that a deep neural network operating in the early 1980s had access to electronic health records with comprehensive clinical tests. It might have identified a subgroup of individuals with blood glucose levels that indicated diabetes as well as auto-antibodies, even though the individuals had never been diagnosed with type 1 diabetes -- the autoimmune form of the disease that arises in young people. Such a neural network would be identifying patients with LADA. As no such computational approach existed, LADA was actually identified by Groop et al. [39]. However, this represents a potential hope for this area. Perhaps deep neural networks, by reevaluating data without the context of our assumptions, can reveal novel classes of treatable conditions.

Alternatively, imagine that a deep neural network is provided with clinical test results gleaned from electronic health records. Because physicians may order certain tests based on their suspected diagnosis, a deep neural network may learn to "diagnose" patients simply based on the tests that are ordered. For some objective functions, such as predicting an International Classification of Diseases (ICD) code, this may offer good performance even though it does not provide insight into the underlying disease beyond physician activity. This challenge is not unique to deep learning approaches; however, it is important for practitioners to be aware of these challenges and the possibility in this domain of constructing highly predictive classifiers of questionable actual utility.

Our goal in this section is to assess the extent to which deep learning is already contributing to the discovery of novel categories. Where it is not, we focus on barriers to achieving these goals. We also highlight approaches that researchers are taking to address challenges within the field, particularly with regards to data availability and labeling.

Imaging applications in healthcare

Deep learning methods have transformed the analysis of natural images and video, and similar examples are beginning to emerge with medical images. Deep learning has been used to classify lesions and nodules; localize organs, regions, landmarks and lesions; segment organs, organ substructures and lesions; retrieve images based on content; generate and enhance images; and combine images with clinical reports [40,41].

Though there are many commonalities with the analysis of natural images, there are also key differences. In all cases that we examined, fewer than one million images were available for training, and datasets are often many orders of magnitude smaller than collections of natural images. Researchers have developed subtask-specific strategies to address this challenge.

The first strategy repurposes features extracted from natural images by deep learning models, such as ImageNet [42], for new purposes. Diagnosing diabetic retinopathy through color fundus images became an area of focus for deep learning researchers after a large labeled image set was made publicly available during a 2015 Kaggle competition [43]. Most participants trained neural networks from scratch [43–45], but Gulshan et al. [46] repurposed a 48-layer Inception-v3 deep architecture pre-trained on natural images and surpassed the state-of-the-art specificity and sensitivity. Such features were also repurposed to detect melanoma, the deadliest form of skin cancer, from dermoscopic [47,48] and non-dermoscopic images of skin lesions [4,49,50] as well as age-related macular degeneration [51]. Pre-training on natural images can enable very deep networks to succeed without overfitting. For the

melanoma task, reported performance was competitive with or better than a board of certified dermatologists [4,47].

Reusing features from natural images is also growing for radiographic images, where datasets are often too small to train large deep neural networks without these techniques [52–55]. Rajkomar et al. [54] showed that a deep CNN trained on natural images boosts performance in radiographic images. However, the target task required either re-training the initial model from scratch with special pre-processing or fine-tuning of the whole network on radiographs with heavy data augmentation to avoid overfitting.

The technique of reusing features from a different task falls into the broader area of transfer learning (see Discussion). Though we've mentioned numerous successes for the transfer of natural image features to new tasks, we expect that a lower proportion of negative results have been published. The analysis of magnetic resonance images (MRIs) is also faced with the challenge of small training sets. In this domain, Amit et al. [56] investigated the tradeoff between pre-trained models from a different domain and a small CNN trained only with MRI images. In contrast with the other selected literature, they found a smaller network trained with data augmentation on few hundred images from a few dozen patients can outperform a pre-trained out-of-domain classifier. Data augmentation is a different strategy to deal with small training sets. The practice is exemplified by a series of papers that analyze images from mammographies [57–61]. To expand the number and diversity of images, researchers constructed adversarial examples [60]. Adversarial examples are constructed by applying a transformation that changes training images but not their content -- for example by rotating an image by a random amount. An alternative in the domain is to train towards human-created features before subsequent fine tuning [58], which can help to sidestep this challenge though it does give up deep learning techniques' strength as feature constructors.

Another way of dealing with limited training data is to divide rich data -- e.g. 3D images -- into numerous reduced projections. Shin et al. [53] compared various deep network architectures, dataset characteristics, and training procedures for computer tomography-based (CT) abnormality detection. They concluded that networks as deep as 22 layers could be useful for 3D data, even though the size of training datasets was limited. However, they noted that choice of architecture, parameter setting, and model fine-tuning needed is very problem- and dataset-specific. Moreover, this type of task often depends on both lesion localization and appearance, which poses challenges for CNN-based approaches. Straightforward attempts to capture useful information from full-size images in all three dimensions simultaneously via standard neural network architectures were computationally unfeasible. Instead, two-dimensional models were used to either process image slices individually (2D), or

aggregate information from a number of 2D projections in the native space (2.5D). Roth et al. compared 2D, 2.5D, and 3D CNNs on a number of tasks for computer-aided detection from CT scans and showed that 2.5D CNNs performed comparably well to 3D analogs, while requiring much less training time, especially on augmented training sets [62]. Another advantage of 2D and 2.5D networks is the wider availability of pre-trained models. But reducing the dimensionality is not always helpful. Nie et al. [63] showed that multimodal, multi-channel 3D deep architecture was successful at learning high-level brain tumor appearance features jointly from MRI, functional MRI, and diffusion MRI images, outperforming single-modality or 2D models. Overall, the variety of modalities, properties and sizes of training sets, the dimensionality of input, and the importance of end goals in medical image analysis are provoking a development of specialized deep neural network architectures, training and validation protocols, and input representations that are not characteristic of widely-studied natural images.

Predictions from deep neural networks can be evaluated for use in workflows that also incorporate human experts. In a large dataset of mammography images, Kooi et al. [64] demonstrated that deep neural networks outperform the traditional computer-aided diagnosis system at low sensitivity and perform comparably at high sensitivity. They also compared network performance to certified screening radiologists on a patch level and found no significant difference between the network and the readers. However, using deep methods for clinical practice is challenged by the difficulty of assigning a level of confidence to each prediction. Leibig et al. [45] estimated the uncertainty of deep networks for diabetic retinopathy diagnosis by linking dropout networks with approximate Bayesian inference. Techniques that assign confidences to each prediction should aid pathologist-computer interactions and improve uptake by physicians.

Systems to aid in the analysis of histology slides are also promising use cases for deep learning [65]. Ciresan et al. [25] developed one of the earliest approaches for histology slides, winning the 2012 International Conference on Pattern Recognition's Contest on Mitosis Detection while achieving human-competitive accuracy. In more recent work, Wang et al. [66] analyzed stained slides of lymph node slices to identify cancers. On this task a pathologist has about a 3% error rate. The pathologist did not produce any false positives, but did have a number of false negatives. The algorithm had about twice the error rate of a pathologist, but the errors were not strongly correlated. In this area, these algorithms may be ready to be incorporated into existing tools to aid pathologists and reduce the false negative rate. Ensembles of deep learning and human experts may help overcome some of the challenges presented by data limitations.

One source of training examples with rich clinical annotations is electronic health records. Recently, Lee et al. [67] developed an approach to distinguish individuals with age-related macular degeneration from control individuals. They trained a deep neural network on approximately 100,000 images extracted from structured electronic health records, reaching greater than 93% accuracy. The authors used their test set to evaluate when to stop training. In other domains, this has resulted in a minimal change in the estimated accuracy [68], but we recommend the use of an independent test set whenever feasible.

Chest X-rays are a common radiological examination for screening and diagnosis of lung diseases. Although hospitals have accumulated a large number of raw radiology images and reports in Picture Archiving and Communication Systems and their related reports in Radiology Information Systems, it is not yet known how to effectively use them to learn the correlation between pathology categories and X-rays. In the last few years, deep learning methods showed remarkable results in chest X-ray image analysis [69,70]. However, it is both costly and time-consuming to annotate a large-scale fully-labeled corpus to facilitate data-intensive deep learning models. As an alternative, Wang et al. [70] proposed to use weakly labeled images. To generate weak labels for X-ray images, they applied a series of natural language processing (NLP) techniques to the associated chest X-ray radiological reports. Specifically, they first extracted all diseases mentioned in the reports using a state-of-the-art NLP tool, then applied a newly-developed negation and uncertainty detection tool (NegBio) to filter negative and equivocal findings in the reports. Evaluation on three independent datasets demonstrated that NegBio is highly accurate for detecting negative and equivocal findings (~90% in F-measure, which balances precision and recall [71]). These highly-accurate results meet the need to generate a corpus with weak labels, which serves as a solid foundation for the later process of image classification. The resulting dataset consists of 108,948 frontal-view chest X-ray images from 32,717 patients, and each image is associated with one or more weakly-labeled pathology category (e.g. pneumonia and cardiomegaly) or "normal" otherwise. Further, Wang et al. [70] used this dataset with a unified weakly-supervised multi-label image classification framework, to detect common thoracic diseases. It showed superior performance over a benchmark using fully-labeled data.

With the exception of natural image-like problems (e.g. melanoma detection), biomedical imaging poses a number of challenges for deep learning. Dataset are typically small, annotations can be sparse, and images are often high-dimensional, multimodal, and multi-channel. Techniques like transfer learning, heavy dataset augmentation, multi-view and multi-stream architectures are more common than in the natural image domain. Furthermore, high model sensitivity and specificity can translate directly into clinical value. Thus,

prediction evaluation, uncertainty estimation, and model interpretation methods are also of great importance in this domain (see Discussion). Finally, there is a need for better pathologist-computer interaction techniques that will allow combining the power of deep learning methods with human expertise and lead to better-informed decisions for patient treatment and care.

Electronic health records

EHR data include substantial amounts of free text, which remains challenging to approach [72]. Often, researchers developing algorithms that perform well on specific tasks must design and implement domain-specific features [73]. These features capture unique aspects of the literature being processed. Deep learning methods are natural feature constructors. In recent work, the authors evaluated the extent to which deep learning methods could be applied on top of generic features for domain-specific concept extraction [74]. They found that performance was in line with, but lower than the best domain-specific method [74]. This raises the possibility that deep learning may impact the field by reducing the researcher time and cost required to develop specific solutions, but it may not always lead to performance increases.

In recent work, Yoon et al[75] analyzed simple features using deep neural networks and found that the patterns recognized by the algorithms could be re-used across tasks. Their aim was to analyze the free text portions of pathology reports to identify the primary site and laterality of tumors. The only features the authors supplied to the algorithms were unigrams and bigrams, counts for single words and two-word combinations in a free text document. They subset the full set of words and word combinations to the 400 most common. The machine learning algorithms that they employed (naïve Bayes, logistic regression, and deep neural networks) all performed relatively similarly on the task of identifying the primary site. However, when the authors evaluated the more challenging task, evaluating the laterality of each tumor, the deep neural network outperformed the other methods. Of particular interest, when the authors first trained a neural network to predict primary site and then repurposed those features as a component of a secondary neural network trained to predict laterality, the performance was higher than a laterality-trained neural network. This demonstrates how deep learning methods can repurpose features across tasks, improving overall predictions as the field tackles new challenges. The Discussion further reviews this type of transfer learning.

Several authors have created reusable feature sets for medical terminologies using natural language processing and neural embedding models, as popularized by Word2vec [76]. A goal of learning terminologies for different entities in the same vector space is to find relationships between different domains (e.g. drugs and the diseases they treat). It is difficult for us to provide

a strong statement on the broad utility of these methods. Manuscripts in this area tend to compare algorithms applied to the same data but lack a comparison against overall best-practices for one or more tasks addressed by these methods. Techniques have been developed for free text medical notes [77], ICD and National Drug Codes, and claims data [78]. Methods for neural embeddings learned from electronic health records have at least some ability to predict disease-disease associations and implicate genes with a statistical association with a disease [79]. However, the evaluations performed did not differentiate between simple predictions (i.e. the same disease in different sites of the body) and non-intuitive ones. While promising, a lack of rigorous evaluations of the real-world utility of these kinds of features makes current contributions in this area difficult to evaluate. To examine the true utility, comparisons need to be performed against leading approaches (i.e. algorithms and data) as opposed to simply evaluating multiple algorithms on the same potentially limited dataset.

Identifying consistent subgroups of individuals and individual health trajectories from clinical tests is also an active area of research. Approaches inspired by deep learning have been used for both unsupervised feature construction and supervised prediction. Early work by Lasko et al. [80], combined sparse autoencoders and Gaussian processes to distinguish gout from leukemia from uric acid sequences. Later work showed that unsupervised feature construction of many features via denoising autoencoder neural networks could dramatically reduce the number of labeled examples required for subsequent supervised analyses [81]. In addition, it pointed towards learned features being useful for subtyping within a single disease. In a concurrent large-scale analysis of EHR data from 700,000 patients, Miotto et al. [82] used a deep denoising autoencoder architecture applied to the number and co-occurrence of clinical events (DeepPatient) to learn a representation of patients. The model was able to predict disease trajectories within one year with over 90% accuracy and patient-level predictions were improved by up to 15% when compared to other methods. Razavian et al. [83] used a set of 18 common lab tests to predict disease onset using both CNN and long short-term memory (LSTM) architectures and demonstrated an improvement over baseline regression models. However, numerous challenges including data integration (patient demographics, family history, laboratory tests, text-based patient records, image analysis, genomic data) and better handling of streaming temporal data with many features, will need to be overcome before we can fully assess the potential of deep learning for this application area.

Still, recent work has also revealed domains in which deep networks have proven superior to traditional methods. Survival analysis models the time leading to an event of interest from a shared starting point, and in the context of EHR data, often associates these events to subject covariates. Exploring

this relationship is difficult, however, given that EHR data types are often heterogeneous, covariates are often missing, and conventional approaches require the covariate-event relationship be linear and aligned to a specific starting point [84]. Early approaches, such as the Faraggi-Simon feed-forward network, aimed to relax the linearity assumption, but performance gains were lacking [85]. Katzman et al. in turn developed a deep implementation of the Faraggi-Simon network that, in addition to outperforming Cox regression, was capable of comparing the risk between a given pair of treatments, thus potentially acting as recommender system [86]. To overcome the remaining difficulties, researchers have turned to deep exponential families, a class of latent generative models that are constructed from any type of exponential family distributions [87]. The result was a deep survival analysis model capable of overcoming challenges posed by missing data and heterogeneous data types, while uncovering nonlinear relationships between covariates and failure time. They showed their model more accurately stratified patients as a function of disease risk score compared to the current clinical implementation.

There is a computational cost for these methods, however, when compared to traditional, non-neural network approaches. For the exponential family models, despite their scalability [88], an important question for the investigator is whether he or she is interested in estimates of posterior uncertainty. Given that these models are effectively Bayesian neural networks, much of their utility simplifies to whether a Bayesian approach is warranted for a given increase in computational cost. Moreover, as with all variational methods, future work must continue to explore just how well the posterior distributions are approximated, especially as model complexity increases [89].

Challenges and opportunities in patient categorization

Generating ground-truth labels can be expensive or impossible

A dearth of true labels is perhaps among the biggest obstacles for EHR-based analyses that employ machine learning. Popular deep learning (and other machine learning) methods are often used to tackle classification tasks and thus require ground-truth labels for training. For EHRs this can mean that researchers must hire multiple clinicians to manually read and annotate individual patients' records through a process called chart review. This allows researchers to assign "true" labels, i.e. those that match our best available knowledge. Depending on the application, sometimes the features constructed by algorithms also need to be manually validated and interpreted by clinicians. This can be time consuming and expensive [90]. Because of these costs, much of this research, including the work cited in this review, skips the process of expert review. Clinicians' skepticism for research without expert review may greatly dampen their enthusiasm for the work and consequently reduce its

impact. To date, even well-resourced large national consortia have been challenged by the task of acquiring enough expert-validated labeled data. For instance, in the eMERGE consortia and PheKB database [91], most samples with expert validation contain only 100 to 300 patients. These datasets are quite small even for simple machine learning algorithms. The challenge is greater for deep learning models with many parameters. While unsupervised and semi-supervised approaches can help with small sample sizes, the field would benefit greatly from large collections of anonymized records in which a substantial number of records have undergone expert review. This challenge is not unique to EHR-based studies. Work on medical images, omics data in applications for which detailed metadata are required, and other applications for which labels are costly to obtain will be hampered as long as abundant curated data are unavailable.

Successful approaches to date in this domain have sidestepped this challenge by making methodological choices that either reduce the need for labeled examples or that use transformations to training data to increase the number of times it can be used before overfitting occurs. For example, the unsupervised and semi-supervised methods that we have discussed reduce the need for labeled examples [81]. The anchor and learn framework [92] uses expert knowledge to identify high-confidence observations from which labels can be inferred. The adversarial training example strategies mentioned above can reduce overfitting, if transformations are available that preserve the meaningful content of the data while transforming irrelevant features [60]. While adversarial training examples can be easily imagined for certain methods that operate on images, it is more challenging to figure out what an equivalent transformation would be for a patient's clinical test results. Consequently, it may be hard to employ adversarial training examples, not to be confused with generative adversarial neural networks, with other applications. Finally, approaches that transfer features can also help use valuable training data most efficiently. Rajkomar et al. trained a deep neural network using generic images before tuning using only radiology images [54]. Datasets that require many of the same types of features might be used for initial training, before fine tuning takes place with the more sparse biomedical examples. Though the analysis has not yet been attempted, it is possible that analogous strategies may be possible with electronic health records. For example, features learned from the electronic health record for one type of clinical test (e.g. a decrease over time in a lab value) may transfer across phenotypes.

Methods to accomplish more with little high-quality labeled data are also being applied in other domains and may also be adapted to this challenge, e.g. data programming [93]. In data programming, noisy automated labeling functions are integrated. Numerous writers have described data as the new oil [94,95]. The idea behind this metaphor is that data are available in large quantities,

valuable once refined, and the underlying resource that will enable a data-driven revolution in how work is done. Contrasting with this perspective, Ratner, Bach, and Ré described labeled training data as "The *New* New Oil" [96]. In this framing, data are abundant and not a scarce resource. Instead, new approaches to solving problems arise when labeled training data become sufficient to enable them. Based on our review of research on deep learning methods to categorize disease, the latter framing rings true.

We expect improved methods for domains with limited data to play an important role if deep learning is going to transform how we categorize states of human health. We don't expect that deep learning methods will replace expert review. We expect them to complement expert review by allowing more efficient use of the costly practice of manual annotation.

Data sharing is hampered by standardization and privacy considerations

To construct the types of very large datasets that deep learning methods thrive on, we need robust sharing of large collections of data. This is in part a cultural challenge. We touch on this challenge in the Discussion section. Beyond the cultural hurdles around data sharing, there are also technological and legal hurdles related to sharing individual health records or deep models built from such records. This subsection deals primarily with these challenges.

EHRs are designed chiefly for clinical, administrative and financial purposes, such as patient care, insurance and billing [97]. Science is at best a tertiary priority, presenting challenges to EHR-based research in general and to deep learning research in particular. Although there is significant work in the literature around EHR data quality and the impact on research [98], we focus on three types of challenges: local bias, wider standards, and legal issues. Note these problems are not restricted to EHRs but can also apply to any large biomedical dataset, e.g. clinical trial data.

Even within the same healthcare system, EHRs can be used differently [99,100]. Individual users have unique documentation and ordering patterns, with different departments and different hospitals having different priorities that code patients and introduce missing data in a non-random fashion [101]. Patient data may be kept across several "silos" within a single health system (e.g. separate nursing documentation, registries, etc.). Even the most basic task of matching patients across systems can be challenging due to data entry issues [102]. The situation is further exacerbated by the ongoing introduction, evolution, and migration of EHR systems, especially where reorganized and acquired healthcare facilities have to merge. Further, even the ostensibly least-biased data type, laboratory measurements, can be biased based by both the healthcare process and patient health state [103]. As a result, EHR data can be

less complete and less objective than expected.

In the wider picture, standards for EHRs are numerous and evolving. Proprietary systems, indifferent and scattered use of health information standards, and controlled terminologies makes combining and comparison of data across systems challenging [104]. Further diversity arises from variation in languages, healthcare practices, and demographics. Merging EHR gathered in different systems (and even under different assumptions) is challenging [105].

Combining or replicating studies across systems thus requires controlling for both the above biases and dealing with mismatching standards. This has the practical effect of reducing cohort size, limiting statistical significance, preventing the detection of weak effects [106], and restricting the number of parameters that can be trained in a model. Further, rules-based algorithms have been popular in EHR-based research, but because these are developed at a single institution and trained with a specific patient population, they do not transfer easily to other healthcare systems [107]. Genetic studies using EHR data are subject to even more bias, as the differences in population ancestry across health centers (e.g. proportion of patients with African or Asian ancestry) can affect algorithm performance. For example, Wiley et al. [108] showed that warfarin dosing algorithms often under-perform in African Americans, illustrating that some of these issues are unresolved even at a treatment best practices level. Lack of standardization also makes it challenging for investigators skilled in deep learning to enter the field, as numerous data processing steps must be performed before algorithms are applied.

Finally, even if data were perfectly consistent and compatible across systems, attempts to share and combine EHR data face considerable legal and ethical barriers. Patient privacy can severely restrict the sharing and use of EHR [109]. Here again, standards are heterogeneous and evolving, but often EHR data can often not be exported or even accessed directly for research purposes without appropriate consent. In the United States, research use of EHR data is subject both to the Common Rule and the Health Insurance Portability and Accountability Act (HIPPA). Ambiguity in the regulatory language and individual interpretation of these rules can hamper use of EHR data [110]. Once again, this has the effect of making data gathering more laborious and expensive, reducing sample size and study power.

Several technological solutions have been proposed in this direction, allowing access to sensitive data satisfying privacy and legal concerns. Software like DataShield [111] and ViPAR [112], although not EHR-specific, allows querying and combining of datasets and calculation of summary statistics across remote sites by "taking the analysis to the data". The computation is carried out at the remote site. Conversely, the EH4CR project [104] allows analysis of private

data by use of an inter-mediation layer that interprets remote queries across internal formats and datastores and returns the results in a de-identified standard form, thus giving real-time consistent but secure access. Continuous Analysis [113] can allow reproducible computing on private data. Using such techniques, intermediate results can be automatically tracked and shared without sharing the original data. While none of these have been used in deep learning, the potential is there.

Even without sharing data, algorithms trained on confidential patient data may present security risks or accidentally allow for the exposure of individual level patient data. Tramer et al. [114] showed the ability to steal trained models via public application programming interfaces (APIs). Dwork and Roth [115] demonstrate the ability to expose individual level information from accurate answers in a machine learning model. Attackers can use similar attacks to find out if a particular data instance was present in the original training set for the machine learning model [116], in this case, whether a person's record was present. This presents a potential hazard for approaches that aim to generate data. Choi et al. propose generative adversarial neural networks as a tool to make sharable EHR data [117]; however, the authors did not take steps to protect the model from such attacks.

There are approaches to protect models, but they pose their own challenges. Training in a differentially private manner provides a limited guarantee that an algorithm's output will be equally likely to occur regardless of the participation of any one individual. The limit is determined by a single parameter which provides a quantification of privacy. Simmons et al. [118] present the ability to perform genome-wide association studies (GWASs) in a differentially private manner, and Abadi et al. [119] show the ability to train deep learning classifiers under the differential privacy framework. Federated learning [120] and secure aggregations [121,122] are complementary approaches that reinforce differential privacy. Both aim to maintain privacy by training deep learning models from decentralized data sources such as personal mobile devices without transferring actual training instances. This is becoming of increasing importance with the rapid growth of mobile health applications. However, the training process in these approaches places constraints on the algorithms used and can make fitting a model substantially more challenging. In our own experience, it can be trivial to train a model without differential privacy, but quite difficult to train one within the differential privacy framework. The problem can be particularly pronounced with small sample sizes.

While none of these problems are insurmountable or restricted to deep learning, they present challenges that cannot be ignored. Technical evolution in EHRs and data standards will doubtless ease -- although not solve -- the problems of data sharing and merging. More problematic are the privacy

issues. Those applying deep learning to the domain should consider the potential of inadvertently disclosing the participants' identities. Techniques that enable training on data without sharing the raw data may have a part to play. Training within a differential privacy framework may often be warranted.

