4-1-2018

Opportunities and obstacles for deep learning in biology and medicine

Travers Ching

S Joshua Swamidass

et al

INTERFACE

rsif.royalsocietypublishing.org

Headline review

Cite this article: Ching T *et al*. 2018 Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**: 20170387.
http://dx.doi.org/10.1098/rsif.2017.0387

Received: 26 May 2017
Accepted: 7 March 2018

Subject Category:
Reviews

Subject Areas:
bioinformatics, computational biology

Keywords:
deep learning, genomics, precision medicine, machine learning

Authors for correspondence:
Anthony Gitter
e-mail: gitter@biostat.wisc.edu
Casey S. Greene
e-mail: greenescientist@gmail.com

†Author order was determined with a randomized algorithm.

THE ROYAL SOCIETY PUBLISHING

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], Michael Zietz[2], Michael M. Hoffman[8,9,10], Wei Xie[11], Gail L. Rosen[12], Benjamin J. Lengerich[13], Johnny Israeli[14], Jack Lanchantin[17], Stephen Woloszynek[12], Anne E. Carpenter[18], Avanti Shrikumar[15], Jinbo Xu[19], Evan M. Cofer[20,21], Christopher A. Lavender[22], Srinivas C. Turaga[23], Amr M. Alexandari[15], Zhiyong Lu[24], David J. Harris[25], Dave DeCaprio[26], Yanjun Qi[17], Anshul Kundaje[15,16], Yifan Peng[24], Laura K. Wiley[27], Marwin H. S. Segler[28], Simina M. Boca[29], S. Joshua Swamidass[30], Austin Huang[31], Anthony Gitter[32,33] and Casey S. Greene[2]

[1]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI, USA
[2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, and
[3]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
[5]Harvard Medical School, Boston, MA, USA
[6]Computational Biology and Stats, Target Sciences, GlaxoSmithKline, Stevenage, UK
[7]Data Science Institute, Imperial College London, London, UK
[8]Princess Margaret Cancer Centre, Toronto, Ontario, Canada
[9]Department of Medical Biophysics and [10]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[11]Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA
[12]Ecological and Evolutionary Signal-processing and Informatics Laboratory, Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA
[13]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[14]Biophysics Program, [15]Department of Computer Science, and [16]Department of Genetics, Stanford University, Stanford, CA, USA
[17]Department of Computer Science, University of Virginia, Charlottesville, VA, USA
[18]Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[19]Toyota Technological Institute at Chicago, Chicago, IL, USA
[20]Department of Computer Science, Trinity University, San Antonio, TX, USA
[21]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
[22]Integrative Bioinformatics, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA
[23]Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA, USA
[24]National Center for Biotechnology Information and National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[25]Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL, USA
[26]ClosedLoop.ai, Austin, TX, USA
[27]Division of Biomedical Informatics and Personalized Medicine, University of Colorado School of Medicine, Aurora, CO, USA
[28]Institute of Organic Chemistry, Westfälische Wilhelms-Universität Münster, Münster, Germany
[29]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA
[30]Department of Pathology and Immunology, Washington University in Saint Louis, St Louis, MO, USA
[31]Department of Medicine, Brown University, Providence, RI, USA
[32]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
[33]Morgridge Institute for Research, Madison, WI, USA

TC, 0000-0002-5577-3516; DSH, 0000-0002-3012-7446;
BKB, 0000-0002-6700-1468; AAK, 0000-0003-4563-3226;
BTD, 0000-0003-4992-2623; GPW, 0000-0002-0503-9348;
EF, 0000-0002-8362-100X; P-MA, 0000-0003-1126-1479;
MZ, 0000-0003-0539-630X; MMH, 0000-0002-4517-1562;
WX, 0000-0002-1871-6846; GLR, 0000-0003-1763-5750;
BJL, 0000-0001-8690-9554; JI, 0000-0003-1633-5780; JL, 0000-0003-0811-0944;
SW, 0000-0003-0568-298X; AEC, 0000-0003-1555-8261;
AS, 0000-0002-6443-4671; JX, 0000-0001-7111-4839;
EMC, 0000-0003-3877-0433; CAL, 0000-0002-7762-1089;
SCT, 0000-0003-3247-6487; AMA, 0000-0001-8655-8109;
ZL, 0000-0001-9998-916X; DJH, 0000-0003-3332-9307;
DD, 0000-0001-8931-9461; YQ, 0000-0002-5796-7453; AK, 0000-0003-3084-2287;
YP, 0000-0001-9309-8331; LKW, 0000-0001-6681-9754;
MHSS, 0000-0001-8008-0546; SMB, 0000-0002-1400-3398;
SJS, 0000-0003-2191-0778; AH, 0000-0003-1349-4030;
AG, 0000-0002-5324-9833; CSG, 0000-0001-8713-9213

Deep learning describes a class of machine learning algorithms that are capable of combining raw inputs into layers of intermediate features. These algorithms have recently shown impressive results across a variety of domains. Biology and medicine are data-rich disciplines, but the data are complex and often ill-understood. Hence, deep learning techniques may be particularly well suited to solve problems of these fields. We examine applications of deep learning to a variety of biomedical problems—patient classification, fundamental biological processes and treatment of patients—and discuss whether deep learning will be able to transform these tasks or if the biomedical sphere poses unique challenges. Following from an extensive literature review, we find that deep learning has yet to revolutionize biomedicine or definitively resolve any of the most pressing challenges in the field, but promising advances have been made on the prior state of the art. Even though improvements over previous baselines have been modest in general, the recent progress indicates that deep learning methods will provide valuable means for speeding up or aiding human investigation. Though progress has been made linking a specific neural network's prediction to input features, understanding how users should interpret these models to make testable hypotheses about the system under study remains an open challenge. Furthermore, the limited amount of labelled data for training presents problems in some domains, as do legal and privacy constraints on work with sensitive health records. Nonetheless, we foresee deep learning enabling changes at both bench and bedside with the potential to transform several areas of biology and medicine.

1. 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 5 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], computational chemistry [4], dermatology [5] and translation among written languages [6]. 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.

Deep learning approaches grew from research on artificial neurons, which were first proposed in 1943 [7] as a model for how the neurons in a biological brain process information. The history of artificial neural networks—referred to as 'neural networks' throughout this article—is interesting in its own right [8]. In neural networks, inputs are fed into the input layer, which feeds into one or more hidden layers, which eventually link to an output layer. A layer consists of a set of nodes, sometimes called 'features' or 'units', which are connected via edges to the immediately earlier and the immediately deeper layers. In some special neural network architectures, nodes can connect to themselves with a delay. The nodes of the input layer generally consist of the variables being measured in the dataset of interest—for example, each node could represent the intensity value of a specific pixel in an image or the expression level of a gene in a specific transcriptomic experiment. 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 machine learning approaches. In particular, deep learning approaches can be used both in *supervised* applications—where the goal is to accurately predict one or more labels or outcomes associated with each data point—in the place of regression approaches, as well as in *unsupervised*, or 'exploratory' applications—where the goal is to summarize, explain or identify interesting patterns in a dataset—as a form of clustering. Deep learning methods may, in fact, combine both of these steps. When sufficient data are available and labelled, these methods construct features tuned to a specific problem and combine those features into a predictor. In fact, if the dataset is 'labelled' with binary classes, a simple neural network with no hidden layers and no cycles between units is equivalent to logistic regression if the output layer is a sigmoid (logistic) function of the input layer. Similarly, for continuous outcomes, linear regression can be seen as a single-layer neural network. Thus, in some ways, supervised deep learning approaches can be seen as an extension of regression models that allow for greater flexibility and are especially well suited for modelling nonlinear relationships among the input features. Recently, hardware improvements and very large training datasets have allowed these deep learning techniques to surpass other machine learning algorithms for many problems. 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 [9] without being told to look for them. What if, more

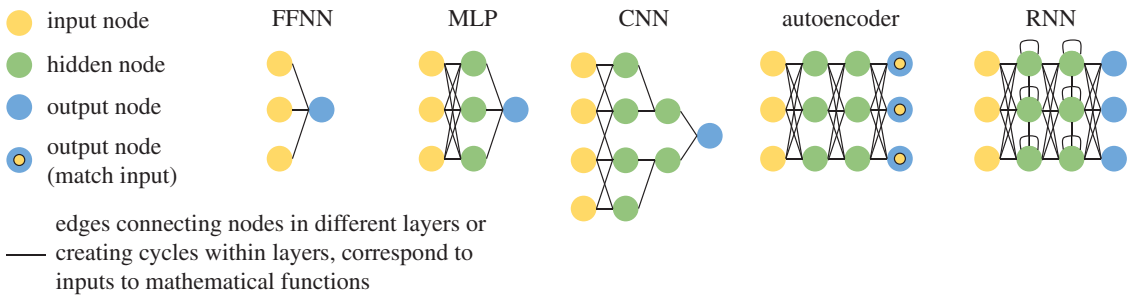

Figure 1. Neural networks come in many different forms. Left: A key for the various types of nodes used in neural networks. Simple FFNN: a feed-forward neural network in which inputs are connected via some function to an output node and the model is trained to produce some output for a set of inputs. MLP: the multi-layer perceptron is a feed-forward neural network in which there is at least one hidden layer between the input and output nodes. CNN: the convolutional neural network is a feed-forward neural network in which the inputs are grouped spatially into hidden nodes. In the case of this example, each input node is only connected to hidden nodes alongside their neighbouring input node. Autoencoder: a type of MLP in which the neural network is trained to produce an output that matches the input to the network. RNN: a deep recurrent neural network is used to allow the neural network to retain memory over time or sequential inputs. This figure was inspired by the Neural Network Zoo by Fjodor Van Veen.

generally, deep learning takes advantage of the growth of data in biomedicine to tackle challenges in this field? 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 [10]. Dropout of nodes, edges and layers makes networks more robust, even when the number of parameters is very large. 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.

This review discusses recent work in the biomedical domain, and most successful applications select neural network architectures that are well suited to the problem at hand. We sketch out a few simple example architectures in figure 1. If data have a natural adjacency structure, a convolutional neural network (CNN) can take advantage of that structure by emphasizing local relationships, especially when convolutional layers are used in early layers of the neural network. Other neural network architectures such as autoencoders 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; an overview of the principal terms used herein is given in table 1. Table 1 also provides select example applications, though in practice each neural network architecture has been broadly applied across multiple types of biomedical data. A recent book from Goodfellow *et al.* [11] covers neural network architectures in detail, and LeCun *et al.* [2] provide a more general introduction.

While deep learning shows increased flexibility over other machine learning approaches, as seen in the remainder of this review, it requires large training sets in order to fit the hidden layers, as well as accurate labels for the supervised learning applications. For these reasons, deep learning has recently become popular in some areas of biology and medicine, while having lower adoption in other areas. At the same time, this highlights the potentially even larger role that it may play in future research, given the increases in data in all biomedical fields. It is also important to see it as a branch of machine learning and acknowledge that it has the same limitations as other approaches in that field. In particular, the results are still dependent on the underlying study design and the usual caveats of correlation versus causation still apply—a more precise answer is only better than a less precise one if it answers the correct question.

1.1. 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'. Grove [12], 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. 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 [13–17], healthcare [18–20] and drug discovery [4,21–23]. 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 judgement. We have divided the large range of topics into three broad classes: disease and patient categorization, fundamental biological study and

Table 1. Glossary.

term	definition	example applications
supervised learning	machine learning approaches with goal of prediction of labels or outcomes	
unsupervised learning	machine learning approaches with goal of data summarization or pattern identification	
neural network (NN)	machine learning approach inspired by biological neurons where inputs are fed into one or more layers, producing an output layer	
deep neural network	NN with multiple hidden layers. Training happens over the network, and consequently such architectures allow for feature construction to occur alongside optimization of the overall training objective	
feed-forward neural network (FFNN)	NN that does not have cycles between nodes in the same layer	most of the examples below are special cases of FFNNs, except recurrent neural networks
MLP	type of FFNN with at least one hidden layer where each deeper layer is a nonlinear function of each earlier layer	MLPs do not impose structure and are frequently used when there is no natural ordering of the inputs (e.g. as with gene expression measurements)
CNN	an NN with layers in which connectivity preserves local structure. *If the data meet the underlying assumptions* performance is often good, and such networks can require fewer examples to train effectively because they have fewer parameters and also provide improved efficiency	CNNs are used for sequence data—such as DNA sequences—or grid data—such as medical and microscopy images
recurrent neural network (RNN)	a neural network with cycles between nodes within a hidden layer.	the RNN architecture is used for sequential data—such as clinical time series and text or genome sequences
LSTM neural network	this special type of RNN has features that enable models to capture longer-term dependencies	LSTMs are gaining a substantial foothold in the analysis of natural language, and may become more widely applied to biological sequence data
autoencoder (AE)	an NN where the training objective is to minimize the error between the output layer and the input layer. Such neural networks are unsupervised and are often used for dimensionality reduction	autoencoders have been used for unsupervised analysis of gene expression data as well as data extracted from the EHR
variational autoencoder (VAE)	this special type of generative AE learns a probabilistic latent variable model	VAEs have been shown to often produce meaningful reduced representations in the imaging domain, and some early publications have used VAEs to analyse gene expression data
denoising autoencoder (DA)	this special type of AE includes a step where noise is added to the input during the training process. The denoising step acts as smoothing and may allow for effective use on input data that is inherently noisy	like AEs, DAs have been used for unsupervised analysis of gene expression data as well as data extracted from the EHR
generative neural network	neural networks that fall into this class can be used to generate data similar to input data. These models can be sampled to produce hypothetical examples	a number of the unsupervised learning neural network architectures that are summarized here can be used in a generative fashion
RBM	a generative NN that forms the building block for many deep learning approaches, having a single input layer and a single hidden layer, with no connections between the nodes within each layer	RBMs have been applied to combine multiple types of omic data (e.g. DNA methylation, mRNA expression and miRNA expression)

(*Continued.*)

Table 1. (*Continued.*)

term	definition	example applications
DBN	generative NN with several hidden layers, which can be obtained from combining multiple RBMs	DBNs can be used to predict new relationships in a drug–target interaction network
generative adversarial network (GAN)	a generative NN approach where two neural networks are trained. One neural network, the generator, is provided with a set of randomly generated inputs and tasked with generating samples. The second, the discriminator, is trained to differentiate real and generated samples. After the two neural networks are trained against each other, the resulting generator can be used to produce new examples	GANs can synthesize new examples with the same statistical properties of datasets that contain individual-level records and are subject to sharing restrictions. They have also been applied to generate microscopy images
adversarial training	a process by which artificial training examples are maliciously designed to fool an NN and then input as training examples to make the resulting NN robust (no relation to GANs)	adversarial training has been used in image analysis
data augmentation	a process by which transformations that do not affect relevant properties of the input data (e.g. arbitrary rotations of histopathology images) are applied to training examples to increase the size of the training set	data augmentation is widely used in the analysis of images because rotation transformations for biomedical images often do not change relevant properties of the image

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.

1.1.1. 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 [24,25]. 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: for instance, denoising autoencoders, an unsupervised approach, can be used to cluster breast cancer patients [26], and CNNs can help count mitotic divisions, a feature that is highly correlated with disease outcome in histological images [27]. 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).

1.1.2. 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 (miRNAs) [28], and CNNs have been applied to predict protein residue–residue contacts and secondary structure [29–31]. Other recent exciting applications of deep learning include recognition of functional genomic elements such as enhancers and promoters [32–34] and prediction of the deleterious effects of nucleotide polymorphisms [35].

1.1.3. 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 [36]. Drug repositioning using deep learning on transcriptomic data is another exciting area of research [37]. Restricted Boltzmann machines (RBMs) can be combined into deep belief networks (DBNs) to predict novel drug–target interactions and formulate drug repositioning hypotheses [38,39]. Finally, deep learning is also prioritizing chemicals in the early stages of drug discovery for new targets [23].

2. Deep learning and patient categorization

In healthcare, 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 and *ad hoc*, potentially impeding the identification of underlying biological mechanisms and their corresponding treatment interventions.

Deep learning methods applied to a large corpus of patient phenotypes may provide a meaningful and more data-driven approach to patient categorization. For example, they may identify new shared mechanisms that would otherwise be obscured due to *ad hoc* historical definitions of disease. Perhaps deep neural networks, by reevaluating data without the context of our assumptions, can reveal novel classes of treatable conditions.

In spite of such optimism, the ability of deep learning models to indiscriminately extract predictive signals must also be assessed and operationalized with care. Imagine a deep neural network is provided with clinical test results gleaned from EHRs. 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 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 labelling.

2.1. 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 [19,40].

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.

Data augmentation provides an effective strategy for working with small training sets. The practice is exemplified by a series of papers that analyse images from mammographies [41–45]. To expand the number and diversity of images, researchers constructed adversarial [44] or augmented [45] examples. Adversarial training examples are constructed by selecting targeted small transformations to input data that cause a model to produce very different outputs. Augmented training applies perturbations to the input data that do not change the underlying meaning, such as rotations for pathology images. An alternative in the domain is to train towards human-created features before subsequent fine-tuning [42], which can help to sidestep this challenge though it does give up deep learning techniques' strength as feature constructors.

A second strategy repurposes features extracted from natural images by deep learning models, such as ImageNet [46], for new purposes. Diagnosing diabetic retinopathy through colour fundus images became an area of focus for deep learning researchers after a large labelled image set was made publicly available during a 2015 Kaggle competition [47]. Most participants trained neural networks from scratch [47–49], but Gulshan *et al*. [50] 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 [51,52] and non-dermoscopic images of skin lesions [5,53,54] as well as age-related macular degeneration [55]. 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 [5,51]. Reusing features from natural images is also an emerging approach for radiographic images, where datasets are often too small to train large deep neural networks without these techniques [56–59]. A deep CNN trained on natural images boosts performance in radiographic images [58]. However, the target task required either re-training the initial model from scratch with special preprocessing 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 have 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 is also faced with the challenge of small training sets. In this domain, Amit *et al*. [60] investigated the trade-off 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 a few hundred images from a few dozen patients can outperform a pre-trained out-of-domain classifier.

Another way of dealing with limited training data is to divide rich data—e.g. 3D images—into numerous reduced projections. Shin *et al*. [57] compared various deep network architectures, dataset characteristics and training procedures for computer tomography (CT)-based abnormality detection. They concluded that networks as deep as 22 layers could be useful for 3D data, despite the limited size of training datasets. 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*. [61] 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 analogues, while requiring much less training time, especially on augmented training sets. Another advantage of 2D and 2.5D networks is the wider availability of pre-trained models. However, reducing the dimensionality is not always helpful. Nie *et al*. [62] showed that multimodal, multi-channel 3D deep architecture was successful at learning high-level brain tumour 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*. [63] demonstrated that deep neural networks outperform a 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*. [49] 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 physician–computer interactions and improve uptake by physicians.

Systems to aid in the analysis of histology slides are also promising use cases for deep learning [64]. Ciresan *et al*. [27] 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*. [65] analysed 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. Combining pre-trained deep network architectures with multiple augmentation techniques enabled accurate detection of breast cancer from a very small set of histology images with less than 100 images per class [66]. 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 phenotypical annotations is the EHR. Billing information in the form of ICD codes are simple annotations but phenotypic algorithms can combine laboratory tests, medication prescriptions and patient notes to generate more reliable phenotypes. 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 EHRs, 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.

Rich clinical information is stored in EHRs. However, manually annotating a large set requires experts and is time-consuming. For chest X-ray studies, a radiologist usually spends a few minutes per example. Generating the number of examples needed for deep learning is infeasibly expensive. Instead, researchers may benefit from using text mining to generate annotations [69], even if those annotations are of modest accuracy. Wang *et al*. [70] proposed to build predictive deep neural network models through the use of images with *weak labels*. Such labels are automatically generated and not verified by humans, so they may be noisy or incomplete. In this case, they applied a series of natural language processing (NLP) techniques to the associated chest X-ray radiological reports. They first extracted all diseases mentioned in the reports using a state-of-the-art NLP tool, then applied a new method, NegBio [71], to filter negative and equivocal findings in the reports. Evaluation of four independent datasets demonstrated that NegBio is highly accurate for detecting negative and equivocal findings (approx. 90% in the F_1 score, which balances precision and recall [72]). The resulting dataset [73] consisted of 112 120 frontal-view chest X-ray images from 30 805 patients, and each image was associated with one or more *text-mined* (weakly labelled) pathology categories (e.g. pneumonia and cardiomegaly) or 'no finding' 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 labelled data.

Another example of semi-automated label generation for hand radiograph segmentation employed positive mining, an iterative procedure that combines manual labelling with automatic processing [74]. First, the initial training set was created by manually labelling 100 of 12 600 unlabelled radiographs that were used to train a model and predict labels for the rest of the dataset. Then, poor-quality predictions were discarded through manual inspection, the initial training set was expanded with the acceptable segmentations, and the process was repeated. This procedure had to be repeated six times to obtain good quality segmentation labelling for all radiographs, except for 100 corner cases that still required manual annotation. These annotations allowed accurate segmentation of all hand images in the test set and boosted the final performance in radiograph classification [74].

With the exception of natural image-like problems (e.g. melanoma detection), biomedical imaging poses a number of challenges for deep learning. Datasets are typically small, annotations can be sparse, and images are often high-dimensional, multimodal and multi-channel. Techniques like transfer learning, heavy dataset augmentation and the use of 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.

2.2. Text applications in healthcare

Owing to the rapid growth of scholarly publications and EHRs, biomedical text mining has become increasingly important in

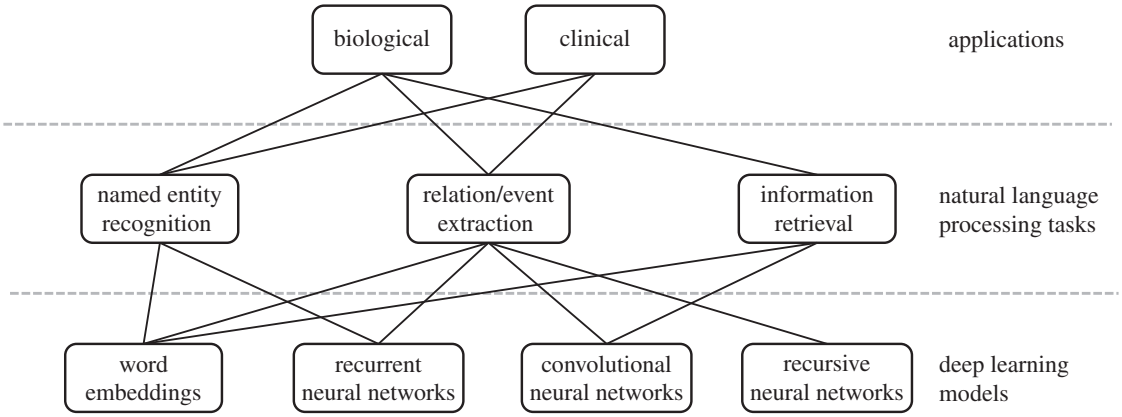

Figure 2. Deep learning applications, tasks and models based on NLP perspectives.

recent years. The main tasks in biological and clinical text mining include, but are not limited to, named entity recognition (NER), relation/event extraction and information retrieval (figure 2). Deep learning is appealing in this domain because of its competitive performance versus traditional methods and ability to overcome challenges in feature engineering. Relevant applications can be stratified by the application domain (biomedical literature versus clinical notes) and the actual task (e.g. concept or relation extraction).

