# Deep Learning, Genomics, and Precision Medicine

For preliminary authorship information, see the contributors on GitHub.

## Abstract

Deep learning, a class of machine learning algorithms, has recently showed impressive results across a variety of domains. Biomedicine and genomics, both data- and feature-rich and yet complex and often ill-understood, present an obvious and potentially valuable target for this new approach. We examine applications of deep learning to a variety of biomedical problems -- classification, fundamental research into biology and patient treatment -- to determine if similar progress can be made there or if the biomedical sphere holds unique challenges. While deep learning has yet to revolutionize or definitively resolve any of these problems, promising (sometimes remarkable) advances have been made on the prior state-of-the-art. Even where improvement over the previous baseline has been modest, there is still the promise of greatly speeding or aiding human investigation. More work is needed in technical directions such as interpretability and how to best model a problem. Further, the limited amount of labeled data for training presents problems in some domains, as can the legal and privacy constraints enforced by working with sensitive health records. Nonetheless, we foresee a growing use of deep learning with potential for transforming several fields of biomedicine, coming to routine use at the bench and bedside.

## Introduction

Biology and medicine are rapidly becoming data-intensive. A recent comparison of genomics with social media, online videos, and other data-intensive scientific disciplines suggested that genomics alone would equal or surpass other fields in data generation and analysis within the next decade [1]. The volume and complexity of these data present not only new opportunities, but also new challenges. Automated algorithms will be crucial in extracting meaningful patterns and actionable knowledge that allow us to better treat, categorize, or study disease, all within data constrained and privacy critical environments.

Over the past five years, a class of machine learning algorithms known as deep learning has revolutionized image classification and speech recognition due to its flexibility and high accuracy [2]. More recently, these algorithms have shown equally promising results in fields as diverse as high-energy physics [3], dermatology [4], and translation among written languages [5]. Across fields, "off-the-shelf" implementations of these algorithms have produced comparable or higher accuracy than previous best-in-class methods that required years of extensive customization, and specialized implementations are now being used at industrial scales.

Deep learning algorithms can also be used in an exploratory, "unsupervised" mode, where the goal is to summarize, explain, or identify interesting patterns in a data set (rather than to accurately predict which labels an expert would assign to each data point). 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 [6], without being told to look for any of them. What if, more generally, deep learning could solve the challenges presented by the growth of data in biomedicine? Could these algorithms identify the "cats" hidden in our data - the patterns unknown to the researcher - and act on them? In this review, we examine whether deep learning's transformation of biomedical science is simply a matter of time or if there are unique challenges posed by biomedical data that render deep learning methods either more challenging or less fruitful to apply.

### Defining deep learning

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. It is built on artificial neural networks, an idea that was first proposed in 1943 [7] as a model for how biological brains process information. Since then, interest in neural networks as computational models has waxed and waned several times. This history is interesting in its own right [8]. In recent years, attention has shifted back to neural networks as processing power has allowed deep learning techniques to surge ahead of other machine learning algorithms. Our focus is primarily on the downstream applications enabled by these advances.

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

### Will deep learning transform the study of human disease?

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

There are numerous examples where deep learning has been applied to biological problems and improved results as well as reviews focused on applications of deep learning in biology [12–16], healthcare [17], and drug discovery [18–21]. We sought cases where deep learning enables researchers to solve challenges that were previously considered infeasible or makes difficult, tedious analyses routine.

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 can 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 are not intended to replace biologists and clinicians but rather cooperate with them, working to prioritize experiments or streamline tasks that do not require expert judgment.

Based on our guiding question, we focus on the application of deep learning to topics of biomedical importance. We have divided the large range of topics into three broad classes: Disease and Patient Categorization, Fundamental Biological Study, and Patient Treatment. Below, we briefly introduce the types of questions, approaches and data that are typical for each class in the application of deep learning.

#### Disease and Patient Categorization

A key challenge in biomedicine is the accurate classification of diseases and disease subtypes. In oncology, current "gold standard" approaches include histology, which requires interpretation by experts, or assessment of molecular markers such as cell surface receptors or gene expression. One example is the PAM50 approach to classifying breast cancer where the expression of 50 marker genes divides breast cancer patients into four subtypes. Significant heterogeneity still remains within these four subtypes [22,23]. Given the increasing wealth of molecular data available, a more comprehensive subtyping seems possible.

Several studies have used deep learning methods in order to better categorize breast cancer patients: denoising autoencoders (DA), an unsupervised approach, can be used to cluster breast cancer patients [24], and convolutional neural networks (CNN) can help count mitotic divisions, a feature that is highly correlated with disease outcome, in histological images [25]. Despite these recent advances, a number of challenges exist in this area of research, most notably the integration of molecular and imaging data with other disparate types of data such as electronic health records (EHR).

#### 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 (RNN) have been used to predict gene targets of microRNAs [26], and CNNs have been applied to predict protein residue-residue contacts and secondary structure on a genome-wide scale [27–29]. Other recent exciting applications of deep learning include recognition of functional genomic elements such as enhancers and promoters [30–32] and prediction of the deleterious effects of nucleotide polymorphisms[33].

#### Patient Treatment

Although the application of deep learning to patient treatment is just beginning, we expect a dramatic increase in methods aiming to recommend patient treatment, predict treatment outcomes, and guide future development of new therapies. Specifically, effort in this area aims to identify drug targets and interactions or predict drug response. One recent approach uses deep learning on protein structures to predict drug interactions and drug bioactivity [34]. Drug repositioning using deep learning on transcriptomic data is another exciting area of research [35]. Interestingly, it was shown that restricted Boltzmann machines (RBMs) can be combined into deep belief networks (DBNs) to predict novel drug-target interactions and formulate drug repositioning hypotheses [36,37]. Finally, deep learning is also being successfully used to prioritize chemicals in the early stages of drug discovery for new targets [21].

## Deep learning and patient categorization

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

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

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

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

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

**Imaging applications in healthcare**

One area where deep learning methods have had substantial success has been in image analysis. Applications in areas of medicine that use imaging extensively are also emerging. To the date, deep learning has been employed for a wide range of tasks in medical imaging, including classification of exams and lesions/nodules, localization of organs, regions, landmarks, and lesions, segmentation of organs, organ substructures, and lesions, medical image registration, content-based image retrieval, image generation and enhancement, and combining image data with clinical reports [40,41].

Closest to natural images are applications of deep learning aimed at detection and recognition of melanoma, the deadliest form of skin cancer. Recent works included applications to both dermoscopy [42,43] and non-dermoscopic clinical photography images of skin lesions [4,44,45]. For both modalities pre-training on natural images appears to be common model initialization that allows the use of very deep networks without overfitting. Reported performance is competitive or better compared to a board of certified dermatologists [4,42]. This approach is known as transfer learning (see Discussion).

Another fast emerging area of deep learning method applications is the detection of ophthalmological diseases such as diabetic retinopathy and age-related macular degeneration. Diagnosis of diabetic retinopathy through color fundus images became of interest for deep learning researchers and practitioners after the large labeled image set was made publicly available during the corresponding 2015 Kaggle competition [46]. Most attempts included training a network from scratch [46,46,47], while Gulshan et al. [48] employed 48-layer Inception-v3 deep architecture pre-trained on natural images and demonstrated substantial increase over the state-of-the-art in both specificity and sensitivity. Interestingly, Leibig et al. [47] proposed a method to estimate the uncertainty of deep networks in diabetic retinopathy diagnosis based on a recent theoretical insight on the link between dropout networks and approximate Bayesian inference. Such developments are important for the whole medical image analysis field, because they have a potential to provide information about a level of confidence for every black-box algorithm's classification result and, thus, improve pathologist-computer interaction. Deep networks were also recently applied to age-related macular degeneration detection, similarly demonstrating the power of transfer learning when training set is limited [49] as well as the efficient use of a deep 16-layer architecture combined with EMR data for training set enrichment.

Mammography has been one area with numerous contributions [50–54]. In most of this work, researchers must work around a challenge typical for the domain - the limited number of well annotated training images. To expand the number and diversity of images, the researchers have employed approaches where they use adversarial examples [53] or first train towards human-created features before subsequent fine tuning [51]. Adaptation to the medical image domain can be further improved by combining in the latter approach with other machine learning techniques, for example, as a cascade of deep learning and random forest models [54]. Using large dataset, Kooi et al. [55] demonstrated that deep neural networks can outperform the traditional computer-aided diagnosis (CAD) 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. Similarly, Geras et al. [56] showed that both using large dataset and using multi-view network architecture help to improve classification performance. The presence of a publicly available large bank of well-annotated mammography images would aid in the application of deep neural networks to this area and shift research focus from model generalization to effective processing of large image sets. Deep network pre-trained on large annotated mammogram set can be helpful for related tasks that do not have as much data using transfer learning [57]. Though this strategy has not yet been employed in this domain, high-quality feature detectors can be constructed from large collections of unlabeled images in an unsupervised context. The small number of labeled examples could be used for subsequent training. Similar strategies have been employed for EHR data where high-quality labeled examples are also difficult to obtain [58].

In radiological image analysis, deep learning techniques are increasingly used even when dataset size is not big enough to train large capacity models from scratch [59–62]. All these studies demonstrate successful transfer of features learnt from natural image datasets, such as ImageNet [63]. Rajkomar et al.[61] showed that a deep CNN trained on natural images can boost performance in radiology image classification. However, the target task required either re-training the initial model from scratch with special pre-processing or fine-tuning of the whole network on radiographs with heavy data augmentation to avoid overfitting. Shin et al. [60] compared various deep network architectures, dataset characteristics, and training procedures for computer tomography-based (CT) abnormality detection. They concluded that in case of three-dimensional data networks as deep as 22 layers can be useful for such problems despite the limited size of training datasets. However, they note, that choice of a specific architecture, parameter setting, and model fine-tuning needed is very problem and dataset-specific. Moreover, this type of tasks often depends on both lesion localization and appearance that pose challenges for CNN-based approaches. Straightforward attempts to capture useful information from full-size images in all three dimensions simultaneously make applications of standard neural network architectures computationally unfeasible due to the curse of dimensionality. Instead, two dimensional models are often used to either process image slices individually (2D), or aggregate information from a number of 2D projections in the native space (2.5D). Roth et al. compared 2D, 2.5D, and 3D CNNs on a number of tasks for computer-aided detection from CT scans and showed that 2.5D CNNs performed comparably well to 3D analogs, while requiring much less training time, especially on augmented training sets [64]. Another advantage of 2D and 2.5D networks is a possibility to use widely available models pre-trained on natural images.

Similarly, in magnetic resonance image (MRI) analysis, limited size of training sets and large dimensionality represent challenges to deep learning applications. For example, Amit et al. [65] investigated the tradeoffs between using pre-trained models from a different domain and retraining a small-size CNN on MRI images. They showed that smaller network trained with sufficient data augmentation on few hundred images from a few dozen patients can outperform a pre-trained out-of-domain classifier. Nie et al. [66] showed that multimodal, multi-channel 3D deep architecture was successful at learning high-level brain tumor appearance features jointly from MRI, functional MRI and diffusion MRI images, outperforming single-modality or 2D models. Overall, the variety of modalities, properties and sizes of training sets, the dimensionality of input, and, finally, 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.