Discrimination and "right to an explanation" laws

In April 2016, the European Union adopted new rules regarding the use of personal information, the General Data Protection Regulation [123]. A component of these rules can be summed up by the phrase "right to an explanation". Those who use machine learning algorithms must be able to explain how a decision was reached. For example, a clinician treating a patient who is aided by a machine learning algorithm may be expected to explain decisions that use the patient's data. The new rules were designed to target categorization or recommendation systems, which inherently profile individuals. Such systems can do so in ways that are discriminatory and unlawful.

As datasets become larger and more complex, we may begin to identify relationships in data that are important for human health but difficult to understand. The algorithms described in this review and others like them may become highly accurate and useful for various purposes, including within medical practice. However, to discover and avoid discriminatory applications it will be important to consider interpretability alongside accuracy. A number of properties of genomic and healthcare data will make this difficult.

First, research samples are frequently non-representative of the general population of interest; they tend to be disproportionately sick [124], male [125], and European in ancestry [126]. One well-known consequence of these biases in genomics is that penetrance is consistently lower in the general population than would be implied by case-control data, as reviewed in [124]. Moreover, real genetic associations found in one population may not hold in other populations with different patterns of linkage disequilibrium (even when population stratification is explicitly controlled for [127]). As a result, many genomic findings are of limited value for people of non-European ancestry [126] and may even lead to worse treatment outcomes for them. Methods have been developed for mitigating some of these problems in genomic studies [124,127], but it is not clear how easily they can be adapted for deep models that are designed specifically to extract subtle effects from high-dimensional data. For example, differences in the equipment that tended to be used for cases versus controls have led to spurious genetic findings (e.g. Sebastiani et al.'s retraction [128]). In some contexts, it may not be possible to correct for all of these differences to the degree that a deep network is unable to use them. Moreover, the complexity of deep networks makes it difficult to determine when their predictions are likely to be based on such nominally-irrelevant features of

the data (called "leakage" in other fields [129]). When we are not careful with our data and models, we may inadvertently say more about the way the data was collected (which may involve a history of unequal access and discrimination) than about anything of scientific or predictive value. This fact can undermine the privacy of patient data [129] or lead to severe discriminatory consequences [130].

There is a small but growing literature on the prevention and mitigation of data leakage [129], as well as a closely-related literature on discriminatory model behavior [131], but it remains difficult to predict when these problems will arise, how to diagnose them, and how to resolve them in practice. There is even disagreement about which kinds of algorithmic outcomes should be considered discriminatory [132]. Despite the difficulties and uncertainties, machine learning practitioners (and particularly those who use deep neural networks, which are challenging to interpret) must remain cognizant of these dangers and make every effort to prevent harm from discriminatory predictions. To reach their potential in this domain, deep learning methods will need to be interpretable. Researchers need to consider the extent to which biases may be learned by the model and whether or not a model is sufficiently interpretable to identify bias. We discuss the challenge of model interpretability more thoroughly in the discussion section.

Applications of deep learning to longitudinal analysis

Longitudinal analysis follows a population across time, for example, prospectively from birth or from the onset of particular conditions. In large patient populations, longitudinal analyses such as the Farmingham Heart Study [133] and the Avon Longitudinal Study of Parents and Children[134] have yielded important discoveries about the development of disease and the factors contributing to health status. Yet, a common practice in EHR-based research is to take a point in time snapshot and convert patient data to a traditional vector for machine learning and statistical analysis. This results in loss of information as timing and order of events can provide insight into a patient's disease and treatment [135]. Efforts to model sequences of events have shown promise [136] but require exceedingly large patient sizes due to discrete combinatorial bucketing. Lasko et al. [80] used autoencoders on longitudinal sequences of serum urine acid measurements to identify population subtypes. More recently, deep learning has shown promise working with both sequences (CNNs) [137] and the incorporation of past and current state (RNNs, LSTMs) [138]. This may be a particular area of opportunity for deep neural networks. The ability to recognize relevant sequences of events from a large number of trajectories requires powerful and flexible feature construction methods -- an area in which deep neural networks excel.

Deep learning to study the fundamental biological processes underlying human disease

The study of cellular structure and core biological processes -- transcription, translation, signaling, metabolism, etc. -- in humans and model organisms will greatly impact our understanding of human disease over the long horizon [139]. Predicting how cellular systems respond to environmental perturbations and are altered by genetic variation remain daunting tasks. Deep learning offers new approaches for modeling biological processes and integrating multiple types of omic data [140], which could eventually help predict how these processes are disrupted in disease. Recent work has already advanced our ability to identify and interpret genetic variants, study microbial communities, and predict protein structures, which also relates to the problems discussed in the drug development section. In addition, unsupervised deep learning has enormous potential for discovering novel cellular states from gene expression, fluorescence microscopy, and other types of data that may ultimately prove to be clinically relevant.

Progress has been rapid in genomics and imaging, fields where important tasks are readily adapted to well-established deep learning paradigms. One-dimensional convolutional and recurrent neural networks are well-suited for tasks related to DNA- and RNA-binding proteins, epigenomics, and RNA splicing. Two dimensional CNNs are ideal for segmentation, feature extraction, and classification in fluorescence microscopy images [16]. Other areas, such as cellular signaling, are biologically important but studied less-frequently to date, with some exceptions [141]. This may be a consequence of data limitations or greater challenges in adapting neural network architectures to the available data. Here, we highlight several areas of investigation and assess how deep learning might move these fields forward.

Gene expression

Gene expression technologies characterize the abundance of many thousands of RNA transcripts within a given organism, tissue, or cell. This characterization can represent the underlying state of the given system and can be used to study heterogeneity across samples as well as how the system reacts to perturbation. While gene expression measurements were traditionally made by quantitative polymerase chain reaction (qPCR), low-throughput fluorescence-based methods, and microarray technologies, the field has shifted in recent years to primarily performing RNA sequencing (RNA-seq) to catalog whole transcriptomes. As RNA-seq continues to fall in price and rise in throughput, sample sizes will increase and training deep models to study gene expression will become even more useful.

Already several deep learning approaches have been applied to gene expression data with varying aims. For instance, many researchers have applied unsupervised deep learning models to extract meaningful representations of gene modules or sample clusters. Denoising autoencoders have been used to cluster yeast expression microarrays into known modules representing cell cycle processes [142] and to stratify yeast strains based on chemical and mutational perturbations [143]. Shallow (one hidden layer) denoising autoencoders have also been fruitful in extracting biological insight from thousands of *Pseudomonas aeruginosa* experiments [144,145] and in aggregating features relevant to specific breast cancer subtypes [24]. These unsupervised approaches applied to gene expression data are powerful methods for identifying gene signatures that may otherwise be overlooked. An additional benefit of unsupervised approaches is that ground truth labels, which are often difficult to acquire or are incorrect, are nonessential. However, the genes that have been aggregated into features must be interpreted carefully. Attributing each node to a single specific biological function risks over-interpreting models. Batch effects could cause models to discover non-biological features, and downstream analyses should take this into consideration.

Deep learning approaches are also being applied to gene expression prediction tasks. For example, a deep neural network with three hidden layers outperformed linear regression in inferring the expression of over 20,000 target genes based on a representative, well-connected set of about 1,000 landmark genes [146]. However, while the deep learning model outperformed existing algorithms in nearly every scenario, the model still displayed poor performance. The paper was also limited by computational bottlenecks that required data to be split randomly into two distinct models and trained separately. It is unclear how much performance would have increased if not for computational restrictions.

Epigenetic data, combined with deep learning, may have sufficient explanatory power to infer gene expression. For instance, a convolutional neural network applied to histone modifications, termed DeepChrome, [147] improved prediction accuracy of high or low gene expression over existing methods. Deep learning can also integrate different data types. For example, Liang et al. combined RBMs to integrate gene expression, DNA methylation, and miRNA data to define ovarian cancer subtypes [148]. While these approaches are promising, many convert gene expression measurements to categorical or binary variables, thus ablating many complex gene expression signatures present in intermediate and relative numbers.

Deep learning applied to gene expression data is still in its infancy, but the future is bright. Many previously untestable hypotheses can now be

interrogated as deep learning enables analysis of increasing amounts of data generated by new technologies. For example, the effects of cellular heterogeneity on basic biology and disease etiology can now be explored by single-cell RNA-seq and high-throughput fluorescence-based imaging, techniques we discuss below that will benefit immensely from deep learning approaches.

Splicing

Pre-mRNA transcripts can be spliced into different isoforms by retaining or skipping subsets of exons or including parts of introns, creating enormous spatiotemporal flexibility to generate multiple distinct proteins from a single gene. This remarkable complexity can lend itself to defects that underlie many diseases [149]. For instance, in Becker muscular dystrophy, a point mutation in dystrophin creates an exon splice silencer that induces skipping of exon 31. A recent study found that quantitative trait loci that affect splicing in lymphoblastoid cell lines are enriched within risk loci for schizophrenia, multiple sclerosis, and other immune diseases, implicating mis-splicing as a more widespread feature of human pathologies than previously thought [150].

Sequencing studies routinely return thousands of unannotated variants, but which cause functional changes in splicing and how are those changes manifested? Prediction of a "splicing code" has been a goal of the field for the past decade. Initial machine learning approaches used a naïve Bayes model and a 2-layer Bayesian neural network with thousands of hand-derived sequence-based features to predict the probability of exon skipping [151,152]. With the advent of deep learning, more complex models were built that provided better predictive accuracy [153,154]. Importantly, these new approaches can take in multiple kinds of epigenomic measurements as well as tissue identity and RNA binding partners of splicing factors. Deep learning is critical in furthering these kinds of integrative studies where different data types and inputs interact in unpredictable (often nonlinear) ways to create higher-order features. Moreover, as in gene expression network analysis, interrogating the hidden nodes within neural networks could potentially illuminate important aspects of splicing behavior. For instance, tissue-specific splicing mechanisms could be inferred by training networks on splicing data from different tissues, then searching for common versus distinctive hidden nodes, a technique employed by Qin et al. for tissue-specific transcription factor (TF) binding predictions [155].

A parallel effort has been to use more data with simpler models. An exhaustive study using readouts of splicing for millions of synthetic intronic sequences uncovered motifs that influence the strength of alternative splice sites [156]. The authors built a simple linear model using hexamer motif frequencies that

successfully generalized to exon skipping. In a limited analysis using single nucleotide polymorphisms (SNPs) from three genes, it predicted exon skipping with three times the accuracy of an existing deep learning-based framework [153]. This case is instructive in that clever sources of data, not just more descriptive models, are still critical.

We already understand how mis-splicing of a single gene can cause diseases such as Duchenne muscular dystrophy. The challenge now is to uncover how genome-wide alternative splicing underlies complex, non-Mendelian diseases such as autism, schizophrenia, Type 1 diabetes, and multiple sclerosis [157]. As a proof of concept, Xiong et al. [153] sequenced five autism spectrum disorder and 12 control samples, each with an average of 42,000 rare variants, and identified mis-splicing in 19 genes with neural functions. Such methods may one day enable scientists and clinicians to rapidly profile thousands of unannotated variants for functional effects on splicing and nominate candidates for further investigation. Moreover, these nonlinear algorithms can deconvolve the effects of multiple variants on a single splice event without the need to perform combinatorial *in vitro* experiments. The ultimate goal is to predict an individual's tissue-specific, exon-specific splicing patterns from their genome sequence and other measurements to enable a new branch of precision diagnostics that also stratifies patients and suggests targeted therapies to correct splicing defects. However, to achieve this we expect that methods to interpret the "black box" of deep neural networks and integrate diverse data sources will be required.

Transcription factors and RNA-binding proteins

Transcription factors and RNA-binding proteins are key components in gene regulation and higher-level biological processes. TFs are regulatory proteins that bind to certain genomic loci and control the rate of mRNA production. While high-throughput sequencing techniques such as chromatin immunoprecipitation and massively parallel DNA sequencing (ChIP-seq) have been able to accurately identify targets for TFs, these experiments are both time consuming and expensive. Thus, there is a need to computationally predict binding sites and understand binding specificities *de novo* from sequence data. In this section we focus on TFs, with the understanding that deep learning methods for TFs are similar to those for RNA-binding proteins, though RNA-specific models do exist [158].

ChIP-seq and related technologies are able to identify highly likely binding sites for a certain TF, and databases such as ENCODE [159] have made freely available ChIP-seq data for hundreds of different TFs across many laboratories. In order to computationally predict transcription factor binding sites (TFBSs) on a DNA sequence, researchers initially used consensus

sequences and position weight matrices to match against a test sequence [160]. Simple neural network classifiers were then proposed to differentiate positive and negative binding sites but did not show meaningful improvements over the weight matrix matching methods [161]. Later, support vector machines (SVMs) outperformed the generative methods by using k-mer features [162,163], but string kernel-based SVM systems are limited by their expensive computational cost, which is proportional to the number of training and testing sequences.

With the advent of deep learning, Alipanahi et al.[164] showed that convolutional neural network models could achieve state of the art results on the TFBS task and are scalable to a large number of genomic sequences. Lanchantin et al. [165] introduced several new convolutional and recurrent neural network models that further improved TFBS predictive accuracy. Due to the motif-driven nature of the TFBS task, most architectures have been convolution-based [166]. While many models for TFBS prediction resemble computer vision and NLP tasks, it is important to note that DNA sequence tasks are fundamentally different. Thus the models should be adapted from traditional deep learning models in order to account for such differences. For example, motifs may appear in either strand of a DNA sequence, resulting in two different forms of the motif (forward and reverse complement) due to complementary base pairing. To handle this issue, specialized reverse complement convolutional models share parameters to find motifs in both directions [167].

Despite these advances, several challenges remain. First, because the inputs (ChIP-seq measurements) are continuous and most current algorithms are designed to produce binary outputs (whether or not there is TF binding at a particular site), false positives or false negatives can result depending on the threshold chosen by the algorithm. Second, most methods predict binding of TFs at sites in isolation, whereas in reality multiple TFs may compete for binding at a single site or act synergistically to co-occupy it. Fortunately, multi-task models are rapidly improving at simultaneous prediction of many TFs' binding at any given site [168]. Third, it is unclear exactly how to define a non-binding or "negative" site in the training data because the number of positive binding sites of a particular TF is relatively small with respect to the total number of base-pairs in a genome (see Discussion).

While deep learning-based models can automatically extract features for TFBS prediction at the sequence level, they often cannot predict binding patterns for cell types or conditions that have not been previously studied. One solution could be to introduce a multimodal model that, in addition to sequence data, incorporates cell-line specific features such as chromatin accessibility, DNA methylation, or gene expression. Without cell-specific features, another

solution could be to use domain adaptation methods where the model trains on a source cell type and uses unsupervised feature extraction methods to predict on a target cell type. TFImpute [155] predicts binding in new cell type-TF pairs, but the cell types must be in the training set for other TFs. This is a step in the right direction, but a more general domain transfer model across cell types would be more useful.

Deep learning can also illustrate TF binding preferences. Lanchantin et al.[165] and Shrikumar et al. [169] developed tools to visualize TF motifs learned from TFBS classification tasks. Alipanahi et al. [164] also introduced mutation maps, where they could easily mutate, add, or delete base pairs in a sequence and see how the model changed its prediction. Though time consuming to assay in a lab, this was easy to simulate with a computational model. As we learn to better visualize and analyze the hidden nodes within deep learning models, our understanding of TF binding motifs and dynamics will likely improve.

Promoters, enhancers, and related epigenomic tasks

Transcriptional control is undoubtedly a vital, early part of the regulation of gene expression. An abundance of sequence and associated functional data (e.g. ENCODE [159] and ExAC [170]) exists across species. At the same time, studies of gene regulation have often focused on the protein (binding) rather than the promoter level [171], perhaps due to the ill-defined nature of cis-regulatory elements (CREs). A promoter itself can be seen as an assemblage of "active" binding sites for transcription factors interspersed by less-characterized and perhaps functionally silent spacer regions. However, the sequence signals that control the start and stop of transcription and translation are still not well understood, compounded by incomplete understanding of alternative transcripts and the context for these alternatives. Sequence similarity is poor even between functionally correlated genes. While homologs might be studied for insight, they may not exist or may be just as poorly characterized.

Recognizing enhancers presents additional challenges. Enhancers may be up to one million base pairs upstream or downstream from the affected promoter on either strand and even within the introns of other genes [172]. They do not necessarily operate on the nearest gene and may affect multiple genes. Their activity is frequently tissue- or context-specific. A substantial fraction of enhancers displays modest or no conservation across species. There is no universal enhancer sequence signal or marker for enhancers, and some literature suggests that enhancers and promoters may be just categories along a spectrum [173]. One study [174] even showed that only 33% of predicted regulatory regions could be validated, while a class of "weak" predicted enhancers were strong drivers of expression. Yet there is growing evidence for

their vast ubiquity, making them possibly the predominant functional non-coding element. Thus, identifying enhancers is critical yet the search space is large.

While prior (non-deep learning) approaches have made steady improvements on promoter prediction, there is little consensus on the best approach and performance is poor. Typically algorithms will recognize only half of all promoters, with an accompanying high false positive rate [175]. Methods with better sensitivity generally do so at the cost of poorer specificity. Conventional identification of enhancers has leaned heavily on simple conservation or laborious experimental techniques, with only moderate sensitivity and specificity. For example, while chromatin accessibility has often been used for identifying enhancers, this also "recognizes" a wide variety of other functional elements, like promoters, silencers, and repressors.

The complex nature of CREs and our lack of understanding makes them a natural candidate for deep learning approaches. Indeed, neural networks were used for promoter recognition as early as 1996, albeit with mixed results [176]. Since then, there has been much work in applying deep learning to this area, although little in the way of comparative studies or formal benchmarks. We therefore focus on a few recent important studies to outline the state of the art and extant problems.

Basset [177] trained CNNs on DNA accessibility datasets, getting a marked improvement on previous methods, albeit still with a high false positive rate. The multi-task architecture resembles DeepSEA [168], which predicted open chromatin regions and histone modifications in addition to TF binding. As noted above, predicting DNA accessibility conflates enhancers with other functional sites. Basset also featured a useful interpretability approach, introducing known protein binding motifs into sequences and measuring the change in predicted accessibility.

Umarov et al. [178] demonstrated the use of CNNs in recognizing promoter sequences, outperforming conventional methods (sensitivity and specificity exceeding 90%). While some results were achieved over bacterial promoters (which are considerably simpler in structure), roughly similar performance was found for human promoters. This work also included a simple method for model interpretation, randomly substituting bases in a recognized promoter region, then checking for a change in recognition (see Discussion).

Xu et al. [179] applied CNNs to the detection of enhancers, achieving incremental improvements in specificity and sensitivity over a previous SVM-based approach, and much better performance for cell-specific enhancers. A massive improvement in speed was also achieved. Additionally, they compared the performance of different CNN architectures, finding that while

layers for batch normalization improved performance, deeper architectures decreased performance.

Singh et al. [180] approached the problem of predicting enhancer-promoter interactions from solely the sequence and location of putative enhancers and promoters in a particular cell type. Performance was comparative to state-of-the-art conventional techniques that used the whole gamut of full functional genomic and non-sequence data.

Given the conflation between different CREs, the study of Li et al[181] is particularly interesting. They used a feed-forward neural network to distinguish classes of CREs and activity states. Active enhancers and promoters could be easily be distinguished, as could active and inactive elements. Perhaps unsurprisingly, it was difficult to distinguish between inactive enhancers and promoters. They also investigated the power of sequence features to drive classification, finding that beyond CpG islands, few were useful.

In summary, deep learning is a promising approach for identifying CREs, able to interrogate sequence features that are complex and ill-understood, already offering marked improvements on the prior state of the art. However, neural network architectures for this task need to be systematically compared. The challenges in predicting TF binding -- such as the lack of large gold standard datasets, model interpretation, and defining negative examples -- are pertinent to CRE identification as well. Furthermore, the quality and meaning of training data needs to be closely considered, given that a "promoter" or "enhancer" may only be putative or dependent on the experimental method or context of identification. Otherwise we risk building detectors not for CREs but putative CREs. Most deep learning studies in this area currently predict the 1D location of enhancers, but modeling 3D chromatin conformations, enhancer-promoter interactions [180], and enhancer-target gene interactions will be critical for understanding transcriptional regulation.

Micro-RNA binding

Prediction of microRNAs (miRNAs) in the genome as well as miRNA targets is of great interest, as they are critical components of gene regulatory networks and are often conserved across great evolutionary distance [182,183]. While many machine learning algorithms have been applied to solve these prediction tasks, they currently require extensive feature selection and optimization. For instance, one of the most widely adopted tools for miRNA target prediction, TargetScan, trained multiple linear regression models on 14 hand-curated features including structural accessibility of the target site on the mRNA, the degree of site conservation, and predicted thermodynamic stability of the miRNA-mRNA complex [184]. Some of these features, including structural

accessibility, are imperfect or empirically derived. In addition, current algorithms suffer from low specificity [185].

As in other applications, deep learning promises to achieve equal or better performance in predictive tasks by automatically engineering complex features to minimize an objective function. Two recently published tools use different recurrent neural network-based architectures to perform miRNA and target prediction with solely sequence data as input [185,186]. Though the results are preliminary and still based on a validation set rather than a completely independent test set, they were able to predict microRNA target sites with higher specificity and sensitivity than TargetScan. Excitingly, these tools seem to show that RNNs can accurately align sequences and predict bulges, mismatches, and wobble base pairing without requiring the user to input secondary structure predictions or thermodynamic calculations. Further incremental advances in deep learning for miRNA and target prediction will likely be sufficient to meet the current needs of systems biologists and other researchers who use prediction tools mainly to nominate candidates that are then tested experimentally.

Protein secondary and tertiary structure

Proteins play fundamental roles in almost all biological processes, and understanding their structure is critical for basic biology and drug development. UniProt currently has about 94 million protein sequences, yet fewer than 100,000 proteins across all species have experimentally-solved structures in Protein Data Bank (PDB). As a result, computational structure prediction is essential for a majority of proteins. However, this is very challenging, especially when similar solved structures, called templates, are not available in PDB. Over the past several decades, many computational methods have been developed to predict aspects of protein structure such as secondary structure, torsion angles, solvent accessibility, inter-residue contact maps, disorder regions, and side-chain packing. In recent years, multiple deep learning architectures have been applied, including deep belief networks, LSTMs, CNNs, and deep convolutional neural fields (DeepCNFs) [29,187].

Here we focus on deep learning methods for two representative sub-problems: secondary structure prediction and contact map prediction. Secondary structure refers to local conformation of a sequence segment, while a contact map contains information on all residue-residue contacts. Secondary structure prediction is a basic problem and an almost essential module of any protein structure prediction package. Contact prediction is much more challenging than secondary structure prediction, but it has a much larger impact on tertiary structure prediction. In recent years, the accuracy of contact prediction has greatly improved [27,188–190].

Protein secondary structure can exhibit three different states (alpha helix, beta strand, and loop regions) or eight finer-grained states. Q3 and Q8 accuracy pertain to 3-state or 8-state predictions, respectively. Several groups [28,191,192] initiated the application of deep learning to protein secondary structure prediction but were unable to achieve significant improvement over the *de facto* standard method PSIPRED [193], which uses two shallow feedforward neural networks. In 2014, Zhou and Troyanskaya demonstrated that they could improve Q8 accuracy by using a deep supervised and convolutional generative stochastic network [194]. In 2016 Wang et al. developed a DeepCNF model that improved Q3 and Q8 accuracy as well as prediction of solvent accessibility and disorder regions [29,187]. DeepCNF achieved a higher Q3 accuracy than the standard maintained by PSIPRED for more than 10 years. This improvement may be mainly due to the ability of convolutional neural fields to capture long-range sequential information, which is important for beta strand prediction. Nevertheless, the improvements in secondary structure prediction from DeepCNF are unlikely to result in a commensurate improvement in tertiary structure prediction since secondary structure mainly reflects coarse-grained local conformation of a protein structure.