NER is a task of identifying text spans that refer to a biological concept of a specific class, such as disease or chemical, in a controlled vocabulary or ontology. NER is often needed as a first step in many complex text mining systems. The current state-of-the-art methods typically reformulate the task as a sequence labelling problem and use conditional random fields [75–77]. In recent years, word embeddings that contain rich latent semantic information of words have been widely used to improve the NER performance. Liu *et al*. [78] studied the effect of word embeddings on drug name recognition and compared them with traditional semantic features. Tang *et al*. [79] investigated word embeddings in the gene, DNA and cell line mention detection tasks. Moreover, Wu *et al*. [80] examined the use of neural word embeddings for clinical abbreviation disambiguation. Liu *et al*. [81] exploited task-oriented resources to learn word embeddings for clinical abbreviation expansion.

Relation extraction involves detecting and classifying semantic relationships between entities from the literature. At present, kernel methods or feature-based approaches are commonly applied [82–84]. Deep learning can relieve the feature sparsity and engineering problems. Some studies focused on jointly extracting biomedical entities and relations simultaneously [85,86], while others applied deep learning on relation classification given the relevant entities. For example, both multi-channel dependency-based CNNs [87] and shortest path-based CNNs [88,89] are well suited for sentence-based protein–protein extraction. Jiang *et al*. [90] proposed a biomedical domain-specific word embedding model to reduce the manual labour of designing semantic representation for the same task. Gu *et al*. [91] employed a maximum-entropy model and a CNN model for chemical-induced disease relation extraction at the inter- and intra-sentence level, respectively. For drug–drug interactions, Zhao *et al*. [92] used a CNN that employs word embeddings with the syntactic information of a sentence as well as features of part-of-speech tags and dependency trees. Asada *et al*. [93] experimented with an attention CNN, and Yi *et al*. [94] proposed an RNN model with multiple attention layers. In both cases, it is a single model with attention mechanism, which allows the decoder to focus on different parts of the source sentence. As a result, it does not require dependency parsing or training multiple models. Both attention CNN and RNN have comparable results, but the CNN model has an advantage in that it can be easily computed in parallel, hence making it faster with recent graphics processing units (GPUs).

For biotopes event extraction, Li *et al*. [95] employed CNNs and distributed representation, while Mehryary *et al*. [96] used long short-term memory (LSTM) networks to extract complicated relations. Li *et al*. [97] applied word embedding to extract complete events from the biomedical text and achieved results comparable to the state-of-the-art systems. There are also approaches that identify event triggers rather than the complete event [98,99]. Taken together, deep learning models outperform traditional kernel methods or feature-based approaches by 1–5% in *f*-score. Among various deep learning approaches, CNNs stand out as the most popular model both in terms of computational complexity and performance, while RNNs have achieved continuous progress.

Information retrieval is a task of finding relevant text that satisfies an information need from within a large document collection. While deep learning has not yet achieved the same level of success in this area as seen in others, the recent surge of interest and work suggest that this may be quickly changing. For example, Mohan *et al*. [100] described a deep learning approach to modelling the relevance of a document's text to a query, which they applied to the entire biomedical literature [100].

To summarize, deep learning has shown promising results in many biomedical text mining tasks and applications. However, to realize its full potential in this domain, either large amounts of labelled data or technical advancements in current methods coping with limited labelled data are required.

2.3. Electronic health records

EHR data include substantial amounts of free text, which remains challenging to approach [101]. Often, researchers developing algorithms that perform well on specific tasks must design and implement domain-specific features [102]. These features capture unique aspects of the literature being processed. Deep learning methods are natural feature constructors. In recent work, Chalapathy *et al*. evaluated the extent to which deep learning methods could be applied on top of generic features for domain-specific concept extraction [103]. They

found that performance was in line with, but lower than the best domain-specific method [103]. 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*. [104] analysed 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 analyse the free text portions of pathology reports to identify the primary site and laterality of tumours. The only features the authors supplied to the algorithms were unigrams (counts for single words) and bigrams (counts for 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 (naive 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 tumour, the deep neural network outperformed the other methods. Of particular interest, when the authors first trained a neural network to predict the 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 NLP and neural embedding models, as popularized by word2vec [105]. Minarro-Giménez *et al*. [106] applied the word2vec deep learning toolkit to medical corpora and evaluated the efficiency of word2vec in identifying properties of pharmaceuticals based on mid-sized, unstructured medical text corpora without any additional background knowledge. 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 [107], ICD and National Drug Codes [108,109] and claims data [110]. Methods for neural embeddings learned from EHRs have at least some ability to predict disease–disease associations and implicate genes with a statistical association with a disease [111], but the evaluations performed did not differentiate between simple predictions (i.e. the same disease in different sites of the body) and non-intuitive ones. Jagannatha & Yu [112] further employed a bidirectional LSTM structure to extract adverse drug events from EHRs, and Lin *et al*. [113] investigated using CNNs to extract temporal relations. While promising, a lack of rigorous evaluation of the real-world utility of these kinds of features makes current contributions in this area difficult to evaluate. Comparisons need to be performed to examine the true utility 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*. [114], combined sparse autoencoders and Gaussian processes to distinguish gout from leukaemia 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 labelled examples required for subsequent supervised analyses [115]. In addition, it pointed towards features learned during unsupervised training being useful for visualizing and stratifying subgroups of patients within a single disease. In a concurrent large-scale analysis of EHR data from 700 000 patients, Miotto *et al*. [116] used a deep denoising autoencoder architecture applied to the number and co-occurrence of clinical events to learn a representation of patients (DeepPatient). The model was able to predict disease trajectories within 1 year with over 90% accuracy, and patient-level predictions were improved by up to 15% when compared to other methods. Choi *et al*. [117] attempted to model the longitudinal structure of EHRs with an RNN to predict future diagnosis and medication prescriptions on a cohort of 260 000 patients followed for 8 years (Doctor AI). Pham *et al*. [118] built upon this concept by using an RNN with an LSTM architecture enabling explicit modelling of patient trajectories through the use of memory cells. The method, DeepCare, performed better than shallow models or plain RNN when tested on two independent cohorts for its ability to predict disease progression, intervention recommendation and future risk prediction. Nguyen *et al*. [119] took a different approach and used word embeddings from EHRs to train a CNN that could detect and pool local clinical motifs to predict unplanned readmission after six months, with performance better than the baseline method (Deepr). Razavian *et al*. [120] used a set of 18 common laboratory tests to predict disease onset using both CNN and 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 [121]. Early approaches, such as the Faraggi–Simon feed-forward network, aimed to relax the linearity assumption, but performance gains were lacking [122]. Katzman *et al*. [123] 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. 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 [124]. 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 [125], 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 [126].

2.4. Challenges and opportunities in patient categorization

2.4.1. 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 [127]. 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 labelled data. For instance, in the eMERGE consortia and PheKB database [128], most samples with expert validation contain only 100–300 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 labelled examples or 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 labelled examples [115]. The anchor and learn framework [129] uses expert knowledge to identify high-confidence observations from which labels can be inferred. If transformations are available that preserve the meaningful content of the data, the adversarial and augmented training techniques discussed above can reduce overfitting. While these can be easily imagined for certain methods that operate on images, it is more challenging to figure out equivalent transformations for a patient's clinical test results. Consequently, it may be hard to employ such training examples with other applications. Finally, approaches that transfer features can also help use valuable training data most efficiently. Rajkomar *et al*. [58] trained a deep neural network using generic images before tuning using only radiology images. 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 EHRs. For example, features learned from the EHR for one type of clinical test (e.g. a decrease over time in a laboratory value) may transfer across phenotypes. Methods to accomplish more with little high-quality labelled data arose in other domains and may also be adapted to this challenge, e.g. data programming [130]. In data programming, noisy automated labelling functions are integrated.

Numerous commentators have described data as the new oil [131,132]. The idea behind this metaphor is that data are available in large quantities, valuable once refined, and this underlying resource will enable a data-driven revolution in how work is done. Contrasting with this perspective, Ratner *et al*. [133] described labelled training data, instead of data, as 'The *New* New Oil'. In this framing, data are abundant and not a scarce resource. Instead, new approaches to solving problems arise when labelled 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 do not 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.

2.4.2. 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 [134]. 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 [135], 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 [136,137]. 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 [138]. 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 [139]. 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. Furthermore, even the ostensibly least-biased data type, laboratory measurements, can be biased based by both the healthcare process and patient health state [140]. 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 [141]. Further diversity arises from variation in languages, healthcare practices and demographics. Merging EHRs gathered in different systems (and even under different assumptions) is challenging [142].

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 [143], and restricting the number of parameters that can be trained in a model. Furthermore, rule-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 [144]. Genetic studies using EHR data are subject to even more bias, as the differences in population ancestry across health centres (e.g. proportion of patients with African or Asian ancestry) can affect algorithm performance. For example, Wiley *et al*. [145] 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 data [146]. Here again, standards are heterogeneous and evolving, but often EHR data cannot be exported or even accessed directly for research purposes without appropriate consent. In the USA, research use of EHR data is subject both to the Common Rule and the Health Insurance Portability and Accountability Act. Ambiguity in the regulatory language and individual interpretation of these rules can hamper use of EHR data [147]. 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 [148] and ViPAR [149], although not EHR-specific, allow 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 [141] 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 [150] 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*. [151] showed the ability to steal trained models via public application programming interfaces (APIs). Dwork & Roth [152] 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 [153], in this case, whether a person's record was present. To protect against these attacks, Simmons *et al*. [154] developed the ability to perform genome-wide association studies in a differentially private manner, and Abadi *et al*. [155] show the ability to train deep learning classifiers under the differential privacy framework.

These attacks also present a potential hazard for approaches that aim to generate data. Choi *et al*. [156] propose generative adversarial neural networks (GANs) as a tool to make sharable EHR data, and Esteban *et al*. [157] showed that recurrent GANs could be used for time-series data. However, in both cases 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 parameters which provide a quantification of privacy. Beaulieu-Jones *et al*. [158] demonstrated the ability to generate data that preserved properties of the SPRINT clinical trial with GANs under the differential privacy framework. Both Beaulieu-Jones *et al*. and Esteban *et al*. train models on synthetic data generated under differential privacy and observe performance from a transfer learning evaluation that is only slightly below models trained on the original, real data. Taken together, these results suggest that differentially private GANs may be an attractive way to generate sharable datasets for downstream reanalysis.

Federated learning [159] and secure aggregations [160] 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. It can be trivial to train a model without

differential privacy, but quite difficult to train one within the differential privacy framework [158]. This 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.

2.4.3. 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 [161]. 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 [162], male [163] and European in ancestry [164]. 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 [162]. 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 [165]). As a result, many genomic findings are of limited value for people of non-European ancestry [164] and may even lead to worse treatment outcomes for them. Methods have been developed for mitigating some of these problems in genomic studies [162,165], 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 [166]). 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 [167]). When we are not careful with our data and models, we may inadvertently say more about the way the data were 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 [167] or lead to severe discriminatory consequences [168].

There is a small but growing literature on the prevention and mitigation of data leakage [167], as well as a closely related literature on discriminatory model behaviour [169], 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 [170]. 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 (see Discussion). 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 Discussion.

2.4.4. Applications of deep learning to longitudinal analysis

The 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 Framingham Heart Study [171] and the Avon Longitudinal Study of Parents and Children [172] 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 snapshot at a point in time 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 [173]. Efforts to model sequences of events have shown promise [174] but require exceedingly large patient sizes due to discrete combinatorial bucketing. Lasko *et al.* [114] used autoencoders on longitudinal sequences of serum uric acid measurements to identify population subtypes. More recently, deep learning has shown promise working with both sequences (CNNs) [175] and the incorporation of past and current state (RNNs, LSTMs) [118]. 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.

3. Deep learning to study the fundamental biological processes underlying human disease

The study of cellular structure and core biological processes—transcription, translation, signalling, metabolism, etc.—in humans and model organisms will greatly impact our understanding of human disease over the long horizon [176]. Predicting how cellular systems respond to environmental perturbations and are altered by genetic variation remain

daunting tasks. Deep learning offers new approaches for modelling biological processes and integrating multiple types of omic data [177], 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 CNNs and RNNs 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 [17]. Other areas, such as cellular signalling, are biologically important but studied less-frequently to date, with some exceptions [178]. 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.

3.1. 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, low-throughput fluorescence-based methods and microarray technologies, the field has shifted in recent years to primarily performing RNA sequencing (RNA-seq) to catalogue 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 [179] and to stratify yeast strains based on chemical and mutational perturbations [180]. Shallow (one hidden layer) denoising autoencoders have also been fruitful in extracting biological insight from thousands of *Pseudomonas aeruginosa* experiments [181,182] and in aggregating features relevant to specific breast cancer subtypes [26]. 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 non-essential. 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 1000 landmark genes [183]. 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.

Epigenomic data, combined with deep learning, may have sufficient explanatory power to infer gene expression. For instance, the DeepChrome CNN [184] improved the prediction accuracy of high or low gene expression from histone modifications over existing methods. AttentiveChrome [185] added a deep attention model to further enhance DeepChrome. Deep learning can also integrate different data types. For example, Liang *et al*. [186] combined RBMs to integrate gene expression, DNA methylation and miRNA data to define ovarian cancer subtypes. 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 aetiology 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.

3.2. Splicing

Pre-mRNA transcripts can be spliced into different isoforms by retaining or skipping subsets of exons or including parts of introns, creating enormous spatio-temporal flexibility to generate multiple distinct proteins from a single gene. This remarkable complexity can lend itself to defects that underlie many diseases. For instance, splicing mutations in the lamin A (*LMNA*) gene can lead to specific variants of dilated cardiomyopathy and limb-girdle muscular dystrophy [187]. 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 [188]. Therapeutic strategies that aim to modulate splicing are also currently being considered for disorders such as Duchenne muscular dystrophy and spinal muscular atrophy [187].

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 naive Bayes model and a two-layer Bayesian neural network with

thousands of hand-derived sequence-based features to predict the probability of exon skipping [189,190]. With the advent of deep learning, more complex models provided better predictive accuracy [191,192]. 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 behaviour. 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*. [193] for tissue-specific transcription factor (TF) binding predictions.

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 [194]. 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 [191]. 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 limb-girdle 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 [195]. As a proof of concept, Xiong *et al*. [191] 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.

3.3. Transcription factors

TFs are proteins that bind regulatory DNA in a sequence-specific manner to modulate the activation and repression of gene transcription. High-throughput *in vitro* experimental assays that quantitatively measure the binding specificity of a TF to a large library of short oligonucleotides [196] provide rich datasets to model the naked DNA sequence affinity of individual TFs in isolation. However, *in vivo* TF binding is affected by a variety of other factors beyond sequence affinity, such as competition and cooperation with other TFs, TF concentration and chromatin state (chemical modifications to DNA and other packaging proteins that DNA is wrapped around) [196]. TFs can thus exhibit highly variable binding landscapes across the same genomic DNA sequence across diverse cell types and states. Several experimental approaches such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) have been developed to profile *in vivo* binding maps of TFs [196]. Large reference compendia of ChIP-seq data are now freely available for a large collection of TFs in a small number of reference cell states in humans and a few other model organisms [197]. Owing to fundamental material and cost constraints, it is infeasible to perform these experiments for all TFs in every possible cellular state and species. Hence, predictive computational models of TF binding are essential to understand gene regulation in diverse cellular contexts.

Several machine learning approaches have been developed to learn generative and discriminative models of TF binding from *in vitro* and *in vivo* TF binding datasets that associate collections of synthetic DNA sequences or genomic DNA sequences to binary labels (bound/unbound) or continuous measures of binding. The most common class of TF binding models in the literature are those that only model the DNA sequence affinity of TFs from *in vitro* and *in vivo* binding data. The earliest models were based on deriving simple, compact, interpretable sequence motif representations such as position weight matrices (PWMs) and other biophysically inspired models [198–200]. These models were outperformed by general k-mer-based models including support vector machines (SVMs) with string kernels [201,202].

In 2015, Alipanahi *et al*. [203] developed DeepBind, the first CNN to classify bound DNA sequences based on *in vitro* and *in vivo* assays against random DNA sequences matched for dinucleotide sequence composition. The convolutional layers learn pattern detectors reminiscent of PWMs from a one-hot encoding of the raw input DNA sequences. DeepBind outperformed several state-of-the-art methods from the DREAM5 *in vitro* TF-DNA motif recognition challenge [200]. Although DeepBind was also applied to RNA-binding proteins, in general, RNA binding is a separate problem [204] and accurate models will need to account for RNA secondary structure. Following DeepBind, several optimized convolutional and recurrent neural network architectures as well as novel hybrid approaches that combine kernel methods with neural networks have been proposed that further improve performance [205–208]. Specialized layers and regularizers have also been proposed to reduce parameters and learn more robust models by taking advantage of specific properties of DNA sequences such as their reverse complement equivalence [209,210].

While most of these methods learn independent models for different TFs, *in vivo* multiple TFs compete or cooperate to occupy DNA binding sites, resulting in complex combinatorial co-binding landscapes. To take advantage of this shared structure in *in vivo* TF binding data, multi-task neural network architectures have been developed that explicitly share parameters across models for multiple TFs [208,211,212]. Some of these multi-task models train and evaluate classification performance relative to an unbound background set of regulatory DNA sequences sampled from the genome rather than using synthetic background sequences with matched dinucleotide composition.

The above-mentioned TF binding prediction models that use only DNA sequences as inputs have a fundamental

limitation. Because the DNA sequence of a genome is the same across different cell types and states, a sequence-only model of TF binding cannot predict different *in vivo* TF binding landscapes in new cell types not used during training. One approach for generalizing TF binding predictions to new cell types is to learn models that integrate DNA sequence inputs with other cell-type-specific data modalities that modulate *in vivo* TF binding such as surrogate measures of TF concentration (e.g. TF gene expression) and chromatin state. Arvey *et al*. [213] showed that combining the predictions of SVMs trained on DNA sequence inputs and cell-type specific DNase-seq data, which measures genome-wide chromatin accessibility, improved *in vivo* TF binding prediction within and across cell types. Several 'footprinting'-based methods have also been developed that learn to discriminate bound from unbound instances of known canonical motifs of a target TF based on high-resolution footprint patterns of chromatin accessibility that are specific to the target TF [214]. However, the genome-wide predictive performance of these methods in new cell types and states has not been evaluated.

Recently, a community challenge known as the 'ENCODE-DREAM *in vivo* TF Binding Site Prediction Challenge' was introduced to systematically evaluate the genome-wide performance of methods that can predict TF binding across cell states by integrating DNA sequence and *in vitro* DNA shape with cell-type-specific chromatin accessibility and gene expression [215]. A deep learning model called FactorNet was among the top three performing methods in the challenge [216]. FactorNet uses a multimodal hybrid convolutional and recurrent architecture that integrates DNA sequence with chromatin accessibility profiles, gene expression and evolutionary conservation of sequence. It is worth noting that FactorNet was slightly outperformed by an approach that does not use neural networks [217]. This top ranking approach uses an extensive set of curated features in a weighted variant of a discriminative maximum conditional likelihood model in combination with a novel iterative training strategy and model stacking. There appears to be significant room for improvement because none of the current approaches for cross cell-type prediction explicitly account for the fact that TFs can co-bind with distinct cofactors in different cell states. In such cases, sequence features that are predictive of TF binding in one cell state may be detrimental to predicting binding in another.

Singh *et al*. [218] developed transfer string kernels for SVMs for cross-context TF binding. Domain adaptation methods that allow training neural networks which are transferable between differing training and test set distributions of sequence features could be a promising avenue going forward [219,220]. These approaches may also be useful for transferring TF binding models across species.

Another class of imputation-based cross cell type *in vivo* TF binding prediction methods leverage the strong correlation between combinatorial binding landscapes of multiple TFs. Given a partially complete panel of binding profiles of multiple TFs in multiple cell types, a deep learning method called TFImpute learns to predict the missing binding profile of a target TF in some target cell type in the panel based on the binding profiles of other TFs in the target cell type and the binding profile of the target TF in other cell types in the panel [193]. However, TFImpute cannot generalize predictions beyond the training panel of cell types and requires TF binding profiles of related TFs.

It is worth noting that TF binding prediction methods in the literature based on neural networks and other machine learning approaches choose to sample the set of bound and unbound sequences in a variety of different ways. These choices and the choice of performance evaluation measures significantly confound systematic comparison of model performance (see Discussion).

Several methods have also been developed to interpret neural network models of TF binding. Alipanahi *et al*. [203] visualize convolutional filters to obtain insights into the sequence preferences of TFs. They also introduced *in silico* mutation maps for identifying important predictive nucleotides in input DNA sequences by exhaustively forward propagating perturbations to individual nucleotides to record the corresponding change in output prediction. Shrikumar *et al*. [221] proposed efficient backpropagation-based approaches to simultaneously score the contribution of all nucleotides in an input DNA sequence to an output prediction. Lanchantin *et al*. [206] developed tools to visualize TF motifs learned from TF binding site classification tasks. These and other general interpretation techniques (see Discussion) will be critical to improve our understanding of the biologically meaningful patterns learned by deep learning models of TF binding.

3.4. Promoters and enhancers

3.4.1. From transcription factor binding to promoters and enhancers

Multiple TFs act in concert to coordinate changes in gene regulation at the genomic regions known as promoters and enhancers. Each gene has an upstream promoter, essential for initiating that gene's transcription. The gene may also interact with multiple enhancers, which can amplify transcription in particular cellular contexts. These contexts include different cell types in development or environmental stresses.

Promoters and enhancers provide a nexus where clusters of TFs and binding sites mediate downstream gene regulation, starting with transcription. The gold standard to identify an active promoter or enhancer requires demonstrating its ability to affect transcription or other downstream gene products. Even extensive biochemical TF binding data has thus far proven insufficient on its own to accurately and comprehensively locate promoters and enhancers. We lack sufficient understanding of these elements to derive a mechanistic 'promoter code' or 'enhancer code'. But extensive labelled data on promoters and enhancers lends itself to probabilistic classification. The complex interplay of TFs and chromatin leading to the emergent properties of promoter and enhancer activity seems particularly apt for representation by deep neural networks.

3.4.2. Promoters

Despite decades of work, computational identification of promoters remains a stubborn problem [222]. Researchers have used neural networks for promoter recognition as early as 1996 [223]. Recently, a CNN recognized promoter sequences with sensitivity and specificity exceeding 90% [224]. Most activity in computational prediction of regulatory regions, however, has moved to enhancer identification. Because one can identify promoters with straightforward biochemical assays [225,226], the direct rewards of promoter prediction alone have decreased. But the reliable ground-truth provided by these

assays makes promoter identification an appealing test bed for deep learning approaches that can also identify enhancers.