Chest X-rays are a common radiological examination for screening and diagnosis of lung diseases. Although hospitals have accumulated a large number of raw radiology images and reports in Picture Archiving and Communication Systems and their related reports in Radiology Information System, it is not yet known how to effectively use them to learn the correlation between pathology categories and X-rays. In the last few years, deep learning methods showed remarkable results in chest X-ray image analysis [67,68]. However, it is both costly and time-consuming to annotate a large-scale fully- labeled corpus to facilitate the data-hungry deep learning models. As an alternative, Wang et al. [68] proposed to use weakly labeled images for training deep learning models. To generate weak labels for X-ray images, they applied a series of Natural Language Processing (NLP) techniques to the associated chest X-ray radiological reports. Specifically, they first extracted all diseases mentioned in the reports using a state-of-the-art NLP tool, then applied a newly developed negation and uncertainty detection tool (NegBio) to filter negative and equivocal findings in the reports. Evaluation on three independent datasets demonstrated that NegBio is highly accurate for detecting negative and equivocal findings (~90% in F-measure overall). These highly accurate results meet the need to generate a corpus with weak labels, which serves as a solid foundation for the later process of image classification. The resulting dataset (CXR-XIV [69]) consists of 108,948 frontal-view chest X-ray images (from 32,717 patients) and each image is associated with one or more weakly-labeled pathology category (e.g. pneumonia and cardiomegaly) or "normal" otherwise. Further, Wang et al. [68] used this dataset with a unified weakly-supervised multi-label image classification framework, to detect common thoracic diseases. It showed superior performance over a benchmark using fully labeled data.

In addition to medical imaging, histology slides are also being analyzed with deep learning approaches[70]. Ciresan et al. [25] developed one of the earliest examples, winning the 2012 International Conference on Pattern Recognition's Contest on Mitosis Detection while achieving human competitive accuracy. Their approach uses what has become a standard convolutional neural network architecture trained on public data. In more recent work, Wang et al. [71] analyzed stained slides to identify cancers within slides of lymph node slices. The approach provided a probability map for each slide. 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. While the algorithm had about twice the error rate of a pathologist, the errors

were not strongly correlated with those of a pathologist, suggesting that the two could be combined, theoretically, reducing the error rate to under 1%. In this area, these algorithms may be ready to incorporate into existing tools to aid pathologists. The authors' work suggests that this could reduce the false negative rate of such evaluations. This theme of an ensemble between deep learning algorithm and human expert may help overcome some of the challenges presented by data limitations.

One source of training examples with rich clinical annotations is the electronic health records. Recently Lee et a[72] developed an approach to distinguish individuals with age-related macular degeneration from control individuals. They extracted approximately 100,000 images from structured electronic health records, which they used to train and evaluate a deep neural network. Combining this data resource with standard deep learning techniques, the authors reach greater than 93% accuracy. One item that is important to note with regards to this work is that the authors used their test set for evaluating when training had concluded. In other domains, this has resulted in a minimal change in the estimated accuracy [73]. However, there is not yet a single accepted standard within the field of biomedical research for such evaluations. We recommend the use of an independent test set wherever it is feasible. Despite this minor limitation, the work clearly illustrates the potential that can be unlocked from images stored in electronic health records.

These examples demonstrate that, except for few natural image-like problems (e.g. melanoma detection), biomedical imaging poses a number of challenges for deep learning applications. Dataset sizes are typically limited, annotations can be sparse, and images are often high-dimensional, multimodal, and multi-channel. Techniques like transfer learning, heavy dataset augmentation, multi-view and multi-stream architectures are used more commonly compared to natural image domain. Furthermore, sensitivity and specificity of a model in this case often can translate directly into a clinical value. Thus, results evaluation, uncertainty estimation, and model interpretation methods are also of great importance in this domain (see Discussion). Finally, there is a need for better pathologist-computer interaction techniques that will allow combining the power of deep learning methods with human expertise and lead to better-informed decisions for patient treatment and care.

**Electronic health records**

EHR data include substantial amounts of free text, which remains challenging to approach[74]. Often, researchers developing algorithms that perform well on specific tasks must design and implement domain- specific features [75]. These features capture unique aspects of the literature being processed. Deep learning methods are natural feature constructors. In recent work, the authors evaluated the extent to which deep learning methods could be applied on top of generic features for domain-specific concept extraction [76]. They found that performance was in line with, but did not exceed, existing state of the art methods. The deep learning method had performance lower than the best performing domain-specific method in their evaluation [76]. This highlights the challenge of predicting the eventual impact of deep learning on the field. This provides support that deep learning may impact the field by reducing the researcher time and cost required to develop specific solutions, but it may not lead to performance increases.

In recent work, Yoon et al[77] analyzed simple features using deep neural networks and found that the patterns recognized by the algorithms could be re-used across tasks. Their aim was to analyze the free text portions of pathology reports to identify the primary site and laterality of tumors. The only features the authors supplied to the algorithms that they evaluated were unigrams and bigrams. These are the counts for single words and two-word combinations in a free text document. They subset the full set of words and word combinations to the 400 most commonly used ones. 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, i.e. evaluating the laterality of each tumor, the deep neural network outperformed the other methods. Of particular interest, when the authors first trained a neural network to predict primary site and then repurposed those features as a component of a secondary neural network trained to predict laterality, the performance was higher than a laterality-trained neural network. This demonstrates how deep learning methods can repurpose features across tasks, improving overall predictions as the field tackles new challenges. For the further review of this type of approaches see Discussion.

Several authors have created reusable feature sets for medical terminologies using natural language processing (NLP) and neural embedding models, as popularized by Word2vec [78]. 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 [79], ICD and NDC, and claims data [80]. Methods for neural embeddings learned from electronic health records have at least some ability to predict disease-disease associations and implicate genes with a statistical association with a disease [81]. However, the evaluations performed did not differentiate between simple predictions (i.e. the same disease in different sites of the body) and non-intuitive ones. While promising, a lack of rigorous evaluations of the real-world utility of these kinds of features makes current contributions in this area difficult to evaluate. To examine the true utility, comparisons need to be performed against leading approaches (i.e. algorithms and data) as opposed to simply evaluating multiple algorithms on the same potentially limited dataset.

Identifying consistent subgroups of individuals and individual health trajectories from clinical tests is also an active area

of research. Approaches inspired by deep learning have been used for both unsupervised feature construction and supervised prediction. Early work by Lasko et al. [82], combined sparse autoencoders and Gaussian processes to distinguish gout from leukemia from uric acid sequences. Later work showed that unsupervised feature construction of many features via denoising autoencoder neural networks could dramatically reduce the number of labeled examples required for subsequent supervised analyses [58,83]. In addition, it pointed towards learned features being useful for subtyping within a single disease. In a concurrent large-scale analysis of EHR data from 700,000 patients, Miotto et al. [84] used a deep denoising autoencoder architecture applied to the number and co-occurrence of clinical events ("DeepPatient") to learn a representation of patients. The model was able to predict disease trajectories within one year with over 90% accuracy and patient-level predictions were improved by up to 15% when compared to other methods. Razavian et al. [85] used a set of 18 common lab tests to predict disease onset using both CNN and LSTM architectures and demonstrated and 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 [86]. Early approaches, such as the Faraggi-Simon feed-forward network, aimed to relax the linearity assumption, but performance gains were lacking [87]. Katzman et al. in turn developed a deep implementation of the Faraggi-Simon network that, in addition to outperforming Cox regression, was capable of comparing the risk between a given pair of treatments, thus potentially acting as recommender system [88]. 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 [89]. 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-network approaches. For the exponential family models, despite their scalability [90], 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 [91].

#### Challenges and opportunities in patient categorization

##### Generating ground-truth labels can be expensive or impossible

A dearth of true labels is perhaps among the biggest obstacles for EHR-based analyses that employ machine learning. Popular deep learning (and 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 [92]. Because of these costs, much of this research, including the work cited in this review, skips the process of expert review. Clinicians' skepticism for research without expert review may greatly dampen their enthusiasm for the work and consequently reduce its impact. To date, even well-resourced large national consortia have been challenged by the task of acquiring enough expert-validated labeled data. For instance, in the eMERGE consortia and PheKB database [93], most samples with expert validation contain only 100 to 300 patients. These datasets are quite small, even for simple machine learning algorithms. The challenge is greater for deep learning models with many parameters. While unsupervised and semi-supervised approaches can help with small sample sizes, the field would benefit greatly from large collections of anonymized records in which a substantial number of records have undergone expert review. This challenge is not unique to EHR-based studies. Work on medical images, omics data in applications for which detailed metadata are required, and other applications for which labels are costly to obtain will be hampered as long as abundant curated data are unavailable.

Successful approaches to date in this domain have sidestepped this challenge by making methodological choices that either reduce the need for labeled examples or that use transformations to training data to increase the number of times it can be used before overfitting occurs. For example, the unsupervised and semi-supervised methods that we have discussed reduce the need for labeled examples [58]. The anchor and learn framework [94] uses expert knowledge to identify high confidence observations from which labels can be inferred. The adversarial training example strategies that we have mentioned can reduce overfitting, if transformations are available that preserve the meaningful content of the data while transforming irrelevant features [53]. While adversarial training examples can be easily imagined for certain

methods that operate on images, it is more challenging to figure out what an equivalent transformation would be for a patient's clinical test results. Consequently, it may be hard to employ adversarial training examples, not to be confused with generative adversarial neural networks, with other applications. Finally, approaches that transfer features can also help use valuable training data most efficiently. Rajkomar et al. trained a deep neural network using generic images before tuning using only radiology images [61]. Datasets that require many of the same types of features might be used for initial training, before fine tuning takes place with the more sparse biomedical examples. Though the analysis has not yet been attempted, it is possible that analogous strategies may be possible with electronic health records. For example, features learned from the electronic health record for one type of clinical test (e.g. a decrease over time in a lab value) may transfer across phenotypes.

Methods to accomplish more with little high-quality labeled data are also being applied in other domains and may also be adapted to this challenge, e.g. data programming [95]. In data programming, noisy automated labeling functions are integrated. Numerous writers have described data as the new oil [96,97]. The idea behind this metaphor is that data are available in large quantities, valuable once refined, and the underlying resource that will enable a data-driven revolution in how work is done. Contrasting with this perspective, Ratner, Bach, and Ré described labeled training data as "The *New* New Oil" [98]. In this framing, data are abundant and not a scarce resource. Instead, new approaches to solving problems arise when labeled training data become sufficient to enable them. Based on our review of research on deep learning methods to categorize disease, the latter framing rings true.