Protein contact prediction and contact-assisted folding (i.e. folding proteins using predicted contacts as restraints) represents a promising new direction for *ab initio* folding of proteins without good templates in PDB. Co-evolution analysis is effective for proteins with a very large number (>1000) of sequence homologs [190], but otherwise fares poorly for proteins without many sequence homologs. By combining co-evolution information with a few other protein features, shallow neural network methods such as MetaPSICOV [188] and CoinDCA-NN [195] have shown some advantage over pure co-evolution analysis for proteins with few sequence homologs, but their accuracy is still far from satisfactory. In recent years, deeper architectures have been explored for contact prediction, such as CMAPpro [196], DNCON [197] and PConsC [198]. However, blindly tested in the well-known CASP competitions, these methods did not show any advantage over MetaPSICOV [188].

Recently, Wang et al. proposed the deep learning method RaptorX-Contact [27], which significantly improves contact prediction over MetaPSICOV and pure co-evolution methods, especially for proteins without many sequence homologs. It employs a network architecture formed by one 1D residual neural network and one 2D residual neural network. Blindly tested in the latest CASP competition (i.e. CASP12 [199]), RaptorX-Contact ranked first in F1 score (a widely-used performance metric combining sensitivity and specificity) on free-modeling targets as well as the whole set of targets. In CAMEO (which can be interpreted as a fully-automated CASP) [200], its predicted contacts were also able to fold proteins with a novel fold and only 65-330 sequence homologs.

This technique also worked well on membrane proteins even when trained on non-membrane proteins [201]. RaptorX-Contact performed better mainly due to introduction of residual neural networks and exploitation of contact occurrence patterns by simultaneously predicting all the contacts in a single protein.

Taken together, *ab initio* folding is becoming much easier with the advent of direct evolutionary coupling analysis and deep learning techniques. We expect further improvements in contact prediction for proteins with fewer than 1000 homologs by studying new deep network architectures. However, it is unclear if there is an effective way to use deep learning to improve prediction for proteins with few or no sequence homologs. Finally, the deep learning methods summarized above also apply to interfacial contact prediction for protein complexes but may be less effective since on average protein complexes have fewer sequence homologs.

Morphological phenotypes

A field poised for dramatic revolution by deep learning is bioimage analysis. Thus far, the primary use of deep learning for biological images has been for segmentation -- that is, for the identification of biologically relevant structures in images such as nuclei, infected cells, or vasculature -- in fluorescence or even brightfield channels [202]. Once so-called regions of interest have been identified, it is often straightforward to measure biological properties of interest, such as fluorescence intensities, textures, and sizes. Given the dramatic successes of deep learning in biological imaging, we simply refer to articles that review recent advancements [16,202,203]. For deep learning to become commonplace for biological image segmentation, user-friendly tools need to be developed.

We anticipate an additional kind of paradigm shift in bioimaging that will be brought about by deep learning: what if images of biological samples, from simple cell cultures to three-dimensional organoids and tissue samples, could be mined for much more extensive biologically meaningful information than is currently standard? For example, a recent study demonstrated the ability to predict lineage fate in hematopoietic cells up to three generations in advance of differentiation [204]. In biomedical research, by far the most common paradigm is for biologists to decide in advance what feature to measure in images from their assay system. Although classical methods of segmentation and feature extraction can produce hundreds of metrics per cell in an image, deep learning is unconstrained by human intuition and can in theory extract more subtle features through its hidden nodes. Already, there is evidence deep learning can surpass the efficacy of classical methods [205], even using generic deep convolutional networks trained on natural images [206], known as transfer learning. Recent work by Johnson et al. [207] demonstrated how the use of a

conditional adversarial autoencoder allows for a probabilistic interpretation of cell and nuclear morphology and structure localization from fluorescence images. The proposed model is able to generalize well to a wide range of subcellular localizations. The generative nature of the model allows it to produce high-quality synthetic images predicting localization of subcellular structures by directly modeling the localization of fluorescent labels. Notably, this approach reduces the modeling time by omitting the subcellular structure segmentation step.

The impact of further improvements on biomedicine could be enormous. Comparing cell population morphologies using conventional methods of segmentation and feature extraction has already proven useful for functionally annotating genes and alleles, identifying the cellular target of small molecules, and identifying disease-specific phenotypes suitable for drug screening [208–210]. Deep learning would bring to these new kinds of experiments -- known as image-based profiling or morphological profiling -- a higher degree of accuracy, stemming from the freedom from human-tuned feature extraction strategies.

Single-cell data

Single-cell methods are generating excitement as biologists recognize the vast heterogeneity within unicellular species and between cells of the same tissue type in the same organism [211]. For instance, tumor cells and neurons can both harbor extensive somatic variation [212]. Understanding single-cell diversity in all its dimensions -- genetic, epigenetic, transcriptomic, proteomic, morphologic, and metabolic -- is key if treatments are to be targeted not only to a specific individual, but also to specific pathological subsets of cells. Single-cell methods also promise to uncover a wealth of new biological knowledge. A sufficiently large population of single cells will have enough representative "snapshots" to recreate timelines of dynamic biological processes. If tracking processes over time is not the limiting factor, single-cell techniques can provide maximal resolution compared to averaging across all cells in bulk tissue, enabling the study of transcriptional bursting with single-cell fluorescence *in situ* hybridization or the heterogeneity of epigenetic patterns with single-cell Hi-C or ATAC-seq [213,214]. Joint profiling of single-cell epigenetic and transcriptional states provides unprecedented views of regulatory processes [215].

However, large challenges exist in studying single cells. Relatively few cells can be assayed at once using current droplet, imaging, or microwell technologies, and low-abundance molecules or modifications may not be detected by chance due to a phenomenon known as dropout. To solve this problem, Angermueller et al. [216] trained a neural network to predict the

presence or absence of methylation of a specific CpG site in single cells based on surrounding methylation signal and underlying DNA sequence, achieving several percentage points of improvement compared to random forests or deep networks trained only on CpG or sequence information. Similar deep learning methods have been applied to impute low-resolution ChIP-seq signal from bulk tissue with great success, and they could easily be adapted to single-cell data [155,217]. Deep learning has also been useful for dealing with batch effects [218].

Examining populations of single cells can reveal biologically meaningful subsets of cells as well as their underlying gene regulatory networks [219]. Unfortunately, machine learning methods generally struggle with imbalanced data -- when there are many more examples of class 1 than class 2 -- because prediction accuracy is usually evaluated over the entire dataset. To tackle this challenge, Arvaniti et al. [220] classified healthy and cancer cells expressing 25 markers by using the most discriminative filters from a CNN trained on the data as a linear classifier. They achieved impressive performance, even for cell types where the subset percentage ranged from 0.1 to 1%, significantly outperforming logistic regression and distance-based outlier detection methods. However, they did not benchmark against random forests, which tend to work better for imbalanced data, and their data was relatively low dimensional. Future work is needed to establish the utility of deep learning in cell subset identification, but the stunning improvements in image classification over the past 5 years [221] suggest transformative potential.

The sheer quantity of omic information that can be obtained from each cell, as well as the number of cells in each dataset, uniquely position single-cell data to benefit from deep learning. In the future, lineage tracing could be revolutionized by using autoencoders to reduce the feature space of transcriptomic or variant data followed by algorithms to learn optimal cell differentiation trajectories [222] or by feeding cell morphology and movement into neural networks [204]. Reinforcement learning algorithms [223] could be trained on the evolutionary dynamics of cancer cells or bacterial cells undergoing selection pressure and reveal whether patterns of adaptation are random or deterministic, allowing us to develop therapeutic strategies that forestall resistance. We are excited to see the creative applications of deep learning to single-cell biology that emerge over the next few years.

Metagenomics

Metagenomics, which refers to the study of genetic material -- 16S rRNA and/or whole-genome shotgun DNA -- from microbial communities, has revolutionized the study of micro-scale ecosystems within and around us. In recent years, machine learning has proved to be a powerful tool for

metagenomic analysis. 16S rRNA has long been used to deconvolve mixtures of microbial genomes, yet this ignores more than 99% of the genomic content. Subsequent tools aimed to classify 300-3000 base pair reads from complex mixtures of microbial genomes based on tetranucleotide frequencies, which differ across organisms [224], using supervised [225,226] or unsupervised methods [227]. Then, researchers began to use techniques that could estimate relative abundances from an entire sample faster than classifying individual reads [228–231]. There is also great interest in identifying and annotating sequence reads [232,233]. However, the focus on taxonomic and functional annotation is just the first step. Several groups have proposed methods to determine host or environment phenotypes from the organisms that are identified [234–237] or overall sequence composition [238]. Also, researchers have looked into how feature selection can improve classification [237,239], and techniques have been proposed that are classifier-independent [240,241].

How have neural networks been of use? Most neural networks are being used for phylogenetic classification or functional annotation from sequence data where there is ample data for training. Neural networks have been applied successfully to gene annotation (e.g. Orphelia [242] and FragGeneScan [243]). Representations (similar to Word2Vec [76] in natural language processing) for protein family classification have been introduced and classified with a skip-gram neural network [244]. Recurrent neural networks show good performance for homology and protein family identification [245,246].

One of the first techniques of *de novo* genome binning used self-organizing maps, a type of neural network [227]. Essinger et al. [247] used Adaptive Resonance Theory to cluster similar genomic fragments and showed that it had better performance than k-means. However, other methods based on interpolated Markov models [248] have performed better than these early genome binners. Neural networks can be slow and therefore have had limited use for reference-based taxonomic classification, with TAC-ELM [249] being the only neural network-based algorithm to taxonomically classify massive amounts of metagenomic data. An initial study successfully applied neural networks to taxonomic classification of 16S rRNA genes, with convolutional networks providing about 10% accuracy genus-level improvement over RNNs and random forests [250]. However, this study evaluated only 3000 sequences.

Neural network uses for classifying phenotype from microbial composition are just beginning. A standard multi-layer perceptron (MLP) was able to classify wound severity from microbial species present in the wound [251]. Recently, Ditzler et al. associated soil samples with pH level using MLPs, DBNs, and RNNs [252]. Besides classifying samples appropriately, internal phylogenetic tree nodes inferred by the networks represented features for low and high pH. Thus, hidden nodes might provide biological insight as well as new features for

future metagenomic sample comparison. Also, an initial study has shown promise of these networks for diagnosing disease [253].

Challenges remain in applying deep neural networks to metagenomics problems. They are not yet ideal for phenotype classification because most studies contain tens of samples and hundreds or thousands of features (species). Such underdetermined, or ill-conditioned, problems are still a challenge for deep neural networks that require many training examples. Also, due to convergence issues [254], taxonomic classification of reads from whole genome sequencing seems out of reach at the moment for deep neural networks. There are only thousands of full-sequenced genomes as compared to hundreds of thousands of 16S rRNA sequences available for training.

However, because RNNs have been applied to base calls for the Oxford Nanopore long-read sequencer with some success [255] (discussed further in the next section), one day the entire pipeline, from denoising of through functional classification, may be combined into one step by using powerful LSTMs [256]. For example, metagenomic assembly usually requires binning then assembly, but could deep neural nets accomplish both tasks in one network? We believe the greatest potential in deep learning is to learn the complete characteristics of a metagenomic sample in one complex network.

Sequencing and variant calling

While we have so far primarily discussed the role of deep learning in analyzing genomic data, deep learning can also substantially improve our ability to obtain the genomic data itself. We discuss two specific challenges: calling SNPs and indels (insertions and deletions) with high specificity and sensitivity and improving the accuracy of new types of data such as nanopore sequencing. These two tasks are critical for studying rare variation, allele-specific transcription and translation, and splice site mutations. In the clinical realm, sequencing of rare tumor clones and other genetic diseases will require accurate calling of SNPs and indels.

Current methods achieve relatively high (>99%) precision at 90% recall for SNPs and indel calls from Illumina short-read data [257], yet this leaves a large number of potentially clinically-important remaining false positives and false negatives. These methods have so far relied on experts to build probabilistic models that reliably separate signal from noise. However, this process is time consuming and fundamentally limited by how well we understand and can model the factors that contribute to noise. Recently, two groups have applied deep learning to construct data-driven unbiased noise models. One of these models, DeepVariant, leverages Inception, a neural network trained for image classification by Google Brain, by encoding reads around a candidate SNP as a

221x100 bitmap image, where each column is a nucleotide and each row is a read from the sample library [257]. The top 5 rows represent the reference, and the bottom 95 rows represent randomly sampled reads that overlap the candidate variant. Each RGBA (red/green/blue/alpha) image pixel encodes the base (A, C, T, G) as a different red value, quality score as a green value, strand as a blue value, and variation from the reference as the alpha value. The neural network outputs genotype probabilities for each candidate variant. They were able to achieve better performance than GATK, a leading genotype caller, even when GATK was given information about population variation for each candidate variant. Another method, still in its infancy, hand-developed 642 features for each candidate variant and fed these vectors into a fully connected deep neural network [258]. Unfortunately, this feature set required at least 15 iterations of software development to fine-tune, which suggests that these models may not generalize.

Going forward, variant calling will benefit more from optimizing neural network architectures than from developing features by hand. An interesting and informative next step would be to rigorously test if encoding raw sequence and quality data as an image, tensor, or some other mixed format produces the best variant calls. Because many of the latest neural network architectures (ResNet, Inception, Xception, and others) are already optimized for and pre-trained on generic, large-scale image datasets [259], encoding genomic data as images could prove to be a generally effective and efficient strategy.

In limited experiments, DeepVariant was robust to sequencing depth, read length, and even species [257]. However, a model built on Illumina data, for instance, may not be optimal for PacBio long-read data or MinION nanopore data, which have vastly different specificity and sensitivity profiles and signal-to-noise characteristics. Recently, Boza et al. used bidirectional recurrent neural networks to infer the *E. coli* sequence from MinION nanopore electric current data with higher per-base accuracy than the proprietary hidden Markov model-based algorithm Metrichor [255]. Unfortunately, training any neural network requires a large amount of data, which is often not available for new sequencing technologies. To circumvent this, one very preliminary study simulated mutations and spiked them into somatic and germline RNA-seq data, then trained and tested a neural network on simulated paired RNA-seq and exome sequencing data [260]. However, because this model was not subsequently tested on ground-truth datasets, it is unclear whether simulation can produce sufficiently realistic data to produce reliable models.

Method development for interpreting new types of sequencing data has historically taken two steps: first, easily implemented hard cutoffs that prioritize specificity over sensitivity, then expert development of probabilistic models with hand-developed inputs [260]. We anticipate that these steps will be replaced by

deep learning, which will infer features simply by its ability to optimize a complex model against data.

The impact of deep learning in treating disease and developing new treatments

Given the need to make better, faster interventions at the point of care -- incorporating the complex calculus of a patients symptoms, diagnostics, and life history -- there have been many attempts to apply deep learning to patient treatment. Success in this area could help to enable personalized healthcare or precision medicine [261,262]. Earlier, we reviewed approaches for patient categorization. Here, we examine the potential for better treatment, which broadly, may divided into methods for improved choices of interventions for patients and those for development of new interventions.

Clinical decision making

In 1996, Tu [263] compared the effectiveness of artificial neural networks and logistic regression, questioning whether these techniques would replace traditional statistical methods for predicting medical outcomes such as myocardial infarction [264] or mortality [265]. He posited that while neural networks have several advantages in representational power, the difficulties in interpretation may limit clinical applications, a limitation that still remains today. In addition, the challenges faced by physicians parallel those encountered by deep learning. For a given patient, the number of possible diseases is very large, with a long tail of rare diseases and patients are highly heterogeneous and may present with very different signs and symptoms for the same disease. Still, in 2006 Lisboa and Taktak [266] examined the use of artificial neural networks in medical journals, concluding that they improved healthcare relative to traditional screening methods in 21 of 27 studies.

While further progress has been made in using deep learning for clinical decision making, it is hindered by a challenge common to many deep learning applications: it is much easier to predict an outcome than to suggest an action to change the outcome. Several attempts [84,86] at recasting the clinical decision-making problem into a prediction problem (i.e. prediction of which treatment will most improve the patient's health) have accurately predicted survival patterns, but technical and medical challenges remain for clinical adoption (similar to those for categorization). In particular, remaining barriers include actionable interpretability of deep learning models, fitting deep models to limited and heterogeneous data, and integrating complex predictive models into a dynamic clinical environment.

A critical challenge in providing treatment recommendations is identifying a causal relationship for each recommendation. Causal inference is often framed in terms of counterfactual question [267]. Johansson et al. [268] use deep neural networks to create representation models for covariates that capture nonlinear effects and show significant performance improvements over existing models. In a less formal approach, Kale et al. [269] first create a deep neural network to model clinical time series and then analyze the relationship of the hidden features to the output using a causal approach.

A common challenge for deep learning is the interpretability of the models and their predictions. The task of clinical decision making is necessarily risk-averse, so model interpretability is key. Without clear reasoning, it is difficult to establish trust in a model. As described above, there has been some work to directly assign treatment plans without interpretability; however, the removal of human experts from the decision-making loop make the models difficult to integrate with clinical practice. To alleviate this challenge, several studies have attempted to create more interpretable deep models, either specifically for healthcare or as a general procedure for deep learning (see Discussion).

Predicting patient trajectories

A common application for deep learning in this domain is the temporal structure of healthcare records. Many studies [270–273] have used RNNs to categorize patients, but most stop short of suggesting clinical decisions. Nemati et al. [274] used deep reinforcement learning to optimize a heparin dosing policy for intensive care patients. However, because the ideal dosing policy is unknown, the model's predictions must be evaluated on counter-factual data. This represents a common challenge when bridging the gap between research and clinical practice. Because the ground-truth is unknown, researchers struggle to evaluate model predictions in the absence of interventional data, but clinical application is unlikely until the model has been shown to be effective. The impressive applications of deep reinforcement learning to other domains [223] have relied on knowledge of the underlying processes (e.g. the rules of the game). Some models have been developed for targeted medical problems [275], but a generalized engine is beyond current capabilities.

Clinical trials efficiency

A clinical deep learning task that has been more successful is the assignment of patients to clinical trials. Ithapu et al. [276] used a randomized denoising autoencoder to learn a multimodal imaging marker that predicts future cognitive and neural decline from positron emission tomography (PET), amyloid florbetapir PET, and structural magnetic resonance imaging. By accurately

predicting which cases will progress to dementia, they were able to efficiently assign patients to a clinical trial and reduced the required sample sizes by a factor of five. Similarly, Artemov et al. [277] applied deep learning to predict which clinical trials were likely to fail and which were likely to succeed. By predicting the side effects and pathway activations of each drug and translating these activations to a success probability, their deep learning-based approach was able to significantly outperform a random forest classifier trained on gene expression changes. These approaches suggest promising directions to improve the efficiency of clinical trials and accelerate drug development.

Drug repositioning

Drug repositioning (or repurposing) is an attractive option for delivering new drugs to the market because of the high costs and failure rates associated with more traditional drug discovery approaches [278,279]. A decade ago, the concept of the Connectivity Map [280] had a sizeable impact on the field. Reverse matching disease gene expression signatures with a large set of reference compound profiles allowed researchers to formulate repurposing hypotheses at scale using a simple non-parametric test. Since then, several advanced computational methods have been applied to formulate and validate drug repositioning hypotheses [281–283]. Using supervised learning and collaborative filtering to tackle this type of problem is proving successful, especially when coupling disease or compound omic data with topological information from protein-protein or protein-compound interaction networks [284–286].

For example, Menden et al. [287] used a shallow neural network to predict sensitivity of cancer cell lines to drug treatment using both cell line and drug features, opening the door to precision medicine and drug repositioning opportunities in cancer. More recently, Aliper et al. [35] used gene- and pathway-level drug perturbation transcriptional profiles from the Library of Network-Based Cellular Signatures [288] to train a fully connected deep neural network to predict drug therapeutic uses and indications. By using confusion matrices and leveraging misclassification, the authors formulated a number of interesting hypotheses, including repurposing cardiovascular drugs such as otenzepad and pinacidil for neurological disorders.

Drug repositioning can also be approached by attempting to predict novel drug-target interactions and then repurposing the drug for the associated indication [289,290]. Wang et al. [291] devised a pairwise input neural network with two hidden layers that takes two inputs, a drug and a target binding site, and predicts whether they interact. Wang et al. [36] trained individual RBMs for each target in a drug-target interaction network and used these models to predict novel interactions pointing to new indications for existing drugs. Wen et

al. [37] extended this concept to deep learning by creating a DBN called DeepDTIs, which is able to predict interactions on the basis of chemical structure and protein sequence features.

Drug repositioning appears to be an obvious candidate for deep learning both because of the large amount of high-dimensional data available and the complexity of the question being asked. However, what is perhaps the most promising piece of work in this space [35] is more of a proof of concept than a real-world hypothesis-generation tool; notably, deep learning was used to predict drug indications but not for the actual repositioning. At present, some of the most popular state-of-the-art methods for signature-based drug repurposing [292] do not use predictive modeling. A mature and production-ready framework for drug repositioning via deep learning is currently missing.

Drug development

Ligand-based prediction of bioactivity

In the biomedical domain, high-throughput chemical screening aims to improve therapeutic options over a long term horizon [20]. The objective is to discover which small molecules (also referred to as chemical compounds or ligands) that specifically affect the activity of a target, such as a kinase, protein-protein interaction, or broader cellular phenotype. This screening is often one of the first steps in a long drug discovery pipeline, where novel molecules are pursued for their ability to inhibit or enhance disease-relevant biological mechanisms [293]. Initial hits are confirmed to eliminate false positives and proceed to the lead generation stage [294], where they are evaluated for absorption, distribution, metabolism, excretion, and toxicity (ADMET) and other properties. It is desirable to advance multiple lead series, clusters of structurally-similar active chemicals, for further optimization by medicinal chemists to protect against unexpected failures in the later stages of drug discovery [293].

Computational work in this domain aims to identify sufficient candidate active compounds without exhaustively screening libraries of hundreds of thousands or millions of chemicals. Predicting chemical activity computationally is known as virtual screening. This task has been treated variously as a classification, regression, or ranking problem. In reality, it does not fit neatly into any of those categories. An ideal algorithm will rank a sufficient number of active compounds before the inactives, but the rankings of actives relative to other actives and inactives are less important [295]. Computational modeling also has the potential to predict ADMET traits for lead generation [296] and how drugs are metabolized [297].

Ligand-based approaches train on chemicals' features without modeling target features (e.g. protein structure). Chemical features may be represented as a list of molecular descriptors such as molecular weight, atom counts, functional groups, charge representations, summaries of atom-atom relationships in the molecular graph, and more sophisticated derived properties [298]. Alternatively, chemicals can be characterized with the fingerprint bit vectors, textual strings, or novel learned representations described below. Neural networks have a long history in this domain [18,21], and the 2012 Merck Molecular Activity Challenge on Kaggle generated substantial excitement about the potential for high-parameter deep learning approaches. The winning submission was an ensemble that included a multi-task multi-layer perceptron network [299]. The sponsors noted drastic improvements over a random forest baseline, remarking "we have seldom seen any method in the past 10 years that could consistently outperform [random forest] by such a margin" [300]. Subsequent work (reviewed in more detail by Goh et al. [19]) explored the effects of jointly modeling far more targets than the Merck challenge [301,302], with Ramsundar et al. [302] showing that the benefits of multi-task networks had not yet saturated even with 259 targets. Although DeepTox [303], a deep learning approach, won another competition, the Toxicology in the 21st Century (Tox21) Data Challenge, it did not dominate alternative methods as thoroughly as in other domains. DeepTox was the top performer on 9 of 15 targets and highly competitive with the top performer on the others. However, for many targets there was little separation between the top two or three methods.

The nuanced Tox21 performance may be more reflective of the practical challenges encountered in ligand-based chemical screening than the extreme enthusiasm generated by the Merck competition. A study of 22 ADMET tasks demonstrated that there are limitations to multi-task transfer learning that are in part a consequence of the degree to which tasks are related [296]. Some of the ADMET datasets showed superior performance in multi-task models with only 22 ADMET tasks compared to multi-task models with over 500 less-similar tasks. In addition, the training datasets encountered in practical applications may be tiny relative to what is available in public datasets and organized competitions. A study of BACE-1 inhibitors included only 1547 compounds [304]. Machine learning models were able to train on this limited dataset, but overfitting was a challenge and the differences between random forests and a deep neural network were negligible, especially in the classification setting. Overfitting is still a problem in larger chemical screening datasets with tens or hundreds of thousands of compounds because the number of active compounds can be very small, on the order of 0.1% of all tested chemicals for a typical target [305]. This is consistent with the strong performance of low-parameter neural networks that emphasize compound-compound similarity, such as influence-relevance voter [295,306], instead of predicting compound activity directly from chemical features.