3.4.3. Enhancers

Recognizing enhancers presents additional challenges. Enhancers may be up to 1 000 000 bp away from the affected promoter, and even within introns of other genes [227]. Enhancers do not necessarily operate on the nearest gene and may affect multiple genes. Their activity is frequently tissue- or context-specific. No biochemical assay can reliably identify all enhancers. Distinguishing them from other regulatory elements remains difficult, and some believe the distinction somewhat artificial [228]. While these factors make the enhancer identification problem more difficult, they also make a solution more valuable.

Several neural network approaches yielded promising results in enhancer prediction. Both Basset [229] and DeepEnhancer [230] used CNNs to predict enhancers. DECRES used a feed-forward neural network [231] to distinguish between different kinds of regulatory elements, such as active enhancers and promoters. DECRES had difficulty distinguishing between inactive enhancers and promoters. They also investigated the power of sequence features to drive classification, finding that beyond CpG islands, few were useful.

Comparing the performance of enhancer prediction methods illustrates the problems in using metrics created with different benchmarking procedures. Both the Basset and DeepEnhancer studies include comparisons to a baseline SVM approach, gkm-SVM [202]. The Basset study reports gkm-SVM attains a mean area under the precision-recall curve (AUPR) of 0.322 over 164 cell types [229]. The DeepEnhancer study reports for gkm-SVM a dramatically different AUPR of 0.899 on nine cell types [230]. This large difference means it is impossible to directly compare the performance of Basset and DeepEnhancer based solely on their reported metrics. DECRES used a different set of metrics altogether. To drive further progress in enhancer identification, we must develop a common and comparable benchmarking procedure (see Discussion).

3.4.4. Promoter–enhancer interactions

In addition to the location of enhancers, identifying enhancer–promoter interactions in three-dimensional space will provide critical knowledge for understanding transcriptional regulation. SPEID used a CNN to predict these interactions with only sequence and the location of putative enhancers and promoters along a one-dimensional chromosome [232]. It compared well to other methods using a full complement of biochemical data from ChIP-seq and other epigenomic methods. Of course, the putative enhancers and promoters used were themselves derived from epigenomic methods. But one could easily replace them with the output of one of the enhancer or promoter prediction methods above.

3.5. MicroRNA binding

Prediction of miRNAs and miRNA targets is of great interest, as they are critical components of gene regulatory networks and are often conserved across great evolutionary distance [233,234]. While many machine learning algorithms have been applied to these 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 [235]. Some of these features, including structural accessibility, are imperfect or empirically derived. In addition, current algorithms suffer from low specificity [236].

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 [236,237]. 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.

3.6. 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 DBNs, LSTMs, CNNs and deep convolutional neural fields [31,238].

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 [29,239–241].

One can represent protein secondary structure with three different states (α-helix, β-strand and loop regions) or eight finer-grained states. The accuracy of a three-state prediction is called Q3, and accuracy of an eight-state prediction is called Q8. Several groups [30,242,243] applied deep learning to protein secondary structure prediction but were unable to achieve significant improvement over the de facto standard

method PSIPRED [244], which uses two shallow feed-forward neural networks. In 2014, Zhou & Troyanskaya [245] demonstrated that they could improve Q8 accuracy by using a deep supervised and convolutional generative stochastic network. 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 [31,238]. 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 β-strand prediction. Nevertheless, the improvements in secondary structure prediction from DeepCNF are unlikely to result in a commensurate improvement in tertiary structure prediction because 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) represent a promising new direction for *ab initio* folding of proteins without good templates in PDB. Coevolution analysis is effective for proteins with a very large number (more than 1000) of sequence homologues [241], but fares poorly for proteins without many sequence homologues. By combining coevolution information with a few other protein features, shallow neural network methods such as MetaPSICOV [239] and CoinDCA-NN [246] have shown some advantage over pure coevolution analysis for proteins with few sequence homologues, but their accuracy is still far from satisfactory. In recent years, deeper architectures have been explored for contact prediction, such as CMAPpro [247], DNCON [248] and PConsC [249]. However, blindly tested in the well-known CASP competitions, these methods did not show any advantage over MetaPSICOV [239].

Recently, Wang *et al*. [29] proposed the deep learning method RaptorX-Contact, which significantly improves contact prediction over MetaPSICOV and pure coevolution methods, especially for proteins without many sequence homologues. It employs a network architecture formed by one one-dimensional residual neural network and one 2D residual neural network. Blindly tested in the latest CASP competition (i.e. CASP12 [250]), RaptorX-Contact ranked first in F_1 score on free-modelling targets as well as the whole set of targets. In CAMEO (which can be interpreted as a fully automated CASP) [251], its predicted contacts were also able to fold proteins with a novel fold and only 65–330 sequence homologues. This technique also worked well on membrane proteins even when trained on non-membrane proteins [252]. RaptorX-Contact performed better mainly due to the 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 homologues by studying new deep network architectures. The deep learning methods summarized above also apply to interfacial contact prediction for protein complexes but may be less effective because on average protein complexes have fewer sequence homologues. Beyond secondary structure and contact maps, we anticipate increased attention to predicting 3D protein structure directly from amino acid sequence and single residue evolutionary information [253].

3.7. Structure determination and cryo-electron microscopy

Complementing computational prediction approaches, cryo-electron microscopy (cryo-EM) allows near-atomic resolution determination of protein models by comparing individual electron micrographs [254]. Detailed structures require tens of thousands of protein images [255]. Technological development has increased the throughput of image capture. New hardware, such as direct electron detectors, has made large-scale image production practical, while new software has focused on rapid, automated image processing.

Some components of cryo-EM image processing remain difficult to automate. For instance, in particle picking, micrographs are scanned to identify individual molecular images that will be used in structure refinement. In typical applications, hundreds of thousands of particles are necessary to determine a structure to near-atomic resolution, making manual selection impractical [255]. Typical selection approaches are semi-supervised; a user will select several particles manually, and these selections will be used to train a classifier [256,257]. Now CNNs are being used to select particles in tools like DeepPicker [258] and DeepEM [259]. In addition to addressing shortcomings from manual selection, such as selection bias and poor discrimination of low-contrast images, these approaches also provide a means of full automation. DeepPicker can be trained by reference particles from other experiments with structurally unrelated macromolecules, allowing for fully automated application to new samples.

Downstream of particle picking, deep learning is being applied to other aspects of cryo-EM image processing. Statistical manifold learning has been implemented in the software package ROME to classify selected particles and elucidate the different conformations of the subject molecule necessary for accurate 3D structures [260]. These recent tools highlight the general applicability of deep learning approaches for image processing to increase the throughput of high-resolution cryo-EM.

3.8. Protein–protein interactions

Protein–protein interactions (PPIs) are highly specific and non-accidental physical contacts between proteins, which occur for purposes other than generic protein production or degradation [261]. Abundant interaction data have been generated in part thanks to advances in high-throughput screening methods, such as yeast two-hybrid and affinity-purification with mass spectrometry. However, because many PPIs are transient or dependent on biological context, high-throughput methods can fail to capture a number of interactions. The imperfections and costs associated with many experimental PPI screening methods have motivated an interest in high-throughput computational prediction.

Many machine learning approaches to PPI have focused on text mining the literature [262,263], but these approaches can fail to capture context-specific interactions, motivating de novo PPI prediction. Early de novo prediction approaches used a variety of statistical and machine learning tools on structural and sequential data, sometimes with reference to the existing body of protein structure knowledge. In the context of PPIs—as in other domains—deep learning shows promise both for exceeding current predictive performance and for circumventing limitations from which other approaches suffer.

One of the key difficulties in applying deep learning techniques to binding prediction is the task of representing peptide and protein sequences in a meaningful way. DeepPPI [264] made PPI predictions from a set of sequence and composition protein descriptors using a two-stage deep neural network that trained two subnetworks for each protein and combined them into a single network. Sun *et al*. [265] applied autocovariances, a coding scheme that returns uniform-size vectors describing the covariance between physico-chemical properties of the protein sequence at various positions. Wang *et al*. [266] used deep learning as an intermediate step in PPI prediction. They examined 70 amino acid protein sequences from each of which they extracted 1260 features. A stacked sparse autoencoder with two hidden layers was then used to reduce feature dimensions and noisiness before a novel type of classification vector machine made PPI predictions.

Beyond predicting whether or not two proteins interact, Du *et al*. [267] employed a deep learning approach to predict the residue contacts between two interacting proteins. Using features that describe how similar a protein's residue is relative to similar proteins at the same position, the authors extracted uniform-length features for each residue in the protein sequence. A stacked autoencoder took two such vectors as input for the prediction of contact between two residues. The authors evaluated the performance of this method with several classifiers and showed that a deep neural network classifier paired with the stacked autoencoder significantly exceeded classical machine learning accuracy.

Because many studies used predefined higher-level features, one of the benefits of deep learning—automatic feature extraction—is not fully leveraged. More work is needed to determine the best ways to represent raw protein sequence information so that the full benefits of deep learning as an automatic feature extractor can be realized.

3.9. Major histocompatibility complex-peptide binding

An important type of PPI involves the immune system's ability to recognize the body's own cells. The major histocompatibility complex (MHC) plays a key role in regulating this process by binding antigens and displaying them on the cell surface to be recognized by T cells. Owing to its importance in immunity and immune response, peptide–MHC binding prediction is a useful problem in computational biology, and one that must account for the allelic diversity in MHC-encoding gene region.

Shallow, feed-forward neural networks are competitive methods and have made progress towards pan-allele and pan-length peptide representations. Sequence alignment techniques are useful for representing variable-length peptides as uniform-length features [268,269]. For pan-allelic prediction, NetMHCpan [270,271] used a pseudo-sequence representation of the MHC class I molecule, which included only polymorphic peptide contact residues. The sequences of the peptide and MHC were then represented using both sparse vector encoding and Blosum encoding, in which amino acids are encoded by matrix score vectors. A comparable method to the NetMHC tools is MHCflurry [272], a method which shows superior performance on peptides of lengths other than nine. MHCflurry adds placeholder amino acids to transform variable-length peptides to length 15 peptides. When training the MHCflurry feed-forward neural network [273], the authors imputed missing MHC-peptide binding affinities using a Gibbs sampling method, showing that imputation improves performance for datasets with roughly 100 or fewer training examples. MHCflurry's imputation method increases its performance on poorly characterized alleles, making it competitive with NetMHCpan for this task. Kuksa *et al*. [274] developed a shallow, higher-order neural network (HONN) comprised both mean and covariance hidden units to capture some of the higher-order dependencies between amino acid locations. Pre-training this HONN with a semi-RBM, the authors found that the performance of the HONN exceeded that of a simple deep neural network, as well as that of NetMHC.

Deep learning's unique flexibility was recently leveraged by Bhattacharya *et al*. [275], who used a gated RNN method called MHCnuggets to overcome the difficulty of multiple peptide lengths. Under this framework, they used smoothed sparse encoding to represent amino acids individually. Because MHCnuggets had to be trained for every MHC allele, performance was far better for alleles with abundant, balanced training data. Vang *et al*. [276] developed HLA-CNN, a method which maps amino acids onto a 15-dimensional vector space based on their context relation to other amino acids before making predictions with a CNN. In a comparison of several current methods, Bhattacharya *et al*. found that the top methods—NetMHC, NetMHCpan, MHCflurry and MHCnuggets—showed comparable performance, but large differences in speed. Convolutional neural networks (in this case, HLA-CNN) showed comparatively poor performance, while shallow networks and RNNs performed the best. They found that MHCnuggets—the recurrent neural network—was by far the fastest-training among the top performing methods.

3.10. Protein–protein interaction networks and graph analysis

Because interacting proteins are more likely to share a similar function, the connectivity of a PPI network itself can be a valuable information source for the prediction of protein function [277]. To incorporate higher-order network information, it is necessary to find a lower-level embedding of network structure that preserves this higher-order structure. Rather than use hand-crafted network features, deep learning shows promise for the automatic discovery of predictive features within networks. For example, Navlakha [278] showed that a deep autoencoder was able to compress a graph to 40% of its original size, while being able to reconstruct 93% of the original graph's edges, improving upon standard dimension reduction methods. To achieve this, each graph was represented as an adjacency matrix with rows sorted in descending node degree order, then flattened into a vector and given as input to the autoencoder. While the activity of some hidden layers correlated with several popular hand-crafted network features such as *k*-core size and graph density, this work showed that deep learning can effectively reduce graph dimensionality while retaining much of its structural information.

An important challenge in PPI network prediction is the task of combining different networks and types of networks. Gligorijevic *et al*. [279] developed a multimodal deep autoencoder, deepNF, to find a feature representation common among several different PPI networks. This common lower-level representation allows for the combination of various PPI data sources towards a single predictive task. An SVM

classifier trained on the compressed features from the middle layer of the autoencoder outperformed previous methods in predicting protein function.

Hamilton *et al*. [280] addressed the issue of large, heterogeneous and changing networks with an inductive approach called GraphSAGE. By finding node embeddings through learned aggregator functions that describe the node and its neighbours in the network, the GraphSAGE approach allows for the generalization of the model to new graphs. In a classification task for the prediction of protein function, Chen & Zhu [281] optimized this approach and enhanced the graph convolutional network with a preprocessing step that uses an approximation to the dropout operation. This preprocessing effectively reduces the number of graph convolutional layers and it significantly improves both training time and prediction accuracy.

3.11. 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 [282]. Once the 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 [17,282,283]. However, user-friendly tools must be developed for deep learning to become commonplace for biological image segmentation.

We anticipate an additional 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 haematopoietic cells up to three generations in advance of differentiation [284]. In biomedical research, most often biologists 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 [285], even using generic deep convolutional networks trained on natural images [286], known as transfer learning. Recent work by Johnson *et al*. [287] 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 modelling the localization of fluorescent labels. Notably, this approach reduces the modelling 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 [288–290]. 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.

3.12. Single-cell data

Single-cell methods are generating excitement as biologists characterize the vast heterogeneity within unicellular species and between cells of the same tissue type in the same organism [291]. For instance, tumour cells and neurons can both harbour extensive somatic variation [292]. Understanding single-cell diversity in all its dimensions—genetic, epigenomic, 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 epigenomic patterns with single-cell Hi-C or ATAC-seq [293,294]. Joint profiling of single-cell epigenomic and transcriptional states provides unprecedented views of regulatory processes [295].

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, not to be confused with the dropout layer of deep learning. To solve this problem, Angermueller *et al*. [296] trained a neural network to predict the presence or the 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 [193,297]. Deep learning has also been useful for dealing with batch effects [298].

Examining populations of single cells can reveal biologically meaningful subsets of cells as well as their underlying gene regulatory networks [299]. 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*. [300] 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 were relatively low dimensional.

Neural networks can also learn low-dimensional representations of single-cell gene expression data for visualization, clustering and other tasks. Both scvis [301] and scVI [302] are unsupervised approaches based on VAEs. Whereas scvis primarily focuses on single-cell visualization as a replacement for t-Distributed Stochastic Neighbour Embedding [303], the scVI model accounts for zero-inflated expression distributions and can impute zero values that are due to technical effects. Beyond VAEs, Lin *et al*. [304] developed a supervised model to predict cell type. Similar to transfer learning approaches for microscopy images [286], they demonstrated that the hidden layer representations were informative in general and could be used to identify cellular subpopulations or match new cells to known cell types. The supervised neural network's representation was better overall at retrieving cell types than alternatives, but all methods struggled to recover certain cell types such as haematopoietic stem cells and inner cell mass cells. As the Human Cell Atlas [305] and related efforts generate more single-cell expression data, there will be opportunities to assess how well these low-dimensional representations generalize to new cell types as well as abundant training data to learn broadly applicable representations.

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 [306] or by feeding cell morphology and movement into neural networks [284]. Reinforcement learning algorithms [307] 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.

3.13. Metagenomics

Metagenomics, which refers to the study of genetic material—16S rRNA 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 bp reads from complex mixtures of microbial genomes based on tetranucleotide frequencies, which differ across organisms [308], using supervised [309,310] or unsupervised methods [311]. Then, researchers began to use techniques that could estimate relative abundances from an entire sample faster than classifying individual reads [312–315]. There is also great interest in identifying and annotating sequence reads [316,317]. 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 [318–321] or overall sequence composition [322]. Also, researchers have looked into how feature selection can improve classification [321,323], and techniques have been proposed that are classifier-independent [324,325].

Most neural networks are 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 [326] and FragGeneScan [327]). Representations (similar to Word2-Vec [105] in NLP) for protein family classification have been introduced and classified with a skip-gram neural network [328]. RNNs show good performance for homology and protein family identification [329,330].

One of the first techniques of de novo genome binning used self-organizing maps, a type of neural network [311]. Essinger *et al*. [331] 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 [332] 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 [333] 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 [334]. However, this study evaluated only 3000 sequences.

Neural network uses for classifying phenotype from the microbial composition are just beginning. A simple multilayer perceptron (MLP) was able to classify wound severity from microbial species present in the wound [335]. Recently, Ditzler *et al*. [336] associated soil samples with pH level using MLPs, DBNs and RNNs. 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 [337].

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 [338], 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 [339] (discussed below), one day the entire pipeline, from denoising to functional classification, may be combined into one step using powerful LSTMs [340]. 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 for deep learning is to learn the complete characteristics of a metagenomic sample in one complex network.

3.14. Sequencing and variant calling

While we have so far primarily discussed the role of deep learning in analysing 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 tumour clones and other genetic diseases will require the accurate calling of SNPs and indels.

Current methods achieve relatively high (greater than 99%) precision at 90% recall for SNPs and indel calls from Illumina short-read data [341], 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 221 × 100 bitmap image, where each column is a nucleotide and each row is a read from the sample library [341]. The top five 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, G, T) 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 [342], 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 62 features for each candidate variant and fed these vectors into a fully connected deep neural network [343]. Unfortunately, this feature set required at least 15 iterations of software development to fine-tune, which suggests that these models may not generalize.

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 [344], 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 [341]. However, a model built on Illumina data, for instance, may not be optimal for Pacific Biosciences long-read data or MinION nanopore data, which have vastly different specificity and sensitivity profiles and signal-to-noise characteristics. Recently, Boža *et al*. [339] used bidirectional RNNs 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. 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 [345]. 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 [345]. 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.

3.15. Neuroscience

Artificial neural networks were originally conceived as a model for computation in the brain [7]. Although deep neural networks have evolved to become a workhorse across many fields, there is still a strong connection between deep networks and the study of the brain. The rich parallel history of artificial neural networks in computer science and neuroscience is reviewed in [346–348].

CNNs were originally conceived as faithful models of visual information processing in the primate visual system, and are still considered so [349]. The activations of hidden units in consecutive layers of deep convolutional networks have been found to parallel the activity of neurons in consecutive brain regions involved in processing visual scenes. Such models of neural computation are called 'encoding' models, as they predict how the nervous system might encode sensory information in the world.

Even when they are not directly modelling biological neurons, deep networks have been a useful computational tool in neuroscience. They have been developed as statistical time-series models of neural activity in the brain. And in contrast to the encoding models described earlier, these models are used for decoding neural activity, for instance, in brain–machine interfaces [350]. They have been crucial to the field of connectomics, which is concerned with mapping the connectivity of biological neural networks in the brain. In connectomics, deep networks are used to segment the shapes of individual neurons and to infer their connectivity from 3D electron microscopic images [351], and they have also been used to infer causal connectivity from optical measurement and perturbation of neural activity [352].

It is an exciting time for neuroscience. Recent rapid progress in deep networks continues to inspire new machine learning-based models of brain computation [346]. And neuroscience continues to inspire new models of artificial intelligence [348].

4. 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 patient's 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 [353,354]. Earlier, we reviewed approaches for patient categorization. Here, we examine the potential for better treatment, which broadly, may be divided into methods for improved choices of interventions for patients and those for development of new interventions.

4.1. Clinical decision-making

In 1996, Tu [355] 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 [356] or mortality [357]. 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 & Taktak [358] 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. Recent applications of deep learning in pharmacogenomics and pharmacoepigenomics show the potential for improving patient treatment response and outcome prediction using patient-specific data, pharmacogenomic targets and pharmacological knowledge bases [20].

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 [121,123] 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 the counterfactual question [359]. Johansson *et al.* [360] 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.* [361] first create a deep neural network to model clinical time series and then analyse 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).

4.1.1. Predicting patient trajectories

A common application for deep learning in this domain is the temporal structure of healthcare records. Many studies [362–365] have used RNNs to categorize patients, but most stop short of suggesting clinical decisions. Nemati *et al.* [366] 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 counterfactual 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 the clinical application is unlikely until the model has been shown to be effective. The impressive applications of deep reinforcement learning to other domains [307] have relied on the knowledge of the underlying processes (e.g. the rules of the game). Some models have been developed for targeted medical problems [367], but a generalized engine is beyond current capabilities.

4.1.2. Clinical trial efficiency

A clinical deep learning task that has been more successful is the assignment of patients to clinical trials. Ithapu *et al.* [368] 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.* [369] 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.

4.2. 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 [370,371]. A decade ago, the Connectivity Map [372] had a sizeable impact. 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 [373–375]. 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 [376–378].

For example, Menden *et al*. [379] 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*. [37] used gene- and pathway-level drug perturbation transcriptional profiles from the Library of Network-Based Cellular Signatures [380] 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 [381,382]. Wang *et al*. [383] 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*. [38] 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*. [39] extended this concept to deep learning by creating a DBN called DeepDTIs, which predicts interactions using chemical structure and protein sequence features.

Drug repositioning appears 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, perhaps the most promising piece of work in this space [37] 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 [384] do not use predictive modelling. A mature and production-ready framework for drug repositioning via deep learning is currently missing.

4.3. Drug development

4.3.1. Ligand-based prediction of bioactivity

High-throughput chemical screening in biomedical research aims to improve therapeutic options over a long-term horizon [22]. The objective is to discover which small molecules (also referred to as chemical compounds or ligands) specifically affect the activity of a target, such as a kinase, PPI 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 [385]. Initial hits are confirmed to eliminate false positives and proceed to the lead generation stage [386], 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 [385].

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. 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 [387]. Computational modelling also has the potential to predict ADMET traits for lead generation [388] and how drugs are metabolized [389].

Ligand-based approaches train on chemicals' features without modelling target features (e.g. protein structure). Neural networks have a long history in this domain [21,23], 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 MLP network [390]. 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' [391], but not all outside experts were convinced [392]. Subsequent work (reviewed in more detail by Goh *et al*. [4]) explored the effects of jointly modelling far more targets than the Merck challenge [393,394], with Ramsundar *et al*. [394] showing that the benefits of multi-task networks had not yet saturated even with 259 targets. Although DeepTox [395], 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 nine 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 [388]. 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 [396]. 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, of the order of 0.1% of all tested chemicals for a typical target [397]. This has motivated low-parameter neural networks that emphasize compound–compound similarity, such as influence-relevance voter [387,398], instead of predicting compound activity directly from chemical features.