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

#### Data sharing is hampered by standardization and privacy considerations

To construct the types of very large datasets that deep learning methods thrive on, we need robust sharing of large collections of data. This is in part a cultural challenge. We touch on this challenge in the Discussion section. Beyond the cultural hurdles around data sharing, there are also technological 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 [99]. Science is at best a tertiary priority, presenting challenges to EHR-based research in general, and to deep learning research particularly. These difficulties can be grouped into three areas: local bias, wider standards and legal issues. Note these problems are not restricted to EHR 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[100,101]. Individual users have unique usage patterns, with different departments and different hospitals having different priorities which code patients and introduce missing data in a non-random fashion [102]. Patient data may be kept across several "silos" within a single health system. Even the most basic task of matching patients across systems can be challenging due to data entry issues [103]. 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. As a result, EHR can be less complete and less objective than expected.

In the wider picture, standards for EHRs are many and evolving. Proprietary systems, indifferent and scattered use of health information standards and controlled terminologies makes combining and comparison of data across systems challenging [104,105]. Further diversity arises from variation in languages, healthcare practices and demographics. Merging EHR gathered in different systems (and even under different assumptions) is challenging [106].

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 [107] and restricting the number of parameters that can be trained in a model. Further, rules-based algorithms have been popular in EHR-based research but because these are developed at a single institution and trained with a specific patient population they do not transfer easily to other populations [108]. For example, Wiley et al. [109] 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 [110]. Here again, standards are heterogeneous and evolving but often EHR data can often not be exported or even accessed directly for research purposes without appropriate consent. Once again, this has the effect of making data gathering more laborious, expansive and reducing sample size and study power.

Several technological solutions have been proposed in this direction, allowing access to sensitive data satisfying privacy and legal concerns. Software like DataShield [111] and ViPAR [112], although not EHR specific, allows querying and

combining of datasets and calculation of summary statistics across remote sites by "taking the analysis to the data". The computation is carried out at the remote site. Conversely, the EH4CR project [105] allows analysis of private data by use of an inter-mediation layer that intreprets remote queries across internal formats and datastores and returns the results in a de-identified standard form, thus giving real-time consistent but secure access. Continuous Analysis [113] can allow reproducible computing on private data. Using such techniques intermediate results can be automatically tracked and shared without sharing the original data. While none of these have been used in deep learning, the potential is there.

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

There are approaches to protect models, but they pose their own challenges. Training in a differentially private manner provides a limited guarantee that an algorithm's output will be equally likely to occur regardless of the participation of any one individual. The limit is determined by a single parameter which provides a quantification of privacy. Simmons et al. [118] present the ability to perform GWASs in a differentially private manner and Abadi et al[119] show the ability to train deep learning classifiers under the differential privacy framework. Federated learning [120] and secure aggregations [121–123] are complementary approaches that reinforce differential privacy. Both aim to maintain privacy by training deep learning models from decentralized data sources such as personal mobile devices without transferring actual training instances. This is becoming of increasing importance with the rapid growth of mobile health applications. However, the training process in these approaches places constraints on the algorithms used and can make fitting a model substantially more challenging. In our own experience it can be trivial to train a model without differential privacy, but quite difficult to train one within the differential privacy framework. The problem can be particularly pronounced with small sample sizes.

While none of these problems are insurmountable or restricted to deep learning, they present challenges that cannot be ignored. Technical evolution in EHRs and data standards will doubtless ease - although not solve - the problems of data sharing and merging. More problematic are the privacy issues. Those applying deep learning to the domain should consider the potential of inadvertently disclosing the participants identity. Techniques that enable training on data without sharing the raw data may have a part to play. Training within a differential privacy framework may often be warranted.

#### Discrimination and "right to an explanation" laws

In April 2016, the European Union adopted new rules regarding the use of personal information, the General Data Protection Regulation (GDPR) [124]. 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
TODO: @traversc citation needed.

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 health care data will make this difficult. First, research samples are frequently non-representative of the general population of interest; they tend to be disproportionately sick [125], male [126], and European in ancestry[127]. 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 [125]. 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; [128]). As a result, many genomic findings are of limited value for people of non-European ancestry[127] and may even lead to worse treatment outcomes for them. Methods have been developed for mitigating some of these problems in genomic studies [125,128], 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. [129]); 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; [130]). When we are not careful with our data and models, we may inadvertently say more about the way the data was collected (which may involve a history of unequal access and discrimination) than about anything of scientific or

predictive value. This fact can undermine the privacy of patient data [130] or lead to severe discriminatory consequences [131]. There is a small but growing literature on the prevention and mitigation of data leakage[130], as well as a closely-related literature on discriminatory model behavior [132], 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 [133]. Despite the difficulties and uncertainties, machine learning practitioners (and particularly those who use deep neural networks, which are challenging to interpret) must remain cognizant of these dangers and make every effort to prevent harm from discriminatory predictions.

To reach their potential in this domain, deep learning methods will need to be interpretable. Researchers need to consider the extent to which biases may be learned by the model and whether or not a model is sufficiently interpretable to identify biases. We discuss the challenge of model interpretability more completely in the discussion section.

#### Temporal Patient Trajectories

Traditionally, physician training programs justified long training hours by citing increased continuity of care and learning by following the progression of a disease over time, despite the known consequences of decreased mental acuity and quality of life [134–137]. Yet, a common practice in EHR-based research is to take a point in time snapshot and convert patient data to a traditional vector for machine learning and statistical analysis. This results in significant signal losses as timing and order of events provide insight into a patient's disease and treatment. Efforts to account for the order of events have shown promise [138] but require exceedingly large patient sizes due to discrete combinatorial bucketing. Lasko et al. [82] used autoencoders on longitudinal sequences of serum urine acid measurements to identify population subtypes. More recently, deep learning has shown promise working with both sequences (Convolutional Neural Networks) [139] and the incorporation of past and current state (Recurrent Neural Networks, Long Short Term Memory Networks)[140]. This may be a particular area of opportunity for deep neural networks. The ability to discover relevant sequences of events from a large number of trajectories requires powerful and flexible feature construction methods - an area at which deep neural networks tend to excel.

## How is deep learning used to study basic biological processes in a manner that may provide future insights into human disease?

*The (awkward) placeholder section title is intended to help define the scope. We do not want this section to become a miscellaneous collection of everything that does not fit in Categorize and Treat.*

*One proposal is that we organize this roughly by what is being predicted, which will generally correspond to the types of data being used. For each sub-section we can quickly introduce the prediction problem and cite some examples of the relevance to disease. Hypothetically, if we had an algorithm that produced perfect predictions on the task, what would we learn and how could those predictions be used?*

*Existing reviews could be mentioned briefly.*

*It may not fit here, but there could be a general discussion of why different neural network architectures are particularly well-suited for different types of input data. For example, CNNs and RNNs for 1-dimensional data are used in several categories below.*

*A few suggestions for sub-sections follow. Some of these could be left out because our goal is not an exhaustive enumeration of methods. Some are important areas of biology, but there may not be much deep learning- specific content to present. Others may be important areas where we lack expertise, in which case we may acknowledge the application area but not dive into merits or weaknesses of individual methods.*

TODO: dropped signaling subsection, decide whether to discuss in intro and why it has received less attention, @rmjDc5rm as example

### 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 have been traditionally made by quantitative polymerase chain reaction (qPCR), low-throughput fluorescence-based methods, and microarray technologies, the field has shifted in recent years to primarily performing RNA sequencing (RNA-seq) to catalog whole transcriptomes. As RNA-seq continues to fall in price and rise in throughput, applying deep learning to study gene expression data is likely to make training deep models more feasible. With increased modeling ability, deep learning approaches are likely to grow in popularity and lead to novel biological insights.

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 [141] and also to stratify yeast strains based on chemical and mutational perturbations [142]. Shallow (one hidden layer) denoising autoencoders have also been fruitful in extracting biological insight from thousands of *Pseudomonas aeruginosa* experiments [143,144] and in aggregating features relevant to specific breast cancer subtypes [24]. These unsupervised approaches applied to gene expression data are powerful methods for aggregating features and identifying gene signatures that may otherwise be overlooked by alternative methods. An additional benefit of unsupervised approaches is that ground truth labels, which are often difficult to acquire or are incorrect, are nonessential. However, careful interpretation must be performed regarding how the genes are aggregated into features. Precisely attributing node activations to specific biological functions risks over-interpreting models and can lead to incorrect conclusions.

Deep learning approaches are also being applied to gene expression prediction tasks. For example, a deep neural network with three hidden layers outperformed linear regression in inferring the expression of over 20,000 target genes based on a representative, well-connected set of about 1,000 landmark genes [145]. 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. Alternatively, epigenetic data may have sufficient explanatory power for inference of gene expression. For instance, a convolutional neural network applied to histone modifications, termed DeepChrome, [146] was shown to improve prediction accuracy of high or low expression over existing methods. Deep learning can also be useful for integrating different data types. For example, Liang et al. combined RBMs to integrate gene expression, DNA methylation, and miRNA data to define ovarian cancer subtypes [147]. While the aforementioned approaches are promising, many convert gene expression measurements to categorical or binary variables, thus ablating many complex gene expression signatures present in intermediate and relative numbers.

Deep learning applied to gene expression data is still in its infancy, but the future is bright. Many previously untestable hypotheses can now be interrogated as deep learning enables analysis of increasing amounts of data generated by new technologies. For example, the effects of cellular heterogeneity on basic biology and disease etiology can now be explored by single-cell RNA-seq and high-throughput fluorescence-based imaging, techniques that will benefit immensely from deep learning approaches.

## Splicing

Pre-mRNA transcripts can be spliced into different isoforms by retaining or skipping subsets of exons, or including parts of introns, creating enormous spatiotemporal flexibility to generate multiple distinct proteins from a single gene. Unfortunately, this remarkable complexity can lend itself to defects that underlie many diseases [148]. For instance, in Becker muscular dystrophy, a point mutation in dystrophin creates an exon splice silencer that induces skipping of exon 31. A recent study found that quantitative trait loci (QTLs) 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 much more widespread feature of human pathologies than previously thought [149].

Sequencing studies routinely return thousands of unannotated variants, but which cause functional changes in splicing, and if so, how? Prediction of a “splicing code” has been a holy grail over the past decade. Initial machine learning approaches used a naive Bayes model and a 2-layer Bayesian neural network with thousands of hand-derived sequence-based features to predict the probability of exon skipping [150,151]. With the advent of deep learning, more complex models were built that provided better predictive accuracy [152,153]. 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”, compared to earlier approaches that often assumed independence of features or required extensive human fine-tuning. Moreover, as in gene expression network analysis, interrogating the hidden nodes within neural networks will likely yield new biological insights into splicing. For instance, tissue-specific splicing mechanisms can be inferred by training networks on splicing data from different tissues, then searching for common versus distinctive nodes, a technique employed by Qin et al. for tissue-specific TF binding predictions [154].

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 [155]. Interestingly, they built a simple linear model using hexamer motif frequencies that successfully generalized to exon skipping: in a limited analysis using SNPs from three genes, it predicted exon skipping with three times the accuracy of Xiong et al.’s deep learning-based framework. This case is instructive in that clever sources of data, not just more descriptive models, are still critical in yielding novel insights.