Much of the recent excitement in this domain has come from what could be considered a creative experimentation phase, in which deep learning has offered novel possibilities for feature representation and modeling of chemical compounds. A molecular graph, where atoms are labeled nodes and bonds are labeled edges, is a natural way to represent a chemical structure. Traditional machine learning approaches relied on preprocessing the graph into a feature vector, such as a fixed-width bit vector fingerprint [307]. The same fingerprints have been used by some drug-target interaction methods discussed above [37]. An overly simplistic but approximately correct view of chemical fingerprints is that each bit represents the presence or absence of a particular chemical substructure in the molecular graph. Modern neural networks can operate directly on the molecular graph as input. Duvenaud et al. [308] generalized standard circular fingerprints by substituting discrete operations in the fingerprinting algorithm with operations in a neural network, producing a real-valued feature vector instead of a bit vector. Other approaches offer trainable networks that can learn chemical feature representations that are optimized for a particular prediction task. Lusci et al. [309] applied recursive neural networks for directed acyclic graphs to undirected molecular graphs by creating an ensemble of directed graphs in which one atom is selected as the root node. Graph convolutions on undirected molecular graphs have eliminated the need to enumerate artificially directed graphs, learning feature vectors for atoms that are a function of the properties of neighboring atoms and local regions on the molecular graph [310,311].

Advances in chemical representation learning have also enabled new strategies for learning chemical-chemical similarity functions. Altae-Tran et al. developed a one-shot learning network [311] to address the reality that most practical chemical screening studies are unable to provide the thousands or millions of training compounds that are needed to train larger multi-task networks. Using graph convolutions to featurize chemicals, the network learns an embedding from compounds into a continuous feature space such that compounds with similar activities in a set of training tasks have similar embeddings. The approach is evaluated in an extremely challenging setting. The embedding is learned from a subset of prediction tasks (e.g. activity assays for individual proteins), and only one to ten labeled examples are provided as training data on a new task. On Tox21 targets, even when trained with *one* task-specific active compound and *one* inactive compound, the model is able to generalize reasonably well because it has learned an informative embedding function from the related tasks. Random forests, which cannot take advantage of the related training tasks, trained in the same setting are only slightly better than a random classifier. Despite the success on Tox21, performance on MUV datasets, which contains assays designed to be challenging for chemical informatics algorithms, is considerably worse. The authors also demonstrate the limitations of transfer learning as embeddings

learned from the Tox21 assays have little utility for a drug adverse reaction dataset.

These novel, learned chemical feature representations may prove to be essential for accurately predicting why some compounds with similar structures yield similar target effects and others produce drastically different results. Currently, these methods are enticing but do not necessarily outperform classic approaches by a large margin. The neural fingerprints [308] were narrowly beaten by regression using traditional circular fingerprints on a drug efficacy prediction task but were superior for predicting solubility or photovoltaic efficiency. In the original study, graph convolutions [310] performed comparably to a multi-task network using standard fingerprints and slightly better than the neural fingerprints [308] on the drug efficacy task but were slightly worse than the influence-relevance voter method on an HIV dataset. [295]. Broader recent benchmarking has shown that relative merits of these methods depends on the dataset and cross validation strategy [312], though evaluation often uses auROC (area under the receiver operating characteristic curve), which has limited utility due to the large active/inactive class imbalance (see Discussion).

We remain optimistic for the potential of deep learning and specifically representation learning in drug discovery and propose that rigorous benchmarking on broad and diverse prediction tasks will be as important as novel neural network architectures to advance the state of the art and convincingly demonstrate superiority over traditional cheminformatics techniques. Fortunately, there has recently been much progress in this direction. The DeepChem software [311,313] and MoleculeNet benchmarking suite [312] built upon it contain chemical bioactivity and toxicity prediction datasets, multiple compound featurization approaches including graph convolutions, and various machine learning algorithms ranging from standard baselines like logistic regression and random forests to recent neural network architectures. Independent research groups have already contributed additional datasets and prediction algorithms to DeepChem. Adoption of common benchmarking evaluation metrics, datasets, and baseline algorithms has the potential to establish the practical utility of deep learning in chemical bioactivity prediction and lower the barrier to entry for machine learning researchers without biochemistry expertise.

One open question in ligand-based screening pertains to the benefits and limitations of transfer learning. Multi-task neural networks have shown the advantages of jointly modeling many targets [301,302]. Other studies have shown the limitations of transfer learning when the prediction tasks are insufficiently related [296,311]. This has important implications for representation learning. The typical approach to improve deep learning models by expanding the dataset size may not be applicable if only "related" tasks are

beneficial, especially because task-task relatedness is ill-defined. The massive chemical state space will also influence the development of unsupervised representation learning methods [314]. Future work will establish whether it is better to train on massive collections of diverse compounds, drug-like small molecules, or specialized subsets.

Structure-based prediction of bioactivity

When protein structure is available, virtual screening has traditionally relied on docking programs to predict how a compound best fits in the target's binding site and score the predicted ligand-target complex [315]. Recently, deep learning approaches have been developed to model protein structure, which is expected to improve upon the simpler drug-target interaction algorithms described above that represent proteins with feature vectors derived from amino acid sequences [37,291].

Structure-based deep learning methods differ in whether they use experimentally-derived or predicted ligand-target complexes and how they represent the 3D structure. The Atomic CNN [316] takes 3D crystal structures from PDBBind [317] as input, ensuring it uses a reliable ligand-target complex. AtomNet [34] samples multiple ligand poses within the target binding site, and DeepVS [318] and Ragoza et al. [319] use a docking program to generate protein-compound complexes. If they are sufficiently accurate, these latter approaches would have wider applicability to a much larger set of compounds and proteins. However, incorrect ligand poses will be misleading during training, and the predictive performance is sensitive to the docking quality [318].

There are two major options for representing a protein-compound complex. A 3D grid can featurize the input complex [34,319]. Each entry in the grid tracks the types of protein and ligand atoms in that region of the 3D space or descriptors derived from those atoms. Both DeepVS [318] and atomic convolutions [316] offer greater flexibility in their convolutions by eschewing the 3D grid. Instead, they each implement techniques for executing convolutions over atoms' neighboring atoms in the 3D space. Gomes et al. demonstrate that currently random forest on a 1D feature vector that describes the 3D ligand-target structure generally outperforms neural networks on the same feature vector as well as atomic convolutions and ligand-based neural networks when predicting the continuous-valued inhibition constant on the PDBBind refined dataset [316]. However, in the long term, atomic convolutions may ultimately overtake grid-based methods, as they provide greater freedom to model atom-atom interactions and the forces that govern binding affinity.

De novo drug design

Whereas the goal of virtual screening is to find active molecules by predicting the biochemical activity of hundreds of thousands to millions of chemicals using existing (virtual) chemical libraries, analogous to robotic high-throughput "wet lab" screening, *de novo* drug design aims to directly *generate* active compounds [320,321].

De novo drug design attempts to model the typical design-synthesize-test cycle of drug discovery [320]. It explores the much larger space of an estimated 10^{60} synthesizable organic molecules with drug-like properties without explicit enumeration [305]. To test or score structures, algorithms like those discussed earlier are used. To "design" and "synthesize", traditional *de novo* design software relied on classical optimizers such as genetic algorithms. Unfortunately, this often leads to overfit, "weird" molecules, which are difficult to synthesize in the lab. Current programs have settled on rule-based virtual chemical reactions to generate molecular structures [321]. Deep learning models that generate realistic, synthesizable molecules have been proposed as an alternative. In contrast to the classical, symbolic approaches, generative models learned from data would not depend on laboriously encoded expert knowledge. The challenge of generating molecules has parallels to the generation of syntactically and semantically correct text [322].

As deep learning models that directly output (molecular) graphs remain under-explored, generative neural networks for drug design typically represent chemicals with the simplified molecular-input line-entry system (SMILES), a standard string-based representation with characters that represent atoms, bonds, and rings [323]. This allows treating molecules as sequences and leveraging recent progress in recurrent neural networks. Gómez-Bombarelli et al. designed a SMILES-to-SMILES autoencoder to learn a continuous latent feature space for chemicals [314]. In this learned continuous space it was possible to interpolate between continuous representations of chemicals in a manner that is not possible with discrete (e.g. bit vector or string) features or in symbolic, molecular graph space. Even more interesting is the prospect of performing gradient-based or Bayesian optimization of molecules within this latent space. The strategy of constructing simple, continuous features before applying supervised learning techniques is reminiscent of autoencoders trained on high-dimensional EHR data [81]. A drawback of the SMILES-to-SMILES autoencoder is that not all SMILES strings produced by the autoencoder's decoder correspond to valid chemical structures. Recently, the Grammar Variational Autoencoder, which takes the SMILES grammar into account and is guaranteed to produce syntactically valid SMILES, has been proposed to alleviate this issue [324].

Another approach to *de novo* design is to train character-based RNNs on large collections of molecules, for example, ChEMBL [325], to first obtain a generic

generative model for drug-like compounds [323]. These generative models successfully learn the grammar of compound representations, with 94% [326] or nearly 98% [323] of generated SMILES corresponding to valid molecular structures. The initial RNN is then fine-tuned to generate molecules that are likely to be active against a specific target by either continuing training on a small set of positive examples [323] or adopting reinforcement learning strategies [326,327]. Both the fine-tuning and reinforcement learning approaches can rediscover known, held-out active molecules. The great flexibility of neural networks, and progress in generative models offers many opportunities for deep architectures in *de novo* design (e.g. the adaptation of Generative Adversarial Networks (GANs) for molecules).

Discussion

Despite the disparate types of data and scientific goals in the learning tasks covered above, several challenges are broadly important for deep learning in the biomedical domain. Here we examine these factors that may impede further progress, ask what steps have already been taken to overcome them, and suggest future research directions.

Evaluation

There are unique challenges to evaluating deep learning predictions in the biomedical domain. We focus on TF binding prediction as a representative task to illustrate some of these issues. The human genome has 3 billion base pairs, and only a small fraction of them are implicated in specific biochemical activities. As a result, classification of genomic regions based on their biochemical activity results in highly imbalanced classification. Class imbalance also arises in other problems we review, such as virtual screening for drug discovery. What are appropriate evaluation metrics that account for the label imbalance? The classification labels are formulated based on continuous value experimental signals. Practitioners must determine an appropriate procedure for formulating binary classification labels based on these signals. In addition, the experimental signals are only partially reproducible across experimental replicates. An appropriate upper bound for classification performance must account for the experimental reproducibility.

Evaluation metrics for imbalanced classification

Less than 1% of the genome can be confidently labeled as bound for most transcription factors. Therefore, it is important to evaluate the genome-wide recall and false discovery rate (FDR) of classification models of biochemical activities. Targeted validation experiments of specific biochemical activities

usually necessitate an FDR of 5-25%. When predicted biochemical activities are used as features in other models, such as gene expression models, a low FDR may not be as critical if the downstream models can satisfy their evaluation criteria. An FDR of 50% in this context may suffice.

What is the correspondence between these metrics and commonly used classification metrics such as auPRC (area under the precision-recall curve) and auROC? auPRC evaluates the average precision, or equivalently, the average FDR across all recall thresholds. This metric provides an overall estimate of performance across all possible use cases, which can be misleading for targeted validation experiments. For example, classification of TF binding sites can exhibit a recall of 0% at 10% FDR and auPRC greater than 0.6. In this case, the auPRC may be competitive, but the predictions are ill-suited for targeted validation that can only examine a few of the highest-confidence predictions. Likewise, auROC evaluates the average recall across all false positive rate (FPR) thresholds, which is often a highly misleading metric in class-imbalanced domains [71,328]. For example, consider a classification model with recall of 0% at FDR less than 25% and 100% recall at FDR greater than 25%. In the context of TF binding predictions where only 1% of genomic regions are bound by the TF, this is equivalent to a recall of 100% for FPR greater than 0.33%. In other words, the auROC would be 0.9967, but the classifier would be useless for targeted validation. It is not unusual to obtain a chromosome-wide auROC greater than 0.99 for TF binding predictions but a recall of 0% at 10% FDR.

Formulation of classification labels

Genome-wide continuous signals are commonly formulated into classification labels through signal peak detection. ChIP-seq peaks are used to identify locations of TF binding and histone modifications. Such procedures rely on thresholding criteria to define what constitutes a peak in the signal. This inevitably results in a set of signal peaks that are close to the threshold, not sufficient to constitute a positive label but too similar to positively labeled examples to constitute a negative label. To avoid an arbitrary label for these example they may be labeled as "ambiguous". Ambiguously labeled examples can then be ignored during model training and evaluation of recall and FDR. The correlation between model predictions on these examples and their signal values can be used to evaluate if the model correctly ranks these examples between positive and negative examples.

Formulation of a performance upper bound

Genome-wide signals across experiments can lead to different sets of positive examples. When experimental replicates do not completely agree, perfect

recall at a low FDR is not possible. The upper bound on the recall is the fraction of positive examples that are in agreement across experiments. This fraction will vary depending on the available experimental data. Reproducibility for experimental replicates from the same lab is typically higher than experimental replicates across multiple labs. One way to handle the range of reproducibility is the use of multiple reproducibility criteria such as reproducibility across technical replicates, biological replicates from the same lab, and biological replicates from multiple labs.

Interpretation

As deep learning models achieve state-of-the-art performance in a variety of domains, there is a growing need to make the models more interpretable. Interpretability matters for two main reasons. First, a model that achieves breakthrough performance may have identified patterns in the data that practitioners in the field would like to understand. However, this would not be possible if the model is a black box. Second, interpretability is important for trust. If a model is making medical diagnoses, it is important to ensure the model is making decisions for reliable reasons and is not focusing on an artifact of the data. A motivating example of this can be found in Ba and Caruana [329], where a model trained to predict the likelihood of death from pneumonia assigned lower risk to patients with asthma, but only because such patients were treated as higher priority by the hospital. In the context of deep learning, understanding the basis of a model's output is particularly important as deep learning models are unusually susceptible to adversarial examples [330] and can output confidence scores over 99.99% for samples that resemble pure noise.

As the concept of interpretability is quite broad, many methods described as improving the interpretability of deep learning models take disparate and often complementary approaches. Some key themes are discussed below.

Assigning example-specific importance scores

Several approaches ascribe importance on an example-specific basis to the parts of the input that are responsible for a particular output. These can be broadly divided into perturbation-based approaches and backpropagation-based approaches.

Pertubration-based approaches change parts of the input and observe the impact on the output of the network. Alipanahi et al. [164] and Zhou & Troyanskaya [168] scored genomic sequences by introducing virtual mutations at individual positions in the sequence and quantifying the change in the output. Umarov et al. [178] used a similar strategy, but with sliding windows

where the sequence within each sliding window was substituted with a random sequence. Kelley et al. [177] inserted known protein-binding motifs into the centers of sequences and assessed the change in predicted accessibility. Ribeiro et al. [331] introduced LIME, which constructs a linear model to locally approximate the output of the network on perturbed versions of the input and assigns importance scores accordingly. For analyzing images, Zeiler and Fergus [332] applied constant-value masks to different input patches. More recently, marginalizing over the plausible values of an input has been suggested as a way to more accurately estimate contributions [333].

A common drawback to perturbation-based approaches is computational efficiency: each perturbed version of an input requires a separate forward propagation through the network to compute the output. As noted by Shrikumar et al. [169], such methods may also underestimate the impact of features that have saturated their contribution to the output, as can happen when multiple redundant features are present. To reduce the computational overhead of perturbation-based approaches, Fong and Vedaldi [334] solve an optimization problem using gradient descent to discover a minimal subset of inputs to perturb in order to decrease the predicted probability of a selected class. Their method converges in many fewer iterations but requires the perturbation to have a differentiable form.

Backpropagation-based methods, in which the signal from a target output neuron is propagated backwards to the input layer, are another way to interpret deep networks that sidestep inefficiencies of the perturbastion-basd methods. A classic example of this is calculating the gradients of the output with respect to the input [335] to compute a "saliency map". Bach et al.[336] proposed a strategy called Layerwise Relevance Propagation, which was shown to be equivalent to the element-wise product of the gradient and input [169,337]. Networks with Rectified Linear Units (ReLUs) create nonlinearities that must be addressed. Several variants exist for handling this [332,338]. Backpropagation-based methods are a highly active area of research. Researchers are still actively identifying weaknesses [339], and new methods are being developed to address them [169,340,341]. Lundberg and Lee [342] noted that several importance scoring methods including integrated gradients and LIME could all be considered approximations to Shapely values [343], which have a long history in game theory for assigning contributions to players in cooperative games.

Matching or exaggerating the hidden representation

Another approach to understanding the network's predictions is to find artificial inputs that produce similar hidden representations to a chosen example. This can elucidate the features that the network uses for prediction and drop the

features that the network is insensitive to. In the context of natural images, Mahendran and Vedaldi [344] introduced the "inversion" visualization, which uses gradient descent and backpropagation to reconstruct the input from its hidden representation. The method required placing a prior on the input to favor results that resemble natural images. For genomic sequence, Finnegan and Song [345] used a Markov chain Monte Carlo algorithm to find the maximum-entropy distribution of inputs that produced a similar hidden representation to the chosen input.

A related idea is "caricaturization", where an initial image is altered to exaggerate patterns that the network searches for [346]. This is done by maximizing the response of neurons that are active in the network, subject to some regularizing constraints. Mordvintsev et al. [347] leveraged caricaturization to generate aesthetically pleasing images using neural networks.

Activation maximization

Activation maximization can reveal patterns detected by an individual neuron in the network by generating images which maximally activate that neuron, subject to some regularizing constraints. This technique was first introduced in Ehran et al. [348] and applied in subsequent work [335,346,347,349]. Lanchantin et al. [165] applied activation maximization to genomic sequence data. One drawback of this approach is that neural networks often learn highly distributed representations where several neurons cooperatively describe a pattern of interest. Thus, visualizing patterns learned by individual neurons may not always be informative.

RNN-specific approaches

Several interpretation methods are specifically tailored to recurrent neural network architectures. The most common form of interpretability provided by RNNs is through attention mechanisms, which have been used in diverse problems such as image captioning and machine translation to select portions of the input to focus on generating a particular output [350,351]. Deming et al. [352] applied the attention mechanism to models trained on genomic sequence. Attention mechanisms provide insight into the model's decision-making process by revealing which portions of the input are used by different outputs. In the clinical domain, Choi et al. [353] leveraged attention mechanisms to highlight which aspects of a patient's medical history were most relevant for making diagnoses. Choi et al. [354] later extended this work to take into account the structure of disease ontologies and found that the concepts represented by the model aligned with medical knowledge. Note that interpretation strategies that rely on an attention mechanism do not provide

insight into the logic used by the attention layer.

Visualizing the activation patterns of the hidden state of a recurrent neural network can also be instructive. Early work by Ghosh and Karamcheti [355] used cluster analysis to study hidden states of comparatively small networks trained to recognize strings from a finite state machine. More recently, Karpathy et al. [356] showed the existence of individual cells in LSTMs that kept track of quotes and brackets in character-level language models. To facilitate such analyses, LSTMVis [357] allows interactive exploration of the hidden state of LSTMs on different inputs.

Another strategy, adopted by Lanchatin et al.[165] looks at how the output of a recurrent neural network changes as longer and longer subsequences are supplied as input to the network, where the subsequences begin with just the first position and end with the entire sequence. In a binary classification task, this can identify those positions which are responsible for flipping the output of the network from negative to positive. If the RNN is bidirectional, the same process can be repeated on the reverse sequence. As noted by the authors, this approach was less effective at identifying motifs compared to the gradient-based backpropagation approach of Simonyan et al. [335], illustrating the need for more sophisticated strategies to assign importance scores in recurrent neural networks.

Murdoch and Szlam [358] showed that the output of an LSTM can be decomposed into a product of factors, where each factor can be interpreted as the contribution at a particular timestep. The contribution scores were then used to identify key phrases from a model trained for sentiment analysis and obtained superior results compared to scores derived via a gradient-based approach.

Miscellaneous approaches

Toward quantifying the uncertainty of predictions, there has been a renewed interest in confidence intervals for deep neural networks. Early work from Chryssolouris et al. [359] provided confidence intervals under the assumption of normally-distributed error. A more recent technique known as test-time dropout [360] can also be used to obtain a probabilistic interpretation of a network's outputs.

It can often be informative to understand how the training data affects model learning. Toward this end, Koh and Liang [361] used influence functions, a technique from robust statistics, to trace a model's predictions back through the learning algorithm to identify the datapoints in the training set that had the most impact on a given prediction. A more free-form approach to interpretability is to visualize the activation patterns of the network on individual inputs and on

subsets of the data. ActiVis and CNNvis [362,363] are two frameworks that enable interactive visualization and exploration of large-scale deep learning models. An orthogonal strategy is to use a knowledge distillation approach to replace a deep learning model with a more interpretable model that achieves comparable performance. Towards this end, Che et al. [364] used gradient boosted trees to learn interpretable healthcare features from trained deep models.

Finally, it is sometimes possible to train the model to provide justifications for its predictions. Lei et al. [365] used a generator to identify "rationales", which are short and coherent pieces of the input text that produce similar results to the whole input when passed through an encoder. The authors applied their approach to a sentiment analysis task and obtained substantially superior results compared to an attention-based method.

Future outlook

While deep learning certainly lags behind most Bayesian models in terms of interpretability, one can safely argue that the interpretability of deep learning is comparable to or exceeds that of many other widely-used machine learning methods such as random forests or SVMs. While it is possible to obtain importance scores for different inputs in a random forest, the same is true for deep learning. Similarly, SVMs trained with a nonlinear kernel are not easily interpretable because the use of the kernel means that one does not obtain an explicit weight matrix. Finally, it is worth noting that some simple machine learning methods are less interpretable in practice than one might expect. A linear model trained on heavily engineered features might be difficult to interpret as the input features themselves are difficult to interpret. Similarly, a decision tree with many nodes and branches may also be difficult for a human to make sense of.

There are several directions that might benefit the development of interpretability techniques. The first is the introduction of gold standard benchmarks that different interpretability approaches could be compared against, similar in spirit to how datasets like ImageNet and CIFAR spurred the development of deep learning for computer vision. It would also be helpful if the community placed more emphasis on domains outside of computer vision. Computer vision is often used as the example application of interpretability methods, but it is arguably not the domain with the most pressing need. Finally, closer integration of interpretability approaches with popular deep learning frameworks would make it easier for practitioners to apply and experiment with different approaches to understanding their deep learning models.

Data limitations

A lack of large-scale, high-quality, correctly labeled training data has impacted deep learning in nearly all applications we have discussed, from healthcare to genomics to drug discovery. The challenges of training complex, high-parameter neural networks from few examples are obvious, but uncertainty in the labels of those examples can be just as problematic. In genomics labeled data may be derived from an experimental assay with known and unknown technical artifacts, biases, and error profiles. It is possible to weight training examples or construct Bayesian models to account for uncertainty or non-independence in the data, as described in the TF binding example above. As another example, Park et al. [366] estimated shared non-biological signal between datasets to correct for non-independence related to assay platform or other factors in a Bayesian integration of many datasets. However, such techniques are rarely placed front and center in any description of methods and may be easily overlooked.

For some types of data, especially images, it is straightforward to augment training datasets by splitting a single labeled example into multiple examples. For example, an image can easily be rotated, flipped, or translated and retain its label [59]. 3D MRI and 4D fMRI (with time as a dimension) data can be decomposed into sets of 2D images [367]. This can greatly expand the number of training examples but artificially treats such derived images as independent instances and sacrifices the structure inherent in the data. CellCnn trains a model to recognize rare cell populations in single-cell data by creating training instances that consist of subsets of cells that are randomly sampled with replacement from the full dataset [220].

Simulated or semi-synthetic training data has been employed in multiple biomedical domains, though many of these ideas are not specific to deep learning. Training and evaluating on simulated data, for instance, generating synthetic TF binding sites with position weight matrices [167] or RNA-seq reads for predicting mRNA transcript boundaries [368], is a standard practice in bioinformatics. This strategy can help benchmark algorithms when the available gold standard dataset is imperfect, but it should be paired with an evaluation on real data, as in the prior examples [167,368]. In rare cases, models trained on simulated data have been successfully applied directly to real data [368].