4.3.2. Chemical featurization and representation learning

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 modelling of chemical compounds. A molecular graph, where atoms are labelled nodes and bonds are labelled edges, is a natural way to represent a chemical structure. Chemical features can 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 [399]. Traditional machine learning approaches relied on preprocessing the graph into a feature vector of molecular descriptors or a fixed-width bit vector known as a fingerprint [400]. The same fingerprints have been used by some drug–target interaction methods discussed above [39]. An overly simplistic but approximately correct view of chemical fingerprints is that each bit represents the presence or the absence of a particular chemical substructure in the molecular graph. Instead of using molecular descriptors or fingerprints as input, modern neural networks can represent chemicals as textual strings [401] or images [402] or operate directly on the molecular graph, which has enabled strategies for learning novel chemical representations.

Virtual screening and chemical property prediction have emerged as one of the major applications areas for graph-based neural networks. Duvenaud *et al*. [403] 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*. [404] 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 neighbouring atoms and local regions on the molecular graph [405–407]. More sophisticated graph algorithms [408,409] addressed limitations of standard graph convolutions that primarily operate on each node's local neighbourhood. We anticipate that these graph-based neural networks could also be applicable in other types of biological networks, such as the PPI networks we discussed previously.

Advances in chemical representation learning have also enabled new strategies for learning chemical–chemical similarity functions. Altae-Tran *et al*. [406] developed a one-shot learning network 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 10 labelled 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 [403] 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 [405] performed comparably to a multi-task network using standard fingerprints and slightly better than the neural fingerprints [403] on the drug efficacy task but were slightly worse than the influence-relevance voter method on an HIV dataset [387]. Broader recent benchmarking has shown that relative merits of these methods depend on the dataset and cross-validation strategy [410], though evaluation in this domain often uses the area under the receiver operating characteristic curve (AUROC) [411], which has limited utility due to the large class imbalance (see Discussion).

We remain optimistic about the potential of deep learning and specifically representation learning in drug discovery. 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 [406,412] and MoleculeNet benchmarking suite [410] 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 modelling many targets [393,394]. Other studies have shown the limitations of transfer learning when the prediction tasks are insufficiently related [388,406]. 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 [401,413]. Future work will establish whether it is better to train on massive collections of diverse compounds, drug-like small molecules or specialized subsets.

4.3.3. 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 [414]. 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 [39,383].

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 [415] and TopologyNet [416] models take 3D structures from PDBBind [417] as input, ensuring the ligand–target complexes are reliable. AtomNet [36] samples multiple ligand poses within the target binding site, and DeepVS [418] and Ragoza *et al*. [419] 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 [418].

There are two established options for representing a protein–compound complex. One option, a 3D grid, can featurize the input complex [36,419]. 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. Alternatively, DeepVS [418] and atomic convolutions [415] offer greater flexibility in their convolutions by eschewing the 3D grid. Instead, they each implement techniques for executing convolutions over atoms' neighbouring atoms in the 3D space. Gomes *et al*. [415] demonstrate that currently random forest on a one-dimensional 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. 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.

4.3.4. De novo drug design

De novo drug design attempts to model the typical design–synthesize–test cycle of drug discovery [420,421]. It explores an estimated 10^{60} synthesizable organic molecules with drug-like properties without explicit enumeration [397]. 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 laboratory. Current programs have settled on rule-based virtual chemical reactions to generate molecular structures [421]. 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 [422].

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 [423]. This allows molecules to be treated as sequences and leveraging recent progress in RNNs. Gómez-Bombarelli *et al*. [401] designed a SMILES-to-SMILES autoencoder to learn a continuous latent feature space for chemicals. 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 [115]. 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 [424].

Another approach to de novo design is to train character-based RNNs on large collections of molecules, for example, ChEMBL [425], to first obtain a generic generative model for drug-like compounds [423]. These generative models successfully learn the grammar of compound representations, with 94% [426] or nearly 98% [423] 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 [423] or adopting reinforcement learning strategies [426,427]. 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 GANs for molecules).

5. 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.

5.1. Customizing deep learning models reflects a trade-off between bias and variance

Some of the challenges in applying deep learning are shared with other machine learning methods. In particular, many problem-specific optimizations described in this review reflect a recurring universal trade-off—controlling the flexibility of a model in order to maximize predictivity. Methods for adjusting the flexibility of deep learning models include dropout, reduced data projections and transfer learning (described below). One way of understanding such model optimizations is that they incorporate external information to limit model flexibility and thereby improve predictions. This balance is formally described as a trade-off between 'bias and variance' [11].

Although the bias-variance trade-off is common to all machine learning applications, recent empirical and theoretical observations suggest that deep learning models may have uniquely advantageous generalization properties [428,429]. Nevertheless, additional advances will be needed to establish a coherent theoretical foundation that enables practitioners to better reason about their models from first principles.

5.1.1. Evaluation metrics for imbalanced classification

Making predictions in the presence of high-class imbalance and differences between training and generalization data are a common feature of many large biomedical datasets, including deep learning models of genomic features, patient classification, disease detection and virtual screening. Prediction of TF binding sites exemplifies the difficulties with learning from highly imbalanced data. The human genome has three billion base pairs, and only a small fraction of them are implicated in specific biochemical activities. Less than 1% of the genome can be confidently labelled as bound for most TFs.

Estimating the false discovery rate (FDR) is a standard method of evaluation in genomics that can also be applied to deep learning model predictions of genomic features. Using deep learning predictions for targeted validation experiments of specific biochemical activities necessitates a more stringent FDR (typically 5–25%). However, when predicted biochemical activities are used as features in other models, such as gene expression models, a low FDR may not be necessary.

What is the correspondence between FDR metrics and commonly used classification metrics such as AUPR and AUROC? AUPR 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 AUPR greater than 0.6. In this case, the AUPR 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 [72,430]. Consider a classification model with the 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. Consequently, practitioners must select the metric most tailored to their subsequent use case to use these methods most effectively.

5.1.2. 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 labelled examples to constitute a negative label. To avoid an arbitrary label for these examples, they may be labelled as 'ambiguous'. Ambiguously labelled 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.

5.1.3. Formulation of a performance upper bound

In assessing the upper bound on the predictive performance of a deep learning model, it is necessary to incorporate inherent between-study variation inherent to biomedical research [431]. Study-level variability limits classification performance and can lead to underestimating prediction error if the generalization error is estimated by splitting a single dataset. Analyses can incorporate data from multiple laboratories and experiments to capture between-study variation within the prediction model mitigating some of these issues.

5.2. Uncertainty quantification

Deep learning-based solutions for biomedical applications could substantially benefit from guarantees on the reliability of predictions and a quantification of uncertainty. Owing to biological variability and precision limits of equipment, biomedical data do not consist of precise measurements but of estimates with noise. Hence, it is crucial to obtain uncertainty measures that capture how noise in input values propagates through deep neural networks. Such measures can be used for reliability assessment of automated decisions in clinical and public health applications, and for guarding against model vulnerabilities in the face of rare or adversarial cases [432]. Moreover, in fundamental biological research, measures of uncertainty help researchers distinguish between true regularities in the data and patterns that are false or merely anecdotal. There are two main uncertainties that one can calculate: epistemic and aleatoric [433]. Epistemic uncertainty describes uncertainty about the model, its structure or its parameters. This uncertainty is caused by insufficient training data or by a difference in the training set and testing set distributions, so it vanishes in the limit of infinite data. On the other hand, aleatoric uncertainty describes uncertainty inherent in the observations. This uncertainty is due to noisy or missing data, so it vanishes with the ability to observe all independent variables with infinite precision. A good way to represent aleatoric uncertainty is to design an appropriate loss function with an uncertainty variable. In the case of data-dependent aleatoric uncertainty, one can train the model to increase its uncertainty when it is incorrect due to noisy or missing data, and in the case of task-dependent aleatoric uncertainty, one can optimize for the best uncertainty parameter for each task [434]. Meanwhile, there are various methods for modelling epistemic uncertainty, outlined below.

In classification tasks, confidence calibration is the problem of using classifier scores to predict class membership probabilities that match the true membership likelihoods. These membership probabilities can be used to assess the uncertainty associated with assigning the example to each of the classes. Guo *et al*. [435] observed that contemporary neural networks are poorly calibrated and provided a simple recommendation

for calibration: temperature scaling, a single parameter special case of Platt scaling [436]. In addition to confidence calibration, there is early work from Chryssolouris *et al*. [437] that described a method for obtaining confidence intervals with the assumption of normally distributed error for the neural network. More recently, Hendrycks & Gimpel [438] discovered that incorrect or out-of-distribution examples usually have lower maximum softmax probabilities than correctly classified examples, allowing for effective detection of misclassified examples. Liang *et al*. [439] used temperature scaling and small perturbations to further separate the softmax scores of correctly classified examples and the scores of out-of-distribution examples, allowing for more effective detection. This approach outperformed the baseline approaches by a large margin, establishing a new state-of-the-art performance.

An alternative approach for obtaining principled uncertainty estimates from deep learning models is to use Bayesian neural networks. Deep learning models are usually trained to obtain the most likely parameters given the data. However, choosing the single most likely set of parameters ignores the uncertainty about which set of parameters (among the possible models that explain the given dataset) should be used. This sometimes leads to uncertainty in predictions when the chosen likely parameters produce high-confidence but incorrect results. On the other hand, the parameters of Bayesian neural networks are modelled as full probability distributions. This Bayesian approach comes with a whole host of benefits, including better calibrated confidence estimates [440] and more robustness to adversarial and out-of-distribution examples [441]. Unfortunately, modelling the full posterior distribution for the model's parameters given the data is usually computationally intractable. One popular method for circumventing this high computational cost is called test-time dropout [442], where an approximate posterior distribution is obtained using variational inference. Gal & Ghahramani [442] showed that a stack of fully connected layers with dropout between the layers is equivalent to approximate inference in a Gaussian process model. The authors interpret dropout as a variational inference method and apply their method to CNNs. This is simple to implement and preserves the possibility of obtaining cheap samples from the approximate posterior distribution. Operationally, obtaining model uncertainty for a given case becomes as straightforward as leaving dropout turned on and predicting multiple times. The spread of the different predictions is a reasonable proxy for model uncertainty. This technique has been successfully applied in an automated system for detecting diabetic retinopathy [443], where uncertainty-informed referrals improved diagnostic performance and allowed the model to meet the National Health Service recommended levels of sensitivity and specificity. The authors also found that entropy performs comparably to the spread obtained via test-time dropout for identifying uncertain cases, and therefore it can be used instead for automated referrals.

Several other techniques have been proposed for effectively estimating predictive uncertainty as uncertainty quantification for neural networks continues to be an active research area. Recently, McClure & Kriegeskorte [444] observed that test-time sampling improved calibration of the probabilistic predictions, sampling weights led to more robust uncertainty estimates than sampling units, and spike-and-slab sampling was superior to Gaussian dropconnect and Bernoulli dropout. Krueger *et al*. [445] introduced Bayesian hypernetworks as another framework for approximate Bayesian inference in deep learning, where an invertible generative hypernetwork maps isotropic Gaussian noise to parameters of the primary network allowing for computationally cheap sampling and efficient estimation of the posterior. Meanwhile, Lakshminarayanan *et al*. [446] proposed using deep ensembles, which are traditionally used for boosting predictive performance, on standard (non-Bayesian) neural networks to obtain well-calibrated uncertainty estimates that are comparable to those obtained by Bayesian neural networks. In cases where model uncertainty is known to be caused by a difference in training and testing distributions, domain adaptation-based techniques can help mitigate the problem [220].

Despite the success and popularity of deep learning, some deep learning models can be surprisingly brittle. Researchers are actively working on modifications to deep learning frameworks to enable them to handle probability and embrace uncertainty. Most notably, Bayesian modelling and deep learning are being integrated with renewed enthusiasm. As a result, several opportunities for innovation arise: understanding the causes of model uncertainty can lead to novel optimization and regularization techniques, assessing the utility of uncertainty estimation techniques on various model architectures and structures can be very useful to practitioners, and extending Bayesian deep learning to unsupervised settings can be a significant breakthrough [447]. Unfortunately, uncertainty quantification techniques are underused in the computational biology communities and largely ignored in the current deep learning for biomedicine literature. Thus, the practical value of uncertainty quantification in biomedical domains is yet to be appreciated.

5.3. 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 artefact of the data. A motivating example of this can be found in Ba & Caruana [448], 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 a 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 [449] 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.

5.3.1. 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- and backpropagation-based approaches.

Perturbation-based approaches change parts of the input and observe the impact on the output of the network. Alipanahi *et al*. [203] and Zhou & Troyanskaya [211] scored genomic sequences by introducing virtual mutations at individual positions in the sequence and quantifying the change in the output. Umarov *et al*. [224] used a similar strategy, but with sliding windows where the sequence within each sliding window was substituted with a random sequence. Kelley *et al*. [229] inserted known protein-binding motifs into the centres of sequences and assessed the change in predicted accessibility. Ribeiro *et al*. [450] 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 analysing images, Zeiler & Fergus [451] 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 [452].

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*. [221], 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 & Vedaldi [453] 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 perturbation-based methods. A classic example of this is calculating the gradients of the output with respect to the input [454] to compute a 'saliency map'. Bach *et al*. [455] proposed a strategy called Layerwise Relevance Propagation, which was shown to be equivalent to the element-wise product of the gradient and input [221,456]. Networks with Rectified Linear Units create nonlinearities that must be addressed. Several variants exist for handling this [451,457]. Backpropagation-based methods are a highly active area of research. Researchers are still actively identifying weaknesses [458], and new methods are being developed to address them [221,459,460]. Lundberg & Lee [461] noted that several importance scoring methods including integrated gradients and LIME could all be considered approximations to Shapely values [462], which have a long history in game theory for assigning contributions to players in cooperative games.

5.3.2. 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 & Vedaldi [463] 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 favour results that resemble natural images. For genomic sequence, Finnegan & Song [464] 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 [465]. This is done by maximizing the response of neurons that are active in the network, subject to some regularizing constraints. Mordvintsev *et al*. [466] leveraged caricaturization to generate aesthetically pleasing images using neural networks.

5.3.3. 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*. [467] and applied in subsequent work [454,465,466,468]. Lanchantin *et al*. [206] applied class-based 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.

5.3.4. 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 [469,470]. Deming *et al*. [471] 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. Singh *et al*. [185] used a hierarchy of attention layers to locate important genome positions and signals for predicting gene expression from histone modifications. In the clinical domain, Choi *et al*. [472] leveraged attention mechanisms to highlight which aspects of a patient's medical history were most relevant for making diagnoses. Choi *et al*. [473] 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 & Karamcheti [474] 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*. [475] 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 [476] allows interactive exploration of the hidden state of LSTMs on different inputs.

Another strategy, adopted by Lanchatin *et al*. [206] 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 that are responsible for flipping the output of the network from negative to positive. If the RNN is bidirectional, the same process can be repeated in the reverse sequence. As noted by the authors, this approach was less effective at identifying motifs compared with the gradient-based backpropagation approach of Simonyan *et al*. [454], illustrating the need for more sophisticated strategies to assign importance scores in RNNs.

Murdoch & Szlam [477] 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 time step. 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.

5.3.5. Latent space manipulation

Interpretation of embedded or latent space features learned through generative unsupervised models can reveal underlying patterns otherwise masked in the original input. Embedded feature interpretation has been emphasized mostly in image- and text-based applications [105,478], but applications to genomic and biomedical domains are increasing.

For example, Way & Greene trained a VAE on gene expression from The Cancer Genome Atlas (TCGA) [479] and use latent space arithmetic to rapidly isolate and interpret gene expression features descriptive of high-grade serous ovarian cancer subtypes [480]. The most differentiating VAE features were representative of biological processes that are known to distinguish the subtypes. Latent space arithmetic with features derived using other compression algorithms were not as informative in this context [481]. Embedding discrete chemical structures with autoencoders and interpreting the learned continuous representations with latent space arithmetic has also facilitated predicting drug-like compounds [401]. Furthermore, embedding biomedical text into lower dimensional latent spaces have improved name entity recognition in a variety of tasks including annotating clinical abbreviations, genes, cell lines and drug names [78–81].

Other approaches have used interpolation through latent space embeddings learned by GANs to interpret unobserved intermediate states. For example, Osokin *et al*. [482] trained GANs on two-channel fluorescent microscopy images to interpret intermediate states of protein localization in yeast cells. Goldsborough *et al*. [483] trained a GAN on fluorescent microscopy images and used latent space interpolation and arithmetic to reveal underlying responses to small molecule perturbations in cell lines.

5.3.6. Miscellaneous approaches

It can often be informative to understand how the training data affects model learning. Towards this end, Koh & Liang [484] 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 [485,486] 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*. [487] 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*. [488] 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.

5.3.7. Future outlook

While deep learning lags behind most Bayesian models in terms of interpretability, 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 the ImageNet [46] and CIFAR [489] datasets 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 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.

5.4. Data limitations

A lack of large-scale, high-quality, correctly labelled training data have impacted deep learning in nearly all applications we have discussed. 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, labelled data may be derived from an experimental assay with known and unknown technical artefacts, 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*. [490] 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 centre 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 labelled example into multiple examples. For example, an image can easily be rotated, flipped or translated and retain its label [43]. 3D MRI and 4D fMRI (with time as a dimension) data can be decomposed into sets of 2D images [491]. 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 [300].

Simulated or semi-synthetic training data have 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 PWMs [209] or RNA-seq reads for predicting mRNA transcript boundaries [492], 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 [209,492]. In rare cases, models trained on simulated data have been successfully applied directly to real data [492].

Data can be simulated to create negative examples when only positive training instances are available. DANN [35] adopts this approach to predict the pathogenicity of genetic variants using semi-synthetic training data from Combined Annotation-Dependent Depletion (CADD) [493]. 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 [345]. This method detected mutations in other semi-synthetic datasets but was not validated on real data.

In settings where the experimental observations are biased towards positive instances, such as MHC protein and peptide ligand binding affinity [273], or the negative instances vastly outnumber the positives, such as high-throughput chemical screening [398], training datasets have been augmented by adding additional instances and assuming they are negative. There is some evidence that this can improve performance [398], but in other cases, it was only beneficial when the real training datasets were extremely small [273]. Overall, training with simulated and semi-simulated data is a valuable idea for overcoming limited sample sizes but one that requires more rigorous evaluation of 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 [494]. 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 the principal component analysis, per class histograms or a Word2vec [105] 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 500 000 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 50 000, a factor of 600.

5.5. 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 [495].

Many have sought to curb these costs, with methods ranging from the very applied (e.g. reduced numerical precision [496–499]) to the exotic and theoretic (e.g. training small networks to mimic large networks and ensembles [448,500]). The largest gains in efficiency have come from computation with GPUs [495,501–505], 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 [502,503,506,507]. However, GPUs also have limited memory, making networks of useful size and complexity difficult to implement on a single GPU or machine [68,501]. This restriction has sometimes forced computational biologists to use workarounds or limit the size of an analysis. Chen *et al.* [183] inferred the expression level of all genes with a single neural network, but due to memory restrictions, they randomly partitioned genes into two separately analysed halves. In other cases, researchers limited the size of their neural network [29] or the total number of training instances [401]. Some have also chosen to use standard central processing unit (CPU) implementations rather than sacrifice network size or performance [508].

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 [448, 496–499,509,510], but there is also growing interest in specialized hardware, such as field-programmable gate arrays (FPGAs) [505,511] and application-specific integrated circuits (ASICs) [512]. Less software is available for such highly specialized hardware [511]. But specialized hardware promises improvements in deep learning at reduced time, energy and memory [505]. 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 [513,514] are not suitable for deep learning [515], but much progress has been made. There now exist a number of algorithms [498,515], tools [516–518] and high-level libraries [519,520] for deep learning in a distributed environment, and it is possible to train very complex networks with limited infrastructure [521]. Besides handling very large networks, distributed or parallelized approaches offer

other advantages, such as improved ensembling [522] or accelerated hyperparameter optimization [523,524].

Cloud computing, which has already seen wide adoption in genomics [525], could facilitate easier sharing of the large datasets common to biology [526,527], and may be key to scaling deep learning. Cloud computing affords researchers flexibility, and enables the use of specialized hardware (e.g. FPGAs, ASICs and GPUs) without major investment. As such, it could be easier to address the different challenges associated with the multitudinous layers and architectures available [528]. Though many are reluctant to store sensitive data (e.g. patient EHRs) in the cloud, secure, regulation-compliant cloud services do exist [529].

5.6. 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 [530]. A field that honours only discoveries and not the hard work of generating useful data will have difficulty encouraging scientists to share their hard-won data. It is 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 [531].

The sharing of high-quality, labelled 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 [229] and variation analysis [343]) and cleaned data (e.g. MoleculeNet [410]) publicly available to catalyse further research. However, there are complex privacy and legal issues involved in sharing patient data that cannot be ignored. Solving these issues will require increased understanding of privacy risks and standards specifying acceptable levels. In some domains, high-quality training data have 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. Recently, Ramsundar *et al.* [532] did this with an open source platform DeepChem, where they released four privately generated datasets.

Code sharing and open source licensing are 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 [533] 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 have been artificially encoded as images in order to benefit from pre-trained image classifiers [341]. 'Model zoos'—collections of pre-trained models—are not yet common in biomedical domains but have started to appear in genomics applications [296,534]. However, it is important to note that sharing models trained on individual data requires great care, because deep learning models can be attacked to identify examples used in training. One possible solution to protect individual samples includes training models under differential privacy [155], which has been used in the biomedical domain [158]. We discussed this issue as well as recent techniques to mitigate these concerns in the patient categorization section.

DeepChem [406,410,412] and DragoNN (Deep RegulAtory GenOmic Neural Networks) [534] exemplify the benefits of sharing pre-trained models and code under an open source licence. 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* TF Binding Site Prediction Challenge [215], DeepChem provides a forum for the fair, critical evaluations that are not always conducted in individual methodological papers, which can be biased towards favouring a new proposed algorithm. Likewise DragoNN 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.

5.7. 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 [535].

In image analysis, previous examples of deep transfer learning applications proved large-scale natural image sets

[46] to be useful for pre-training models that serve as generic feature extractors for various types of biological images [15,286,536,537]. More recently, deep learning models predicted protein subcellular localization for proteins not originally present in a training set [538]. Moreover, learned features performed reasonably well even when applied to images obtained using different fluorescent labels, imaging techniques and different cell types [539]. 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 [58]). Fine-tuning of specific biological datasets enables more focused predictions.

In genomics, the Basset package [229] 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*. [230] 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 [423,426].

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) [540]. Deep graphical models such as DBNs are 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 unlabelled data and then fine-tuned on a smaller number of labelled examples. When labels are available, CNNs are ubiquitously used because they can be trained end-to-end with backpropagation and demonstrate state-of-the-art performance in many discriminative tasks [15].