We already understand how mis-splicing of a single gene can cause diseases such as Duchenne muscular dystrophy. The challenge now is to uncover how genome-wide alternative splicing underlies complex, non-Mendelian diseases such as autism, schizophrenia, Type 1 diabetes, and multiple sclerosis [156]. As a proof of concept, Xiong et al.[152] sequenced five ASD and 12 control samples, each with an average of 42,000 rare variants, and identified mis-splicing in 19 genes with neural functions. Deep learning will allow 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.

Our end goal is to predict an individual's tissue-specific, exon-specific splicing patterns from their genome sequence and other measurements. Knowing exactly which genes are mis-spliced in each tissue could enable a new branch of precision diagnostics that also stratifies patients and suggests targeted therapies to correct splicing defects. A continued focus on interpreting the "black box" of deep neural networks, along with integrating diverse data sources, will help us better understand the basic determinants of splicing and its links to complex disease, which will lead to novel diagnostics and therapeutics.

## Transcription factors and RNA-binding proteins

Transcription Factor and RNA-binding proteins are key components for gene regulation, making them very important to understand for higher level biological processes. While high-throughput sequencing techniques such as chromatin immunoprecipitation and massively parallel DNA sequencing (ChIP-seq) have been able to accurately identify binding regions for DNA and RNA proteins, these experiments are both time consuming and expensive. In addition, the sequencing methods do not provide any sort of analysis on the proteins which would lead to a better understanding of the underlying process. Thus, there is a need to computationally predict and understand these binding regions de novo from sequences.

### Transcription factors

Transcription factors (TFs) are regulatory proteins that bind to certain locations on a DNA sequence and control the rate of mRNA production. ChIP-seq and related technologies are able to identify highly likely binding sites for a certain TF, and databases such as ENCODE [157] have provided ChIP-seq data for hundreds of different TFs across many laboratories. However, ChIP-seq experiments are expensive and time consuming. Since the data that scientists have discovered is available, *in silico* methods to predict binding sites are of great interest, thus eliminating the need to do new ChIP-seq experiments every time analysis is done on a new sequence.

In order to computationally predict transcription factor binding sites (TFBSs) on a DNA sequence, researchers initially used consensus sequences and position weight matrices to match against a test sequence [158]. Simple neural network classifiers were then proposed to differentiate positive and negative binding sites but did not show significant improvements over the weight matrix matching methods [159]. Later, SVM techniques outperformed the generative methods by using k-mer features [160,161], but string kernel based SVM systems are limited by expensive computational cost proportional to the number of training and testing sequences. More recently, [162] showed that convolutional neural network models could achieve state of the art results on the TFBS task and are scalable to a large number of genomic sequences.

Since the work by [162] was published, there have been a multitude of deep learning works on the TFBS task. Due to the motif-driven nature of the TFBS task, most architectures have been convolutional-based, as summarized in [163]. [164] introduced several new convolutional and recurrent neural network models for predicting TFBSs, which showed improvements over other deep learning models on the dataset from [162]. While many models for TFBS prediction resemble computer vision and natural language processing (NLP) tasks, it is important to note that DNA sequence tasks are fundamentally different than NLP tasks, and thus the models should be adapted from traditional deep learning models in order to account for such differences. For example, motifs may appear in either strand of a DNA sequence, resulting in two different forms of the motif (forward and reverse complement) due to complementary base pairing. To handle this issue, [165] created a convolutional model which can find motifs in both directions. Since deep learning for protein binding prediction is still in early stages, we expect to see an increase in domain-specific architectures for this task.

Deep learning models have shown great accuracy on TFBS prediction, but the results are not fully convincing for several reasons. First, ChIP-seq experiments give a continuous value of binding likelihood at a certain location based on many experiments at that location in DNA. Based on these values, the TFBS task is usually converted into a binary classification task based on a certain threshold of the ChIP-seq value. However, the ChIP-seq values are often noisy, resulting in a target class that may be incorrect dependent on the defined threshold. Converting the task into a binary classification isn't completely accurate since the model may give a high probability of a binding site, but the ChIP-seq binding signal may be just barely over the binary classification threshold, resulting in potential false positive (or the signal may be just below the threshold, resulting in potential false negatives).

Second, most datasets for TFBS prediction are separated by TF, requiring a separate model for each TF (i.e. binary classification on each TF sub-dataset). In reality, there may be multiple TFs binding at the same location and TF binding may be dependent on other TFs, thus requiring both a dataset which gives all TF binding values at every location as well as a multi-task model. [162] use multiple TFs at once, but TF binding prediction is an intermediate step for predicting effects of noncoding variants in their model, so TFBS prediction is not heavily analyzed.

A third issue is that it is unclear exactly how to include non-binding or "negative" sites in the datasets. Since the number

of positive binding sites of a particular TF is relatively small with respect to the total number of base-pairs in DNA, we must choose a small subset of the total non-binding sites, resulting in some sort of bias over all of the actual negative sites. Regardless of the negative site selection, most datasets evenly balance the positive and negative binding sites and report auROC for the metric. This is very misleading in a task where the binding sites are very unevenly balanced in the real world (see Discussion). Thus, we need datasets which more accurately model real TFBS data.

At the model level, while deep learning models have shown that it is possible to automatically extract features for TFBS prediction at the sequence level, these basic models cannot predict the binding on new unseen cell types or conditions. Since ChIP-seq experiments have been performed on only a small subset of cell lines, these prediction models are not of use for new or rare cell lines. To handle this issue and create models which can be used across cell lines, there are several options. The most prominent would be to introduce a multimodal model that, in addition to sequence data, incorporates cell-line specific features such as chromatin accessibility, DNA methylation, or gene expression. Without cell-specific features, the other option is to use domain adaptation methods where we train our model on a source cell type and use unsupervised feature extraction methods to predict on a target cell type. [154] predicts binding in new cell type-TF pairs, but the cell types must be in the training set for other TFs besides the target TF. A more general domain transfer model across cell types would be more useful.

While neural architectures are rapidly changing and producing better results, it is clear that deep learning can be efficiently and effectively used to do functional prediction on the genome given raw data.

Accurately predicting transcription factors computationally is useful, but it is also important to understand how these computational models make their predictions. To handle this, several papers have focused on understanding machine learning models [162,164,166]. [162] was the first to introduce a visualization method for a deep learning model on the TFBS task, and they did so by visualizing the learned convolution filters which were informative for the model's prediction of a specific sample. However, this approach was specific to visualizing certain samples fed through their particular model. [164] introduced a suite of state-of-the-art deep learning models and new visualizations techniques for a more in-depth analysis of TFBSs. Furthermore, [166] introduced an advanced visualization method and toolbox for analyzing possible TFBS sequences. [162] also introduced mutation maps, where they could easily mutate, add, or delete base pairs in a sequence and see how the model changed its prediction. This is something that would be very time consuming in a lab setting, but easy to simulate using their model. Visualization techniques on deep learning models are important because they can provide new insights on regulatory mechanisms and can lead biologists to test and verify in a lab setting, leading to new biomedical knowledge. Since the "linguistics" of DNA are unclear, interpretability of models is crucial to pushing our understanding forward.

TODO: cut RNA-binding proteins from above section, refer to representative papers

### Promoters, enhancers, and related epigenomic tasks

TODO: There's inevitably a lot of overlap been this and the transcription binding section. Maybe in the long term think about restructuring to one big "r

Identification of promoters and other cis-regulatory elements (CREs) presents an obvious use case for deep learning. Transcriptional control is undoubtedly a vital - and early - part of the regulation of gene expression. An abundance of sequence and associated functional data (e.g. ENCODE, ExAC) exists across species. At the same time, studies of gene regulation have often focused on the protein (binding) rather than promoter level [167], perhaps due to the ill-defined nature of CREs. A promoter itself can be seen as an assemblage of "active" binding sites for transcription factors interspersed by less-characterized and perhaps functionally silent spacer regions. However, the sequence signals that control the start and stop of transcription and translation are still not well understood, compounded by incomplete understanding of alternative transcripts and the context for these alternatives. Sequence similarity is poor even between functionally correlated genes. While homologs might be studied for insight, they may not exist or may be just as poorly characterized.

Recognizing enhancers presents additional challenges. Enhancers may be up to 1 Mbp upstream or downstream from the target promoter, on either strand, even within the introns of other genes [168]. They do not necessarily operate on the nearest gene and may in fact effect multiple genes. Their activity is frequently tissue- or context-specific. A substantial fraction of enhancers displays modest or no conservation across species. There is no universal enhancer sequence signal or marker for enhancers, and some literature suggests that enhancers and promoters may be just categories along a spectrum [169]. One study [170] even showed that only 33% of predicted regulatory regions could be validated, while a class of "weak" predicted enhancers were strong drivers of expression. Yet there is growing evidence for their vast ubiquity, making them possibly the predominant functional non-coding element. Thus, identifying enhancers is critical yet the search space is embarrassingly large.

While prior (non-deep learning) approaches have made steady improvements on promoter prediction, there is little consensus on the best approach and performance is poor. Typically algorithms will recognize only half of all promoters, with an accompanying high false positive rate [171]. Methods with better sensitivity generally do so at the cost of poorer specificity. Conventional identification of enhancers has leaned heavily on simple conservation or laborious experimental techniques, with only moderate sensitivity and specificity. For example, while chromatin accessibility has often been used for identifying enhancers, this also "recognizes" a wide variety of other functional elements, like

promoters, silencers and repressors.

The complex nature of CREs (and our ignorance at to what are the important features of them) therefore seems a good subject for deep learning approaches. Indeed, neural networks were used for promoter recognition as early as 1996, albeit with mixed results [172]. Since then, there has been much work in applying deep learning to this area, although little in the way of comparative studies or reviews. We will therefore focus on a few recent characteristic studies to outline the state of the art and extant problems.

Most broadly, Kelley et al. [173] trained CNNs on DNA accessibility datasets, getting a marked improvement on previous methods, albeit still with a high false positive rate. (Note as above, using DNA accessibility conflates enhancers with other functional sites.) This study also featured a useful interpretability approach (analogous to in silico mutagenesis [174]) introducing known protein binding motifs into sequences and measuring the change in predicted accessibility.

Umarov et al. [175] demonstrated the use of Convolutional Neural Networks in recognizing promoter sequences, achieving markedly better performance than conventional methods (sensitivity and specificity exceeding 90%). While some results were achieved over bacterial promoters (which are considerably simpler in structure), surprisingly roughly similar performance was found for human promoters. This work also included a promising and simple method for interpretability (randomly substituting bases in a recognized promoter region, then checking for a change in recognition) that would be useful to exploit more widely.

Xu et al. [176] applied CNNs to the detection of enhancers, achieving incremental improvements in specificity and sensitivity over previous SVM-based approach, and markedly better performance for cell-specific enhancers. A massive improvement in speed was also achieved. Additionally, they compared the performance of different CNN architectures, finding that while layers for batch normalization improved performance, surprisingly deeper architectures decreased performance.