Data can be simulated to create negative examples when only positive training instances are available. DANN [33] adopts this approach to predict the pathogenicity of genetic variants using semi-synthetic training data from Combined Annotation-Dependent Depletion [369]. Though our emphasis here is on the training strategy, it should be noted that logistic regression outperformed DANN when distinguishing known pathogenic mutations from likely benign variants in real data. Similarly, a somatic mutation caller has been

trained by injecting mutations into real sequencing datasets [260]. This method detected mutations in other semi-synthetic datasets but was not validated on real data.

In settings where the experimental observations are biased toward positive instances, such as MHC protein and peptide ligand binding affinity [370], or the negative instances vastly outnumber the positives, such as high-throughput chemical screening [306], training datasets have been augmented by adding additional instances and assuming they are negative. There is some evidence that this can improve performance [306], but in other cases it was only beneficial when the real training datasets were extremely small [370]. Overall, training with simulated and semi-simulated data is a valuable idea for overcoming limited sample sizes but one that requires more rigorous evaluation on real ground-truth datasets before we can recommend it for widespread use. There is a risk that a model will easily discriminate synthetic examples but not generalize to real data.

Multimodal, multi-task, and transfer learning, discussed in detail below, can also combat data limitations to some degree. There are also emerging network architectures, such as Diet Networks for high-dimensional SNP data [371]. These use multiple networks to drastically reduce the number of free parameters by first flipping the problem and training a network to predict parameters (weights) for each input (SNP) to learn a feature embedding. This embedding (e.g. from principal component analysis, per class histograms, or a Word2vec [76] generalization) can be learned directly from input data or take advantage of other datasets or domain knowledge. Additionally, in this task the features are the examples, an important advantage when it is typical to have 500 thousand or more SNPs and only a few thousand patients. Finally, this embedding is of a much lower dimension, allowing for a large reduction in the number of free parameters. In the example given, the number of free parameters was reduced from 30 million to 50 thousand, a factor of 600.

Hardware limitations and scaling

Efficiently scaling deep learning is challenging, and there is a high computational cost (e.g. time, memory, and energy) associated with training neural networks and using them to make predictions. This is one of the reasons why neural networks have only recently found widespread use [372].

Many have sought to curb these costs, with methods ranging from the very applied (e.g. reduced numerical precision [373–376]) to the exotic and theoretic (e.g. training small networks to mimic large networks and ensembles [329,377]). The largest gains in efficiency have come from computation with graphics processing units (GPUs) [372,378–382], which excel at the matrix and

vector operations so central to deep learning. The massively parallel nature of GPUs allows additional optimizations, such as accelerated mini-batch gradient descent [379,380,383,384]. However, GPUs also have limited memory, making networks of useful size and complexity difficult to implement on a single GPU or machine [68,378]. This restriction has sometimes forced computational biologists to use workarounds or limit the size of an analysis. Chen et al. [146] inferred the expression level of all genes with a single neural network, but due to memory restrictions they randomly partitioned genes into two separately analyzed halves. In other cases, researchers limited the size of their neural network [27] or the total number of training instances[314]. Some have also chosen to use standard central processing unit (CPU) implementations rather than sacrifice network size or performance [385].

While steady improvements in GPU hardware may alleviate this issue, it is unclear whether advances will occur quickly enough to keep pace with the growing biological datasets and increasingly complex neural networks. Much has been done to minimize the memory requirements of neural networks [329,373–376,386,387], but there is also growing interest in specialized hardware, such as field-programmable gate arrays (FPGAs) [382,388] and application-specific integrated circuits (ASICs) [389]. Less software is available for such highly specialized hardware [388]. But specialized hardware promises improvements in deep learning at reduced time, energy, and memory [382]. Specialized hardware may be a difficult investment for those not solely interested in deep learning, but for those with a deep learning focus these solutions may become popular.

Distributed computing is a general solution to intense computational requirements and has enabled many large-scale deep learning efforts. Some types of distributed computation [390,391] are not suitable for deep learning [392], but much progress has been made. There now exist a number of algorithms [375,392,393], tools [394–396], and high-level libraries [397,398] for deep learning in a distributed environment, and it is possible to train very complex networks with limited infrastructure [399]. Besides handling very large networks, distributed or parallelized approaches offer other advantages, such as improved ensembling [400] or accelerated hyperparameter optimization [401,402].

Cloud computing, which has already seen wide adoption in genomics[403], could facilitate easier sharing of the large datasets common to biology [404,405], and may be key to scaling deep learning. Cloud computing affords researchers flexibility, and enables the use of specialized hardware (e.g. FPGAs, ASICs, GPUs) without major investment. As such, it could be easier to address the different challenges associated with the multitudinous layers and architectures available [406]. Though many are reluctant to store sensitive data

(e.g. patient electronic health records) in the cloud, secure, regulation-compliant cloud services do exist [407].

Data, code, and model sharing

A robust culture of data, code, and model sharing would speed advances in this domain. The cultural barriers to data sharing in particular are perhaps best captured by the use of the term "research parasite" to describe scientists who use data from other researchers [408]. A field that honors only discoveries and not the hard work of generating useful data will have difficulty encouraging scientists to share their hard-won data. Unfortunately, it's precisely those data that would help to power deep learning in the domain. Efforts are underway to recognize those who promote an ecosystem of rigorous sharing and analysis [409].

The sharing of high-quality, labeled datasets will be especially valuable. In addition, researchers who invest time to preprocess datasets to be suitable for deep learning can make the preprocessing code (e.g. Basset [177] and variationanalysis [258]) and cleaned data (e.g. MoleculeNet [312]) publicly available to catalyze further research. However, there are complex privacy and legal issues involved in sharing patient data that cannot be ignored. In some domains high-quality training data has been generated privately, i.e. high-throughput chemical screening data at pharmaceutical companies. One perspective is that there is little expectation or incentive for this private data to be shared. However, data are not inherently valuable. Instead, the insights that we glean from them are where the value lies. Private companies may establish a competitive advantage by releasing data sufficient for improved methods to be developed.

Code sharing and open source licensing is essential for continued progress in this domain. We strongly advocate following established best practices for sharing source code, archiving code in repositories that generate digital object identifiers, and open licensing [410] regardless of the minimal requirements, or lack thereof, set by journals, conferences, or preprint servers. In addition, it is important for authors to share not only code for their core models but also scripts and code used for data cleaning (see above) and hyperparameter optimization. These improve reproducibility and serve as documentation of the detailed decisions that impact model performance but may not be exhaustively captured in a manuscript's methods text.

Because many deep learning models are often built using one of several popular software frameworks, it is also possible to directly share trained predictive models. The availability of pre-trained models can accelerate research, with image classifiers as an apt example. A pre-trained neural

network can be quickly fine-tuned on new data and used in transfer learning, as discussed below. Taking this idea to the extreme, genomic data has been artificially encoded as images in order to benefit from pre-trained image classifiers [257]. "Model zoos" -- collections of pre-trained models -- are not yet common in biomedical domains but have started to appear in genomics applications [216,411]. Sharing models for patient data requires great care because deep learning models can be attacked to identify examples used in training. We discuss this issue as well as recent techniques to mitigate these concerns in the patient categorization section.

DeepChem [311–313] and DragoNN [411] exemplify the benefits of sharing pre-trained models and code under an open source license. DeepChem, which targets drug discovery and quantum chemistry, has actively encouraged and received community contributions of learning algorithms and benchmarking datasets. As a consequence, it now supports a large suite of machine learning approaches, both deep learning and competing strategies, that can be run on diverse test cases. This realistic, continual evaluation will play a critical role in assessing which techniques are most promising for chemical screening and drug discovery. Like formal, organized challenges such as the ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge[412], DeepChem provides a forum for the fair, critical evaluations that are not always conducted in individual methodological papers, which can be biased toward favoring a new proposed algorithm. Likewise DragoNN (Deep RegulAtory GenOmic Neural Networks) offers not only code and a model zoo but also a detailed tutorial and partner package for simulating training data. These resources, especially the ability to simulate datasets that are sufficiently complex to demonstrate the challenges of training neural networks but small enough to train quickly on a CPU, are important for training students and attracting machine learning researchers to problems in genomics and healthcare.

Multimodal, multi-task, and transfer learning

The fact that biomedical datasets often contain a limited number of instances or labels can cause poor performance of deep learning algorithms. These models are particularly prone to overfitting due to their high representational power. However, transfer learning techniques, also known as domain adaptation, enable transfer of extracted patterns between different datasets and even domains. This approach consists of training a model for the base task and subsequently reusing the trained model for the target problem. The first step allows a model to take advantage of a larger amount of data and/or labels to extract better feature representations. Transferring learned features in deep neural networks improves performance compared to randomly initialized features even when pre-training and target sets are dissimilar. However,

transferability of features decreases as the distance between the base task and target task increases [413].

In image analysis, previous examples of deep transfer learning applications proved large-scale natural image sets [42] to be useful for pre-training models that serve as generic feature extractors for various types of biological images [14,206,414,415]. More recently, deep learning models predicted protein sub-cellular localization for proteins not originally present in a training set [416]. Moreover, learned features performed reasonably well even when applied to images obtained using different fluorescent labels, imaging techniques, and different cell types [417]. However, there are no established theoretical guarantees for feature transferability between distant domains such as natural images and various modalities of biological imaging. Because learned patterns are represented in deep neural networks in a layer-wise hierarchical fashion, this issue is usually addressed by fixing an empirically chosen number of layers that preserve generic characteristics of both training and target datasets. The model is then fine-tuned by re-training top layers on the specific dataset in order to re-learn domain-specific high level concepts (e.g. fine-tuning for radiology image classification [54]). Fine-tuning on specific biological datasets enables more focused predictions.

In genomics, the Basset package[177] for predicting chromatin accessibility was shown to rapidly learn and accurately predict on new data by leveraging a model pre-trained on available public data. To simulate this scenario, authors put aside 15 of 164 cell type datasets and trained the Basset model on the remaining 149 datasets. Then, they fine-tuned the model with one training pass of each of the remaining datasets and achieved results close to the model trained on all 164 datasets together. In another example, Min et al. [179] demonstrated how training on the experimentally validated FANTOM5 permissive enhancer dataset followed by fine-tuning on ENCODE enhancer datasets improved cell type-specific predictions, outperforming state-of-the-art results. In drug design, general RNN models trained to generate molecules from the ChEMBL database have been fine-tuned to produce drug-like compounds for specific targets [323,326].

Related to transfer learning, multimodal learning assumes simultaneous learning from various types of inputs, such as images and text. It can capture features that describe common concepts across input modalities. Generative graphical models like RBMs, deep Boltzmann machines, and DBNs, demonstrate successful extraction of more informative features for one modality (images or video) when jointly learned with other modalities (audio or text) [418]. Deep graphical models such as DBNs are considered to be well-suited for multimodal learning tasks because they learn a joint probability distribution from inputs. They can be pre-trained in an unsupervised fashion on

large unlabeled data and then fine-tuned on a smaller number of labeled examples. When labels are available, convolutional neural networks are ubiquitously used because they can be trained end-to-end with backpropagation and demonstrate state-of-the-art performance in many discriminative tasks [14].

Jha et al. [154] showed that integrated training delivered better performance than individual networks. They compared a number of feed-forward architectures trained on RNA-seq data with and without an additional set of CLIP-seq, knockdown, and over-expression based input features. The integrative deep model generalized well for combined data, offering a large performance improvement for alternative splicing event estimation. Chaudhary et al. [419] trained a deep autoencoder model jointly on RNA-seq, miRNA-seq, and methylation data from The Cancer Genome Atlas to predict survival subgroups of hepatocellular carcinoma patients. This multimodal approach that treated different omic data types as different modalities outperformed both traditional methods (principal component analysis) and single-omic models. Interestingly, multi-omic model performance did not improve when combined with clinical information, suggesting that the model was able to capture redundant contributions of clinical features through their correlated genomic features. Chen et al. [141] used deep belief networks to learn phosphorylation states of a common set of signaling proteins in primary cultured bronchial cells collected from rats and humans treated with distinct stimuli. By interpreting species as different modalities representing similar high-level concepts, they showed that DBNs were able to capture cross-species representation of signaling mechanisms in response to a common stimuli. Another application used DBNs for joint unsupervised feature learning from cancer datasets containing gene expression, DNA methylation, and miRNA expression data [148]. This approach allowed for the capture of intrinsic relationships in different modalities and for better clustering performance over conventional k-means.

Multimodal learning with CNNs is usually implemented as a collection of individual networks in which each learns representations from single data type. These individual representations are further concatenated before or within fully-connected layers. FIDDLE [420] is an example of a multimodal CNN that represents an ensemble of individual networks that take NET-seq, MNase-seq, ChIP-seq, RNA-seq, and raw DNA sequence as input to predict Transcription Start Site-seq. The combined model radically improves performance over separately trained datatype-specific networks, suggesting that it learns the synergistic relationship between datasets.

Multi-task learning is an approach related to transfer learning. In a multi-task learning framework, a model learns a number of tasks simultaneously such that

features are shared across them. DeepSEA [168] implemented multi-task joint learning of diverse chromatin factors from raw DNA sequence. This allowed a sequence feature that was effective in recognizing binding of a specific TF to be simultaneously used by another predictor for a physically interacting TF. Similarly, TFImpute [155] learned information shared across transcription factors and cell lines to predict cell-specific TF binding for TF-cell line combinations. Yoon et al. [75] demonstrated that predicting the primary cancer site from cancer pathology reports together with its laterality substantially improved the performance for the latter task, indicating that multi-task learning can effectively leverage the commonality between two tasks using a shared representation. Many studies employed multi-task learning to predict chemical bioactivity [299,302] and drug toxicity [303,421]. Kearnes et al. [296] systematically compared single-task and multi-task models for ADMET properties and found that multi-task learning generally improved performance. Smaller datasets tended to benefit more than larger datasets.

Multi-task learning is complementary to multimodal and transfer learning. All three techniques can be used together in the same model. For example, Zhang et al. [414] combined deep model-based transfer and multi-task learning for cross-domain image annotation. One could imagine extending that approach to multimodal inputs as well. A common characteristic of these methods is better generalization of extracted features at various hierarchical levels of abstraction, which is attained by leveraging relationships between various inputs and task objectives.

Despite demonstrated improvements, transfer learning approaches pose challenges. There are no theoretically sound principles for pre-training and fine-tuning. Best practice recommendations are heuristic and must account for additional hyper-parameters that depend on specific deep architectures, sizes of the pre-training and target datasets, and similarity of domains. However, similarity of datasets and domains in transfer learning and relatedness of tasks in multi-task learning is difficult to access. Most studies address these limitations by empirical evaluation of the model. Unfortunately, negative results are typically not reported. Rajkomar et al. [54] showed that a deep CNN trained on natural images can boost radiology image classification performance. However, due to differences in imaging domains, the target task required either re-training the initial model from scratch with special pre-processing or fine-tuning of the whole network on radiographs with heavy data augmentation to avoid overfitting. Exclusively fine-tuning top layers led to much lower validation accuracy (81.4 versus 99.5). Fine-tuning the aforementioned Basset model with more than one pass resulted in overfitting [177]. DeepChem successfully improved results for low-data drug discovery with one-shot learning for related tasks. However, it clearly demonstrated the limitations of cross-task generalization across unrelated tasks in one-shot models, specifically nuclear

receptor assays and patient adverse reactions [311].

In the medical domain, multimodal, multi-task and transfer learning strategies not only inherit most methodological issues from natural image, text, and audio domains, but also pose domain-specific challenges. There is a compelling need for the development of privacy-preserving transfer learning algorithms, such as Private Aggregation of Teacher Ensembles [122]. We suggest that these types of models deserve deeper investigation to establish sound theoretical guarantees and determine limits for the transferability of features between various closely related and distant learning tasks.

Conclusions

Deep learning-based methods now match or surpass the previous state of the art in a diverse array of tasks in patient and disease categorization, fundamental biological study, genomics, and treatment development. Returning to our central question: given this rapid progress, has deep learning transformed the study of human disease? Though the answer is highly dependent on the specific domain and problem being addressed, we conclude that deep learning has not yet realized its transformative potential or induced a strategic inflection point. Despite its dominance over competing machine learning approaches in many of the areas reviewed here and quantitative improvements in predictive performance, deep learning has not yet definitively "solved" these problems.

As an analogy, consider recent progress in conversational speech recognition. Since 2009 there have been drastic performance improvements with error rates dropping from more than 20% to less than 6% [422] and finally approaching or exceeding human performance in the past year [423,424]. The phenomenal improvements on benchmark datasets are undeniable, but greatly reducing the error rate on these benchmarks did not fundamentally transform the domain. Widespread adoption of conversational speech technologies will require solving the problem, i.e. methods that surpass human performance, and persuading users to adopt them [422]. We see parallels in healthcare, where achieving the full potential of deep learning will require outstanding predictive performance as well as acceptance and adoption by biologists and clinicians. These experts will rightfully demand rigorous evidence that deep learning has impacted their respective disciplines -- elucidated new biological mechanisms and improved patient outcomes -- to be convinced that the promises of deep learning are more substantive than those of previous generations of artificial intelligence.

Some of the areas we have discussed are closer to surpassing this lofty bar than others, generally those that are more similar to the non-biomedical tasks

that are now monopolized by deep learning. In medical imaging, diabetic retinopathy [46], diabetic macular edema [46], tuberculosis [55], and skin lesion [4] classifiers are highly accurate and comparable to clinician performance.

In other domains, perfect accuracy will not be required because deep learning will primarily prioritize experiments and assist discovery. For example, in chemical screening for drug discovery, a deep learning system that successfully identifies dozens or hundreds of target-specific, active small molecules from a massive search space would have immense practical value even if its overall precision is modest. In medical imaging, deep learning can point an expert to the most challenging cases that require manual review [55], though the risk of false negatives must be addressed. In protein structure prediction, errors in individual residue-residue contacts can be tolerated when using the contacts jointly for 3D structure modeling. Improved contact map predictions [27] have led to notable improvements in fold and 3D structure prediction for some of the most challenging proteins, such as membrane proteins [201].

Conversely, the most challenging tasks may be those in which predictions are used directly for downstream modeling or decision-making, especially in the clinic. As an example, errors in sequence variant calling will be amplified if they are used directly for GWAS. In addition, the stochasticity and complexity of biological systems implies that for some problems, for instance predicting gene regulation in disease, perfect accuracy will be unattainable.

We are witnessing deep learning models achieving human-level performance across a number of biomedical domains. However, machine learning algorithms, including deep neural networks, are also prone to mistakes that humans are much less likely to make, such as misclassification of adversarial examples [425,426], a reminder that these algorithms do not understand the semantics of the objects presented. It may be impossible to guarantee that a model is not susceptible to adversarial examples, but work in this area is continuing [427,428]. Cooperation between human experts and deep learning algorithms addresses many of these challenges and can achieve better performance than either individually [66]. For sample and patient classification tasks, we expect deep learning methods to augment clinicians and biomedical researchers.

We are extremely optimistic about the future of deep learning in biology and medicine. It is by no means inevitable that deep learning will revolutionize these domains, but given how rapidly the field is evolving, we are confident that its full potential in biomedicine has not been explored. We have highlighted numerous challenges beyond improving training and predictive accuracy, such as preserving patient privacy and interpreting models. Ongoing research has begun to address these problems and shown that they are not insurmountable.

Deep learning offers the flexibility to model data in its most natural form, for example, longer DNA sequences instead of k-mers for transcription factor binding prediction and molecular graphs instead of pre-computed bit vectors for drug discovery. These flexible input feature representations have spurred creative modeling approaches that would be infeasible with other machine learning techniques. Unsupervised methods are currently less-developed than their supervised counterparts, but they may have the most potential because of how expensive and time-consuming it is to label large amounts of biomedical data. If future deep learning algorithms can summarize very large collections of input data into interpretable models that spur scientists to ask questions that they did not know how to ask, it will be clear that deep learning has transformed biology and medicine.

Methods

Continuous collaborative manuscript drafting

We recognized that deep learning in precision medicine is a rapidly developing area. Hence, diverse expertise was required to provide a forward-looking perspective. Accordingly, we collaboratively wrote this review in the open, enabling anyone with expertise to contribute. We wrote the manuscript in markdown and tracked changes using git. Contributions were handled through GitHub, with individuals submitting "pull requests" to suggest additions to the manuscript.

To facilitate citation, we defined a markdown citation syntax. We supported citations to the following identifier types (in order of preference): DOIs, PubMed IDs, arXiv IDs, and URLs. References were automatically generated from citation metadata by querying APIs to generate Citation Style Language (CSL) JSON items for each reference. Pandoc and pandoc-citeproc converted the markdown to HTML and PDF, while rendering the formatted citations and references. In total, referenced works consisted of 280 DOIs, 5 PubMed records, 108 arXiv manuscripts, and 39 URLs (webpages as well as manuscripts lacking standardized identifiers).

We implemented continuous analysis so the manuscript was automatically regenerated whenever the source changed [113]. We configured Travis CI -- a continuous integration service -- to fetch new citation metadata and rebuild the manuscript for every commit. Accordingly, formatting or citation errors in pull requests would cause the Travis CI build to fail, automating quality control. In addition, the build process renders templated variables, such as the reference counts mentioned above, to automate the updating of dynamic content. When contributions were merged into the `master` branch, Travis CI deployed the built manuscript by committing back to the GitHub repository. As a result, the latest

manuscript version is always available at https://greenelab.github.io/deep-review. To ensure a consistent software environment, we defined a versioned conda environment of the software dependencies.

In addition, we instructed the Travis CI deployment script to perform blockchain timestamping [429,430]. Using OpenTimestamps, we submitted hashes for the manuscript and the source git commit for timestamping in the Bitcoin blockchain [431]. These timestamps attest that a given version of this manuscript (and its history) existed at a given point in time. The ability to irrefutably prove manuscript existence at a past time could be important to establish scientific precedence and enforce an immutable record of authorship.

Author contributions

We created an open repository on the GitHub version control platform (greenelab/deep-review) [432]. Here, we engaged with numerous authors from papers within and outside of the area. The manuscript was drafted via GitHub commits by 27 individuals who met the ICMJE standards of authorship. These were individuals who contributed to the review of the literature; drafted the manuscript or provided substantial critical revisions; approved the final manuscript draft; and agreed to be accountable in all aspects of the work. Individuals who did not contribute in all of these ways, but who did participate, are acknowledged below. We grouped authors into the following four classes of approximately equal contributions and randomly ordered authors within each contribution class. Drafted multiple sub-sections along with extensive editing, pull request reviews, or discussion: A.A.K., B.K.B., B.T.D., D.S.H., E.F., G.P.W., P.A., T.C. Drafted one or more sub-sections: A.E.C., A.S., B.J.L., E.M.C., G.L.R., J.I., J.L., J.X., S.W., W.X. Revised specific sub-sections or supervised drafting one or more sub-sections: A.K., D.D., D.J.H., L.K.W., M.H.S.S., Y.P., Y.Q. Drafted sub-sections, edited the manuscript, reviewed pull requests, and coordinated co-authors: A.G., C.S.G..

Competing Interests

A.K. is on the Advisory Board of Deep Genomics Inc. E.F. is a full-time employee of GlaxoSmithKline. The remaining authors have no competing interests to declare.

Acknowledgements

We gratefully acknowledge Christof Angermueller, Kumardeep Chaudhary, Gökcen Eraslan, Michael M. Hoffman, Mikael Huss, Bharath Ramsundar and Xun Zhu for their discussion of the manuscript and reviewed papers on GitHub. We would like to thank Zhiyong Lu for revisions to the text that were not

captured on GitHub as well as GitHub users aaronsheldon and swamidass who contributed text but did not formally approve the manuscript. Finally, we acknowledge funding from the Gordon and Betty Moore Foundation awards GBMF4552 (C.S.G. and D.S.H.) and GBMF4563 (D.J.H.); the National Institutes of Health awards DP2GM123485 (A.K.), R01AI116794 (B.K.B.), R01GM089652 (A.E.C.), R01GM089753 (J.X.), T32GM007753 (B.T.D.), and U54AI117924 (A.G.); the National Science Foundation awards 1245632 (G.L.R.), 1531594 (E.M.C.), and 1564955 (J.X.); and the National Institutes of Health Intramural Research Program and National Library of Medicine (Y.P.).