Jha *et al*. [192] 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*. [541] trained a deep autoencoder model jointly on RNA-seq, miRNA-seq and methylation data from TCGA 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*. [178] used DBNs to learn phosphorylation states of a common set of signalling 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 signalling 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 [186]. 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 the single data type. These individual representations are further concatenated before or within fully connected layers. FIDDLE [542] 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 sites. 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 [211] 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 [193] learned information shared across TFs and cell lines to predict cell-specific TF binding for TF-cell line combinations. Yoon *et al*. [104] 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 [390,394] and drug toxicity [395,543]. Kearnes *et al*. [388] 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*. [536] combined deep model-based transfer and multi-task learning for cross-domain image annotation. One could imagine extending that approach also to multimodal inputs. A common characteristic of these methods is a 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, the similarity of datasets and domains in transfer learning and relatedness of tasks in multi-task learning are difficult to access. Most studies address these limitations by empirical evaluation of the model. Unfortunately, negative results are typically not reported. A deep CNN trained on natural images boosts performance in radiographic images [58]. However, due to differences in imaging domains, the target task required either re-training the initial model from scratch with special preprocessing 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 [229]. 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 [406].

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 [544]. 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.

6. Conclusion

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% [545] and finally approaching or exceeding human performance in the past year [546,547]. 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 [545]. 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 [50], diabetic macular oedema [50], tuberculosis [59] and skin lesion [5] 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 [59], 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 modelling. Improved contact map predictions [29] have led to notable improvements in fold and 3D structure prediction for some of the most challenging proteins, such as membrane proteins [252].

Conversely, the most challenging tasks may be those in which predictions are used directly for downstream modelling or decision-making, especially in the clinic. As an example, errors in sequence variant calling will be amplified if they are used directly for genome-wide association studies. In addition, the stochasticity and complexity of biological systems imply 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 [548,549], 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 [550,551]. Cooperation between human experts and deep learning algorithms addresses many of these challenges and can achieve better performance than either individually [65]. For sample and patient classification tasks, we expect deep learning methods to augment clinicians and biomedical researchers.

We are 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 accuracies, 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 TF binding prediction and molecular graphs instead of pre-computed bit vectors for drug discovery. These flexible input feature representations have spurred creative modelling 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.

7. Methods

7.1. 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 Central IDs, PubMed IDs, arXiv IDs and URLs. References were automatically generated from citation metadata by querying APIs to generate Citation Style Language 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 372 DOIs, six PubMed Central records, 129 arXiv manuscripts and 48 URLs (webpages as well as manuscripts lacking standardized identifiers).

We implemented continuous analysis so the manuscript was automatically regenerated whenever the source changed [150]. 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 [552,553]. Using OpenTimestamps, we submitted hashes for the manuscript and the source git commit for timestamping in the Bitcoin blockchain [554]. 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.

Data accessibility. This article has no additional data.

Authors' contributions. We created an open repository on the GitHub version control platform (greenelab/deep-review) [555]. Here, we engaged with numerous authors from papers within and outside of the area. The manuscript was drafted via GitHub commits by 36 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., M.M.H., M.Z., P.A. and T.C. Drafted one or more sub-sections: A.E.C., A.M.A., A.S., B.J.L., C.A.L., E.M.C., G.L.R., J.I., J.L., J.X., S.C.T., S.W., W.X. and Z.L. Revised specific sub-sections or supervised drafting one or more sub-sections: A.H., A.K., D.D., D.J.H., L.K.W., M.H.S.S., S.J.S., S.M.B., Y.P. and Y.Q. Drafted sub-sections, edited the manuscript, reviewed pull requests and coordinated co-authors: A.G. and 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.

Funding. 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 Howard Hughes Medical Institute (S.C.T.); the National Institutes of Health awards DP2GM123485 (A.K.), P30CA051008 (S.M.B.), R01AI116794 (B.K.B.), R01GM089652 (A.E.C.), R01GM089753 (J.X.), R01LM012222 (S.J.S.), R01LM012482 (S.J.S.), R21CA220398 (S.M.B.), T32GM007753 (B.T.D.), T32HG000046 (G.P.W.) and U54AI117924 (A.G.); the National Institutes of Health Intramural Research Program and National Library of Medicine (Y.P. and Z.L.); the National Science Foundation awards 1245632 (G.L.R.), 1531594 (E.M.C.) and 1564955 (J.X.); the Natural Sciences and Engineering Research Council of Canada award RGPIN-2015-3948 (M.M.H.) and the Roy and Diana Vagelos Scholars Program in the Molecular Life Sciences (M.Z.).

Acknowledgements. We gratefully acknowledge Christof Angermueller, Kumardeep Chaudhary, Gökcen Eraslan, Mikael Huss, Bharath Ramsundar and Xun Zhu for their discussion of the manuscript and reviewed papers on GitHub. We thank Aaron Sheldon, who contributed text but did not formally approve the manuscript; Anna Greene for a careful proofreading of the manuscript in advance of the first submission; Sebastian Raschka for clarifying edits to the abstract and introduction and Robert Gieseke, Ruibang Luo, Stephen Ra, Sourav Singh and GitHub user snikumbh for correcting typos, formatting and references.

References

1. Stephens ZD *et al.* 2015 Big data: astronomical or genomical? *PLoS Biol.* **13**, e1002195. (doi:10.1371/journal.pbio.1002195)
2. LeCun Y, Bengio Y, Hinton G. 2015 Deep learning. *Nature* **521**, 436–444. (doi:10.1038/nature14539)
3. Baldi P, Sadowski P, Whiteson D. 2014 Searching for exotic particles in high-energy physics with deep learning. *Nat. Comm.* **5**, 1. (doi:10.1038/ncomms5308)
4. Goh GB, Hodas NO, Vishnu A. 2017 Deep learning for computational chemistry. *J. Comput. Chem.* **38**, 1291–1307. (doi:10.1002/jcc.24764)
5. 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. (doi:10.1038/nature21056)
6. Wu Y *et al.* 2016 Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv*. (https://arxiv.org/abs/1609.08144v2)

7. McCulloch WS, Pitts W. 1943 A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133. (doi:10.1007/bf02478259)
8. Block HD, Knight BW, Rosenblatt F. 1962 Analysis of a four-layer series-coupled perceptron. II. *Rev. Mod. Phys.* **34**, 135–142. (doi:10.1103/revmodphys.34.135)
9. Google Research Publication: Building High-level Features Using Large Scale Unsupervised Learning. 2016. See http://research.google.com/archive/unsupervised_icml2012.html.
10. Niu F, Recht B, Re C, Wright SJ. 2011 HOGWILD!: a lock-free approach to parallelizing stochastic gradient descent. arXiv, 1106.5730. See https://arxiv.org/abs/1106.5730v2.
11. Goodfellow I, Bengio Y, Courville A. 2016 Deep learning. See http://www.deeplearningbook.org/.
12. Grove AS. 1998 Academy of management. See http://www.intel.com/pressroom/archive/speeches/ag080998.htm.
13. Park Y, Kellis M. 2015 Deep learning for regulatory genomics. *Nat. Biotechnol.* **33**, 825–826. (doi:10.1038/nbt.3313)
14. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. 2016 Applications of deep learning in biomedicine. *Mol. Pharm.* **13**, 1445–1454. (doi:10.1021/acs.molpharmaceut.5b00982)
15. Angermueller C, Pärnamaa T, Parts L, Stegle O. 2016 Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878. (doi:10.15252/msb.20156651)
16. Min S, Lee B, Yoon S. 2016 Deep learning in bioinformatics. *Brief. Bioinform.* **31**, bbw068. (doi:10.1093/bib/bbw068)
17. Kraus OZ, Frey BJ. 2016 Computer vision for high content screening. *Crit. Rev. Biochem. Mol. Biol.* **51**, 102–109. (doi:10.3109/10409238.2015.1135868)
18. Miotto R, Wang F, Wang S, Jiang Z, Dudley JT. 2017 Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **375**, 4. (doi:10.1093/bib/bbx044)
19. Litjens G *et al.* 2017 A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88. (doi:10.1016/j.media.2017.07.005)
20. Kalinin AA, Higgins GA, Reamaroon N, Reza Soroushmehr SM, Allyn-Feuer A, Dinov ID, Najarian K, Athey BD. 2018 Deep learning in pharmacogenomics: from gene regulation to patient stratification. arXiv, 1801.08570 (https://arxiv.org/abs/1801.08570v1)
21. Gawehn E, Hiss JA, Schneider G. 2015 Deep learning in drug discovery. *Mol. Inform.* **35**, 3–14. (doi:10.1002/minf.201501008)
22. Pérez-Sianes J, Pérez-Sánchez H, Díaz F. 2016 Virtual screening: a challenge for deep learning. *Adv. Int. Syst. Comput.* **477**, 13–22. (doi:10.1007/978-3-319-40126-3_2)
23. Baskin II, Winkler D, Tetko IV. 2016 A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discovery* **11**, 785–795. (doi:10.1080/17460441.2016.1201262)
24. Parker JS *et al.* 2009 Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167. (doi:10.1200/jco.2008.18.1370)
25. Mayer IA, Abramson VG, Lehmann BD, Pietenpol JA. 2014 New strategies for triple-negative breast cancer—deciphering the heterogeneity. *Clin. Cancer Res.* **20**, 782–790. (doi:10.1158/1078-0432.ccr-13-0583)
26. 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. *Pac Symp Biocomput.* **20**, 132–143. (doi:10.1142/9789814644730_0014)
27. 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.
28. Zurada J. 1994 End effector target position learning using feedforward with error back-propagation and recurrent neural networks. In *Proc. of 1994 IEEE Int. Conf. on Neural Networks* (*ICNN'94*), *Orlando, FL, USA, 28 June–2 July 1994*, vol. 4, pp. 2633–2638.
29. 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 Comput. Biol.* **13**, e1005324. (doi:10.1371/journal.pcbi.1005324)
30. Spencer M, Eickholt J, Cheng J. 2015 A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **12**, 103–112. (doi:10.1109/tcbb.2014.2343960)
31. Wang S, Peng J, Ma J, Xu J. 2016 Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* **6**, 18962. (doi:10.1038/srep18962)
32. Liu F, Li H, Ren C, Bo X, Shu W. 2016 PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *bioRxiv* (doi:10.1101/036129)
33. Li Y, Chen C-Y, Wasserman WW. 2015 Deep feature selection: theory and application to identify enhancers and promoters. In (ed. T Przytycka) *Research in computational molecular biology. RECOMB 2015*. Lecture Notes in Computer Science, vol. 9029. Cham, Switzerland: Springer.
34. Kleftogiannis D, Kalnis P, Bajic VB. 2015 DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* **43**, e6. (doi:10.1093/nar/gku1058)
35. Quang D, Chen Y, Xie X. 2015 DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763. (doi:10.1093/bioinformatics/btu703)
36. Wallach I, Dzamba M, Heifets A. 2015 AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv, 1510.02855 (https://arxiv.org/abs/1510.02855v1)
37. 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. *Mol. Pharm.* **13**, 2524–2530. (doi:10.1021/acs.molpharmaceut.6b00248)
38. Wang Y, Zeng J. 2013 Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* **29**, i126–i134. (doi:10.1093/bioinformatics/btt234)
39. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H. 2017 Deep-learning-based drug–target interaction prediction. *J. Proteome Res.* **16**, 1401–1409. (doi:10.1021/acs.jproteome.6b00618)
40. Shen D, Wu G, Suk H. 2017 Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248. (doi:10.1146/annurev-bioeng-071516-044442)
41. Dhungel N, Carneiro G, Bradley AP. 2015 Deep learning and structured prediction for the segmentation of mass in mammograms. In *18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Munich, Germany, October*. Lecture Notes in Computer Science, vol. 9349. Cham, Switzerland: Springer.
42. 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*. Lecture Notes in Computer Science, vol. 9901. Cham: Springer.
43. Zhu W, Lou Q, Scott Vang Y, Xie X. 2016 Deep multi-instance networks with sparse label assignment for whole mammogram classification. *bioRxiv*. (doi:10.1101/095794)
44. Zhu W, Xie X. 2016 Adversarial deep structural networks for mammographic mass segmentation. *bioRxiv* (doi:10.1101/095786)
45. Dhungel N, Carneiro G, Bradley AP. 2017 A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med. Image Anal.* **37**, 114–128. (doi:10.1016/j.media.2017.01.009)
46. Russakovsky O *et al.* 2015 ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252. (doi:10.1007/s11263-015-0816-y)
47. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. 2016 Convolutional neural networks for diabetic retinopathy. *Procedia Comp. Sci.* **90**, 200–205. (doi:10.1016/j.procs.2016.07.014)
48. Abràmoff DM, 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. *Invest. Opthalmol. Vis. Sci.* **57**, 5200. (doi:10.1167/iovs.16-19964)
49. Leibig C, Allken V, Seckin Ayhan M, Berens P, Wahl S. 2016 Leveraging uncertainty information from deep neural networks for disease detection. *bioRxiv*. (doi:10.1101/084210)
50. 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. (doi:10.1001/jama.2016.17216)
51. 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. *arXiv*, 1610.04662 (https://arxiv.org/abs/1610.04662v2)

52. Yu L, Chen H, Dou Q, Qin J, Heng P-A. 2017 Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* **36**, 994–1004. (doi:10.1109/tmi.2016.2642839)
53. Hossein Jafari M, Nasr-Esfahani E, Karimi N, Reza Soroushmehr SM, Samavi S, Najarian K. 2017 Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 1021–1030. (doi:10.1007/s11548-017-1567-8)
54. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, Najarian K. 2016 Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), *Orlando, FL, USA, 16–20 August 2016*, pp. 1373–1376.
55. Burlina P, Freund DE, Joshi N, Wolfson Y, Bressler NM. 2016 Detection of age-related macular degeneration via deep learning. In *2016 IEEE 13th Int. Symp. on Biomedical Imaging* (*ISBI*), *Czech Republic, 13–16 April 2016*, pp. 184–188.
56. 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, Orlando, FL, USA, 21–26 February 2015*, pp. 94140V.
57. Shin H-C, Roth HR, Mingchen Gao LL, 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 Trans. Med. Imaging* **35**, 1285–1298. (doi:10.1109/tmi.2016.2528162)
58. Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. 2017 High-throughput classification of radiographs using deep convolutional neural networks. *J. Digit. Imaging* **30**, 95–101. (doi:10.1007/s10278-016-9914-9)
59. Lakhani P, Sundaram B. 2017 Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582. (doi:10.1148/radiol.2017162326)
60. 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, Orlando, FL, USA, 11–16 February 2017*, pp. 101341H. SPIE.
61. Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, Kim L, Summers RM. 2016 Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans. Med. Imaging* **35**, 1170–1181. (doi:10.1109/tmi.2015.2482920)
62. 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* (eds S Ourselin, L Joskowicz, M Sabuncu, G Unal, W Wells), Lecture Notes in Computer Science, vol. 9901, October 17–21, 2016, Istanbul, Turkey. Cham: Springer.
63. 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. *Med. Image Anal.* **35**, 303–312. (doi:10.1016/j.media.2016.07.007)
64. Litjens G *et al.* 2016 Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286. (doi:10.1038/srep26286)
65. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. 2016 Deep learning for identifying metastatic breast cancer. *arXiv*, 1606.05718 (https://arxiv.org/abs/1606.05718v1)
66. Rakhlin A, Shvets A, Iglovikov V, Kalinin A. 2018 Deep convolutional neural networks for breast cancer histology image analysis. *bioRxiv*. (doi:10.1101/259911)
67. Lee CS, Baughman DM, Lee AY. 2016 Deep learning is effective for the classification of OCT oimages of normal versus age-related macular degeneration. *bioRxiv*. (doi:10.1101/094276)
68. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of the 25th Int. Conf. on Neural Information Processing Systems*. See http://dl.acm.org/citation.cfm?id=2999134.2999257.
69. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Bretonnel Cohen K, Duch W. 2007 A shared task involving multi-label classification of clinical free text. In *Proc. of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing—BioNLP '07, Prague, Czech Republic, 29 June 2007*, Association for Computational Linguistics Stroudsburg, PA, USA.
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. *arXiv*, 1705.02315 (https://arxiv.org/abs/1705.02315v4)
71. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. 2017 NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *arXiv*, 1712.05898 (https://arxiv.org/abs/1712.05898v2)
72. Lever J, Krzywinski M, Altman N. 2016 Points of significance: classification evaluation. *Nat. Methods* **13**, 603–604. (doi:10.1038/nmeth.3945)
73. NIH Clinical Center. 2017 NIH chest X-ray dataset. See https://nihcc.app.box.com/v/ChestXray-NIHCC.
74. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. 2017 Pediatric bone age assessment using deep convolutional neural networks. *bioRxiv* (doi:10.1101/234120)
75. Leaman R, Lu Z. 2016 TaggerOne: joint named entity recognition and normalization with semi-Markov models. *Bioinformatics* **32**, 2839–2846. (doi:10.1093/bioinformatics/btw343)
76. Wei C-H, Harris BR, Kao H-Y, Lu Z. 2013 tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* **29**, 1433–1439. (doi:10.1093/bioinformatics/btt156)
77. Leaman R, Islamaj Dogan R, Lu Z. 2013 DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* **29**, 2909–2917. (doi:10.1093/bioinformatics/btt474)
78. Liu S, Tang B, Chen Q, Wang X. 2015 Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information* **6**, 848–865. (doi:10.3390/info6040848)
79. Tang B, Cao H, Wang X, Chen Q, Xu H. 2014 Evaluating word representation features in biomedical named entity recognition tasks. *BioMed Res. Int.* **2014**, 1–6. (doi:10.1155/2014/240403)
80. Wu Y, Xu J, Zhang Y, Xu H. 2015 Clinical abbreviation disambiguation using neural word embeddings. In *Proc. of the 2015 Workshop on Biomedical Natural Language Processing* (*BioNLP 2015*), *Beijing, China, 30 July 2015*. Association for Computational Linguistics, Stroudsburg, PA, pp. 171–176.
81. Liu Y, Ge T, Mathews K, Ji H, McGuinness D. 2015 Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In *Proc. of the 2015 Workshop on Biomedical Natural Language Processing* (*BioNLP 2015*), *Beijing, China, 30 July 2015*. Association for Computational Linguistics, Stroudsburg, PA, pp. 92–97.
82. Tikk D, Thomas P, Palaga P, Hakenberg J, Leser U. 2010 A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput. Biol.* **6**, e1000837. (doi:10.1371/journal.pcbi.1000837)
83. Peng Y, Wei C-H, Lu Z. 2016 Improving chemical disease relation extraction with rich features and weakly labeled data. *J. Cheminformatics* **8**, 1. (doi:10.1186/s13321-016-0165-z)
84. Niu Y, Otasek D, Jurisica I. 2010 Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics* **26**, 111–119. (doi:10.1093/bioinformatics/btp602)
85. Li F, Zhang Y, Zhang M, Ji D. 2016 Joint models for extracting adverse drug events from biomedical text. In *Proc. Twenty-Fifth Int. Joint Conf. on Artificial Intelligence* See http://dl.acm.org/citation.cfm?id=3060832.3061018.
86. Li F, Zhang M, Fu G, Ji D. 2017 A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* **18**, 1. (doi:10.1186/s12859-017-1609-9)
87. Peng Y, Lu Z. 2017 Deep learning for extracting protein-protein interactions from biomedical literature. In *Proc. of the BioNLP 2017 Workshop, Vancouver, Canada, 4 August 2017*, pp. 29–38. Stroudsburg, PA: Association for Computational Linguistics.
88. Hua L, Quan C. 2016 A shortest dependency path based convolutional neural network for protein-protein relation extraction. *BioMed Res. Int.* **2016**, 1–9. (doi:10.1155/2016/8479587)