Singh et al. [177] also used batch normalization, approaching the problem of predicting enhancer-promoter interactions from solely the sequence and location of putative enhancers and promoters in a particular cell type. Performance was comparative to state of the art conventional techniques that used the whole gamut of full functional genomic and non-sequence data.

Given the conflation between different CREs, the study of Li et al[178] is particularly interesting. They used a feed-forward neural network to distinguish classes of CRES and activity states. Active enhancers and promoters could be easily be distinguished, as could active and inactive elements. Perhaps unsurprisingly, it was difficult to distinguish between inactive enhancers and promoters. They also investigated the power of sequence features to drive classification, finding that beyond CpG islands, few were useful.

In summary, deep learning is a promising approach for identifying CREs, able to interrogate sequence features that are complex and ill-understood, already offering marked improvements on the prior state of the art. However, the exact methodology is up for debate and needs examination and more comparative study. Work needs to be done in understanding the best architecture to be used. Further concern surrounds the training data. CNNs require a large training set to avoid over-fitting, yet in many cases we do not have a large "gold standard" dataset to train against. Further more, the quality and meaning of training data needs to be closely considered, given that a "promoter" or "enhancer" may only be putative or dependent on the experimental method or context of identification. Else we risk building detectors not for CREs but putative CREs. The best model of negative sample needs also to be considered, as does the problem of imbalanced data. In a sentence, these methods can only be as good as their training data. Finally, interpretability is a problem but one that appears to be solvable given empirical approaches.

TODO: discuss enhancer-promoter and enhancer-target prediction

## Micro-RNA binding

Prediction of microRNAs (miRNAs) in the genome as well as miRNA targets is of great interest, as they are critical components of gene regulatory networks and are often conserved across great evolutionary distance [179,180]. While many machine learning algorithms have been applied to solve these prediction tasks, they currently require extensive feature selection and optimization. For instance, one of the most widely adopted tools for miRNA target prediction, TargetScan, trained multiple linear regression models on 14 hand-curated features including structural accessibility of the target site on the mRNA, the degree of site conservation, and predicted thermodynamic stability of the miRNA:mRNA complex [181]. Some of these features, including structural accessibility, are imperfect or empirically derived. In addition, current algorithms suffer from low specificity [182].

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 [182,183]. 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 15-25% 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 neural-network approaches 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. Similar to other applications, the major contribution of deep learning will be to deliver deep insights into the biology of miRNA targeting as we learn to interrogate the hidden nodes within neural networks.

## 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, various computational methods have been developed to predict aspects of protein structure such as torsion angles, solvent accessibility, inter-residue contact maps, disorder regions, and side-chain packing. Since 2012, various deep learning architectures have been utilized, including deep belief networks, LSTM (long short-term memory), deep convolutional neural networks (DCNN), and deep convolutional neural fields (CNF) [29,184].

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 significantly improved [27,185–187].

Protein secondary structure can exhibit three different states (alpha helix, beta strand, and loop regions) or eight finer-grained states. Q3 and Q8 accuracy pertain to 3-state or 8-state predictions, respectively. Initially, several groups [28,188,189] made important technical advances in adapting deep learning to protein structure prediction, but were unable to achieve significant improvement over the industry standard PSIPRED [190], which uses two shallow feedforward neural networks. In 2014, Zhou and Troyanskaya demonstrated that they could improve Q8 accuracy by using a deep supervised and convolutional generative stochastic network [191]. In 2016 Wang et al. developed a deep CNF model (DeepCNF) that significantly improved Q3 and Q8 accuracy as well as prediction of solvent accessibility and disorder regions [29,184]. DeepCNF was the first tool to achieve Q3 accuracy of 84-85%, much higher than the 80% accuracy standard maintained by PSIPRED for more than 10 years. This improvement may be mainly due to the ability of convolutional neural fields to capture long-range sequential information, which is important for beta strand prediction. Nevertheless, improving secondary structure prediction from 80% to 84-85% is unlikely to result in a commensurate improvement in tertiary structure prediction since secondary structure mainly reflects coarse-grained local conformation of a protein structure.

Protein contact prediction and contact-assisted folding (i.e. folding proteins using predicted contacts as restraints) represents a promising new direction for ab initio folding of proteins without good templates in PDB. Evolutionary coupling analysis (ECA) is effective for proteins with a very large number (>1000) of sequence homologs [187], but otherwise fares poorly for proteins without many sequence homologs. By combining ECA with a few other protein features, shallow neural network-based methods such as MetaPSICOV [185] and CoinDCA-NN [192] have shown some advantage over ECA for proteins with few sequence homologs, but their accuracy is far from optimal. In recent years, deeper architectures have been explored for contact prediction. For example, Di Lena et al. introduced a deep spatio-temporal neural network (up to 100 layers) that utilizes both spatial and temporal features to predict protein contacts [193]. Eickholt and Cheng combined deep belief networks and boosting techniques to predict protein contacts[194] and trained deep networks by layer-wise unsupervised learning followed by fine-tuning of the entire network. Skwark and Elofsson et al. developed an iterative deep learning technique for contact prediction by stacking a series of Random Forests [195]. However, blindly tested in the well-known CASP competitions, these methods did not show any advantage over MetaPSICOV [185].

Recently, Wang et al. proposed the deep learning method RaptorX-Contact[27], which significantly improves contact prediction over MetaPSICOV, especially for proteins without many sequence homologs. It employs a network architecture formed by one 1D residual neural network and one 2D residual neural network. Blindly tested in the latest CASP competition (i.e. CASP12 [196]), RaptorX-Contact ranked first in F1 score (a widely-used performance metric combining sensitivity and specificity) on free-modeling targets as well as the whole set of targets. In CAMEO (which can be interpreted as a fully-automated CASP) [197], its predicted contacts were also able to fold proteins with a novel fold and only 65-330 sequence homologs. This technique also worked well on membrane protein contact prediction even when trained mostly on non-membrane proteins [198]. RaptorX-Contact performed better mainly due to introduction of residual neural networks and exploitation of contact occurrence patterns by simultaneously predicting of all the contacts in a single protein.

Taken together, we believe it is still possible to further improve contact prediction for proteins with fewer than 1000

homologs by studying new deep network architectures. However, it is unclear whether there is an effective way to use deep learning to improve prediction for proteins with almost no homologs. Finally, the deep learning methods summarized above also apply to interfacial contact prediction for protein complexes, but may be less effective since on average protein complexes have fewer sequence homologs.

## Morphological phenotypes

A field poised for dramatic revolution by deep learning is bioimage analysis. Thus far, the primary use of deep learning for biological images has been for segmentation - that is, for the identification of biologically relevant structures in images such as nuclei, infected cells, or vasculature, in fluorescence or even brightfield channels [199]. Once so-called regions of interest have been identified, it is often straightforward to measure biological properties of interest, such as fluorescence intensities, textures, and sizes. Given the dramatic successes of deep learning in biological imaging, we simply refer to articles that review recent advancements [16,199,200]. We believe deep learning will become a commonplace tool for biological image segmentation once user-friendly tools exist.

We anticipate an additional kind of paradigm shift in bioimaging that will be brought about by deep learning: what if images of biological samples, from simple cell cultures to three-dimensional organoids and tissue samples, could be mined for much more extensive biologically meaningful information than is currently standard? For example, a recent study demonstrated the ability to predict lineage fate in hematopoietic cells up to three generations in advance of differentiation [201]. In biomedical research, by far the most common paradigm is for biologists to decide in advance what feature to measure in images from their assay system. Although classical methods of segmentation and feature extraction can produce hundreds of metrics per cell in an image, deep learning is unconstrained by human intuition and can in theory extract more subtle features through its hidden nodes. Already, there is evidence deep learning can surpass the efficacy of classical methods [202], even using generic deep convolutional networks trained on natural images [203], known as transfer learning.

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 [204–206]. 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. Perhaps most excitingly, focused characterization of these higher-level features will likely lead to new and valuable biological insights.

## Single-cell

Single-cell methods are generating extreme excitement as biologists recognize the vast heterogeneity within unicellular species and between cells of the same tissue type in the same organism [56]. For instance, tumor cells and neurons can both harbor extensive somatic variation [207]. Understanding single-cell diversity in all its dimensions — genetic, epigenetic, transcriptomic, proteomic, morphologic, and metabolic — is key if precision medicine is 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 rapid 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 FISH or the heterogeneity of epigenetic patterns with single-cell Hi-C or ATAC-seq [208,209].

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 in a phenomenon known as dropout. To solve this problem, Angermueller et al. [210] trained a neural network to predict the presence or absence of methylation of a specific CpG site in single cells based on surrounding methylation signal and underlying DNA sequence, achieving several percentage points of improvement compared to random forests or deep networks trained only on CpG or sequence information. Similar deep learning methods have been applied to impute low-resolution ChIP-seq signal from bulk tissue with great success, and they could easily be adapted to single cell data [154,211].

Examining populations of single cells can reveal biologically meaningful subsets of cells as well as their underlying gene regulatory networks [212]. Unfortunately, machine learning generally struggles with unbalanced data — when there are many more inputs of class 1 than class 2 — because prediction accuracy is usually evaluated over the entire dataset. To tackle this challenge, Arvaniti et al. [213] 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 an impressive precision of 50% to 90% with 80% recall on cells where the subset percentage ranged from 0.1 to 1%, which significantly outperformed logistic regression and distance-based outlier detection methods. However, they did not benchmark against random forests, which tend to be better with unbalanced data, or against the neural network itself, and their data was fairly low dimensional. Future work will be needed to establish the utility of deep learning in cell subset identification, but the stunning improvements in image classification over the past 5 years [214] suggest that this goal will be achievable.

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 [215], or by feeding cell morphology and movement into neural networks [201]. Reinforcement learning algorithms [216] 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. It will be exciting to see the creative applications of deep learning to single-cell biology that emerge over the next few years.