References

1. Stephens ZD *et al.* 2015 Big Data: Astronomical or Genomical? *PLOS Biology* **13**, e1002195. See https://doi.org/10.1371/journal.pbio.1002195.

2. LeCun Y, Bengio Y, Hinton G. 2015 Deep learning. *Nature* **521**, 436–444. See https://doi.org/10.1038/nature14539.

3. Baldi P, Sadowski P, Whiteson D. 2014 Searching for exotic particles in high-energy physics with deep learning. *Nature Communications* **5**. See https://doi.org/10.1038/ncomms5308.

4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. 2017 Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. See https://doi.org/10.1038/nature21056.

5. Wu Y *et al.* 2016 Google's neural machine translation system: Bridging the gap between human and machine translation. See https://arxiv.org/abs/1609.08144v2.

6. McCulloch WS, Pitts W. 1943 A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* **5**, 115–133. See https://doi.org/10.1007/bf02478259.

7. Block HD, Knight BW, Rosenblatt F. 1962 Analysis of a Four-Layer Series-Coupled Perceptron. II. *Reviews of Modern Physics* **34**, 135–142. See https://doi.org/10.1103/revmodphys.34.135.

8. 2016 Google Research Publication: Building High-level Features Using Large Scale Unsupervised Learning. See http://research.google.com/archive/unsupervised_icml2012.html.

9. Niu F, Recht B, Re C, Wright SJ. 2011 HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. See https://arxiv.org/abs/1106.5730v2.

10. Goodfellow I, Bengio Y, Courville A. 2016 Deep Learning. See http://www.deeplearningbook.org/.

11. Grove AS. 1998 Academy of Management: Andrew S. Grove. See http://www.intel.com/pressroom/archive/speeches/ag080998.htm.

12. Park Y, Kellis M. 2015 Deep learning for regulatory genomics.*Nature Biotechnology* **33**, 825–826. See https://doi.org/10.1038/nbt.3313.

13. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. 2016 Applications of Deep Learning in Biomedicine. *Molecular Pharmaceutics* **13**, 1445–1454. See https://doi.org/10.1021/acs.molpharmaceut.5b00982.

14. Angermueller C, Pärnamaa T, Parts L, Stegle O. 2016 Deep learning for computational biology. *Molecular Systems Biology* **12**, 878. See https://doi.org/10.15252/msb.20156651.

15. Min S, Lee B, Yoon S. 2016 Deep learning in bioinformatics.*Briefings in Bioinformatics*, bbw068. See https://doi.org/10.1093/bib/bbw068.

16. Kraus OZ, Frey BJ. 2016 Computer vision for high content screening. *Critical Reviews in Biochemistry and Molecular Biology* **51**, 102–109. See https://doi.org/10.3109/10409238.2015.1135868.

17. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. 2017 Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* See https://doi.org/10.1093/bib/bbx044.

18. Gawehn E, Hiss JA, Schneider G. 2015 Deep Learning in Drug Discovery. *Molecular Informatics* **35**, 3–14. See https://doi.org/10.1002/minf.201501008.

19. Goh GB, Hodas NO, Vishnu A. 2017 Deep learning for computational chemistry. *Journal of Computational Chemistry* **38**, 1291–1307. See https://doi.org/10.1002/jcc.24764.

20. Pérez-Sianes J, Pérez-Sánchez H, Díaz F. 2016 Virtual Screening: A Challenge for Deep Learning. In *Advances in Intelligent Systems and Computing*, pp. 13–22. Springer International Publishing. See https://doi.org/10.1007/978-3-319-40126-3_2.

21. Baskin II, Winkler D, Tetko IV. 2016 A renaissance of neural networks in drug discovery. *Expert Opinion on Drug Discovery* **11**, 785–795. See https://doi.org/10.1080/17460441.2016.1201262.

22. Parker JS *et al.* 2009 Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* **27**, 1160–1167. See

https://doi.org/10.1200/jco.2008.18.1370.

23. Mayer IA, Abramson VG, Lehmann BD, Pietenpol JA. 2014 New Strategies for Triple-Negative Breast Cancer–Deciphering the Heterogeneity. *Clinical Cancer Research* **20**, 782–790. See https://doi.org/10.1158/1078-0432.ccr-13-0583.

24. TAN J, UNG M, CHENG C, GREENE CS. 2014 UNSUPERVISED FEATURE CONSTRUCTION AND KNOWLEDGE EXTRACTION FROM GENOME-WIDE ASSAYS OF BREAST CANCER WITH DENOISING AUTOENCODERS. In *Biocomputing 2015*, WORLD SCIENTIFIC. See https://doi.org/10.1142/9789814644730_0014.

25. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. 2013 Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pp. 411–418. Springer Berlin Heidelberg. See https://doi.org/10.1007/978-3-642-40763-5_51.

26. Zurada J. 1994 End effector target position learning using feedforward with error back-propagation and recurrent neural networks. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, IEEE. See https://doi.org/10.1109/icnn.1994.374637.

27. Wang S, Sun S, Li Z, Zhang R, Xu J. 2017 Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology* **13**, e1005324. See https://doi.org/10.1371/journal.pcbi.1005324.

28. Spencer M, Eickholt J, Cheng J. 2015 A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, 103–112. See https://doi.org/10.1109/tcbb.2014.2343960.

29. Wang S, Peng J, Ma J, Xu J. 2016 Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports* **6**. See https://doi.org/10.1038/srep18962.

30. Liu F, Li H, Ren C, Bo X, Shu W. 2016 PEDLA: predicting enhancers with a deep learning-based algorithmic framework. See https://doi.org/10.1101/036129.

31. Li Y, Chen C-Y, Wasserman WW. 2015 Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. In *Lecture Notes in Computer Science*, pp. 205–217. Springer International Publishing. See https://doi.org/10.1007/978-3-319-16706-0_20.

32. Kleftogiannis D, Kalnis P, Bajic VB. 2014 DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research* **43**, e6–e6. See https://doi.org/10.1093/nar/gku1058.

33. Quang D, Chen Y, Xie X. 2014 DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763. See https://doi.org/10.1093/bioinformatics/btu703.

34. Wallach I, Dzamba M, Heifets A. 2015 AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. See https://arxiv.org/abs/1510.02855v1.

35. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. 2016 Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular Pharmaceutics* **13**, 2524–2530. See https://doi.org/10.1021/acs.molpharmaceut.6b00248.

36. Wang Y, Zeng J. 2013 Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* **29**, i126–i134. See https://doi.org/10.1093/bioinformatics/btt234.

37. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H. 2017 Deep-Learning-Based Drug–Target Interaction Prediction. *Journal of Proteome Research* **16**, 1401–1409. See https://doi.org/10.1021/acs.jproteome.6b00618.

38. Stenstrom G, Gottsater A, Bakhtadze E, Berger B, Sundkvist G. 2005 Latent Autoimmune Diabetes in Adults: Definition, Prevalence, -Cell Function, and Treatment. *Diabetes* **54**, S68–S72. See https://doi.org/10.2337/diabetes.54.suppl_2.s68.

39. Groop LC, Bottazzo GF, Doniach D. 1986 Islet Cell Antibodies Identify Latent Type I Diabetes in Patients Aged 35-75 Years at Diagnosis. *Diabetes* **35**, 237–241. See https://doi.org/10.2337/diab.35.2.237.

40. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Laak JAWM van der, Ginneken B van, Sánchez CI. 2017 A survey on deep learning in medical image analysis. See https://arxiv.org/abs/1702.05747v1.

41. Shen D, Wu G, Suk H-I. 2016 Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering* **19**. See https://doi.org/10.1146/annurev-bioeng-071516-044442.

42. Russakovsky O *et al.* 2015 ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**, 211–252. See https://doi.org/10.1007/s11263-015-0816-y.

43. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. 2016 Convolutional Neural Networks for Diabetic Retinopathy. *Procedia Computer Science* **90**, 200–205. See https://doi.org/10.1016/j.procs.2016.07.014.

44. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, Niemeijer M. 2016 Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Investigative Opthalmology & Visual Science* **57**, 5200. See https://doi.org/10.1167/iovs.16-19964.

45. Leibig C, Allken V, Berens P, Wahl S. 2016 Leveraging uncertainty information from deep neural networks for disease detection. See https://doi.org/10.1101/084210.

46. Gulshan V *et al.* 2016 Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402. See https://doi.org/10.1001/jama.2016.17216.

47. Codella N, Nguyen Q-B, Pankanti S, Gutman D, Helba B, Halpern A, Smith JR. 2016 Deep learning ensembles for melanoma recognition in dermoscopy images. See https://arxiv.org/abs/1610.04662v2.

48. Yu L, Chen H, Dou Q, Qin J, Heng P-A. 2017 Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging* **36**, 994–1004. See https://doi.org/10.1109/tmi.2016.2642839.

49. Jafari MH, Nasr-Esfahani E, Karimi N, Soroushmehr SMR, Samavi S, Najarian K. 2016 Extraction of skin lesions from non-dermoscopic images using deep learning. See https://arxiv.org/abs/1609.02374v1.

50. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr S, Jafari M, Ward K, Najarian K. 2016 Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE. See https://doi.org/10.1109/embc.2016.7590963.

51. Burlina P, Freund DE, Joshi N, Wolfson Y, Bressler NM. 2016 Detection of age-related macular degeneration via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, IEEE. See https://doi.org/10.1109/isbi.2016.7493240.

52. Bar Y, Diamant I, Wolf L, Greenspan H. 2015 Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015: Computer-Aided Diagnosis* (eds LM Hadjiiski, GD Tourassi), SPIE. See https://doi.org/10.1117/12.2083124.

53. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. 2016 Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging* **35**, 1285–1298. See https://doi.org/10.1109/tmi.2016.2528162.

54. Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. 2016 High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *Journal of Digital Imaging* **30**, 95–101. See https://doi.org/10.1007/s10278-016-9914-9.

55. Lakhani P, Sundaram B. 2017 Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 162326. See https://doi.org/10.1148/radiol.2017162326.

56. Amit G, Ben-Ari R, Hadad O, Monovich E, Granot N, Hashoul S. 2017 Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches. In *Medical Imaging 2017: Computer-Aided Diagnosis* (eds SG Armato, NA Petrick), SPIE. See https://doi.org/10.1117/12.2249981.

57. Dhungel N, Carneiro G, Bradley AP. 2015 Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms. In *Lecture Notes in Computer Science*, pp. 605–612. Springer International Publishing. See https://doi.org/10.1007/978-3-319-24553-9_74.

58. Dhungel N, Carneiro G, Bradley AP. 2016 The Automated Learning of Deep Features for Breast Mass Classification from Mammograms. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 106–114. Springer International Publishing. See https://doi.org/10.1007/978-3-319-46723-8_13.

59. Zhu W, Lou Q, Vang YS, Xie X. 2016 Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. See https://doi.org/10.1101/095794.

60. Zhu W, Xie X. 2016 Adversarial Deep Structural Networks for Mammographic Mass Segmentation. See https://doi.org/10.1101/095786.

61. Dhungel N, Carneiro G, Bradley AP. 2017 A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis* **37**, 114–128. See https://doi.org/10.1016/j.media.2017.01.009.

62. Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, Kim L, Summers RM. 2016 Improving Computer-Aided Detection Using_newlineConvolutional Neural

Networks and Random View Aggregation. *IEEE Transactions on Medical Imaging* **35**, 1170–1181. See https://doi.org/10.1109/tmi.2015.2482920.

63. Nie D, Zhang H, Adeli E, Liu L, Shen D. 2016 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 212–220. Springer International Publishing. See https://doi.org/10.1007/978-3-319-46723-8_25.

64. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N. 2017 Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis* **35**, 303–312. See https://doi.org/10.1016/j.media.2016.07.007.

65. Litjens G *et al.* 2016 Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports* **6**. See https://doi.org/10.1038/srep26286.

66. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. 2016 Deep learning for identifying metastatic breast cancer. See https://arxiv.org/abs/1606.05718v1.

67. Lee CS, Baughman DM, Lee AY. 2016 Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. See https://doi.org/10.1101/094276.

68. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105. USA: Curran Associates Inc. See http://dl.acm.org/citation.cfm?id=2999134.2999257.

69. Wang X, Lu L, Shin H-c, Kim L, Bagheri M, Nogues I, Yao J, Summers RM. 2017 Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. See https://arxiv.org/abs/1701.06599v1.

70. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 2017 ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. See https://arxiv.org/abs/1705.02315v2.

71. Lever J, Krzywinski M, Altman N. 2016 Points of Significance: Classification evaluation. *Nature Methods* **13**, 603–604. See https://doi.org/10.1038/nmeth.3945.

72. Ohno-Machado L. 2011 Realizing the full potential of electronic health

records: the role of natural language processing. *Journal of the American Medical Informatics Association* **18**, 539–539. See https://doi.org/10.1136/amiajnl-2011-000501.

73. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. 2011 Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association* **18**, 557–562. See https://doi.org/10.1136/amiajnl-2011-000150.

74. Chalapathy R, Borzeshi EZ, Piccardi M. 2016 Bidirectional lstm-crf for clinical concept extraction. See https://arxiv.org/abs/1611.08373v1.

75. Yoon H-J, Ramanathan A, Tourassi G. 2016 Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports. In *Advances in Big Data*, pp. 195–204. Springer International Publishing. See https://doi.org/10.1007/978-3-319-47898-2_21.

76. Mikolov T, Chen K, Corrado G, Dean J. 2013 Efficient estimation of word representations in vector space. See https://arxiv.org/abs/1301.3781v3.

77. De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. 2014 Medical Semantic Similarity with a Neural Language Model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, ACM Press. See https://doi.org/10.1145/2661829.2661974.

78. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J. 2016 Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, ACM Press. See https://doi.org/10.1145/2939672.2939823.

79. Gligorijevic D, Stojanovic J, Djuric N, Radosavljevic V, Grbovic M, Kulathinal RJ, Obradovic Z. 2016 Large-Scale Discovery of Disease-Disease and Disease-Gene Associations. *Scientific Reports* **6**. See https://doi.org/10.1038/srep32404.

80. Lasko TA, Denny JC, Levy MA. 2013 Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE* **8**, e66341. See https://doi.org/10.1371/journal.pone.0066341.

81. Beaulieu-Jones BK, Greene CS. 2016 Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics* **64**, 168–178. See https://doi.org/10.1016/j.jbi.2016.10.007.

82. Miotto R, Li L, Kidd BA, Dudley JT. 2016 Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports* **6**. See https://doi.org/10.1038/srep26094.

83. Razavian N, Marcus J, Sontag D. 2016 Multi-task prediction of disease onsets from longitudinal lab tests. See https://arxiv.org/abs/1608.00647v3.

84. Ranganath R, Perotte A, Elhadad N, Blei D. 2016 Deep survival analysis. See https://arxiv.org/abs/1608.02158v2.

85. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S. 2000 Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational Statistics & Data Analysis* **34**, 243–257. See https://doi.org/10.1016/s0167-9473(99)00098-5.

86. Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, Kluger Y. 2016 Deep survival: A deep cox proportional hazards network. See https://arxiv.org/abs/1606.00931v2.

87. Ranganath R, Tang L, Charlin L, Blei DM. 2014 Deep exponential families. See https://arxiv.org/abs/1411.2581v1.

88. Hoffman M, Blei DM, Wang C, Paisley J. 2012 Stochastic variational inference. See https://arxiv.org/abs/1206.7051v3.

89. Ranganath R, Tran D, Blei DM. 2015 Hierarchical variational models. See https://arxiv.org/abs/1511.02386v2.

90. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. 2017 A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics* **97**, 120–127. See https://doi.org/10.1016/j.ijmedinf.2016.09.014.

91. 2017 Implementations by Phenotype | PheKB. See https://phekb.org/implementations.

92. Halpern Y, Horng S, Choi Y, Sontag D. 2016 Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association* **23**, 731–740. See https://doi.org/10.1093/jamia/ocw011.

93. Ratner A, Sa CD, Wu S, Selsam D, Ré C. 2016 Data programming: Creating large training sets, quickly. See https://arxiv.org/abs/1605.07723v3.

94. Palmer M. 2006 Data is the New Oil. *ANA Marketing Maestros*. See http://ana.blogs.com/maestros/2006/11/data_is_the_new.html.

95. Haupt M. 2016 'Data is the New Oil' — A Ludicrous Proposition – Twenty One Hundred – Medium. *Medium*. See https://medium.com/twenty-one-hundred/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294.

96. Ratner A, Bach S, Ré C. 2016 Data Programming: Machine Learning with Weak Supervision. See http://hazyresearch.github.io/snorkel/blog/weak_supervision.html.

97. Jensen PB, Jensen LJ, Brunak S. 2012 Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13**, 395–405. See https://doi.org/10.1038/nrg3208.

98. Weiskopf NG, Weng C. 2013 Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20**, 144–151. See https://doi.org/10.1136/amiajnl-2011-000681.

99. Bowman S. 2013 Impact of Electronic Health Record Systems on Information Integrity: Quality and Safety Implications. *Perspect Health Inf Manag* **10**, 1c. See https://www.ncbi.nlm.nih.gov/pubmed/24159271.

100. Botsis T, Hartvigsen G, Chen F, Weng C. 2010 Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translat Bioinforma* **2010**, 1–5. See https://www.ncbi.nlm.nih.gov/pubmed/21347133.

101. Serdén L, Lindqvist R, Rosén M. 2003 Have DRG-based prospective payment systems influenced the number of secondary diagnoses in health care administrative data? *Health Policy* **65**, 101–107. See https://doi.org/10.1016/s0168-8510(02)00208-7.

102. Just BH, Marc D, Munns M, Sandefer R. 2016 Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields. *Perspect Health Inf Manag* **13**, 1e. See https://www.ncbi.nlm.nih.gov/pubmed/27134610.

103. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. 2014 Identifying and mitigating biases in EHR laboratory tests. *Journal of Biomedical Informatics* **51**, 24–34. See https://doi.org/10.1016/j.jbi.2014.03.016.

104. De Moor G *et al.* 2015 Using electronic health records for clinical research: The case of the EHR4CR project. *Journal of Biomedical Informatics* **53**, 162–173. See https://doi.org/10.1016/j.jbi.2014.10.006.

105. Oemig F, Snelick R. 2016 *Healthcare Interoperability Standards Compliance Handbook*. Springer International Publishing. See https://doi.org/10.1007/978-3-319-44839-8.

106. Faber J, Fonseca LM. 2014 How sample size influences research outcomes. *Dental Press Journal of Orthodontics* **19**, 27–29. See https://doi.org/10.1590/2176-9451.19.4.027-029.ebo.

107. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. 2014 A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* **21**, 221–230. See https://doi.org/10.1136/amiajnl-2013-001935.

108. WILEY LK, VANHOUTEN JP, SAMUELS DC, ALDRICH MC, RODEN DM, PETERSON JF, DENNY JC. 2016 STRATEGIES FOR EQUITABLE PHARMACOGENOMIC-GUIDED WARFARIN DOSING AMONG EUROPEAN AND AFRICAN AMERICAN INDIVIDUALS IN A CLINICAL POPULATION. In *Biocomputing 2017*, WORLD SCIENTIFIC. See https://doi.org/10.1142/9789813207813_0050.

109. Rahu M, McKee M. 2008 Epidemiological research labelled as a violation of privacy: the case of Estonia. *International Journal of Epidemiology* **37**, 678–682. See https://doi.org/10.1093/ije/dyn022.

110. Wiley LK, Tarczy-Hornoch P, Denny JC, Freimuth RR, Overby CL, Shah N, Martin RD, Sarkar IN. 2016 Harnessing next-generation informatics for personalizing medicine: a report from AMIA's 2014 Health Policy Invitational Meeting. *Journal of the American Medical Informatics Association* **23**, 413–419. See https://doi.org/10.1093/jamia/ocv111.

111. Gaye A *et al.* 2014 DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology* **43**, 1929–1944. See https://doi.org/10.1093/ije/dyu188.

112. Carter KW *et al.* 2015 ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data. *International Journal of Epidemiology* **45**, 408–416. See https://doi.org/10.1093/ije/dyv193.

113. Beaulieu-Jones BK, Greene CS. 2017 Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology* **35**, 342–346. See https://doi.org/10.1038/nbt.3780.

114. Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. 2016 Stealing machine learning models via prediction apis. See https://arxiv.org/abs/1609.02943v2.

115. Dwork C, Roth A. 2013 The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* **9**, 211–407. See https://doi.org/10.1561/0400000042.

116. Shokri R, Stronati M, Song C, Shmatikov V. 2016 Membership inference attacks against machine learning models. See https://arxiv.org/abs/1610.05820v2.

117. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. 2017 Generating multi-label discrete electronic health records using generative adversarial networks. See https://arxiv.org/abs/1703.06490v1.

118. Simmons S, Sahinalp C, Berger B. 2016 Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations. *Cell Systems* **3**, 54–61. See https://doi.org/10.1016/j.cels.2016.04.013.

119. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. 2016 Deep learning with differential privacy. See https://arxiv.org/abs/1607.00133v2.

120. McMahan B, Moore E, Ramage D, Hampson S, Aguera y Arcas B. 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data. See http://proceedings.mlr.press/v54/mcmahan17a.html.

121. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K. 2017 Practical Secure Aggregation for Privacy Preserving Machine Learning. See https://eprint.iacr.org/2017/281.

122. Papernot N, Abadi M, Erlingsson Ú, Goodfellow I, Talwar K. 2016 Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. See https://openreview.net/forum?id=HkwoSDPgg.

123. Goodman B, Flaxman S. 2016 European union regulations on algorithmic decision-making and a 'right to explanation'. See https://arxiv.org/abs/1606.08813v3.

124. Zöllner S, Pritchard JK. 2007 Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data. *The American Journal of Human Genetics* **80**, 605–615. See https://doi.org/10.1086/512821.

125. Beery AK, Zucker I. 2011 Sex bias in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews* **35**, 565–572. See https://doi.org/10.1016/j.neubiorev.2010.07.002.

126. Carlson CS *et al.* 2013 Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biology* **11**, e1001661. See https://doi.org/10.1371/journal.pbio.1001661.

127. Price AL, Zaitlen NA, Reich D, Patterson N. 2010 New approaches to

population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463. See https://doi.org/10.1038/nrg2813.

128. Sebastiani P *et al.* 2011 Retraction. *Science* **333**, 404–404. See https://doi.org/10.1126/science.333.6041.404-a.

129. Kaufman S, Rosset S, Perlich C, Stitelman O. 2012 Leakage in data mining. *ACM Transactions on Knowledge Discovery from Data* **6**, 1–21. See https://doi.org/10.1145/2382577.2382579.

130. Lum K, Isaac W. 2016 To predict and serve? *Significance* **13**, 14–19. See https://doi.org/10.1111/j.1740-9713.2016.00960.x.

131. Hardt M, Price E, Srebro N. 2016 Equality of opportunity in supervised learning. See https://arxiv.org/abs/1610.02413v1.

132. Joseph M, Kearns M, Morgenstern J, Neel S, Roth A. 2016 Rawlsian fairness for machine learning. See https://arxiv.org/abs/1610.09559v3.

133. Mahmood SS, Levy D, Vasan RS, Wang TJ. 2014 The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet* **383**, 999–1008. See https://doi.org/10.1016/s0140-6736(13)61752-3.

134. Pearson H. 2012 Children of the 90s: Coming of age.*Nature* **484**, 155–158. See https://doi.org/10.1038/484155a.

135. Kaplan EL, Meier P. 1958 Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association***53**, 457–481. See https://doi.org/10.1080/01621459.1958.10501452.

136. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, Jensen PB, Jensen LJ, Brunak S. 2014 Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications* **5**. See https://doi.org/10.1038/ncomms5022.

137. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. 2016 Deepr: A convolutional net for medical records. See https://arxiv.org/abs/1607.07519v1.

138. Pham T, Tran T, Phung D, Venkatesh S. 2016 DeepCare: A deep dynamic memory model for predictive medicine. See https://arxiv.org/abs/1602.00357v2.

139. NIH. 2012 Curiosity Creates Cures: The Value and Impact of Basic Research. See https://www.nigms.nih.gov/Education/Documents/curiosity.pdf.

140. Kim M, Rai N, Zorraquino V, Tagkopoulos I. 2016 Multi-omics integration

accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nature Communications* **7**, 13090. See https://doi.org/10.1038/ncomms13090.