89. Quan C, Hua L, Sun X, Bai W. 2016 Multichannel convolutional neural network for biological relation extraction. *BioMed Res. Int.* **2016**, 1–10. (doi:10.1155/2016/1850404)
90. Jiang Z, Li S, Huang D. 2016 A general protein-protein interaction extraction architecture based on word representation and feature selection. *Int. J. Data Mining Bioinf.* **14**, 276. (doi:10.1504/ijdmb.2016.074878)
91. Gu J, Sun F, Qian L, Zhou G. 2017 Chemical-induced disease relation extraction via convolutional neural network. *Database* **2017**, bax024. (doi:10.1093/database/bax024)
92. Zhao Z, Yang Z, Luo L, Lin H, Wang J. 2016 Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* **3**, btw486. (doi:10.1093/bioinformatics/btw486)
93. Asada M, Miwa M, Sasaki S. 2017 Extracting drug–drug interactions with attention CNNs. In *Proc. of the BioNLP 2017 Workshop, Vancouver, Canada, 4 August 2017*, pp. 9–18. Stroudsburg, PA: Association for Computational Linguistics.
94. Yi Z, Li S, Yu J, Wu Q. 2017 Drug-drug interaction extraction via recurrent neural network with multiple attention layers. *arXiv* (https://arxiv.org/abs/1705.03261v2)
95. Li H, Zhang J, Wang J, Lin H, Yang Z. 2016 DUTIR in BioNLP-ST 2016: utilizing convolutional network and distributed representation to extract complicate relations. In *Proc. of the 4th BioNLP Shared Task Workshop, 13 August 2016, Berlin, Germany*, pp. 93–100. Stroudsburg, PA: Association for Computational Linguistics
96. Mehryary F, Björne J, Pyysalo S, Salakoski T, Ginter F. 2016 Deep learning with minimal training data: TurkuNLP entry in the BioNLP shared task 2016. In *Proc. of the 4th BioNLP Shared Task Workshop, 13 August 2016, Berlin, Germany*, pp. 73–81. Stroudsburg, PA: Association for Computational Linguistics.
97. Li C, Song R, Liakata M, Vlachos A, Seneff S, Zhang X. 2015 Using word embedding for bio-event extraction. In *Proc. of the 2015 Workshop on Biomedical Natural Language Processing* (*BioNLP 2015*), *Beijing, China, 30 July 2015*, pp. 121–126. Stroudsburg, PA: Association for Computational Linguistics.
98. Nie Y, Rong W, Zhang Y, Ouyang Y, Xiong Z. 2015 Embedding assisted prediction architecture for event trigger identification. *J. Bioinform. Comput. Biol.* **13**, 1541001. (doi:10.1142/s0219720015410012)
99. Rahul PVSS, Sahu SK, Anand A. 2017 Biomedical event trigger identification using bidirectional recurrent neural network based models. *arXiv* (https://arxiv.org/abs/1705.09516v1)
100. Mohan S, Fiorini N, Kim S, Lu Z. 2017 Deep learning for biomedical information retrieval: learning textual relevance from click logs. In *Proc. of the BioNLP 2017 Workshop, Vancouver , Canada, 4 August 2017*, pp. 222–231. Stroudsburg, PA: Association for Computational Linguistics.
101. Ohno-Machado L. 2011 Realizing the full potential of electronic health records: the role of natural language processing. *J. Am. Med. Inform. Assoc.* **18**, 539. (doi:10.1136/amiajnl-2011-000501)
102. Bruijn Bd, 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. *J. Am. Med. Inform. Assoc.* **18**, 557–562. (doi:10.1136/amiajnl-2011-000150)
103. Chalapathy R, Borzeshi EZ, Piccardi M. 2016 Bidirectional LSTM-CRF for clinical concept extraction. *arXiv* (https://arxiv.org/abs/1611.08373v1)
104. 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, INNS 2016, 23–25 October 2016, Thessaloniki, Greece* (eds P Angelov, Y Manolopoulos, L Iliadis, A Roy, M Vellasco). Advances in Intelligent Systems and Computing, vol. 529. Cham: Springer.
105. Mikolov T, Chen K, Corrado G, Dean J. 2013 Efficient estimation of word representations in vector space. *arXiv* (https://arxiv.org/abs/1301.3781v3)
106. Antonio M-GJ, Oscar M-A, Matthias S. 2014 Exploring the application of deep learning techniques on medical text corpora. *Stud. Health Technol. Inform.* **206**, 584–588. (doi:10.3233/978-1-61499-432-9-584)
107. De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. 2014 Medical semantic similarity with a neural language model. In *Proc. of the 23rd ACM Int. Conf. on Information and Knowledge Management—CIKM '14, 3–7 November 2014, Shanghai, China*, pp. 1819–1822. New York, NY, USA: ACM.
108. Karimi S, Dai X, Hassanzadeh H, Nguyen A. 2017 Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In *Proc. of the BioNLP 2017 Workshop, 4 August 2017, Vancouver, Canada*, pp. 328–332. Stroudsburg, PA: Association for Computational Linguistics.
109. International Classification of Diseases. 2017 (http://www.who.int/classifications/icd/en/)
110. 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 *Proc. of the 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining - KDD '16' 13–17 August 2016, San Francisco, CA, USA*, pp. 1495–1504. New York, NY, USA: ACM.
111. 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. *Sci. Rep.* **6**, 32404. (doi:10.1038/srep32404)
112. Jagannatha AN, Yu H. 2016 Bidirectional RNN for Medical Event Detection in Electronic Health Records. In *Proc. of the Conf. Association for Computational Linguistics. North American Chapter. Meeting*. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5119627/.
113. Lin C, Miller T, Dligach D, Bethard S, Savova G. 2017 Representations of time expressions for temporal relation extraction with convolutional neural networks. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017*, vol. 2, short papers, pp. 746–751. Stroudsburg, PA: Association for Computational Linguistics.
114. 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. (doi:10.1371/journal.pone.0066341)
115. Beaulieu-Jones BK, Greene CS. 2016 Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* **64**, 168–178. (doi:10.1016/j.jbi.2016.10.007)
116. 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. *Sci. Rep.* **6**, 26094. (doi:10.1038/srep26094)
117. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. 2015 Doctor AI: predicting clinical events via recurrent neural networks. *arXiv* (https://arxiv.org/abs/1511.05942v11)
118. Pham T, Tran T, Phung D, Venkatesh S. 2016 DeepCare: a deep dynamic memory model for predictive medicine. *arXiv* (https://arxiv.org/abs/1602.00357v2)
119. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. 2017 Deepr: a convolutional net for medical records. *IEEE J. Biomed. Health Inform.* **21**, 22–30. (doi:10.1109/jbhi.2016.2633963)
120. Razavian N, Marcus J, Sontag D. 2016 Multi-task prediction of disease onsets from longitudinal lab tests. *arXiv* (https://arxiv.org/abs/1608.00647v3)
121. Ranganath R, Perotte A, Elhadad N, Blei D. 2016 Deep survival analysis. *arXiv* (https://arxiv.org/abs/1608.02158v2)
122. 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. *Comput. Stat. Data Anal.* **34**, 243–257. (doi:10.1016/s0167-9473(99)00098-5)
123. Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, Kluger Y. 2016 DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *arXiv* (https://arxiv.org/abs/1606.00931v3)
124. Ranganath R, Tang L, Charlin L, Blei DM. 2014 Deep exponential families. *arXiv* (https://arxiv.org/abs/1411.2581v1)
125. Hoffman M, Blei DM, Wang C, Paisley J. 2012 Stochastic variational inference. *arXiv* (https://arxiv.org/abs/1206.7051v3)
126. Ranganath R, Tran D, Blei DM. 2015 Hierarchical variational models. *arXiv* (https://arxiv.org/abs/1511.02386v2)
127. 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. *Int. J. Med. Inform.* **97**, 120–127. (doi:10.1016/j.ijmedinf.2016.09.014)

128. Implementations by Phenotype | PheKB. 2017 (https://phekb.org/implementations)
129. Halpern Y, Horng S, Choi Y, Sontag D. 2016 Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Inform. Assoc.* **23**, 731–740. (doi:10.1093/jamia/ocw011)
130. Ratner A, De Sa C, Wu S, Selsam D, Ré C. 2016 Data programming: creating large training sets, quickly. *arXiv* (https://arxiv.org/abs/1605.07723v3)
131. Palmer M. 2006 Data is the new oil. *ANA marketing maestros*. (http://ana.blogs.com/maestros/2006/11/data_is_the_new.html)
132. Haupt M. 2016 'Data is the New Oil'—A ludicrous proposition. *Medium*. See https://medium.com/twenty-one-hundred/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294.
133. Ratner A, Bach S, Ré C. 2016 Data programming: machine learning with weak supervision. See http://hazyresearch.github.io/snorkel/blog/weak_supervision.html.
134. Jensen PB, Jensen LJ, Brunak S. 2012 Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405. (doi:10.1038/nrg3208)
135. Weiskopf NG, Weng C. 2013 Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **20**, 144–151. (doi:10.1136/amiajnl-2011-000681)
136. Bowman S. 2013 Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect. Health Inf. Manag.* **10**, 1c.
137. Botsis T, Hartvigsen G, Chen F, Weng C. 2010 Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Trans. Bioinform.* **2010**, 1–5.
138. 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. (doi:10.1016/s0168-8510(02)00208-7)
139. 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.
140. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. 2014 Identifying and mitigating biases in EHR laboratory tests. *J. Biomed. Inform.* **51**, 24–34. (doi:10.1016/j.jbi.2014.03.016)
141. De Moor G *et al.* 2015 Using electronic health records for clinical research: the case of the EHR4CR project. *J. Biomed. Inform.* **53**, 162–173. (doi:10.1016/j.jbi.2014.10.006)
142. Oemig F, Snelick R. 2016 *Healthcare interoperability standards compliance handbook*. Cham, The Netherlands: Springer International Publishing.
143. Faber J, Fonseca LM. 2014 How sample size influences research outcomes. *Dental Press J. Orthod.* **19**, 27–29. (doi:10.1590/2176-9451.19.4.027-029.ebo)
144. 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. *J. Am. Med. Inform. Assoc.* **21**, 221–230. (doi:10.1136/amiajnl-2013-001935)
145. Wiley LK, Vanhouten JP, Samuels DC, Aldrich MC, Roden DM, Peterson JF, Dennyl JC. 2017 Strategies for equitable pharmacogenomic-guided warfarin dosing among European and African American individuals in a clinical population. *Pac Symp Biocomput.* **22**, 545–556. (doi:10.1142/9789813207813_0050)
146. Rahu M, McKee M. 2008 Epidemiological research labelled as a violation of privacy: the case of Estonia. *Int. J. Epidemiol.* **37**, 678–682. (doi:10.1093/ije/dyn022)
147. 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. *J. Am. Med. Inform. Assoc.* **23**, 413–419. (doi:10.1093/jamia/ocv111)
148. Gaye A *et al.* 2014 DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int. J. Epidemiol.* **43**, 1929–1944. (doi:10.1093/ije/dyu188)
149. Carter KW *et al.* 2016 ViPAR: a software platform for the virtual pooling and analysis of research data. *Int. J. Epidemiol.* **45**, 408–416. (doi:10.1093/ije/dyv193)
150. Beaulieu-Jones BK, Greene CS. 2017 Reproducibility of computational workflows is automated using continuous analysis. *Nat. Biotechnol.* **35**, 342–346. (doi:10.1038/nbt.3780)
151. Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. 2016 Stealing machine learning models via prediction APIs. *arXiv* (https://arxiv.org/abs/1609.02943v2)
152. Dwork C, Roth A. 2013 The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407. (doi:10.1561/0400000042)
153. Shokri R, Stronati M, Song C, Shmatikov V. 2016 Membership inference attacks against machine learning models. *arXiv* (https://arxiv.org/abs/1610.05820v2)
154. Simmons S, Sahinalp C, Berger B. 2016 Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Syst.* **3**, 54–61. (doi:10.1016/j.cels.2016.04.013)
155. Abadi M, Chu A, Goodfellow I, Brendan McMahan H, Mironov I, Talwar K, Zhang L. 2016 Deep Learning with Differential Privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security - CCS'16, 24–28 October 2016, Vienna, Austria*, pp. 308–318. New York, NY, USA: ACM.
156. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. 2017 Generating multi-label discrete electronic health records using generative adversarial networks. *arXiv* (https://arxiv.org/abs/1703.06490v1)
157. Esteban C, Hyland SL, Rätsch G. 2017 Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv* (https://arxiv.org/abs/1706.02633v1)
158. Beaulieu-Jones BK, Wu ZS, Williams C, Byrd JB, Greene CS. 2017 Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv* (doi:10.1101/159756)
159. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. 2017 Communication-efficient learning of deep networks from decentralized data. See http://proceedings.mlr.press/v54/mcmahan17a.html.
160. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, Brendan McMahan H, Patel S, Ramage D, Segal A, Seth K. 2017 Practical secure aggregation for privacy preserving machine learning. See https://eprint.iacr.org/2017/281.
161. Goodman B, Flaxman S. 2016 European Union regulations on algorithmic decision-making and a 'right to explanation'. *arXiv* (https://arxiv.org/abs/1606.08813v3)
162. Zöllner S, Pritchard JK. 2007 Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**, 605–615. (doi:10.1086/512821)
163. Beery AK, Zucker I. 2011 Sex bias in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* **35**, 565–572. (doi:10.1016/j.neubiorev.2010.07.002)
164. Carlson CS *et al.* 2013 Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661. (doi:10.1371/journal.pbio.1001661)
165. Price AL, Zaitlen NA, Reich D, Patterson N. 2010 New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463. (doi:10.1038/nrg2813)
166. Sebastiani P *et al.* 2011 Retraction. *Science* **333**, 404. (doi:10.1126/science.333.6041.404-a)
167. Kaufman S, Rosset S, Perlich C, Stitelman O. 2012 Leakage in data mining. *ACM Trans. Knowl. Discov. Data* **6**, 1–21. (doi:10.1145/2382577.2382579)
168. Lum K, Isaac W. 2016 To predict and serve? *Significance* **13**, 14–19. (doi:10.1111/j.1740-9713.2016.00960.x)
169. Hardt M, Price E, Srebro N. 2016 Equality of opportunity in supervised learning. *arXiv* (https://arxiv.org/abs/1610.02413v1)
170. Joseph M, Kearns M, Morgenstern J, Neel S, Roth A. 2016 Fair algorithms for infinite and contextual bandits. *arXiv* (https://arxiv.org/abs/1610.09559v4)
171. Mahmood SS, Levy D, Vasan RS, Wang TJ. 2014 The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008. (doi:10.1016/s0140-6736(13)61752-3)
172. Pearson H. 2012 Children of the 90s: coming of age. *Nature* **484**, 155–158. (doi:10.1038/484155a)
173. Kaplan EL, Meier P. 1958 Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457. (doi:10.2307/2281868)
174. 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. *Nat. Commun.* **5**, 1769. (doi:10.1038/ncomms5022)

175. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. 2016 Deepr: a convolutional net for medical records. *arXiv* (https://arxiv.org/abs/1607.07519v1)
176. NIH. 2012 Curiosity creates cures: the value and impact of basic research. See https://www.nigms.nih.gov/Education/Documents/curiosity.pdf.
177. Kim M, Rai N, Zorraquino V, Tagkopoulos I. 2016 Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* **7**, 13090. (doi:10.1038/ncomms13090)
178. 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. (doi:10.1093/bioinformatics/btv315)
179. Gupta A, Wang H, Ganapathiraju M. 2015 Learning structure in gene expression data using deep architectures, with an application to gene clustering. *bioRxiv* (doi:10.1101/031906)
180. 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**, 51. (doi:10.1186/s12859-015-0852-1)
181. Tan J, Hammond JH, Hogan DA, Greene CS. 2016 ADAGE-based integration of publicly available *Pseudomonas aeruginosa* gene expression data with denoising autoencoders illuminates microbe–host interactions. *mSystems* **1**, e00025-15. (doi:10.1128/msystems.00025-15)
182. Tan J *et al.* 2016 Unsupervised extraction of stable expression signatures from public compendia with eADAGE. *bioRxiv*. (doi:10.1101/078659)
183. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. 2016 Gene expression inference with deep learning. *Bioinformatics* **32**, 1832–1839. (doi:10.1093/bioinformatics/btw074)
184. Singh R, Lanchantin J, Robins G, Qi Y. 2016 DeepChrome: deep-learning for predicting gene expression from histone modifications. *arXiv* (https://arxiv.org/abs/1607.02078v1)
185. Singh R, Lanchantin J, Sekhon A, Qi Y. 2017 Attend and predict: understanding gene regulation by selective attention on chromatin. *arXiv* (https://arxiv.org/abs/1708.00339v3)
186. 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 Trans. Comput. Biol. Bioinform.* **12**, 928–937. (doi:10.1109/tcbb.2014.2377729)
187. Scotti MM, Swanson MS. 2016 RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32. (doi:10.1038/nrg.2015.3)
188. 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. (doi:10.1126/science.aad9417)
189. 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. (doi:10.1038/nature09000)
190. Xiong HY, Barash Y, Frey BJ. 2011 Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–2562. (doi:10.1093/bioinformatics/btr444)
191. Xiong HY *et al.* 2015 The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806. (doi:10.1126/science.1254806)
192. Jha A, Gazzara MR, Barash Y. 2017 Integrative deep models for alternative splicing. *bioRxiv*. (doi:10.1101/104869)
193. Qin Q, Feng J. 2017 Imputation for transcription factor binding predictions based on deep learning. *PLoS Comput. Biol.* **13**, e1005403. (doi:10.1371/journal.pcbi.1005403)
194. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015 Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711. (doi:10.1016/j.cell.2015.09.054)
195. Juan-Mateu J, Villate O, Eizirik DL. 2016 Mechanisms in endocrinology: alternative splicing: the new frontier in diabetes research. *Eur. J. Endocrinol.* **174**, R225–R238. (doi:10.1530/eje-15-0916)
196. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. 2014 Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399. (doi:10.1016/j.tibs.2014.07.002)
197. Dunham I *et al.* 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. (doi:10.1038/nature11247)
198. Stormo GD. 2000 DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23. (doi:10.1093/bioinformatics/16.1.16)
199. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208. (doi:10.1093/nar/gkp335)
200. Weirauch MT *et al.* 2013 Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134. (doi:10.1038/nbt.2486)
201. Agius P, Arvey A, Chang W, Noble WS, Leslie C. 2010 High resolution models of transcription factor-DNA affinities improve *in vitro* and *in vivo* binding predictions. *PLoS Comput. Biol.* **6**, e1000916. (doi:10.1371/journal.pcbi.1000916)
202. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014 Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711. (doi:10.1371/journal.pcbi.1003711)
203. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015 Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838. (doi:10.1038/nbt.3300)
204. 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**, 106. (doi:10.1186/s12859-017-1561-8)
205. Zeng H, Edwards MD, Liu G, Gifford DK. 2016 Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* **32**, i121–i127. (doi:10.1093/bioinformatics/btw255)
206. Lanchantin J, Singh R, Wang B, Qi Y. 2016 Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. *arXiv* (https://arxiv.org/abs/1608.03644v4)
207. Morrow A, Shankar V, Petersohn D, Joseph A, Recht B, Yosef N. 2017 Convolutional kitchen sinks for transcription factor binding site prediction. *arXiv* (https://arxiv.org/abs/1706.00125v1)
208. Chen D, Jacob L, Mairal J. 2017 Predicting transcription factor binding sites with convolutional kernel networks. *bioRxiv*. (doi:10.1101/217257)
209. Shrikumar A, Greenside P, Kundaje A. 2017 Reverse-complement parameter sharing improves deep learning models for genomics. *bioRxiv*. (doi:10.1101/103663)
210. Alexandari AM, Shrikumar A, Kundaje A. 2017 Separable fully connected layers improve deep learning models for genomics. *bioRxiv*. (doi:10.1101/146431)
211. Zhou J, Troyanskaya OG. 2015 Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934. (doi:10.1038/nmeth.3547)
212. Quang D, Xie X. 2016 DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107. (doi:10.1093/nar/gkw226)
213. Arvey A, Agius P, Noble WS, Leslie C. 2012 Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22**, 1723–1734. (doi:10.1101/gr.127712.111)
214. Gusmao EG, Allhoff M, Zenke M, Costa IG. 2016 Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* **13**, 303–309. (doi:10.1038/nmeth.3772)
215. ENCODE-DREAM *in vivo* TRANSCRIPTION FACTOR BINDING SITE PREDICTION CHALLENGE. 2017 See https://www.synapse.org/#!Synapse:syn6131484/wiki/402026.
216. Quang D, Xie X. 2017 FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *bioRxiv*. (doi:10.1101/151274)
217. Keilwagen J, Posch S, Grau J. 2017 Learning from mistakes: accurate prediction of cell type-specific transcription factor binding. *bioRxiv*. (doi:10.1101/230011)
218. Singh R, Lanchantin J, Robins G, Qi Y. 2016 Transfer string kernel for cross-context DNA-protein binding prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **PP**, 1. (doi:10.1109/tcbb.2016.2609918)
219. Long M, Cao Y, Wang J, Jordan MI. 2015 Learning transferable features with deep adaptation networks. *arXiv* (https://arxiv.org/abs/1502.02791v2)
220. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. 2015

Domain-adversarial training of neural networks. *arXiv* (https://arxiv.org/abs/1505.07818v4)

221. Shrikumar A, Greenside P, Kundaje A. 2017 Learning important features through propagating activation differences. *arXiv* (https://arxiv.org/abs/1704.02685v1)
222. Werner T. 2003 The state of the art of mammalian promoter recognition. *Brief. Bioinform.* **4**, 22–30. (doi:10.1093/bib/4.1.22)
223. Matis S, Xu Y, Shah M, Guan X, Ralph Einstein J, Mural R, Uberbacher E. 1996 Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Comput. Chem.* **20**, 135–140. (doi:10.1016/s0097-8485(96)80015-5)
224. Umarov RK, Solovyev VV. 2017 Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* **12**, e0171410. (doi:10.1371/journal.pone.0171410)
225. Shiraki T *et al.* 2003 Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15 776–15 781. (doi:10.1073/pnas.2136655100)
226. Yamashita R *et al.* 2011 Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.* **21**, 775–789. (doi:10.1101/gr.110254.110)
227. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. 2013 Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295. (doi:10.1038/nrg3458)
228. Andersson R, Sandelin A, Danko CG. 2015 A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433. (doi:10.1016/j.tig.2015.05.007)
229. Kelley DR, Snoek J, Rinn JL. 2016 Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999. (doi:10.1101/gr.200535.115)
230. Min X, Chen N, Chen T, Jiang R. 2016 DeepEnhancer: predicting enhancers by convolutional neural networks. In *2016 IEEE Int. Conf. on Bioinformatics and Biomedicine* (*BIBM*), *December 15–18 2016, Shenzhen, China*, pp. 637–644. IEEE.
231. Li Y, Shi W, Wasserman WW. 2016 Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *bioRxiv*. (doi:10.1101/041616)
232. Singh S, Yang Y, Poczos B, Ma J. 2016 Predicting enhancer–promoter interaction from genomic sequence with deep neural networks. *bioRxiv*. (doi:10.1101/085241)
233. Bracken CP, Scott HS, Goodall GJ. 2016 A network-biology perspective of microRNA function and dysfunction in cancer. *Nat. Rev. Genet.* **17**, 719–732. (doi:10.1038/nrg.2016.134)
234. Berezikov E. 2011 Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* **12**, 846–860. (doi:10.1038/nrg3079)
235. Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015 Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, 101. (doi:10.7554/elife.05005)
236. Lee B, Baek J, Park S, Yoon S. 2016 deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. *arXiv* (https://arxiv.org/abs/1603.09123v2)
237. Park S, Min S, Choi H, Yoon S. 2016 DeepMiRGene: deep neural network based precursor microrna prediction. *arXiv* (https://arxiv.org/abs/1605.00017v1)
238. 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, ECML PKDD 2016, 19–23 September 2016* (eds P Frasconi, N Landwehr, G Manco, J Vreeken). Lecture Notes in Computer Science, vol. 9852. Cham/Riva del Garda: Springer.
239. Jones DT, Singh T, Kosciolek T, Tetchner S. 2015 MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999–1006. (doi:10.1093/bioinformatics/btu791)
240. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009 Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72. (doi:10.1073/pnas.0805923106)
241. 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. (doi:10.1371/journal.pone.0028766)
242. Qi Y, Oja M, Weston J, Noble WS. 2012 A unified multitask architecture for predicting local protein properties. *PLoS ONE* **7**, e32235. (doi:10.1371/journal.pone.0032235)
243. 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. *Sci. Rep.* **5**, 11476. (doi:10.1038/srep11476)
244. Jones DT. 1999 Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202. (doi:10.1006/jmbi.1999.3091)
245. Zhou J, Troyanskaya OG. 2014 Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *arXiv* (https://arxiv.org/abs/1403.1347v1)
246. 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. (doi:10.1093/bioinformatics/btv472)
247. Lena PD, Nagata K, Baldi P. 2012 Deep architectures for protein contact map prediction. *Bioinformatics* **28**, 2449–2457. (doi:10.1093/bioinformatics/bts475)
248. Eickholt J, Cheng J. 2012 Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics* **28**, 3066–3072. (doi:10.1093/bioinformatics/bts598)
249. Skwark MJ, Raimondi D, Michel M, Elofsson A. 2014 Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput. Biol.* **10**, e1003889. (doi:10.1371/journal.pcbi.1003889)
250. RR Results - CASP12. 2017. See http://www.predictioncenter.org/casp12/rrc_avrg_results.cgi.
251. CAMEO - Continuous Automated Model Evaluation. 2017. See http://www.cameo3d.org/.
252. Li Z, Wang S, Yu Y, Xu J. 2017 Predicting membrane protein contacts from non-membrane proteins by deep transfer learning. *arXiv* (https://arxiv.org/abs/1704.07207v1)
253. AlQuraishi M. 2018 End-to-end differentiable learning of protein structure. *bioRxiv*. (doi:10.1101/265231)
254. Cheng Y. 2015 Single-particle Cryo-EM at crystallographic resolution. *Cell* **161**, 450–457. (doi:10.1016/j.cell.2015.03.049)
255. Cheng Y, Grigorieff N, Penczek PA, Walz T. 2015 A primer to single-particle cryo-electron microscopy. *Cell* **161**, 438–449. (doi:10.1016/j.cell.2015.03.050)
256. Woolford D *et al.* 2007 SwarmPS: rapid, semi-automated single particle selection software. *J. Struct. Biol.* **157**, 174–188. (doi:10.1016/j.jsb.2006.04.006)
257. Scheres SHW. 2015 Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122. (doi:10.1016/j.jsb.2014.11.010)
258. Wang F, Gong H, Liu G, Li M, Yan C, Xia T, Li X, Zeng J. 2016 DeepPicker: a deep learning approach for fully automated particle picking in cryo-EM. *J. Struct. Biol.* **195**, 325–336. (doi:10.1016/j.jsb.2016.07.006)
259. Zhu Y, Ouyang Q, Mao Y. 2017 A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics* **18**, 29. (doi:10.1186/s12859-017-1757-y)
260. Wu J, Ma Y-B, Congdon C, Brett B, Chen S, Xu Y, Ouyang Q, Mao Y. 2017 Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning. *PLoS ONE* **12**, e0182130. (doi:10.1371/journal.pone.0182130)
261. De Las Rivas J, Fontanillo C. 2010 Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* **6**, e1000807. (doi:10.1371/journal.pcbi.1000807)
262. Zhou D, He Y. 2008 Extracting interactions between proteins from the literature. *J. Biomed. Inform.* **41**, 393–407. (doi:10.1016/j.jbi.2007.11.008)
263. Peng Y, Lu Z. 2017 Deep learning for extracting protein–protein interactions from biomedical literature. *arXiv* (https://arxiv.org/abs/1706.01556v2)
264. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. 2017 DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J. Chem. Inf. Model.* **57**, 1499–1510. (doi:10.1021/acs.jcim.7b00028)
265. Sun T, Zhou B, Lai L, Pei J. 2017 Sequence-based prediction of protein–protein interaction using a deep-learning algorithm. *BMC Bioinformatics* **18**, 1. (doi:10.1186/s12859-017-1700-2)
266. Wang Y-B, You Z-H, Li X, Jiang T-H, Chen X, Zhou X, Wang L. 2017 Predicting protein–protein interactions from protein sequences by a stacked

sparse autoencoder deep neural network. *Mol. Biosyst.* **13**, 1336–1344. (doi:10.1039/c7mb00188f)