TODO: https://github.com/greenelab/deep-review/issues/153

## Metagenomics

TODO: Add reference tags to this section Metagenomics (which refers to the study of genetic material, 16S rRNA and/or whole-genome shotgun DNA, from microbial communities) has revolutionized the study of micro-scale ecosystems within us and around us. There is increasing literature of applying machine learning in general to metagenomic analysis. In the late 2000’s, a plethora of machine learning methods were applied to classifying DNA sequencing reads to the thousands of species within a sample. An important problem is genome assembly from these mixed-organism samples. And to do that, the organisms should be “binned” before assembling. Binning methods began with many k-mer techniques [217] and then delved into other clustering algorithms, such as self-organizing maps (SOM)[218]. Then came the taxonomic classification problem, with researchers naturally using BLAST [219], followed by other machine learning techniques such as SVMs [220], naive Bayesian classifiers [221], etc. to classify each read. Then, researchers began to use techniques that could be used to estimate relative abundances of an entire sample, instead of the precise but painstakingly slow read-by-read classification. Relative abundance estimators (a.k.a diversity profilers) are MetaPhlan [222], (WGS)Quikr [223], and some configurations of tools like OneCodex[224] and LMAT [225]. While one cannot identify which reads were mapped back to an organism using relative abundance estimators, they can be useful for faster comparative and other downstream analyses. Newer methods hope to classify reads and estimate relative abundances at faster rates [226] and as of this writing, there are more than 70 metagenomic taxonomic classifiers in existence. Besides binning and classification of species, there is functional identification and annotation of sequence reads [227,228]. However, the focus on taxonomic/functional annotation is just the first step. Once organisms are identified, there is the interest in understanding the interrelationship between these organisms and host/environment phenotypes [229]. One of the first attempts was a survey of supervised classification methods for microbes->phenotype classification [230], followed by similar studies that are more massive in scale[231,232]. There have been techniques that bypass the taxonomic classification step altogether [233], (sequence composition to phenotype classification). Also, researchers have looked into how feature selection can improve classification [232,234], and techniques have been proposed that are classifier-independent [235,236].

So, how have neural networks (NNs) been of use? Most neural networks are being used for short sequence->taxa/function classification, where there is a lot of data for training (and thus suitable for NNs). Neural networks have been applied successfully to gene annotation (e.g. Orphelia [237] and FragGeneScan [238]), which usually has plenty of training examples. Representations (similar to Word2Vec [78] in natural language processing) for protein family classification has been introduced and classified with a skip-gram neural network [239]. Recurrent neural networks show good performance for homology and protein family identification [240,241]. Interestingly, Hochreiter, who invented Long Short Term Memory, delved into homology/protein family classification in 2007, and therefore, deep learning is deeply rooted in functional classification methods.

One of the first techniques of “de novo” genome binning used self-organizing maps, a type of NN[218]. Essinger et al. use ART, a neural network algorithm called Adaptive Resonance Theory, to cluster similar genomic fragments and showed that it has better performance than K-means. However, other methods based on interpolated Markov models [242] have performed better than these early genome binners. Also, neural networks can be slow, and therefore, have had limited use for reference-based taxonomic classification, with TAC-ELM [243] being the only NN-based algorithm to taxonomically classify massive amounts of metagenomic data. Also, neural networks can fail to perform if there are not enough training examples, which is the case with taxonomic classification (since only ~10% of estimated species have been sequenced). An initial study shows that deep neural networks have been successfully applied to taxonomic classification of 16S rRNA genes, with convolutional networks provide about 10% accuracy genus-level improvement over RNNs and even random forests [244]. However, this study performed 10-fold cross-validation on 3000 sequences in total.

Due to the traditionally small numbers of metagenomic samples in studies, neural network uses for classifying phenotype from microbial composition are just beginning. A standard MLP was able to classify wound severity from microbial species present in the wound [245]. Recently, multi-layer, recurrent networks (and convolutional networks) have been applied to microbiome genotype-phenotype, with Ditzler et al. being the first to associate soil samples with pH level using multi-layer perceptrons, deep-belief networks, and recursive neural networks (RNNs) [246]. Besides classifying the samples appropriately, Ditzler shows that internal phylogenetic tree nodes inferred by the networks are appropriate features representing low/high pH, which can provide additional useful information and new features for future metagenomic sample comparison. Also, an initial study has show promise of these networks for diagnosing disease [247].

There are still a lot of challenges with applying deep neural networks to metagenomics problems. They are not ideal for microbial/functional composition->phenotype classification because most studies contain tens of samples (~20->40) and hundreds/thousands of features (aka species). Such underdetermined/ill-conditioned problems are still a challenge for deep neural networks that require many more training examples than features to sufficiently converge the weights on the hidden layers. Also, due to convergence issues (slowness and instability due to large neural networks modeling very large datasets [248]), taxonomic classification of reads from whole genome sequencing seems out of reach at the moment for deep neural networks -- due to only thousands of full-sequenced genomes as compared to hundreds of thousands of 16S rRNA sequences available for training.

However, because recurrent neural networks are showing success for base-calling (and thus removing the large error in the measurement of a pore's current signal) for the relatively new Oxford Nanopore sequencer [249], there is hope that the process of denoising->organism/function classification can be combined into one step in using powerful LSTM's. LSTM's are working miracles in raw speech signal->meaning translation [250], and combining steps in metagenomics are not out of the question. For example, metagenomic assembly usually requires binning then assembly, but could deep neural nets accomplish both tasks in one network? Does functional/taxonomic classification need to be separate processes? The largest potential in deep learning is to learn "everything" in one complex network, with a plethora of labeled (reference) data and unlabeled (microbiome experiments) examples.

## Sequencing and variant calling

While we have so far discussed the role of deep learning in analyzing genomic data, deep learning approaches can also substantially improve our ability to obtain the genomic data itself. We will discuss two specific challenges: calling SNPs (single nucleotide polymorphisms) 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, among others. In the clinical realm, sequencing of rare tumor clones and other genetic diseases will require accurate calling of SNP and indels.

Current methods achieve relatively high (>99%) precision at 90% recall for SNPs and indel calls from Illumina short-read data [251], 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, more importantly, 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 and, therefore, unbiased noise models. One of these models, DeepVariant, leverages Inception, a neural network trained for image classification by Google Brain, by encoding reads around a candidate SNP as a 221x100 bitmap image, where each column is a nucleotide and each row is a read from the sample library [251]. The top 5 rows represent the reference, and the bottom 95 rows represent randomly sampled reads that overlap the candidate variant. Each RGBA (red/green/blue/alpha) image pixel encodes the base (A, C, T, G) as a different R value, quality score as a G value, strand as a B value, and variation from the reference as the alpha value. The neural network outputs genotype probabilities for each candidate variant. They were able to achieve better performance than GATK, a leading genotype caller, even when GATK was given information about population variation for each candidate variant. Another method, still in its infancy, hand-developed 642 features for each candidate variant and fed these vectors into a fully connected deep neural network [252]. Unfortunately, this feature set required at least 15 iterations of software development to fine-tune, which will likely not be generalizable.

Going forward, we believe that 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 whether 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 [253], 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[251]. However, a model built on Illumina data, for instance, may not be applicable to PacBio long-read data or MinION nanopore data, which have vastly different specificity and sensitivity profiles and signal-to-noise characteristics. Recently, Boza et al. used bidirectional recurrent neural networks to infer the *E. coli* sequence from MinION nanopore electric current data with 2% higher per-base accuracy than the proprietary hidden Markov model-based algorithm Metrichor (86% to 88%) [249]. 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 [254]. 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 [254]. We anticipate that these steps will be replaced by deep learning, which will infer features simply by its ability to optimize a complex model against data.

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

Given the ever-present need to make better interventions faster at the point of care -- incorporating the complex calculus of a patients symptoms, diagnostics and life history -- there is a long history of attempts to apply deep learning to patient treatment. Success in this area would also be very useful for directions like personalized healthcare or precision medicine [255,256]. Earlier, we have written of the possibilities for patient categorization. Here, we examine the potential for better treatment, which broadly, may divided into methods for improved choices of interventions for patients and those for development of new interventions.

## Categorizing patients for clinical decision making

As long ago as 1996, Tu [257] compared the effectiveness of artificial neural networks and logistic regression, questioning whether deep learning would replace traditional statistical methods for predicting medical outcomes such as myocardial infarction [258] or mortality [259]. He posited that while neural networks have several advantages in representational power, the difficulties in interpretation may limit clinical applications. Similarly, in 2006 Lisboa and Taktak [260] 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.

TODO: could really use some references for following paragraph

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 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 prescription habits, but technical and medical challenges remain for clinical adoption (similar to those for categorization). In particular, remaining challenges 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 moving from prediction to treatment recommendations is the necessity to establish a causal relationship for a recommendation. Causal inference is often framed in terms of counterfactual question [261]. Johansson et al. [262] 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. [263] first create a deep neural network to model clinical time series and then analyze the relationship of the hidden features to the output using a causal approach.

### Applications

#### Trajectory Prediction for Treatment

A common application for deep learning techniques in this domain is to leverage the temporal structure of healthcare records. As previously discussed, many studies [264–267] have used deep recurrent networks to categorize patients but most stop short of suggesting clinical decisions. Nemati et al. [268] used deep reinforcement learning to optimize a heparin dosing policy for intensive care patients. However, because the ideal dosing policy is unknown, the model's predictions must be evaluated on counter-factual data. This represents a common challenge when bridging the gap between research and clinical practice: because the ground-truth is unknown, researchers struggle to evaluate model predictions in the absence of interventional data, but clinical application is unlikely until the model has been shown to be effective. The impressive applications of deep reinforcement learning to other domains [216] have relied on knowledge of the underlying processes (e.g. the rules of the game). Some models have been developed for targeted medical problems [269], but a generalized engine is beyond current capabilities. Further development of the rules underlying biological processes could unleash deep learning methods that are currently hampered by the difficulties of counter-factual inference.

#### Efficient Clinical trials

A clinical task to deep learning which has been more successfully applied is the assignment of patients to clinical trials. Ithapu et al [270] 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 [271] 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 then 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.

### Challenges

#### Actionable Interpretability

A common challenge in many applied deep learning problems is the consideration of deep learning models as uninterpretable "black boxes". Without human- intelligible reasoning for the model's predictions, it is difficult to trust the model. This presents a major challenge for the risk-averse task of clinical decision making. 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. Further work in interpreting predictions and understanding the knowledge learned by deep neural networks seem necessary for transformative impact in clinical practice. Interpretability in deep learning is reviewed more extensively in the Discussion.

#### Integrating Deep Learning with Clinical Practice

As deep learning models are difficult to interpret, many current models have been designed to replace aspects of clinical practice rather than to assist trained clinicians. This makes it difficult to integrate deep learning with clinical decision making. In addition, the challenges that physicians face are largely similar to those faced by machine learning models. For a given patient, the number of possible diseases is very large, with a long tail of rare diseases. Furthermore, patients are highly heterogeneous and may present with very different signs and symptoms for the same disease. Physicians are experienced in treating patients with common diseases, but rare diseases are extremely challenging. Unfortunately, machine learning methods also struggle for rare diseases. Directly applying current deep learning models to diagnose patients with rare diseases would require prohibitively large datasets to provide sufficient training data to capture the rare instances. Focused effort in reducing the data requirements of deep learning by integrating pre-existing knowledge or compiling large datasets of patient records may unlock the power of deep learning for clinical practice.

## 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 [272,273]. A decade ago, the concept of the Connectivity Map [274] had a sizeable impact on the field: reverse matching disease gene expression signatures with a large set of reference compound profiles allowed 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 [275–277]. Using supervised learning and collaborative filtering to tackle this type of problems is proving successful in different scenarios, especially when coupling disease or compound omic data with topological information from protein-protein or protein-compound interaction networks [278–280].