141. Chen L, Cai C, Chen V, Lu X. 2015 Trans-species learning of cellular signaling systems with bimodal deep belief networks. *Bioinformatics* **31**, 3008–3015. See https://doi.org/10.1093/bioinformatics/btv315.

142. Gupta A, Wang H, Ganapathiraju M. 2015 Learning structure in gene expression data using deep architectures, with an application to gene clustering. See https://doi.org/10.1101/031906.

143. Chen L, Cai C, Chen V, Lu X. 2016 Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17**. See https://doi.org/10.1186/s12859-015-0852-1.

144. Tan J, Hammond JH, Hogan DA, Greene CS. 2016 ADAGE-Based Integration of Publicly AvailablePseudomonas aeruginosaGene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems* **1**, e00025–15. See https://doi.org/10.1128/msystems.00025-15.

145. Tan J *et al.* 2016 Unsupervised extraction of stable expression signatures from public compendia with eADAGE. See https://doi.org/10.1101/078659.

146. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. 2016 Gene expression inference with deep learning. *Bioinformatics* **32**, 1832–1839. See https://doi.org/10.1093/bioinformatics/btw074.

147. Singh R, Lanchantin J, Robins G, Qi Y. 2016 DeepChrome: Deep-learning for predicting gene expression from histone modifications. See https://arxiv.org/abs/1607.02078v1.

148. Liang M, Li Z, Chen T, Zeng J. 2015 Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, 928–937. See https://doi.org/10.1109/tcbb.2014.2377729.

149. Scotti MM, Swanson MS. 2015 RNA mis-splicing in disease.*Nature Reviews Genetics* **17**, 19–32. See https://doi.org/10.1038/nrg.2015.3.

150. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. 2016 RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604. See https://doi.org/10.1126/science.aad9417.

151. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010 Deciphering the splicing code. *Nature* **465**, 53–59. See https://doi.org/10.1038/nature09000.

152. Xiong HY, Barash Y, Frey BJ. 2011 Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–2562. See https://doi.org/10.1093/bioinformatics/btr444.

153. Xiong HY *et al.* 2014 The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806–1254806. See https://doi.org/10.1126/science.1254806.

154. Jha A, Gazzara MR, Barash Y. 2017 Integrative Deep Models for Alternative Splicing. See https://doi.org/10.1101/104869.

155. Qin Q, Feng J. 2017 Imputation for transcription factor binding predictions based on deep learning. *PLOS Computational Biology* **13**, e1005403. See https://doi.org/10.1371/journal.pcbi.1005403.

156. Rosenberg A, Patwardhan R, Shendure J, Seelig G. 2015 Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **163**, 698–711. See https://doi.org/10.1016/j.cell.2015.09.054.

157. Juan-Mateu J, Villate O, Eizirik DL. 2015 MECHANISMS IN ENDOCRINOLOGY: Alternative splicing: the new frontier in diabetes research. *European Journal of Endocrinology* **174**, R225–R238. See https://doi.org/10.1530/eje-15-0916.

158. Pan X, Shen H-B. 2017 RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics* **18**. See https://doi.org/10.1186/s12859-017-1561-8.

159. Dunham I *et al.* 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. See https://doi.org/10.1038/nature11247.

160. Stormo GD. 2000 DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23. See https://doi.org/10.1093/bioinformatics/16.1.16.

161. Horton PB, Kanehisa M. 1992 An assessment of neural network and statistical approaches for prediction of E.coli Promoter sites. *Nucleic Acids Research* **20**, 4331–4338. See https://doi.org/10.1093/nar/20.16.4331.

162. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014 Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology* **10**, e1003711. See https://doi.org/10.1371/journal.pcbi.1003711.

163. Setty M, Leslie CS. 2015 SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology* **11**, e1004271. See https://doi.org/10.1371/journal.pcbi.1004271.

164. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015 Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831–838. See https://doi.org/10.1038/nbt.3300.

165. Lanchantin J, Singh R, Wang B, Qi Y. 2016 Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. See https://arxiv.org/abs/1608.03644v4.

166. Zeng H, Edwards MD, Liu G, Gifford DK. 2016 Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* **32**, i121–i127. See https://doi.org/10.1093/bioinformatics/btw255.

167. Shrikumar A, Greenside P, Kundaje A. 2017 Reverse-complement parameter sharing improves deep learning models for genomics. See https://doi.org/10.1101/103663.

168. Zhou J, Troyanskaya OG. 2015 Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**, 931–934. See https://doi.org/10.1038/nmeth.3547.

169. Shrikumar A, Greenside P, Kundaje A. 2017 Learning important features through propagating activation differences. See https://arxiv.org/abs/1704.02685v1.

170. Lek M *et al.* 2016 Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291. See https://doi.org/10.1038/nature19057.

171. Werner T. 2003 The state of the art of mammalian promoter recognition. *Briefings in Bioinformatics* **4**, 22–30. See https://doi.org/10.1093/bib/4.1.22.

172. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. 2013 Enhancers: five essential questions. *Nature Reviews Genetics* **14**, 288–295. See https://doi.org/10.1038/nrg3458.

173. Andersson R, Sandelin A, Danko CG. 2015 A unified architecture of transcriptional regulatory elements. *Trends in Genetics* **31**, 426–433. See https://doi.org/10.1016/j.tig.2015.05.007.

174. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014 High-throughput functional testing of ENCODE segmentation predictions. *Genome Research* **24**, 1595–1602. See https://doi.org/10.1101/gr.173518.114.

175. Fickett JW, Hatzigeorgiou AG. 1997 Eukaryotic Promoter Recognition. *Genome Research* **7**, 861–878. See https://doi.org/10.1101/gr.7.9.861.

176. Matis S, Xu Y, Shah M, Guan X, Einstein J, Mural R, Uberbacher E. 1996

Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Computers & Chemistry* **20**, 135–140. See https://doi.org/10.1016/s0097-8485(96)80015-5.

177. Kelley DR, Snoek J, Rinn JL. 2016 Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **26**, 990–999. See https://doi.org/10.1101/gr.200535.115.

178. Umarov RK, Solovyev VV. 2017 Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLOS ONE* **12**, e0171410. See https://doi.org/10.1371/journal.pone.0171410.

179. Xu Min, Ning Chen, Ting Chen, Rui Jiang. 2016 DeepEnhancer: Predicting enhancers by convolutional neural networks. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE. See https://doi.org/10.1109/bibm.2016.7822593.

180. Singh S, Yang Y, Poczos B, Ma J. 2016 Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. See https://doi.org/10.1101/085241.

181. Li Y, Shi W, Wasserman WW. 2016 Genome-Wide Prediction of cis-Regulatory Regions Using Supervised Deep Learning Methods. See https://doi.org/10.1101/041616.

182. Bracken CP, Scott HS, Goodall GJ. 2016 A network-biology perspective of microRNA function and dysfunction in cancer. *Nature Reviews Genetics* **17**, 719–732. See https://doi.org/10.1038/nrg.2016.134.

183. Berezikov E. 2011 Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics* **12**, 846–860. See https://doi.org/10.1038/nrg3079.

184. Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015 Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**. See https://doi.org/10.7554/elife.05005.

185. Lee B, Baek J, Park S, Yoon S. 2016 DeepTarget: End-to-end learning framework for microRNA target prediction using deep recurrent neural networks. See https://arxiv.org/abs/1603.09123v2.

186. Park S, Min S, Choi H, Yoon S. 2016 DeepMiRGene: Deep neural network based precursor microRNA prediction. See https://arxiv.org/abs/1605.00017v1.

187. Wang S, Sun S, Xu J. 2016 AUC-Maximized Deep Convolutional Neural

Fields for Protein Sequence Labeling. In *Machine Learning and Knowledge Discovery in Databases*, pp. 1–16. Springer International Publishing. See https://doi.org/10.1007/978-3-319-46227-1_1.

188. Jones DT, Singh T, Kosciolek T, Tetchner S. 2014 MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006. See https://doi.org/10.1093/bioinformatics/btu791.

189. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2008 Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67–72. See https://doi.org/10.1073/pnas.0805923106.

190. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011 Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE* **6**, e28766. See https://doi.org/10.1371/journal.pone.0028766.

191. Qi Y, Oja M, Weston J, Noble WS. 2012 A Unified Multitask Architecture for Predicting Local Protein Properties. *PLoS ONE* **7**, e32235. See https://doi.org/10.1371/journal.pone.0032235.

192. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports* **5**. See https://doi.org/10.1038/srep11476.

193. Jones DT. 1999 Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195–202. See https://doi.org/10.1006/jmbi.1999.3091.

194. Zhou J, Troyanskaya OG. 2014 Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. See https://arxiv.org/abs/1403.1347v1.

195. Ma J, Wang S, Wang Z, Xu J. 2015 Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* **31**, 3506–3513. See https://doi.org/10.1093/bioinformatics/btv472.

196. Di Lena P, Nagata K, Baldi P. 2012 Deep architectures for protein contact map prediction. *Bioinformatics* **28**, 2449–2457. See https://doi.org/10.1093/bioinformatics/bts475.

197. Eickholt J, Cheng J. 2012 Predicting protein residue-residue contacts

using deep networks and boosting. *Bioinformatics* **28**, 3066–3072. See https://doi.org/10.1093/bioinformatics/bts598.

198. Skwark MJ, Raimondi D, Michel M, Elofsson A. 2014 Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology* **10**, e1003889. See https://doi.org/10.1371/journal.pcbi.1003889.

199. 2017 RR Results - CASP12. See http://www.predictioncenter.org/casp12/rrc_avrg_results.cgi.

200. 2017 CAMEO - Continuous Automated Model Evaluation. See http://www.cameo3d.org/.

201. Li Z, Wang S, Yu Y, Xu J. 2017 Predicting membrane protein contacts from non-membrane proteins by deep transfer learning. See https://arxiv.org/abs/1704.07207v1.

202. Van Valen DA *et al.* 2016 Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLOS Computational Biology* **12**, e1005177. See https://doi.org/10.1371/journal.pcbi.1005177.

203. Ronneberger O, Fischer P, Brox T. 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*, pp. 234–241. Springer International Publishing. See https://doi.org/10.1007/978-3-319-24574-4_28.

204. Buggenthin F *et al.* 2017 Prospective identification of hematopoietic lineage choice by deep learning. *Nature Methods* **14**, 403–406. See https://doi.org/10.1038/nmeth.4182.

205. Eulenberg P, Koehler N, Blasi T, Filby A, Carpenter AE, Rees P, Theis FJ, Wolf FA. 2016 Deep Learning for Imaging Flow Cytometry: Cell Cycle Analysis of Jurkat Cells. See https://doi.org/10.1101/081364.

206. Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. 2016 Automating Morphological Profiling with Generic Deep Convolutional Networks. See https://doi.org/10.1101/085118.

207. Johnson GR, Donovan-Maiye RM, Maleckar MM. 2017 Generative modeling with conditional autoencoders: Building an integrated cell. See https://arxiv.org/abs/1705.00092v1.

208. Caicedo JC, Singh S, Carpenter AE. 2016 Applications in image-based profiling of perturbations. *Current Opinion in Biotechnology* **39**, 134–142. See

https://doi.org/10.1016/j.copbio.2016.04.003.

209. Bougen-Zhukov N, Loh SY, Lee HK, Loo L-H. 2016 Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry Part A* **91**, 115–125. See https://doi.org/10.1002/cyto.a.22909.

210. Grys BT, Lo DS, Sahin N, Kraus OZ, Morris Q, Boone C, Andrews BJ. 2016 Machine learning and computer vision approaches for phenotypic profiling. *The Journal of Cell Biology* **216**, 65–71. See https://doi.org/10.1083/jcb.201610026.

211. Gawad C, Koh W, Quake SR. 2016 Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**, 175–188. See https://doi.org/10.1038/nrg.2015.16.

212. Lodato MA *et al.* 2015 Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98. See https://doi.org/10.1126/science.aab1785.

213. Liu S, Trapnell C. 2016 Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* See https://doi.org/10.12688/f1000research.7223.1.

214. Vera M, Biswas J, Senecal A, Singer RH, Park HY. 2016 Single-Cell and Single-Molecule Analysis of Gene Expression Regulation. *Annual Review of Genetics* **50**, 267–291. See https://doi.org/10.1146/annurev-genet-120215-034854.

215. Clark SJ *et al.* 2017 Joint Profiling Of Chromatin Accessibility, DNA Methylation And Transcription In Single Cells. See https://doi.org/10.1101/138685.

216. Angermueller C, Lee HJ, Reik W, Stegle O. 2017 DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology* **18**. See https://doi.org/10.1186/s13059-017-1189-z.

217. Koh PW, Pierson E, Kundaje A. 2016 Denoising genome-wide histone ChIP-seq with convolutional neural networks. See https://doi.org/10.1101/052118.

218. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. 2016 Removal of batch effects using distribution-matching residual networks. See https://arxiv.org/abs/1610.04181v5.

219. Gaublomme J *et al.* 2015 Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* **163**, 1400–1412. See

https://doi.org/10.1016/j.cell.2015.11.009.

220. Arvaniti E, Claassen M. 2016 Sensitive detection of rare disease-associated cell subsets via representation learning. See https://doi.org/10.1101/046508.

221. He K, Zhang X, Ren S, Sun J. 2015 Deep residual learning for image recognition. See https://arxiv.org/abs/1512.03385v1.

222. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H, Trapnell C. 2017 Reversed graph embedding resolves complex single-cell developmental trajectories. See https://doi.org/10.1101/110668.

223. Silver D *et al.* 2016 Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489. See https://doi.org/10.1038/nature16961.

224. Karlin S, Mrázek J, Campbell AM. 1997 Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology* **179**, 3899–3913. See https://doi.org/10.1128/jb.179.12.3899-3913.1997.

225. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2006 Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4**, 63–72. See https://doi.org/10.1038/nmeth976.

226. Rosen GL, Reichenberger ER, Rosenfeld AM. 2010 NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* **27**, 127–129. See https://doi.org/10.1093/bioinformatics/btq619.

227. Abe T. 2003 Informatics for Unveiling Hidden Genome Signatures. *Genome Research* **13**, 693–702. See https://doi.org/10.1101/gr.634603.

228. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012 Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**, 811–814. See https://doi.org/10.1038/nmeth.2066.

229. Koslicki D, Foucart S, Rosen G. 2014 WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification. *PLoS ONE* **9**, e91784. See https://doi.org/10.1371/journal.pone.0091784.

230. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. 2013 Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* **29**, 2253–2260. See https://doi.org/10.1093/bioinformatics/btt389.

231. Vervier K, Mahé P, Tournoud M, Veyrieras J-B, Vert J-P. 2015 Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* **32**, 1023–1032. See https://doi.org/10.1093/bioinformatics/btv683.

232. Yok NG, Rosen GL. 2011 Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* **12**, 20. See https://doi.org/10.1186/1471-2105-12-20.

233. Soueidan H, Nikolski M. 2017 Machine learning for metagenomics: methods and tools. *Metagenomics* **1**. See https://doi.org/10.1515/metgen-2016-0001.

234. Guetterman H, Auvil L, Russell N, Welge M, Berry M, Gatzke L, Bushell C, Holscher H. 2016 Utilizing Machine Learning Approaches to Understand the Interrelationship of Diet, the Human Gastrointestinal Microbiome, and Health. *The FASEB Journal*. See http://www.fasebj.org/content/30/1_Supplement/406.3.

235. Knights D, Costello EK, Knight R. 2011 Supervised classification of human microbiota. *FEMS Microbiology Reviews* **35**, 343–359. See https://doi.org/10.1111/j.1574-6976.2010.00251.x.

236. Statnikov A *et al.* 2013 A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1**, 11. See https://doi.org/10.1186/2049-2618-1-11.

237. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016 Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology* **12**, e1004977. See https://doi.org/10.1371/journal.pcbi.1004977.

238. Ding X, Cheng F, Cao C, Sun X. 2015 DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC Bioinformatics* **16**. See https://doi.org/10.1186/s12859-015-0753-3.

239. Liu Z, Chen D, Sheng L, Liu AY. 2014 Correction: Class Prediction and Feature Selection with Linear Optimization for Metagenomic Count Data. *PLoS ONE* **9**, e97958. See https://doi.org/10.1371/journal.pone.0097958.

240. Ditzler G, Morrison JC, Lan Y, Rosen GL. 2015 Fizzy: feature subset selection for metagenomics. *BMC Bioinformatics* **16**. See https://doi.org/10.1186/s12859-015-0793-8.

241. Ditzler G, Polikar R, Rosen G. 2015 A Bootstrap Based Neyman-Pearson

Test for Identifying Variable Importance. *IEEE Transactions on Neural Networks and Learning Systems* **26**, 880–886. See https://doi.org/10.1109/tnnls.2014.2320415.

242. Hoff KJ, Lingner T, Meinicke P, Tech M. 2009 Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research* **37**, W101–W105. See https://doi.org/10.1093/nar/gkp327.

243. Rho M, Tang H, Ye Y. 2010 FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research* **38**, e191–e191. See https://doi.org/10.1093/nar/gkq747.

244. Asgari E, Mofrad MRK. 2015 Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE* **10**, e0141287. See https://doi.org/10.1371/journal.pone.0141287.

245. Hochreiter S, Heusel M, Obermayer K. 2007 Fast model-based protein homology detection without alignment. *Bioinformatics* **23**, 1728–1736. See https://doi.org/10.1093/bioinformatics/btm247.

246. Sønderby SK, Sønderby CK, Nielsen H, Winther O. 2015 Convolutional lstm networks for subcellular localization of proteins. See https://arxiv.org/abs/1503.01919v1.

247. Essinger SD, Polikar R, Rosen GL. 2010 Neural network-based taxonomic clustering for metagenomics. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, IEEE. See https://doi.org/10.1109/ijcnn.2010.5596644.

248. Kelley DR, Salzberg SL. 2010 Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* **11**, 544. See https://doi.org/10.1186/1471-2105-11-544.

249. RASHEED Z, RANGWALA H. 2012 METAGENOMIC TAXONOMIC CLASSIFICATION USING EXTREME LEARNING MACHINES. *Journal of Bioinformatics and Computational Biology* **10**, 1250015. See https://doi.org/10.1142/s0219720012500151.

250. Mrzelj N. 2016 Globoko ucenje na genomskih in filogenetskih podatkih. See https://repozitorij.uni-lj.si/IzpisGradiva.php?id=85515.

251. Chudobova D *et al.* 2015 Influence of microbiome species in hard-to-heal wounds on disease severity and treatment duration. *The Brazilian Journal of Infectious Diseases* **19**, 604–613. See https://doi.org/10.1016/j.bjid.2015.08.013.

252. Ditzler G, Polikar R, Rosen G. 2015 Multi-Layer and Recursive Neural Networks for Metagenomic Classification. *IEEE Transactions on NanoBioscience* **14**, 608–616. See https://doi.org/10.1109/tnb.2015.2461219.

253. Faruqi AA. 2016 TensorFlow vs. scikit-learn : The Microbiome Challenge. *Ali A. Faruqi*. See http://alifar76.github.io/sklearn-metrics/.

254. Bengio Y, Boulanger-Lewandowski N, Pascanu R. 2012 Advances in optimizing recurrent networks. See https://arxiv.org/abs/1212.0901v2.

255. Boža V, Brejová B, Vinař T. 2016 DeepNano: Deep recurrent neural networks for base calling in minion nanopore reads. See https://arxiv.org/abs/1603.09195v1.

256. Sutskever I, Vinyals O, Le QV. 2014 Sequence to sequence learning with neural networks. See https://arxiv.org/abs/1409.3215v3.

257. Poplin R, Newburger D, Dijamco J, Nguyen N, Loy D, Gross SS, McLean CY, DePristo MA. 2016 Creating a universal SNP and small indel variant caller with deep neural networks. See https://doi.org/10.1101/092890.

258. Torracinta R, Campagne F. 2016 Training Genotype Callers with Neural Networks. See https://doi.org/10.1101/097469.

259. Chollet F. 2016 Xception: Deep learning with depthwise separable convolutions. See https://arxiv.org/abs/1610.02357v3.

260. Torracinta R, Mesnard L, Levine S, Shaknovich R, Hanson M, Campagne F. 2016 Adaptive Somatic Mutations Calls with Deep Learning and Semi-Simulated Data. See https://doi.org/10.1101/079087.

261. Hamburg MA, Collins FS. 2010 The Path to Personalized Medicine.*New England Journal of Medicine* **363**, 301–304. See https://doi.org/10.1056/nejmp1006304.

262. Belle A, Kon MA, Najarian K. 2013 Biomedical Informatics for Computer-Aided Decision Support Systems: A Survey. *The Scientific World Journal* **2013**, 1–8. See https://doi.org/10.1155/2013/769639.

263. Tu JV. 1996 Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* **49**, 1225–1231. See https://doi.org/10.1016/s0895-4356(96)00002-9.

264. Baxt WG. 1991 Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction. *Annals of Internal Medicine* **115**, 843. See

https://doi.org/10.7326/0003-4819-115-11-843.

265. Wasson JH, Sox HC, Neff RK, Goldman L. 1985 Clinical Prediction Rules. *New England Journal of Medicine* **313**, 793–799. See https://doi.org/10.1056/nejm198509263131306.

266. Lisboa PJ, Taktak AF. 2006 The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks* **19**, 408–415. See https://doi.org/10.1016/j.neunet.2005.10.007.

267. Rubin DB. 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701. See https://doi.org/10.1037/h0037350.

268. Johansson FD, Shalit U, Sontag D. 2016 Learning representations for counterfactual inference. See https://arxiv.org/abs/1605.03661v2.

269. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. 2015 Causal Phenotype Discovery via Deep Networks. *AMIA Annu Symp Proc* **2015**, 677–686. See https://www.ncbi.nlm.nih.gov/pubmed/26958203.

270. Lipton ZC, Kale DC, Wetzel R. 2016 Modeling missing data in clinical time series with rnns. See https://arxiv.org/abs/1606.04130v5.

271. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. 2016 Recurrent neural networks for multivariate time series with missing values. See https://arxiv.org/abs/1606.01865v2.

272. Huddar V, Desiraju BK, Rajan V, Bhattacharya S, Roy S, Reddy CK. 2016 Predicting Complications in Critical Care Using Heterogeneous Clinical Data. *IEEE Access* **4**, 7988–8001. See https://doi.org/10.1109/access.2016.2618775.

273. Lipton ZC, Kale DC, Wetzel RC. 2015 Phenotyping of clinical time series with lstm recurrent neural networks. See https://arxiv.org/abs/1510.07641v2.

274. Nemati S, Ghassemi MM, Clifford GD. 2016 Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE. See https://doi.org/10.1109/embc.2016.7591355.

275. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. 2014 From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association* **21**, 315–325. See https://doi.org/10.1136/amiajnl-2013-001815.

276. Ithapu VK, Singh V, Okonkwo OC, Chappell RJ, Dowling NM, Johnson SC. 2015 Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer's & Dementia* **11**, 1489–1499. See https://doi.org/10.1016/j.jalz.2015.01.010.

277. Artemov AV, Putin E, Vanhaelen Q, Aliper A, Ozerov IV, Zhavoronkov A. 2016 Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. See https://doi.org/10.1101/095653.

278. DiMasi JA, Grabowski HG, Hansen RW. 2016 Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33. See https://doi.org/10.1016/j.jhealeco.2016.01.012.

279. Waring MJ *et al.* 2015 An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **14**, 475–486. See https://doi.org/10.1038/nrd4609.

280. Lamb J. 2006 The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313**, 1929–1935. See https://doi.org/10.1126/science.1132939.

281. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. 2015 A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics* **17**, 2–12. See https://doi.org/10.1093/bib/bbv020.

282. Musa A, Ghoraie LS, Zhang S-D, Galzko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F. 2017 A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics*, bbw112. See https://doi.org/10.1093/bib/bbw112.

283. Brown AS, Patel CJ. 2016 A review of validation strategies for computational drug repositioning. *Brief Bioinform* See https://academic.oup.com/bib/article/doi/10.1093/bib/bbw110/2562646/A-review-of-validation-strategies-for.

284. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D. 2013 Drug Repositioning: A Machine-Learning Approach through Data Integration. *Journal of Cheminformatics* **5**, 30. See https://doi.org/10.1186/1758-2946-5-30.