267. Du T, Liao L, Wu CH, Sun B. 2016 Prediction of residue-residue contact matrix for protein–protein interaction with Fisher score features and deep learning. *Methods* **110**, 97–105. (doi:10.1016/j.ymeth.2016.06.001)
268. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. 2003 Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017. (doi:10.1110/ps.0239403)
269. Andreatta M, Nielsen M. 2016 Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517. (doi:10.1093/bioinformatics/btv639)
270. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M. 2009 NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13. (doi:10.1007/s00251-008-0341-z)
271. Nielsen M, Andreatta M. 2016 NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 51. (doi:10.1186/s13073-016-0288-x)
272. O'Donnell T, Rubinsteyn A, Bonsack M, Riemer A, Hammerbacher J. 2017 MHCflurry: open-source class I MHC binding affinity prediction. *bioRxiv*. (doi:10.1101/174243)
273. Rubinsteyn A, O'Donnell T, Damaraju N, Hammerbacher J. 2016 Predicting peptide-MHC binding affinities with imputed training data. *bioRxiv*. (doi:10.1101/054775)
274. Kuksa PP, Min MR, Dugar R, Gerstein M. 2015 High-order neural networks and kernel methods for peptide–MHC binding prediction. *Bioinformatics* **400–401**, btv371. (doi:10.1093/bioinformatics/btv371)
275. Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, Karchin R. 2017 Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. *bioRxiv*. (doi:10.1101/154757)
276. Vang YS, Xie X. 2017 HLA class I binding prediction via convolutional neural networks. *Bioinformatics* **33**, 2658–2665. (doi:10.1093/bioinformatics/btx264)
277. Sharan R, Ulitsky I, Shamir R. 2007 Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 1021. (doi:10.1038/msb4100129)
278. Navlakha S. 2017 Learning the structural vocabulary of a network. *Neural Comput.* **29**, 287–312. (doi:10.1162/neco_a_00924)
279. Gligorijević V, Barot M, Bonneau R. 2017 deepNF: deep network fusion for protein function prediction. *bioRxiv*. (doi:10.1101/223339)
280. Hamilton WL, Ying R, Leskovec J. 2017 Inductive representation learning on large graphs. *arXiv* (https://arxiv.org/abs/1706.02216v2)
281. Chen J, Zhu J. 2017 Stochastic training of graph convolutional networks. *arXiv* (https://arxiv.org/abs/1710.10568v1)
282. Van Valen DA *et al.* 2016 Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS Comput. Biol.* **12**, e1005177. (doi:10.1371/journal.pcbi.1005177)
283. Ronneberger O, Fischer P, Brox T. 2015 U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015* (eds N Navab, J Hornegger, W Wells, A Frangi). Lecture Notes in Computer Science, vol 9351. Cham, Switzerland: Springer.
284. Buggenthin F *et al.* 2017 Prospective identification of hematopoietic lineage choice by deep learning. *Nat. Methods* **14**, 403–406. (doi:10.1038/nmeth.4182)
285. Eulenberg P, Koehler N, Blasi T, Filby A, Carpenter AE, Rees P, Theis FJ, Wolf FA. 2016 Reconstructing cell cycle and disease progression using deep learning. *bioRxiv*. (doi:10.1101/081364)
286. Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. 2016 Automating morphological profiling with generic deep convolutional networks. *bioRxiv*. (doi:10.1101/085118)
287. Johnson GR, Donovan-Maiye RM, Maleckar MM. 2017 Generative modeling with conditional autoencoders: building an integrated cell. *arXiv* (https://arxiv.org/abs/1705.00092v1)
288. Caicedo JC, Singh S, Carpenter AE. 2016 Applications in image-based profiling of perturbations. *Curr. Opin Biotechnol.* **39**, 134–142. (doi:10.1016/j.copbio.2016.04.003)
289. Bougen-Zhukov N, Loh SY, Lee HK, Loo L-H. 2017 Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry Part A* **91**, 115–125. (doi:10.1002/cyto.a.22909)
290. Grys BT, Lo DS, Sahin N, Kraus OZ, Morris Q, Boone C, Andrews BJ. 2017 Machine learning and computer vision approaches for phenotypic profiling. *J. Cell Biol.* **216**, 65–71. (doi:10.1083/jcb.201610026)
291. Gawad C, Koh W, Quake SR. 2016 Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188. (doi:10.1038/nrg.2015.16)
292. Lodato MA *et al.* 2015 Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98. (doi:10.1126/science.aab1785)
293. Liu S, Trapnell C. 2016 Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* **5**, 182. (doi:10.12688/f1000research.7223.1)
294. Vera M, Biswas J, Senecal A, Singer RH, Park HY. 2016 Single-cell and single-molecule analysis of gene expression regulation. *Annu. Rev. Genet.* **50**, 267–291. (doi:10.1146/annurev-genet-120215-034854)
295. Clark SJ *et al.* 2017 Joint profiling of chromatin accessibility, DNA methylation and transcription in single cells. *bioRxiv*. (doi:10.1101/138685)
296. Angermueller C, Lee HJ, Reik W, Stegle O. 2017 DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 597–600. (doi:10.1186/s13059-017-1189-z)
297. Koh PW, Pierson E, Kundaje A. 2016 Denoising genome-wide histone ChIP-seq with convolutional neural networks. *bioRxiv*. (doi:10.1101/052118)
298. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. 2017 Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33**, 2539–2546. (doi:10.1093/bioinformatics/btx196)
299. Gaublomme JT *et al.* 2015 Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* **163**, 1400–1412. (doi:10.1016/j.cell.2015.11.009)
300. Arvaniti E, Claassen M. 2016 Sensitive detection of rare disease-associated cell subsets via representation learning. *bioRxiv*. (doi:10.1101/046508)
301. Ding J, Condon AE, Shah SP. 2017 Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *bioRxiv*. (doi:10.1101/178624)
302. Lopez R, Regier J, Cole M, Jordan M, Yosef N. 2017 A deep generative model for gene expression profiles from single-cell RNA sequencing. *arXiv* (https://arxiv.org/abs/1709.02082v3)
303. van der Maaten L, Hinton G. 2008 Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.
304. Lin C, Jain S, Kim H, Bar-Joseph Z. 2017 Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* **45**, e156. (doi:10.1093/nar/gkx681)
305. Regev A *et al.* 2017 Science forum: the human cell atlas. *eLife* **6**, 503. (doi:10.7554/elife.27041)
306. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H, Trapnell C. 2017 Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv*. (doi:10.1101/110668)
307. Silver D *et al.* 2016 Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489. (doi:10.1038/nature16961)
308. Karlin S, Mrázek J, Campbell AM. 1997 Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**, 3899–3913. (doi:10.1128/jb.179.12.3899-3913.1997)
309. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007 Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* **4**, 63–72. (doi:10.1038/nmeth976)
310. Rosen GL, Reichenberger ER, Rosenfeld AM. 2011 NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* **27**, 127–129. (doi:10.1093/bioinformatics/btq619)
311. Abe T. 2003 Informatics for unveiling hidden genome signatures. *Genome Res.* **13**, 693–702. (doi:10.1101/gr.634603)
312. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012 Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814. (doi:10.1038/nmeth.2066)
313. Koslicki D, Foucart S, Rosen G. 2014 WGSQuikr: fast whole-genome shotgun metagenomic classification. *PLoS ONE* **9**, e91784. (doi:10.1371/journal.pone.0091784)

314. Ames SK, Hysom DA, Gardner SN, Scott Lloyd G, Gokhale MB, Allen JE. 2013 Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* **29**, 2253–2260. (doi:10.1093/bioinformatics/btt389)
315. Vervier K, Mahé P, Tournoud M, Veyrieras J-B, Vert J-P. 2016 Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* **32**, 1023–1032. (doi:10.1093/bioinformatics/btv683)
316. Yok NG, Rosen GL. 2011 Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* **12**, 20. (doi:10.1186/1471-2105-12-20)
317. Soueidan H, Nikolski M. 2017 Machine learning for metagenomics: methods and tools. *Metagenomics* **1**, 1396. (doi:10.1515/metgen-2016-0001)
318. 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. *FASEB J.* Abstr. 406.3. See http://www.fasebj.org/doi/abs/10.1096/fasebj.30.1_supplement.406.3.
319. Knights D, Costello EK, Knight R. 2011 Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359. (doi:10.1111/j.1574-6976.2010.00251.x)
320. Statnikov A *et al.* 2013 A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1**, 11. (doi:10.1186/2049-2618-1-11)
321. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016 Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977. (doi:10.1371/journal.pcbi.1004977)
322. 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**, R245. (doi:10.1186/s12859-015-0753-3)
323. 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. (doi:10.1371/journal.pone.0053253)
324. Ditzler G, Calvin Morrison J, Lan Y, Rosen GL. 2015 Fizzy: feature subset selection for metagenomics. *BMC Bioinformatics* **16**, 59. (doi:10.1186/s12859-015-0793-8)
325. Ditzler G, Polikar R, Rosen G. 2015 A bootstrap based Neyman–Pearson test for identifying variable importance. *IEEE Trans. Neural Networks Learn. Syst.* **26**, 880–886. (doi:10.1109/tnnls.2014.2320415)
326. Hoff KJ, Lingner T, Meinicke P, Tech M. 2009 Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* **37**, W101–W105. (doi:10.1093/nar/gkp327)
327. Rho M, Tang H, Ye Y. 2010 FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191. (doi:10.1093/nar/gkq747)
328. Asgari E, Mofrad MRK. 2015 Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* **10**, e0141287. (doi:10.1371/journal.pone.0141287)
329. Hochreiter S, Heusel M, Obermayer K. 2007 Fast model-based protein homology detection without alignment. *Bioinformatics* **23**, 1728–1736. (doi:10.1093/bioinformatics/btm247)
330. Sønderby SK, Sønderby CK, Nielsen H, Winther O. 2015 Convolutional LSTM networks for subcellular localization of proteins. In *AlCoB 2015, Proc. of the Second Int. Conf. on Algorithms for Computational Biology, 4–5 August 2015, Mexico City, Mexico*, vol. 9199, pp. 68–80. New York, NY: Springer.
331. Essinger SD, Polikar R, Rosen GL. 2010 Neural network-based taxonomic clustering for metagenomics. In *IEEE 2010 Int. Joint Conf. on Neural Networks* (*IJCNN*), *Barcelona, Spain*, pp. 1–7.
332. Kelley DR, Salzberg SL. 2010 Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* **11**, 544. (doi:10.1186/1471-2105-11-544)
333. Rasheed Z, Rangwala H. 2012 Metagenomic taxonomic classification using extreme learning machines. *J. Bioinform. Comput. Biol.* **10**, 1250015. (doi:10.1142/s0219720012500151)
334. Nina Mrzelj. 2016 *Globoko ucenje na genomskih in filogenetskih podatkih*. Univerza v Ljubljani, Fakulteta za racunalništvo in informatiko. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=85515
335. Chudobova D *et al.* 2015 Influence of microbiome species in hard-to-heal wounds on disease severity and treatment duration. *Braz. J. Infect. Dis.* **19**, 604–613. (doi:10.1016/j.bjid.2015.08.013)
336. Ditzler G, Polikar R, Rosen G. 2015 Multi-layer and recursive neural networks for metagenomic classification. *IEEE Trans. Nanobioscience* **14**, 608–616. (doi:10.1109/tnb.2015.2461219)
337. Faruqi AA. 2016 TensorFlow vs. scikit-learn: the microbiome challenge. See http://alifar76.github.io/sklearn-metrics/.
338. Bengio Y, Boulanger-Lewandowski N, Pascanu R. 2012 Advances in optimizing recurrent networks. *arXiv* (https://arxiv.org/abs/1212.0901v2)
339. Boža V, Brejová B, Vinař T. 2017 DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE* **12**, e0178751. (doi:10.1371/journal.pone.0178751)
340. Sutskever I, Vinyals O, Le QV. 2014 Sequence to sequence learning with neural networks. *arXiv* (https://arxiv.org/abs/1409.3215v3)
341. 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. *bioRxiv*. (doi:10.1101/092890)
342. DePristo MA *et al.* 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498. (doi:10.1038/ng.806)
343. Torracinta R, Campagne F. 2016 Training genotype callers with neural networks. *bioRxiv*. (doi:10.1101/097469)
344. Chollet F. 2016 Xception: deep learning with depthwise separable convolutions. *arXiv* (https://arxiv.org/abs/1610.02357v3)
345. Torracinta R, Mesnard L, Levine S, Shaknovich R, Hanson M, Campagne F. 2016 Adaptive somatic mutations calls with deep learning and semi-simulated data. *bioRxiv*. (doi:10.1101/079087)
346. Marblestone AH, Wayne G, Kording KP. 2016 Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 406. (doi:10.3389/fncom.2016.00094)
347. Kietzmann TC, McClure P, Kriegeskorte N. 2017 Deep neural networks in computational neuroscience. *bioRxiv*. (doi:10.1101/133504)
348. Hassabis D, Kumaran D, Summerfield C, Botvinick M. 2017 Neuroscience-Inspired artificial intelligence. *Neuron* **95**, 245–258. (doi:10.1016/j.neuron.2017.06.011)
349. Yamins DL, DiCarlo JJ. 2016 Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365. (doi:10.1038/nn.4244)
350. Pandarinath C *et al.* 2017 Inferring single-trial neural population dynamics using sequential auto-encoders. *bioRxiv*. (doi:10.1101/152884)
351. Jain V, Sebastian Seung H, Turaga SC. 2010 Machines that learn to segment images: a crucial technology for connectomics. *Curr. Opin Neurobiol.* **20**, 653–666. (doi:10.1016/j.conb.2010.07.004)
352. Aitchison L, Russell L, Packer AM, Yan J, Castonguay P, Hausser M, Turaga SC. 2017 Model-based Bayesian inference of neural activity and connectivity from all-optical interrogation of a neural circuit. See http://papers.nips.cc/paper/6940-model-based-bayesian-inference-of-neural-activity-and-connectivity-from-all-optical-interrogation-of-a-neural-circuit.
353. Hamburg MA, Collins FS. 2010 The path to personalized medicine. *N. Engl. J. Med.* **363**, 301–304. (doi:10.1056/nejmp1006304)
354. Belle A, Kon MA, Najarian K. 2013 Biomedical informatics for computer-aided decision support systems: a survey. *Scient. World J.* **2013**, 1–8. (doi:10.1155/2013/769639)
355. Tu JV. 1996 Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231. (doi:10.1016/s0895-4356(96)00002-9)
356. Baxt WG. 1991 Use of an artificial neural network for the diagnosis of myocardial infarction. *Ann. Intern. Med.* **115**, 843. (doi:10.7326/0003-4819-115-11-843)
357. Wasson JH, Sox HC, Neff RK, Goldman L. 1985 Clinical prediction rules. *N. Engl. J. Med.* **313**, 793–799. (doi:10.1056/nejm198509263131306)
358. Lisboa PJ, Taktak AFG. 2006 The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw.* **19**, 408–415. (doi:10.1016/j.neunet.2005.10.007)
359. Rubin DB. 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701. (doi:10.1037/h0037350)
360. Johansson FD, Shalit U, Sontag D. 2016 Learning representations for counterfactual inference. *arXiv* (https://arxiv.org/abs/1605.03661v2)

361. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. 2015 Causal phenotype discovery via deep networks. In *AMIA Ann. Symp. Proc*. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765623/.
362. Lipton ZC, Kale DC, Wetzel R. 2016 Modeling missing data in clinical time series with RNNs. *arXiv* (https://arxiv.org/abs/1606.04130v5)
363. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. 2016 Recurrent neural networks for multivariate time series with missing values. *arXiv* (https://arxiv.org/abs/1606.01865v2)
364. 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. (doi:10.1109/access.2016.2618775)
365. Lipton ZC, Kale DC, Wetzel RC. 2015 Phenotyping of clinical time series with LSTM recurrent neural networks. *arXiv* (https://arxiv.org/abs/1510.07641v2)
366. Nemati S, Ghassemi MM, Clifford GD. 2016 Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In *2016 38th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), *16–20 August 2016, Orlando, FL, USA*, pp. 2978–2981.
367. 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. *J. Am. Med. Inform. Assoc.* **21**, 315–325. (doi:10.1136/amiajnl-2013-001815)
368. Ithapu VK, Singh V, Okonkwo OC, Chappell RJ, Maritza Dowling N, Johnson SC. 2015 Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimers Dement.* **11**, 1489–1499. (doi:10.1016/j.jalz.2015.01.010)
369. 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. *bioRxiv*. (doi:10.1101/095653)
370. DiMasi JA, Grabowski HG, Hansen RW. 2016 Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33. (doi:10.1016/j.jhealeco.2016.01.012)
371. Waring MJ *et al.* 2015 An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discovery* **14**, 475–486. (doi:10.1038/nrd4609)
372. Lamb J. 2006 The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935. (doi:10.1126/science.1132939)
373. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. 2016 A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **17**, 2–12. (doi:10.1093/bib/bbv020)
374. 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. *Brief. Bioinform.* **97**, bbw112. (doi:10.1093/bib/bbw112)
375. Brown AS, Patel CJ. 2016 A review of validation strategies for computational drug repositioning. *Brief. Bioinform.* **19**, 174–177. (doi:10.1093/bib/bbw110)
376. 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. *J. Cheminform.* **5**, 30. (doi:10.1186/1758-2946-5-30)
377. 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. *J. Chem. Inf. Model.* **54**, 2562–2569. (doi:10.1021/ci500340n)
378. 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**, 1178. (doi:10.1186/s12859-015-0845-0)
379. 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. (doi:10.1371/journal.pone.0061318)
380. Vidovic D, Koleti A, Schürer 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. *Front. Genet.* **5**, e77521. (doi:10.3389/fgene.2014.00342)
381. Coelho ED, Arrais JP, Oliveira JL. 2016 Computational discovery of putative leads for drug repositioning through drug–target interaction prediction. *PLoS Comput. Biol.* **12**, e1005219. (doi:10.1371/journal.pcbi.1005219)
382. Lim H, Poleksic A, Yao Y, Hanghang Tong DH, 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 Comput. Biol.* **12**, e1005135. (doi:10.1371/journal.pcbi.1005135)
383. 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 Int. Conf. on Bioinformatics and Biomedicine* (*BIBM*), *2–5 November 2014, Belfast*, pp. 67–70.
384. Duan Q *et al.* 2016 L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Syst. Biol. Appl.* **2**, 257. (doi:10.1038/npjsba.2016.15)
385. Bleicher KH, Böhm H-J, Müller K, Alanine AI. 2003 A guide to drug discovery: hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2**, 369–378. (doi:10.1038/nrd1086)
386. Keserű GM, Makara GM. 2006 Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **11**, 741–748. (doi:10.1016/j.drudis.2006.06.016)
387. 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. *J. Chem. Inf. Model.* **49**, 756–766. (doi:10.1021/ci8004379)
388. Kearnes S, Goldman B, Pande V. 2016 Modeling industrial ADMET data with multitask networks. *arXiv* (https://arxiv.org/abs/1606.08793v3)
389. Zaretzki J, Matlock M, Swamidass SJ. 2013 XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J. Chem. Inf. Model.* **53**, 3373–3383. (doi:10.1021/ci400518g)
390. Dahl GE, Jaitly N, Salakhutdinov R. 2014 Multi-task neural networks for QSAR predictions. *arXiv* (https://arxiv.org/abs/1406.1231v1)
391. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. 2015 Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* **55**, 263–274. (doi:10.1021/ci500747n)
392. Lowe D. 2012 Did Kaggle predict drug candidate activities? Or not? *In the Pipeline*. See http://blogs.sciencemag.org/pipeline/archives/2012/12/11/did_kaggle_predict_drug_candidate_activities_or_not.
393. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner JK, Ceulemans H, Hochreiter S. 2014 Deep learning as an opportunity in virtual screening. In *Neural Information Processing Systems 2014: deep Learning and Representation Learning Workshop*. See http://www.dlworkshop.org/23.pdf?attredirects=0.
394. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. 2015 Massively multitask networks for drug discovery. *arXiv* (https://arxiv.org/abs/1502.02072v1)
395. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. 2016 DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 24. (doi:10.3389/fenvs.2015.00080)
396. Subramanian G, Ramsundar B, Pande V, Denny RA. 2016 Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* **56**, 1936–1949. (doi:10.1021/acs.jcim.6b00290)
397. Reymond J-L, Ruddigkeit L, Blum L, Deursen Rv. 2012 The enumeration of chemical space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 717–733. (doi:10.1002/wcms.1104)
398. Lusci A, Fooshee D, Browning M, Swamidass J, Baldi P. 2015 Accurate and efficient target prediction using a potency-sensitive influence–relevance voter. *J. Cheminform.* **7**, 361. (doi:10.1186/s13321-015-0110-6)
399. Todeschini R, Consonni V. 2009 *Molecular descriptors for chemoinformatics*. Hoboken, NJ: Wiley.
400. Rogers D, Hahn M. 2010 Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754. (doi:10.1021/ci100050t)
401. 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. *arXiv* (https://arxiv.org/abs/1610.02415v1)
402. Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. 2017 Chemception: a deep neural network with

minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv* (https://arxiv.org/abs/1706.06689v1)