For example, Menden et al. [281] used a shallow neural network to predict sensitivity of cancer cell lines to drug treatment using both cell line and drug features, opening the door to precision medicine and drug repositioning opportunities in cancer. More recently, Aliper et al [35] used gene- and pathway-level drug perturbation transcriptional profiles from the Library of Network-Based Cellular Signatures (LINCS) [282] to train a fully connected deep neural network able to predict drug therapeutic uses and indications. By using confusion matrices and leveraging misclassification, the authors formulate 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 [283,284]. Wang et al. [285] devised a pairwise input neural network with two hidden layers that takes two inputs, a drug and a target binding site, and predicts whether they interact. Wang et al [36] trained individual RBMs for each target in a drug-target interaction network and used these models to predict novel interactions pointing to new indications for existing drugs. Wen et al. [37] extended this concept to deep learning by creating a DBN of stacked RBMs called DeepDTIs, which is able to predict interactions on the basis of chemical structure and protein sequence features.

Drug repositioning appears to be an obvious candidate for the development of deep learning applications both because of the large amount of high-dimensional data available and the complexity of the question being asked. However, what is perhaps the most promising piece of work in this space [35] is more 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 [286] do not use predictive modelling. While this might change in the future, we believe that a mature and production-ready framework where deep learning is directly applied to the problem of drug repositioning is currently missing.

## Ligand-Based Prediction of Bioactivity

In the biomedical domain, high-throughput chemical screening aims to improve therapeutic options over a long term horizon [20]. The objective is to discover which small molecules (also referred to as chemical compounds or ligands)

effectively and specifically affect the activity of a target, such as a kinase, protein-protein interaction, or broader cellular phenotype. This screening process can serve as one of the first steps in the long drug discovery pipeline, where novel chemicals are pursued for their ability to inhibit or enhance disease-relevant biological mechanisms [287]. Initial hits are confirmed to eliminate false positives and proceed to the lead generation stage [288], 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 [287].

The appeal of machine learning in this domain is the need to improve the efficiency of the initial high-throughput screens such that sufficient candidate active compounds can be identified without exhaustively screening libraries of hundreds of thousands or millions of chemicals. Predicting chemical activity computationally is known as virtual screening. This task has been treated variously as a classification, regression, or ranking problem. In reality, it does not fit neatly into any of those categories. An ideal algorithm will rank a sufficient number of active compounds before the inactives, but the rankings of actives relative to other actives and inactives are less important [289]. Computational modeling also has the potential to predict ADMET traits for lead generation [290] and how drugs are metabolized [291].

Here we primarily focus on ligand-based approaches that train on chemicals' features without requiring prior knowledge of the target. Chemical features may be represented as a list of molecular descriptors such as molecular weight, atom counts, charge representations, summaries of atom-atom relationships in the molecular graph, and more sophisticated derived properties [292]. Alternatively, chemicals can be characterized with the fingerprint bit vectors, textual strings, or novel learned representations described below. Neural networks have a long history in this domain [21], and the 2012 Merck Molecular Activity Challenge on Kaggle generated substantial excitement about the potential for high-parameter deep learning approaches. The winning submission was an ensemble that included a multitask multilayer perceptron network [293], and the sponsors noted drastic improvements over a random forest (RF) baseline, remarking "we have seldom seen any method in the past 10 years that could consistently outperform RF by such a margin" [294]. Subsequent work (reviewed in more detail by Goh et al. [19]) explored the effects of jointly modeling far more targets than the Merck challenge [295,296], with [296] showing that the benefits of multi-task networks had not yet saturated even with 259 targets. Although a deep learning approach, DeepTox [297], was also the overall winner of another competition, the Toxicology in the 21st Century (Tox21) Data Challenge, it did not dominate alternative methods as thoroughly as in other domains. DeepTox was the top performer on 9 of 15 targets and highly competitive with the top performer on the others. However, for many targets there was little separation between the top two or three methods.

The nuanced Tox21 performance may be more reflective of the practical challenges encountered in ligand-based chemical screening than the extreme enthusiasm generated by the Merck competition. A study of 22 ADMET tasks demonstrated that there are limitations to multi-task transfer learning that are in part a consequence of the degree to which tasks are related [290]. 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, 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 [298]. 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 were negligible, especially in the classification setting. Overfitting is still a problem in larger chemical screening datasets with tens or hundreds of thousands of compounds because the number of active compounds can be very small, on the order of 0.1% of all tested chemicals for a typical target [299]. This is consistent with the strong performance of low-parameter neural networks that emphasize compound-compound similarity, such as influence-relevance voter, [289,300] instead of predicting compound activity directly from chemical features.

Much of the recent excitement in this domain has come from what could be considered a creative experimentation phase, in which deep learning has offered novel possibilities for feature representation and modeling of chemical compounds. A molecular graph, where atoms are nodes and bonds are edges, is a natural way to represent a chemical structure. Traditional machine learning approaches relied on preprocessing the graph into a feature vector, such as a fixed-width bit vector fingerprint [301]. The same fingerprints have been used by some drug-target interaction methods discussed above [37]. An overly simplistic but approximately correct view of chemical fingerprints is that each bit represents the presence of absence of a particular chemical substructure in the molecular graph. Modern neural networks can operate directly on the molecular graph as input. Duvenaud et al. [302] 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 in theory learn chemical feature representations that are optimized for a particular prediction task. Lusci et al. [303] adapted recursive neural networks for directed acyclic graphs for undirected molecular graphs by creating an ensemble of directed graphs in which one atom is selected as the root node. A single feature vector is obtained by summing over all feature vectors for all directed graphs in the ensemble. Graph convolutions on undirected molecular graphs have eliminated the need to enumerate artificial directed graphs, learning feature vectors for atoms that are a function of the properties of neighboring atoms and local regions on the molecular graph [304,305].

Advances in chemical representation learning have also enabled new strategies for learning chemical-chemical similarity functions. Altae-Tran et al. developed a one-shot learning network [305] 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 multitask 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 where the embedding is learned from a subset of prediction tasks (e.g. activity assays for individual proteins) and only one to ten labeled examples are provided as training data on a new task. On Tox21 targets, even when trained with *one* task-specific active compound and *one* inactive compound, the model is able to generalize reasonably well because it has learned an informative embedding function from the related tasks. Random forests, which cannot take advantage of the related training tasks, trained in the same setting are only slightly better than a random classifier. Despite the success on Tox21, performance on MUV datasets, which contains assays designed to be challenging for chemical informatics algorithms, is considerably worse. The authors also demonstrate the limitations of transfer learning as embeddings learned from the Tox21 assays have little utility for a drug adverse reaction dataset.

These novel, learned chemical feature representations may prove to be essential for accurately predicting why some compounds with similar structures yield similar target effects and others produce drastically different results. Currently, these methods are enticing but do not necessarily outperform classic approaches by a large margin. The neural fingerprints [302] 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 [304] performed comparably to a multitask network using standard fingerprints and slightly better than the neural fingerprints [302] on the drug efficacy task but were slightly worse than the influence-relevance voter method on an HIV dataset. [289]. Broader recent benchmarking has shown that relative merits of these methods depends on the dataset and cross validation strategy [306], though we caution against over-interpreting AUC ROC-based results, a popular metric in this domain despite the active/inactive class imbalance (see Discussion).

We remain optimistic for the potential of deep learning and specifically representation learning in this domain and propose that rigorous benchmarking on broad and diverse prediction tasks will be as important as novel neural network architectures to advance the state of the art and convincingly demonstrate superiority over traditional cheminformatics techniques. Fortunately, there has recently been much progress in this direction. The DeepChem software [305,307] and MoleculeNet benchmarking suite [306] 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, and 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. Multitask neural networks have shown the advantages of jointly modeling many targets [295,296]. Other studies have shown the limitations of transfer learning when the prediction tasks are insufficiently related [290,305]. 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 [308]. Future work will establish whether it is better to train on massive collections of diverse compounds, drug-like small molecules, or specialized subsets.

Two emerging areas that we anticipate will be increasingly important in deep learning for drug discovery are *de novo* drug design and protein structure-based models. Whereas the goal of virtual screening is to prediction the biochemical activity of hundreds of thousands to millions of chemicals, *de novo* drug design explores the much larger space of at least $10^{60}$ organic molecules with drug-like properties that could be chemically synthesized[299]. Generative neural networks for drug design typically represent chemicals with the simplified molecular-input line-entry system (SMILES), a standard way string-based representation with characters that represent atoms, bonds, and rings [309]. Gómez-Bombarelli et al. designed a SMILES-to-SMILES autoencoder to learn a continuous latent feature space for chemicals [308]. In this learned continuous space it was possible to train some types of supervised learning algorithms and interpolate between continuous representations of chemicals in a manner that is not possible with discrete (e.g. bit vector or string) features. The strategy of constructing simple, continuous features before applying supervised learning techniques is reminiscent of autoencoders trained on high-dimensional EHR data [58]. 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. More recent approaches train RNNs on compounds from ChEMBL [310] to first obtain a generic generative model for drug-like compounds [309,311]. These generative models successfully learn the grammar of compound representations, with 94% [311] or nearly 98% [309] 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 [309] or adopting reinforcement learning strategies [311].

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 [312]. Recently, deep learning approaches have been developed to model protein structure, which is expected to improve upon the simpler drug-target interaction algorithms described above that represent proteins with feature vectors derived from amino acid sequences [37,285]. 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 Convolutional Neural Network [313] takes 3D crystal structures from PDBBind[314] as input, ensuring it uses the correct ligand-target complex. Alternatively, AtomNet [34] samples multiple ligand poses within the target binding site, and DeepVS[315] and Ragoza et al. [316] 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 [315]. A 3D grid can be used to represent a protein-compound complex [34,316]. Each entry in the grid tracks the types of protein and ligand atoms in that region of the 3D space or descriptors derived from those atoms. Both DeepVS [315] and atomic convolutions [313] offer greater flexibility in their convolutions by eschewing the 3D grid. Instead, they each implement techniques for executing convolutions over atoms' neighboring atoms in the 3D space. Gomes et al. demonstrate that currently random forest on a 1D feature vector that describes the 3D ligand-target structure generally outperforms neural networks on the same feature vector, 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.

# Discussion

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

## Evaluation

*What are the challenges in evaluating deep learning models that are specific to this domain? This can include a discussion of ROC versus precision-recall curves for the imbalanced classes often encountered in biomedical datasets. It could also mention alternative metrics that are used in specific sub-areas such as enrichment factors in virtual screening. A lack of true gold standard data for some problems complicates both training and evaluation. Confidence-weighted labels are valuable when available.*

*Is progress in some biomedical areas slowed when new predictions (e.g. from generative models) cannot be assessed by any human expert and require experimental testing? For example, contrast a painting or song generated by a GAN versus a novel chemical compound. Related is the idea that on some tasks (e.g. the recent wave of deep learning versus MD image classification papers) it is easy to tell when deep learning has produced a breakthrough because human-level performance is an impressive baseline. In many tasks we reviewed, human-level performance is irrelevant.*

TODO: draft coming May 4 or 5

## Interpretation

As deep learning models achieve state-of-the-art performance in a variety of domains, there is a growing need to make the models more interpretable. Interpretability matters for two main reasons. First, a model that achieves breakthrough performance may have identified patterns in the data that practitioners in the field would like to understand. However, this would not be possible if the model is a black box. Second, interpretability is important for trust. If a model is making medical diagnoses, it is important to ensure the model is making decisions for reliable reasons and is not focusing on an artifact of the data. A motivating example of this can be found in Caruana et al. [317], where a model trained to predict the likelihood of death from pneumonia assigned lower risk to patients with asthma, but only because such patients were treated as higher priority by the hospital. In the context of deep learning, understanding the basis of a model's output is particularly important as deep learning models are unusually susceptible to adversarial examples [318] and can output confidence scores over 99.99% for samples that resemble pure noise.