285. Yang J, Li Z, Fan X, Cheng Y. 2014 Drug–Disease Association and Drug-Repositioning Predictions in Complex Diseases Using Causal Inference–Probabilistic Matrix Factorization. *Journal of Chemical Information and Modeling* **54**, 2562–2569. See https://doi.org/10.1021/ci500340n.

286. Huang C-H, Chang PM-H, Hsu C-W, Huang C-YF, Ng K-L. 2016 Drug

repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory. *BMC Bioinformatics* **17**. See https://doi.org/10.1186/s12859-015-0845-0.

287. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J. 2013 Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE* **8**, e61318. See https://doi.org/10.1371/journal.pone.0061318.

288. VidoviÄ‡ D, Koleti A, SchÃ¼rer SC. 2014 Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. *Frontiers in Genetics* **5**. See https://doi.org/10.3389/fgene.2014.00342.

289. Coelho ED, Arrais JP, Oliveira JL. 2016 Computational Discovery of Putative Leads for Drug Repositioning through Drug-Target Interaction Prediction. *PLOS Computational Biology* **12**, e1005219. See https://doi.org/10.1371/journal.pcbi.1005219.

290. Lim H, Poleksic A, Yao Y, Tong H, He D, Zhuang L, Meng P, Xie L. 2016 Large-Scale Off-Target Identification Using Fast and Accurate Dual Regularized One-Class Collaborative Filtering and Its Application to Drug Repurposing. *PLOS Computational Biology* **12**, e1005135. See https://doi.org/10.1371/journal.pcbi.1005135.

291. Wang C, Liu J, Luo F, Tan Y, Deng Z, Hu Q-N. 2014 Pairwise input neural network for target-ligand interaction prediction. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE. See https://doi.org/10.1109/bibm.2014.6999129.

292. Duan Q *et al.* 2016 L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Systems Biology and Applications* **2**. See https://doi.org/10.1038/npjsba.2016.15.

293. Bleicher KH, Böhm H-J, Müller K, Alanine AI. 2003 A guide to drug discovery: Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery* **2**, 369–378. See https://doi.org/10.1038/nrd1086.

294. Keserű GM, Makara GM. 2006 Hit discovery and hit-to-lead approaches. *Drug Discovery Today* **11**, 741–748. See https://doi.org/10.1016/j.drudis.2006.06.016.

295. Swamidass SJ, Azencott C-A, Lin T-W, Gramajo H, Tsai S-C, Baldi P. 2009 Influence Relevance Voting: An Accurate And Interpretable Virtual High Throughput Screening Method. *Journal of Chemical Information and Modeling*

49, 756–766. See https://doi.org/10.1021/ci8004379.

296. Kearnes S, Goldman B, Pande V. 2016 Modeling industrial admet data with multitask networks. See https://arxiv.org/abs/1606.08793v3.

297. Zaretzki J, Matlock M, Swamidass SJ. 2013 XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *Journal of Chemical Information and Modeling* **53**, 3373–3383. See https://doi.org/10.1021/ci400518g.

298. Todeschini R, Consonni V, editors. 2009 *Molecular Descriptors for Chemoinformatics*. Wiley-VCH Verlag GmbH & Co. KGaA. See https://doi.org/10.1002/9783527628766.

299. Dahl GE, Jaitly N, Salakhutdinov R. 2014 Multi-task neural networks for qsar predictions. See https://arxiv.org/abs/1406.1231v1.

300. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. 2015 Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling* **55**, 263–274. See https://doi.org/10.1021/ci500747n.

301. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner JK, Ceulemans H, Hochreiter S. 2014 Deep learning as an opportunity in virtual screening. *Neural Information Processing Systems 2014: Deep Learning and Representation Learning Workshop* See http://www.dlworkshop.org/23.pdf?attredirects=0.

302. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. 2015 Massively multitask networks for drug discovery. See https://arxiv.org/abs/1502.02072v1.

303. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. 2016 DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **3**. See https://doi.org/10.3389/fenvs.2015.00080.

304. Subramanian G, Ramsundar B, Pande V, Denny RA. 2016 Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *Journal of Chemical Information and Modeling* **56**, 1936–1949. See https://doi.org/10.1021/acs.jcim.6b00290.

305. Reymond J-L, Ruddigkeit L, Blum L, van Deursen R. 2012 The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 717–733. See https://doi.org/10.1002/wcms.1104.

306. Lusci A, Fooshee D, Browning M, Swamidass J, Baldi P. 2015 Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. *Journal of Cheminformatics* **7**. See https://doi.org/10.1186/s13321-015-0110-6.

307. Rogers D, Hahn M. 2010 Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754. See https://doi.org/10.1021/ci100050t.

308. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP. 2015 Convolutional Networks on Graphs for Learning Molecular Fingerprints. See http://papers.nips.cc/paper/5954-convolutional-networks-on-graphs-for-learning-molecular-fingerprints.

309. Lusci A, Pollastri G, Baldi P. 2013 Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *Journal of Chemical Information and Modeling* **53**, 1563–1575. See https://doi.org/10.1021/ci400187y.

310. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. 2016 Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **30**, 595–608. See https://doi.org/10.1007/s10822-016-9938-8.

311. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. 2017 Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* **3**, 283–293. See https://doi.org/10.1021/acscentsci.6b00367.

312. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. 2017 MoleculeNet: A benchmark for molecular machine learning. See https://arxiv.org/abs/1703.00564v1.

313. 2017 deepchem/deepchem. *GitHub*. See https://github.com/deepchem/deepchem.

314. Gómez-Bombarelli R, Duvenaud D, Hernández-Lobato JM, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. 2016 Automatic chemical design using a data-driven continuous representation of molecules. See https://arxiv.org/abs/1610.02415v2.

315. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. 2012 Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal* **14**, 133–141. See https://doi.org/10.1208/s12248-012-9322-0.

316. Gomes J, Ramsundar B, Feinberg EN, Pande VS. 2017 Atomic convolutional networks for predicting protein-ligand binding affinity. See

https://arxiv.org/abs/1703.10603v1.

317. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. 2005 The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry* **48**, 4111–4119. See https://doi.org/10.1021/jm048957q.

318. Pereira JC, Caffarena ER, dos Santos CN. 2016 Boosting Docking-Based Virtual Screening with Deep Learning. *Journal of Chemical Information and Modeling* **56**, 2495–2506. See https://doi.org/10.1021/acs.jcim.6b00355.

319. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. 2016 Protein-ligand scoring with convolutional neural networks. See https://arxiv.org/abs/1612.02751v1.

320. Hartenfeller M, Schneider G. 2011 Enabling future drug discovery by de novo design. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 742–759. See https://doi.org/10.1002/wcms.49.

321. Schneider P, Schneider G. 2016 De Novo Design at the Edge of Chaos. *Journal of Medicinal Chemistry* **59**, 4077–4086. See https://doi.org/10.1021/acs.jmedchem.5b01849.

322. Graves A. 2013 Generating sequences with recurrent neural networks. See https://arxiv.org/abs/1308.0850v5.

323. Segler MHS, Kogej T, Tyrchan C, Waller MP. 2017 Generating focussed molecule libraries for drug discovery with recurrent neural networks. See https://arxiv.org/abs/1701.01329v1.

324. Kusner MJ, Paige B, Hernández-Lobato JM. 2017 Grammar variational autoencoder. See https://arxiv.org/abs/1703.01925v1.

325. Gaulton A *et al.* 2011 ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107. See https://doi.org/10.1093/nar/gkr777.

326. Olivecrona M, Blaschke T, Engkvist O, Chen H. 2017 Molecular de novo design through deep reinforcement learning. See https://arxiv.org/abs/1704.07555v1.

327. Jaques N, Gu S, Bahdanau D, Hernández-Lobato JM, Turner RE, Eck D. 2016 Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. See https://arxiv.org/abs/1611.02796v8.

328. Davis J, Goadrich M. 2006 The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine*

learning - ICML '06, ACM Press. See https://doi.org/10.1145/1143844.1143874.

329. Ba LJ, Caruana R. 2013 Do deep nets really need to be deep? See https://arxiv.org/abs/1312.6184v7.

330. Nguyen A, Yosinski J, Clune J. 2014 Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. See https://arxiv.org/abs/1412.1897v4.

331. Ribeiro MT, Singh S, Guestrin C. 2016 'Why should i trust you?': Explaining the predictions of any classifier. See https://arxiv.org/abs/1602.04938v3.

332. Zeiler MD, Fergus R. 2013 Visualizing and understanding convolutional networks. See https://arxiv.org/abs/1311.2901v3.

333. Zintgraf LM, Cohen TS, Adel T, Welling M. 2017 Visualizing deep neural network decisions: Prediction difference analysis. See https://arxiv.org/abs/1702.04595v1.

334. Fong R, Vedaldi A. 2017 Interpretable explanations of black boxes by meaningful perturbation. See https://arxiv.org/abs/1704.03296v1.

335. Simonyan K, Vedaldi A, Zisserman A. 2013 Deep inside convolutional networks: Visualising image classification models and saliency maps. See https://arxiv.org/abs/1312.6034v2.

336. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. 2015 On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **10**, e0130140. See https://doi.org/10.1371/journal.pone.0130140.

337. Kindermans P-J, Schütt K, Müller K-R, Dähne S. 2016 Investigating the influence of noise and distractors on the interpretation of neural networks. See https://arxiv.org/abs/1611.07270v1.

338. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. 2014 Striving for simplicity: The all convolutional net. See https://arxiv.org/abs/1412.6806v3.

339. Mahendran A, Vedaldi A. 2016 Salient Deconvolutional Networks. In *Computer Vision – ECCV 2016*, pp. 120–135. Springer International Publishing. See https://doi.org/10.1007/978-3-319-46466-4_8.

340. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2016 Grad-cam: Visual explanations from deep networks via gradient-based

localization. See https://arxiv.org/abs/1610.02391v3.

341. Sundararajan M, Taly A, Yan Q. 2017 Axiomatic attribution for deep networks. See https://arxiv.org/abs/1703.01365v1.

342. Lundberg S, Lee S-I. 2016 An unexpected unity among methods for interpreting model predictions. See https://arxiv.org/abs/1611.07478v3.

343. Shapley LS. 1953 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, Princeton University Press. See https://doi.org/10.1515/9781400881970-018.

344. Mahendran A, Vedaldi A. 2014 Understanding deep image representations by inverting them. See https://arxiv.org/abs/1412.0035v1.

345. Finnegan AI, Song JS. 2017 Maximum Entropy Methods for Extracting the Learned Features of Deep Neural Networks. See https://doi.org/10.1101/105957.

346. Mahendran A, Vedaldi A. 2016 Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. *International Journal of Computer Vision* **120**, 233–255. See https://doi.org/10.1007/s11263-016-0911-8.

347. Mordvintsev A, Olah C, Tyka M. 2015 Inceptionism: Going Deeper into Neural Networks. *Google Research Blog*. See http://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html.

348. Erhan D, Bengio Y, Courville A, Vincent P. 2009 Visualizing Higher-Layer Features of a Deep Network. See http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/247.

349. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. 2015 Understanding neural networks through deep visualization. See https://arxiv.org/abs/1506.06579v1.

350. Bahdanau D, Cho K, Bengio Y. 2014 Neural machine translation by jointly learning to align and translate. See https://arxiv.org/abs/1409.0473v7.

351. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y. 2015 Show, attend and tell: Neural image caption generation with visual attention. See https://arxiv.org/abs/1502.03044v3.

352. Deming L, Targ S, Sauder N, Almeida D, Ye CJ. 2016 Genetic architect: Discovering genomic structure with learned neural architectures. See https://arxiv.org/abs/1605.07156v1.

353. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. 2016 RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. See https://arxiv.org/abs/1608.05745v4.

354. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. 2016 GRAM: Graph-based attention model for healthcare representation learning. See https://arxiv.org/abs/1611.07012v3.

355. Ghosh J, Karamcheti V. 1992 Sequence learning with recurrent networks: analysis of internal representations. In *Science of Artificial Neural Networks* (ed DW Ruck), SPIE. See https://doi.org/10.1117/12.140112.

356. Karpathy A, Johnson J, Fei-Fei L. 2015 Visualizing and understanding recurrent networks. See https://arxiv.org/abs/1506.02078v2.

357. Strobelt H, Gehrmann S, Huber B, Pfister H, Rush AM. 2016 Visual analysis of hidden state dynamics in recurrent neural networks. See https://arxiv.org/abs/1606.07461v1.

358. Murdoch WJ, Szlam A. 2017 Automatic rule extraction from long short term memory networks. See https://arxiv.org/abs/1702.02540v2.

359. Chryssolouris G, Lee M, Ramsey A. 1996 Confidence interval prediction for neural network models. *IEEE Transactions on Neural Networks* **7**, 229–232. See https://doi.org/10.1109/72.478409.

360. Gal Y, Ghahramani Z. 2015 Dropout as a bayesian approximation: Representing model uncertainty in deep learning. See https://arxiv.org/abs/1506.02142v6.

361. Koh PW, Liang P. 2017 Understanding black-box predictions via influence functions. See https://arxiv.org/abs/1703.04730v1.

362. Kahng M, Andrews P, Kalro A, Chau DH. 2017 ActiVis: Visual exploration of industry-scale deep neural network models. See https://arxiv.org/abs/1704.01942v1.

363. Liu M, Shi J, Li Z, Li C, Zhu J, Liu S. 2016 Towards better analysis of deep convolutional neural networks. See https://arxiv.org/abs/1604.07043v3.

364. Che Z, Purushotham S, Khemani R, Liu Y. 2015 Distilling knowledge from deep networks with applications to healthcare domain. See https://arxiv.org/abs/1512.03542v1.

365. Lei T, Barzilay R, Jaakkola T. 2016 Rationalizing neural predictions. See https://arxiv.org/abs/1606.04155v2.

366. Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG. 2013 Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes. *PLoS Computational Biology* **9**, e1002957. See https://doi.org/10.1371/journal.pcbi.1002957.

367. Sarraf S, DeSouza DD, Anderson J, Tofighi G,. 2016 DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. See https://doi.org/10.1101/070441.

368. Shao M, Ma J, Wang S. 2017 DeepBound: Accurate Identification of Transcript Boundaries via Deep Convolutional Neural Fields. See https://doi.org/10.1101/125229.

369. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014 A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310–315. See https://doi.org/10.1038/ng.2892.

370. Rubinsteyn A, O'Donnell T, Damaraju N, Hammerbacher J. 2016 Predicting Peptide-MHC Binding Affinities With Imputed Training Data. See https://doi.org/10.1101/054775.

371. Romero A *et al.* 2016 Diet Networks: Thin Parameters for Fat Genomics. *ICLR 2017* See https://openreview.net/forum?id=Sk-oDY9ge&noteId=Sk-oDY9ge.

372. Schmidhuber J. 2015 Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117. See https://doi.org/10.1016/j.neunet.2014.09.003.

373. Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. 2015 Deep learning with limited numerical precision. See https://arxiv.org/abs/1502.02551v1.

374. Courbariaux M, Bengio Y, David J-P. 2014 Training deep neural networks with low precision multiplications. See https://arxiv.org/abs/1412.7024v5.

375. Sa CD, Zhang C, Olukotun K, Ré C. 2015 Taming the wild: A unified analysis of hogwild!-style algorithms. See https://arxiv.org/abs/1506.06438v2.

376. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. 2016 Quantized neural networks: Training neural networks with low precision weights and activations. See https://arxiv.org/abs/1609.07061v1.

377. Hinton G, Vinyals O, Dean J. 2015 Distilling the knowledge in a neural network. See https://arxiv.org/abs/1503.02531v1.

378. Raina R, Madhavan A, Ng AY. 2009 Large-scale deep unsupervised

learning using graphics processors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, ACM Press. See https://doi.org/10.1145/1553374.1553486.

379. Vanhoucke V, Senior A, Mao MZ. 2011 Improving the speed of neural networks on CPUs. See https://research.google.com/pubs/pub37631.html.

380. Seide F, Fu H, Droppo J, Li G, Yu D. 2014 On parallelizability of stochastic gradient descent for speech DNNS. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. See https://doi.org/10.1109/icassp.2014.6853593.

381. Hadjis S, Abuzaid F, Zhang C, Ré C. 2015 Caffe con troll: Shallow ideas to speed up deep learning. See https://arxiv.org/abs/1504.04343v2.

382. Edwards C. 2015 Growing pains for deep learning. *Communications of the ACM* **58**, 14–16. See https://doi.org/10.1145/2771283.

383. Su H, Chen H. 2015 Experiments on parallel training of deep neural network using model averaging. See https://arxiv.org/abs/1507.01239v2.

384. Li M, Zhang T, Chen Y, Smola AJ. 2014 Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, ACM Press. See https://doi.org/10.1145/2623330.2623612.

385. Hamanaka M, Taneishi K, Iwata H, Ye J, Pei J, Hou J, Okuno Y. 2016 CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Molecular Informatics* **36**, 1600045. See https://doi.org/10.1002/minf.201600045.

386. Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, Shelhamer E. 2014 CuDNN: Efficient primitives for deep learning. See https://arxiv.org/abs/1410.0759v3.

387. Chen W, Wilson JT, Tyree S, Weinberger KQ, Chen Y. 2015 Compressing neural networks with the hashing trick. See https://arxiv.org/abs/1504.04788v1.

388. Lacey G, Taylor GW, Areibi S. 2016 Deep learning on fpgas: Past, present, and future. See https://arxiv.org/abs/1602.04283v1.

389. Jouppi NP *et al.* 2017 In-datacenter performance analysis of a tensor processing unit. See https://arxiv.org/abs/1704.04760v1.

390. Dean J, Ghemawat S. 2008 MapReduce. *Communications of the ACM*

51, 107. See https://doi.org/10.1145/1327452.1327492.

391. Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A, Hellerstein JM. 2012 Distributed GraphLab. *Proceedings of the VLDB Endowment***5**, 716–727. See https://doi.org/10.14778/2212351.2212354.

392. Dean J *et al.* 2012 Large Scale Distributed Deep Networks. *Neural Information Processing Systems 2012*. See http://research.google.com/archive/large_deep_networks_nips2012.html.

393. De Sa C, Zhang C, Olukotun K, Ré C. 2015 Taming the Wild: A Unified Analysis of Hogwild!-Style Algorithms. *Adv Neural Inf Process Syst***28**, 2656–2664. See https://www.ncbi.nlm.nih.gov/pubmed/27330264.

394. Moritz P, Nishihara R, Stoica I, Jordan MI. 2015 SparkNet: Training deep networks in spark. See https://arxiv.org/abs/1511.06051v4.

395. Meng X *et al.* 2015 MLlib: Machine learning in apache spark. See https://arxiv.org/abs/1505.06807v1.

396. Abadi M *et al.* 2016 TensorFlow: Large-scale machine learning on heterogeneous distributed systems. See https://arxiv.org/abs/1603.04467v2.

397. 2017 fchollet/keras. *GitHub*. See https://github.com/fchollet/keras.

398. 2017 maxpumperla/elephas. *GitHub*. See https://github.com/maxpumperla/elephas.

399. Coates A, Huval B, Wang T, Wu D, Catanzaro B, Andrew N. 2013 Deep learning with COTS HPC systems. See http://www.jmlr.org/proceedings/papers/v28/coates13.html.

400. Sun S, Chen W, Liu T-Y. 2016 Ensemble-compression: A new method for parallel training of deep neural networks. See https://arxiv.org/abs/1606.00575v1.

401. Bergstra J, Bardenet R, Bengio Y, Kégl B. 2011 Algorithms for Hyper-parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 2546–2554. USA: Curran Associates Inc. See http://dl.acm.org/citation.cfm?id=2986459.2986743.

402. Bergstra J, Bengio Y. 2012 Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* See http://www.jmlr.org/papers/v13/bergstra12a.html.

403. Schatz MC, Langmead B, Salzberg SL. 2010 Cloud computing and the DNA data race. *Nature Biotechnology***28**, 691–693. See

https://doi.org/10.1038/nbt0710-691.

404. Muir P *et al.* 2016 The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* **17**. See https://doi.org/10.1186/s13059-016-0917-0.

405. Stein LD. 2010 The case for cloud computing in genome informatics. *Genome Biology* **11**, 207. See https://doi.org/10.1186/gb-2010-11-5-207.

406. Krizhevsky A. 2014 One weird trick for parallelizing convolutional neural networks. See https://arxiv.org/abs/1404.5997v2.

407. Armbrust M *et al.* 2010 A view of cloud computing. *Communications of the ACM* **53**, 50. See https://doi.org/10.1145/1721654.1721672.

408. Longo DL, Drazen JM. 2016 Data Sharing. *New England Journal of Medicine* **374**, 276–277. See https://doi.org/10.1056/nejme1516564.

409. Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. 2017 Celebrating parasites. *Nature Genetics* **49**, 483–484. See https://doi.org/10.1038/ng.3830.

410. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JPA, Taufer M. 2016 Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241. See https://doi.org/10.1126/science.aah6168.

411. 2016 DragoNN. See http://kundajelab.github.io/dragonn/.

412. 2017 ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. See https://www.synapse.org/#!Synapse:syn6131484/wiki/402026.

413. Yosinski J, Clune J, Bengio Y, Lipson H. 2014 How transferable are features in deep neural networks? See https://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.

414. Zhang W, Li R, Zeng T, Sun Q, Kumar S, Ye J, Ji S. 2015 Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, ACM Press. See https://doi.org/10.1145/2783258.2783304.

415. Zeng T, Li R, Mukkamala R, Ye J, Ji S. 2015 Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics* **16**. See https://doi.org/10.1186/s12859-015-0553-9.

416. Pärnamaa T, Parts L. 2017 Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3: Genes|Genomes|Genetics* **7**, 1385–1392. See https://doi.org/10.1534/g3.116.033654.

417. Kraus OZ, Grys BT, Ba J, Chong Y, Frey BJ, Boone C, Andrews BJ. 2017 Automated analysis of high-content microscopy data with deep learning. *Molecular Systems Biology* **13**, 924. See https://doi.org/10.15252/msb.20177551.

418. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. 2011 Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning*, See https://ccrma.stanford.edu/~juhan/pubs/NgiamKhoslaKimNamLeeNg2011.pdf.

419. Chaudhary K, Poirion OB, Lu L, Garmire L. 2017 Deep Learning based multi-omics integration robustly predicts survival in liver cancer. See https://doi.org/10.1101/114892.

420. Eser U, Churchman LS. 2016 FIDDLE: An integrative deep learning framework for functional genomic data inference. See https://doi.org/10.1101/081380.

421. Hughes TB, Dang NL, Miller GP, Swamidass SJ. 2016 Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. *ACS Central Science* **2**, 529–537. See https://doi.org/10.1021/acscentsci.6b00162.

422. BI Intelligence. 2017 IBM edges closer to human speech recognition. *Business Insider*. See http://www.businessinsider.com/ibm-edges-closer-to-human-speech-recognition-2017-3.

423. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G. 2016 Achieving human parity in conversational speech recognition. See https://arxiv.org/abs/1610.05256v2.

424. Saon G *et al.* 2017 English conversational telephone speech recognition by humans and machines. See https://arxiv.org/abs/1703.02136v1.

425. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. 2013 Intriguing properties of neural networks. See https://arxiv.org/abs/1312.6199v4.

426. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and harnessing adversarial examples. See https://arxiv.org/abs/1412.6572v3.

427. Papernot N, McDaniel P, Sinha A, Wellman M. 2016 Towards the science

of security and privacy in machine learning. See https://arxiv.org/abs/1611.03814v1.

428. Xu W, Evans D, Qi Y. 2017 Feature squeezing: Detecting adversarial examples in deep neural networks. See https://arxiv.org/abs/1704.01155v1.

429. Carlisle BG. 2014 The Grey Literature — Proof of prespecified endpoints in medical research with the bitcoin blockchain. See https://www.bgcarlisle.com/blog/2014/08/25/proof-of-prespecified-endpoints-in-medical-research-with-the-bitcoin-blockchain/.

430. Himmelstein D. 2017 The most interesting case of scientific irreproducibility? *Satoshi Village*. See http://blog.dhimmel.com/irreproducible-timestamps/.

431. 2017 OpenTimestamps: a timestamping proof standard. See https://opentimestamps.org/.

432. 2017 greenelab/deep-review. *GitHub*. See https://github.com/greenelab/deep-review.