403. 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.
404. Lusci A, Pollastri G, Baldi P. 2013 Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53**, 1563–1575. (doi:10.1021/ci400187y)
405. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. 2016 Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des* **30**, 595–608. (doi:10.1007/s10822-016-9938-8)
406. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. 2017 Low data drug discovery with one-shot learning. *ACS Central Sci.* **3**, 283–293. (doi:10.1021/acscentsci.6b00367)
407. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. 2017 Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* **57**, 1757–1772. (doi:10.1021/acs.jcim.6b00601)
408. Matlock MK, Dang NL, Swamidass SJ. 2018 Learning a local-variable model of aromatic and conjugated systems. *ACS Central Sci.* **4**, 52–62. (doi:10.1021/acscentsci.7b00405)
409. Kondor R, Son HT, Pan H, Anderson B, Trivedi S. 2018 Covariant compositional networks for learning graphs. *arXiv* (https://arxiv.org/abs/1801.02144v1)
410. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. 2018 MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530. (doi:10.1039/c7sc02664a)
411. Nicholls A. 2008 What do we know and when do we know it? *J. Comput. Aided Mol. Des* **22**, 239–255. (doi:10.1007/s10822-008-9170-2)
412. deepchem/deepchem GitHub. 2017 See https://github.com/deepchem/deepchem.
413. Jaeger S, Fulle S, Turk S. 2018 Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35. (doi:10.1021/acs.jcim.7b00616)
414. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. 2012 Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **14**, 133–141. (doi:10.1208/s12248-012-9322-0)
415. Gomes J, Ramsundar B, Feinberg EN, Pande VS. 2017 Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv* (https://arxiv.org/abs/1703.10603v1)
416. Cang Z, Wei G-W. 2017 TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**, e1005690. (doi:10.1371/journal.pcbi.1005690)
417. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. 2005 The PDBbind database: methodologies and updates. *J. Med. Chem.* **48**, 4111–4119. (doi:10.1021/jm048957q)
418. Pereira JC, Caffarena ER, dos Santos CN. 2016 Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* **56**, 2495–2506. (doi:10.1021/acs.jcim.6b00355)
419. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. 2016 Protein-ligand scoring with convolutional neural networks. *arXiv* (https://arxiv.org/abs/1612.02751v1)
420. Hartenfeller M, Schneider G. 2011 Enabling future drug discovery by denovo design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 742–759. (doi:10.1002/wcms.49)
421. Schneider P, Schneider G. 2016 De novo design at the edge of chaos. *J. Med. Chem.* **59**, 4077–4086. (doi:10.1021/acs.jmedchem.5b01849)
422. Graves A. 2013 Generating sequences with recurrent neural networks. *arXiv* (https://arxiv.org/abs/1308.0850v5)
423. Segler MHS, Kogej T, Tyrchan C, Waller MP. 2017 Generating focussed molecule libraries for drug discovery with recurrent neural networks. *arXiv* (https://arxiv.org/abs/1701.01329v1)
424. Kusner MJ, Paige B, Hernández-Lobato JM. 2017 Grammar variational autoencoder. *arXiv* (https://arxiv.org/abs/1703.01925v1)
425. Gaulton A *et al.* 2012 ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107. (doi:10.1093/nar/gkr777)
426. Olivecrona M, Blaschke T, Engkvist O, Chen H. 2017 Molecular de novo design through deep reinforcement learning. *arXiv* (https://arxiv.org/abs/1704.07555v2)
427. 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. *arXiv* (https://arxiv.org/abs/1611.02796v9)
428. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. 2016 Understanding deep learning requires rethinking generalization. *arXiv* (https://arxiv.org/abs/1611.03530v2)
429. Lin HW, Tegmark M, Rolnick D. 2016 Why does deep and cheap learning work so well? (https://arxiv.org/abs/1608.08225v3)
430. Davis J, Goadrich M. 2006 The relationship between Precision-Recall and ROC curves. In *Proc. of the 23rd Int. Conf. on Machine Learning - ICML '06, 25–29 June 2006, Pittsburgh, Pennsylvania, USA*, pp. 233–240. New York, NY, USA: ACM.
431. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. 2014 An open investigation of the reproducibility of cancer biology research. *eLife* **3**, 5773. (doi:10.7554/elife.04333)
432. Bradshaw J, de G. Matthews AG, Ghahramani Z. 2017 Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv* (https://arxiv.org/abs/1707.02476v1)
433. Kendall A, Gal Y. 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *arXiv* (https://arxiv.org/abs/1703.04977v2)
434. Kendall A, Gal Y, Cipolla R. 2017 Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv* (https://arxiv.org/abs/1705.07115v1)
435. Guo C, Pleiss G, Sun Y, Weinberger KQ. 2017 On calibration of modern neural networks. *arXiv* (https://arxiv.org/abs/1706.04599v2)
436. Platt JC. 1999 Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (eds A Smola, P Bartlett, B Schölkopf, D Schuurmans) pp. 61–74. Cambridge, MA: MIT Press.
437. Chryssolouris G, Lee M, Ramsey A. 1996 Confidence interval prediction for neural network models. *IEEE Trans. Neural Netw.* **7**, 229–232. (doi:10.1109/72.478409)
438. Hendrycks D, Gimpel K. 2016 A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* (https://arxiv.org/abs/1610.02136v2)
439. Liang S, Li Y, Srikant R. 2017 Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv* (https://arxiv.org/abs/1706.02690v3)
440. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. 2016 Concrete problems in AI safety. *arXiv* (https://arxiv.org/abs/1606.06565v2)
441. Carlini N, Wagner D. 2017 Adversarial examples are not easily detected: bypassing ten detection methods. *arXiv* (https://arxiv.org/abs/1705.07263v2)
442. Gal Y, Ghahramani Z. 2015 Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *arXiv* (https://arxiv.org/abs/1506.02142v6)
443. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. 2017 Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* **7**, 17816. (doi:10.1038/s41598-017-17876-z)
444. McClure P, Kriegeskorte N. 2016 Robustly representing inferential uncertainty in deep neural networks through sampling. *arXiv* (https://arxiv.org/abs/1611.01639v6)
445. Krueger D, Huang C-W, Islam R, Turner R, Lacoste A, Courville A. 2017 Bayesian hypernetworks. *arXiv* (https://arxiv.org/abs/1710.04759v1)
446. Lakshminarayanan B, Pritzel A, Blundell C. 2016 Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv* (https://arxiv.org/abs/1612.01474v3)
447. Gal Y. 2016 Uncertainty in deep learning. PhD thesis, University of Cambridge, Cambridge, UK.
448. Ba LJ, Caruana R. 2013 Do deep nets really need to be deep? *arXiv* (https://arxiv.org/abs/1312.6184v7)
449. Nguyen A, Yosinski J, Clune J. 2014 Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *arXiv* (https://arxiv.org/abs/1412.1897v4)
450. Ribeiro MT, Singh S, Guestrin C. 2016 'Why should I trust you?': explaining the predictions of any classifier. *arXiv* (https://arxiv.org/abs/1602.04938v3)
451. Zeiler MD, Fergus R. 2013 Visualizing and understanding convolutional networks. *arXiv* (https://arxiv.org/abs/1311.2901v3)

452. Zintgraf LM, Cohen TS, Adel T, Welling M. 2017 Visualizing deep neural network decisions: prediction difference analysis. *arXiv* (https://arxiv.org/abs/1702.04595v1)
453. Fong RC, Vedaldi A. 2017 Interpretable explanations of black boxes by meaningful perturbation. In *Proc. of the 2017 IEEE Int. Conf. on Computer Vision (ICCV), 22–29 October 2017, Venice, Italy*.
454. Simonyan K, Vedaldi A, Zisserman A. 2013 Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv* (https://arxiv.org/abs/1312.6034v2)
455. 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. (doi:10.1371/journal.pone.0130140)
456. Kindermans P-J, Schütt K, Müller K-R, Dähne S. 2016 Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv* (https://arxiv.org/abs/1611.07270v1)
457. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. 2014 Striving for simplicity: the all convolutional net. *arXiv* (https://arxiv.org/abs/1412.6806v3)
458. Mahendran A, Vedaldi A. 2016 Salient deconvolutional networks. In *Computer Vision—ECCV 2016, 8–16 October 2016, Amsterdam* (eds B Leibe, J Matas, N Sebe, M Welling). Lecture Notes in Computer Science, vol. 9910. Cham: Springer.
459. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2016 Grad-CAM: visual explanations from deep networks via gradient-based localization. *arXiv* (https://arxiv.org/abs/1610.02391v3)
460. Sundararajan M, Taly A, Yan Q. 2017 Axiomatic attribution for deep networks. *arXiv* (https://arxiv.org/abs/1703.01365v2)
461. Lundberg S, Lee S-I. 2016 An unexpected unity among methods for interpreting model predictions. *arXiv* (https://arxiv.org/abs/1611.07478v3)
462. Shapley LS. 1953 A value for n-person games. In *Contributions to the theory of games*. Annals of Mathematics, vol. 2, pp. 307–317. Princeton, NJ: Princeton University Press.
463. Mahendran A, Vedaldi A. 2014 Understanding deep image representations by inverting them. *arXiv* (https://arxiv.org/abs/1412.0035v1)
464. Finnegan AI, Song JS. 2017 Maximum entropy methods for extracting the learned features of deep neural networks. *bioRxiv*. (doi:10.1101/105957)
465. Mahendran A, Vedaldi A. 2016 Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vision* **120**, 233–255. (doi:10.1007/s11263-016-0911-8)
466. 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.
467. Erhan D, Bengio Y, Courville A, Vincent P. 2009 Visualizing higher-layer features of a deep network. Montreal, Canada: University of Montreal. See http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/247.
468. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. 2015 Understanding neural networks through deep visualization. *arXiv* (https://arxiv.org/abs/1506.06579v1)
469. Bahdanau D, Cho K, Bengio Y. 2014 Neural machine translation by jointly learning to align and translate. *arXiv* (https://arxiv.org/abs/1409.0473v7)
470. 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. *arXiv* (https://arxiv.org/abs/1502.03044v3)
471. Deming L, Targ S, Sauder N, Almeida D, Ye CJ. 2016 Genetic architect: discovering genomic structure with learned neural architectures. *arXiv* (https://arxiv.org/abs/1605.07156v1)
472. 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. *arXiv* (https://arxiv.org/abs/1608.05745v4)
473. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. 2016 GRAM: graph-based attention model for healthcare representation learning. *arXiv* (https://arxiv.org/abs/1611.07012v3)
474. Ghosh J, Karamcheti V. 1992 Sequence learning with recurrent networks: analysis of internal representations. In *Science of Artificial Neural Networks, SPIE 1710, Aerospace sensing, 1 July 1992, Orlando, FL, USA*. (doi:10.1117/12.140112)
475. Karpathy A, Johnson J, Fei-Fei L. 2015 Visualizing and understanding recurrent networks. *arXiv* (https://arxiv.org/abs/1506.02078v2)
476. Strobelt H, Gehrmann S, Pfister H, Rush AM. 2016 LSTMVis: a tool for visual analysis of hidden state dynamics in recurrent neural networks. *arXiv* (https://arxiv.org/abs/1606.07461v2)
477. Murdoch WJ, Szlam A. 2017 Automatic rule extraction from long short term memory networks. *arXiv* (https://arxiv.org/abs/1702.02540v2)
478. Radford A, Metz L, Chintala S. 2015 Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* (https://arxiv.org/abs/1511.06434v2)
479. Chang K *et al.* 2013 The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120. (doi:10.1038/ng.2764)
480. Way GP, Greene CS. 2017 Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *bioRxiv*. (doi:10.1101/174474)
481. Way GP, Greene CS. 2017 Evaluating deep variational autoencoders trained on pan-cancer gene expression. *arXiv* (https://arxiv.org/abs/1711.04828v1)
482. Osokin A, Chessel A, Carazo Salas RE, Vaggi F. 2017 GANs for biological image synthesis. *arXiv* (https://arxiv.org/abs/1708.04692v2)
483. Goldsborough P, Pawlowski N, Caicedo JC, Singh S, Carpenter A. 2017 CytoGAN: generative modeling of cell images. *bioRxiv* (doi:10.1101/227645)
484. Koh PW, Liang P. 2017 Understanding black-box predictions via influence functions. *arXiv* (https://arxiv.org/abs/1703.04730v2)
485. Kahng M, Andrews PY, Kalro A, Chau DH. 2017 ActiVis: visual exploration of industry-scale deep neural network models. *arXiv* (https://arxiv.org/abs/1704.01942v2)
486. Liu M, Shi J, Li Z, Li C, Zhu J, Liu S. 2016 Towards better analysis of deep convolutional neural networks. *arXiv* (https://arxiv.org/abs/1604.07043v3)
487. Che Z, Purushotham S, Khemani R, Liu Y. 2015 Distilling knowledge from deep networks with applications to healthcare domain. *arXiv* (https://arxiv.org/abs/1512.03542v1)
488. Lei T, Barzilay R, Jaakkola T. 2016 Rationalizing neural predictions. *arXiv* (https://arxiv.org/abs/1606.04155v2)
489. Krizhevsky A. 2009 Learning multiple layers of features from tiny images. *arXiv* (https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)
490. 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 Comput. Biol.* **9**, e1002957. (doi:10.1371/journal.pcbi.1002957)
491. Sarraf S, DeSouza DD, Anderson J, Tofighi G. 2016 DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*. (doi:10.1101/070441)
492. Shao M, Ma J, Wang S. 2017 DeepBound: accurate identification of transcript boundaries via deep convolutional neural fields. *bioRxiv*. (doi:10.1101/125229)
493. 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. *Nat. Genet.* **46**, 310–315. (doi:10.1038/ng.2892)
494. Romero A *et al.* 2016 Diet networks: thin parameters for fat genomics. In *Int. Conf. on Learning Representations 2017*. See https://openreview.net/forum?id=Sk-oDY9ge&noteId=Sk-oDY9ge.
495. Schmidhuber J. 2015 Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117. (doi:10.1016/j.neunet.2014.09.003)
496. Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. 2015 Deep learning with limited numerical precision. *arXiv* (https://arxiv.org/abs/1502.02551v1)
497. Courbariaux M, Bengio Y, David J-P. 2014 Training deep neural networks with low precision multiplications. *arXiv* (https://arxiv.org/abs/1412.7024v5)
498. De Sa C, Zhang C, Olukotun K, Ré C. 2015 Taming the wild: a unified analysis of Hogwild!-Style Algorithms. In *Advances in neural information processing systems*. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907892/.
499. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. 2016 Quantized neural networks: training neural networks with low precision weights and

activations. *arXiv* (https://arxiv.org/abs/1609.07061v1)

500. Hinton G, Vinyals O, Dean J. 2015 Distilling the knowledge in a neural network. *arXiv* (https://arxiv.org/abs/1503.02531v1)
501. Raina R, Madhavan A, Ng AY. 2009 Large-scale deep unsupervised learning using graphics processors. In *Proc. of the 26th Ann. Int. Conf. on Machine Learning—ICML '09,14–18 June 2009, Montreal, Quebec, Canada,* pp. 873–880. New York, NY: ACM.
502. Vanhoucke V, Senior A, Mao MZ. 2011 Improving the speed of neural networks on CPUs. See https://research.google.com/pubs/pub37631.html.
503. Seide F, Fu H, Droppo J, Li G, Yu D. 2014 On parallelizability of stochastic gradient descent for speech DNNS. In *2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP*), *4–9 May 2014, Florence, Italy*, pp. 235–239.
504. Hadjis S, Abuzaid F, Zhang C, Ré C. 2015 Caffe con troll: shallow ideas to speed up deep learning. *arXiv* (https://arxiv.org/abs/1504.04343v2)
505. Edwards C. 2015 Growing pains for deep learning. *Commun. ACM* **58**, 14–16. (doi:10.1145/2771283)
506. Su H, Chen H. 2015 Experiments on parallel training of deep neural network using model averaging. *arXiv* (https://arxiv.org/abs/1507.01239v2)
507. Li M, Zhang T, Chen Y, Smola AJ. 2014 Efficient mini-batch training for stochastic optimization. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining—KDD '14, 24–27 August 2014, New York, NY, USA*, pp. 661–670. New York, NY: ACM.
508. Hamanaka M, Taneishi K, Iwata H, Ye J, Pei J, Hou J, Okuno Y. 2017 CGBVS-DNN: prediction of compound-protein interactions based on deep learning. *Mol. Inform.* **36**, 1600045. (doi:10.1002/minf.201600045)
509. Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, Shelhamer E. 2014 cuDNN: efficient primitives for deep learning. *arXiv* (https://arxiv.org/abs/1410.0759v3)
510. Chen W, Wilson JT, Tyree S, Weinberger KQ, Chen Y. 2015 Compressing neural networks with the hashing trick. *arXiv* (https://arxiv.org/abs/1504.04788v1)
511. Lacey G, Taylor GW, Areibi S. 2016 Deep learning on FPGAs: past, present, and future. *arXiv* (https://arxiv.org/abs/1602.04283v1)
512. Jouppi NP *et al.* 2017 In-datacenter performance analysis of a tensor processing unit. *arXiv* (https://arxiv.org/abs/1704.04760v1)
513. Dean J, Ghemawat S. 2008 MapReduce. *Commun. ACM* **51**, 107. (doi:10.1145/1327452.1327492)
514. Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A, Hellerstein JM. 2012 Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proc. VLDB Endowment* **5**, 716–727. (doi:10.14778/2212351.2212354)
515. Dean J *et al.* 2012 Large scale distributed deep networks. In *Neural Information Processing Systems 2012*. See http://research.google.com/archive/large_deep_networks_nips2012.html.
516. Moritz P, Nishihara R, Stoica I, Jordan MI. 2015 SparkNet: training deep networks in Spark. *arXiv* (https://arxiv.org/abs/1511.06051v4)
517. Meng X *et al.* 2015 MLlib: machine learning in Apache Spark. *arXiv* (https://arxiv.org/abs/1505.06807v1)
518. Abadi M *et al.* 2016 TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv* (https://arxiv.org/abs/1603.04467v2)
519. fchollet/keras GitHub. 2017 See https://github.com/fchollet/keras.
520. maxpumperla/elephas GitHub. 2017 See https://github.com/maxpumperla/elephas.
521. 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.
522. Sun S, Chen W, Bian J, Liu X, Liu T-Y. 2016 Ensemble-compression: a new method for parallel training of deep neural networks. *arXiv* (https://arxiv.org/abs/1606.00575v2)
523. Bergstra J, Bardenet R, Bengio Y, Kégl B. 2011 Algorithms for hyper-parameter optimization. In *Proc. of the 24th Int. Conf. on Neural Information Processing Systems*. See http://dl.acm.org/citation.cfm?id=2986459.2986743.
524. Bergstra J, Bengio Y. 2012 Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305.
525. Schatz MC, Langmead B, Salzberg SL. 2010 Cloud computing and the DNA data race. *Nat. Biotechnol.* **28**, 691–693. (doi:10.1038/nbt0710-691)
526. Muir P *et al.* 2016 The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 4731. (doi:10.1186/s13059-016-0917-0)
527. Stein LD. 2010 The case for cloud computing in genome informatics. *Genome Biol.* **11**, 207. (doi:10.1186/gb-2010-11-5-207)
528. Krizhevsky A. 2014 One weird trick for parallelizing convolutional neural networks. *arXiv* (https://arxiv.org/abs/1404.5997v2)
529. Armbrust M *et al.* 2010 A view of cloud computing. *Commun. ACM* **53**, 50. (doi:10.1145/1721654.1721672)
530. Longo DL, Drazen JM. 2016 Data sharing. *N. Engl. J. Med.* **374**, 276–277. (doi:10.1056/nejme1516564)
531. Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. 2017 Celebrating parasites. *Nat. Genet.* **49**, 483–484. (doi:10.1038/ng.3830)
532. Ramsundar B, Liu B, Wu Z, Verras A, Tudor M, Sheridan RP, Pande V. 2017 Is multitask deep learning practical for pharma? *J. Chem. Inf. Model.* **57**, 2068–2076. (doi:10.1021/acs.jcim.7b00146)
533. 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. (doi:10.1126/science.aah6168)
534. DragoNN. 2016 See http://kundajelab.github.io/dragonn/.
535. 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.
536. 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 *IEEE transactions on Big Data*, vol. PP, pp. 1–1. (doi:10.1109/TBDATA.2016.2573280)
537. 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 Bioinform.* **16**, 309. (doi:10.1186/s12859-015-0553-9)
538. Pärnamaa T, Parts L. 2017 Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *G3-Genes Genom. Genet.* **7**, 1385–1392. (doi:10.1534/g3.116.033654)
539. 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. *Mol. Syst. Biol.* **13**, 924. (doi:10.15252/msb.20177551)
540. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. 2011 Multimodal deep learning. In *Proc. of the 28th Int. Conf. on Machine Learning*. See https://ccrma.stanford.edu/~juhan/pubs/NgiamKhoslaKimNamLeeNg2011.pdf.
541. Chaudhary K, Poirion OB, Lu L, Garmire LX. 2017 Deep learning based multi-omics integration robustly predicts survival in liver cancer. *bioRxiv*. (doi:10.1101/114892)
542. Eser U, Stirling Churchman L. 2016 FIDDLE: an integrative deep learning framework for functional genomic data inference. *bioRxiv*. (doi:10.1101/081380)
543. Hughes TB, Dang NL, Miller GP, Swamidass SJ. 2016 Modeling reactivity to biological macromolecules with a deep multitask network. *ACS Central Sci.* **2**, 529–537. (doi:10.1021/acscentsci.6b00162)
544. 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.
545. 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.
546. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G. 2016 Achieving human parity in conversational speech recognition. *arXiv* (https://arxiv.org/abs/1610.05256v2)
547. Saon G *et al.* 2017 English conversational telephone speech recognition by humans and machines. *arXiv* (https://arxiv.org/abs/1703.02136v1)
548. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. 2013 Intriguing properties of neural networks. *arXiv* (https://arxiv.org/abs/1312.6199v4)
549. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and harnessing adversarial examples. *arXiv* (https://arxiv.org/abs/1412.6572v3)

550. Papernot N, McDaniel P, Sinha A, Wellman M. 2016 Towards the science of security and privacy in machine learning. *arXiv* (https://arxiv.org/abs/1611.03814v1)
551. Xu W, Evans D, Qi Y. 2017 Feature squeezing: detecting adversarial examples in deep neural networks. *arXiv* (https://arxiv.org/abs/1704.01155v1)
552. 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/.
553. Himmelstein D. 2017 The most interesting case of scientific irreproducibility? Satoshi Village. See http://blog.dhimmel.com/irreproducible-timestamps/.
554. 2017 OpenTimestamps: a timestamping proof standard. See https://opentimestamps.org/.
555. 2017 greenelab/deep-review GitHub. See https://github.com/greenelab/deep-review.