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

### Assigning example-specific importance scores

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

#### Perturbation-based approaches

These approaches make perturbations to individual inputs and observes the impact on the output of the network. Zhou and Troyanskaya [162] scored genomic sequences by introducing virtual mutations at each position and quantifying the change in the output. Ribeiro et al. [319] introduced LIME which constructs a linear model to locally approximate the

output of the network on perturbed versions of the input and assigned importance scores accordingly. For analyzing images, Zeiler and Fergus [320] applied constant-value masks to different input patches and studied the changes in the activations of later layers. As an alternative to using masks, which can produce misleading results, Zintgraf et al. [321] proposed a novel strategy based on marginalizing over plausible values of an input patch to more accurately estimate its contribution. TODO: do we need an example of what is misleading?

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. [322], such methods may also underestimate the impact of features that have saturated their contribution to the output, as can happen when multiple redundant features are present.

To reduce the computational overhead of perturbation-based approaches, Fong and Vedaldi[323] 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. When tested on image data, their method took about 300 iterations to converge, compared to the ~5000 iterations used by LIME. One drawback of this approach is that the use of gradient descent requires the perturbation to have a differentiable form.

TODO: tag:Alipanahi2015_predicting (DeepBind) was in the original draft. Does that still fit somewhere? It is okay to leave out, it is already cited in th

#### Backpropagation-based approaches

A second strategy for addressing the computational inefficiency of perturbation-based approaches is to propagate an important signal from a target output neuron backwards through the layers to the input layer in a single backpropagation-like pass. A classic example of this calculating the gradients of the output w.r.t. the input [324] to compute a "saliency map". Bach et al. [325] proposed a strategy called Layerwise Relevance Propagation, which was shown to be equivalent to the elementwise product of the gradient and input [166,326]. Several variants of gradients exist which differ in their handling of the ReLU nonlinearity. While gradients zero-out the importance signal at ReLUs if the input to the ReLU is negative, deconvolutional networks [320] zero-out the importance signal if the signal itself is negative. Guided backpropagation [327] combines the two strategies to zero-out the importance signal if either the input to ReLU is negative or the importance signal is negative, in effect discarding negative gradients. However, Mahendran and Vedaldi [328] showed that while guided backpropagation excelled at identifying salient features in the input image, these features showed little class-specificity, producing very similar saliency maps regardless of the class under consideration. Selvaraju et al. [329] attempted to alleviate this by combining gradients and guided backpropagation in Guided Grad-CAM (Class Activation Mapping). Feature maps in the last convolutional layer were associated with classes using gradients, and the weighted activation of these feature maps was multiplied with the result of guided backpropagation to introduce more class specificity. Note that these approaches still would not highlight features that have saturated their contribution to the output, as the gradients with respect to such features would be zero at the input.

To address the saturation failure mode, strategies have been developed to consider how the output changes between some reference input and the actual input, where the reference input represents a "null" input that it is informative to measure differences against. Sundararajan et al. [330] integrated the gradients as the input was linearly increased from the reference to its actual value (in their examples, which were on image-like data, they used a reference of all zeros). While the numerical integration adds computational overhead, the method is still more efficient on average than perturbation approaches. Further, by relying only on the gradients, the method is a fully black-box approach that is guaranteed to give the same answer for functionally equivalent networks. Shrikumar et al. [322] developed DeepLIFT, a strategy that used the difference between a neuron's activation on the reference input compared to its activation on the actual input to improve the backpropagation of importance scores. DeepLIFT is a white box method that requires knowledge of the network architecture, but it is more computationally efficient than integrated gradients. Lundberg and Lee [331] noted that several importance scoring methods, including DeepLIFT, integrated gradients and LIME, could all be considered approximations to the Shapely values, which have a long history in game theory for assigning contributions to players in cooperative games. TODO: Add reference for Shapely values and/or explain that term DeepLIFT introduced a modification which treated positive and negative contributions separately to address some failure cases of integrated gradients; the modification can be understood as an improved approximation of the Shapely values.

### Matching or exaggerating the hidden representation

Another approach to understanding the network's predictions is to find artificial inputs that produce similar hidden representations to a chosen example. This can elucidate the features that the network uses for prediction and drop the features that the network is insensitive to. In the context of natural images, Mahendran and Vedaldi [332] introduced the "inversion" visualization which uses gradient descent and backpropagation to reconstruct the input from its hidden representation. The method required placing a prior on the input to favor results which resemble natural images. For genomic sequence, Finnegan and Song [333] 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 net searches for [334]. This is done by maximizing the response of neurons that are active in the network, subject to some regularizing constraints. Mordvintsev et al. [335] leveraged caricaturization to generate aesthetically pleasing images using neural

networks.

**Activation maximization**

Activation maximization can reveal patterns detected by an individual neuron in the network by generating images which maximally activate that neuron, subject to some regularizing constraints. This technique was first introduced in Ehran et al. [336] and applied in Simonyan et al.[324], Mordvintsev et al.[335], Yosinksi et al.[337] and Mahendran and Vedaldi [334]. Lanchantin et al.[164] applied activation maximization to genomic sequence. One drawback of this approach is that neural networks often learn highly distributed representations where several neurons cooperatively describe a pattern of interest - thus, visualizing patterns learned by individual neurons may not always be informative.

**RNN-specific approaches**

Several interpretation methods are specifically tailored to recurrent neural network architectures. A few key approaches are summarized below.

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 for generating a particular output [338,339]. Deming et al.[340] applied the attention mechanism to models trained on genomic sequence. Attention mechanisms provide insight into the model's decision-making process by revealing which portions of the input are used by different outputs. In the clinical domain, Choi et al. [341] leveraged attention mechanisms to highlight which aspects of a patient's medical history were most relevant for making diagnoses. Choi et al. [342] later extended this work to take into account the structure of disease ontologies and found that the concepts represented by the model were aligned with medical knowledge. Note that interpretation strategies that rely on an attention mechanism do not provide insight into the internal logic used by the attention layer to decide which inputs to attend to.

Visualizing the activation patterns of the hidden state of a recurrent neural network can also be instructive. Early work by Ghosh and Karamcheti [343] 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. [344] showed the existence of individual cells in LSTMs that kept track of quotes and brackets in character-level language models. To facilitate such analyses, LSTM vis [345] allows interactive exploration of the hidden state of LSTMs on different inputs.

Another strategy, adopted by Lanchatin et al.[164] looks at how the output of a recurrent neural network changes as longer and longer subsequences are supplied as input to the network, where the subsequences begin with just the first position and end with the entire sequence. In a binary classification task, this can identify those positions which are responsible for flipping the output of the network from negative to positive. If the RNN is bidirectional, the same process can be repeated on the reverse sequence. As noted by the authors, this approach was less effective at identifying motifs compared to the gradient-based backpropagation approach of Simonyan et al. [324] illustrating the need for more sophisticated strategies to assign importance scores in recurrent neural networks.

Murdoch and Szlam [346] showed that the output of an LSTM can be decomposed into a product of factors where each factor can be interpreted as the contribution at a particular timestep. The contribution scores were then used to identify key phrases from a model trained to do sentiment analysis and obtained superior results compared to scores derived via a gradient-based approach.

**Miscellaneous approaches**

Toward quantifying the uncertainty of predictions, there has been a renewed interest in confidence intervals for deep neural networks. Early work from Chryssolouris et al. [347] provided confidence intervals under the assumption of normally distributed error. A more recent technique known as test-time dropout [348] can also be used to obtain a probabilistic interpretation of a network's outputs.

It can often be informative to understand how the training data affects the learning of a model. Toward this end, Koh and Liang [349] 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 [350,351] 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. [352] 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[353] 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.

TODO: Are there any brief final thoughts you would like to add? This paper is part review, part perspective piece so we have the opportunity to specul

## Data limitations

A lack of large-scale, high-quality, correctly labeled training data has impacted deep learning in nearly all applications we have discussed, from healthcare to genomics to drug discovery. The challenges of training complex, high-parameter neural networks from few examples are obvious, but uncertainty in the labels of those examples can be just as problematic. For example, in genomics labeled data may be derived from an experimental assay with known and unknown technical artifacts, biases, and error profiles. It is possible to weight training examples or construct Bayesian models to account for uncertainty or non-independence in the data. To this end, Park et al. [354] estimated shared non-biological signal between datasets to correct for non-independence related to assay platform or other factors in a Bayesian integration of many datasets. However, such techniques are rarely placed front and center in any description of methods, and so may be easily overlooked.

For some types of data, especially images, it is straightforward to augment training datasets by splitting a single labeled example into multiple examples. For example, an image can easily be rotated, flipped, or translated and retain its label [52]. 3D MRI and 4D fMRI (with time as a dimension) data can be decomposed into sets of 2D images [355]. 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 random subsets of cells that are randomly sampled with replacement from the full dataset [213].

Simulated or semi-synthetic training data has also been employed in multiple biomedical domains.
TODO: simulated data: #5 #99 #293, maybe #117 and #197. There is a counter-example from drug discovery to include as well that is related to #55

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 [356]. 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 (i.e. PCA, per class histograms, or a Word2vec [78] generalization) can be learned directly from input data or take advantage of other datasets or domain knowledge. Additionally, in this task the features are the examples, an important advantage when it is typical to have 500 thousand or more SNPs and only a few thousand patients. Finally, this embedding is of a much lower dimension, allowing for a large reduction in the number of free parameters. In the example given, the number of free parameters from was reduced from 30 million to 50 thousand, a factor of 600.

## Hardware limitations and scaling

Efficiently scaling deep learning is challenging, and there is a high computational cost (e.g. time, memory, energy) associated with training neural networks and using them for classification. This is one of the reasons why neural networks have only recently found widespread use [357].

Many have sought to curb these costs, with methods ranging from the very applied (e.g. reduced numerical precision [358–361]) to the exotic and theoretic (e.g. training small networks to mimic large networks and ensembles [317,362]). The largest gains in efficiency have come from computation with graphics processing units (GPUs) [357,363–367], 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 [364,365,368,369]. However, GPUs also have a limited quantity of memory, making it difficult to implement networks of useful size and complexity on a single GPU or machine [73,363]. This restriction has sometimes forced computational biologists to use workarounds or limit the size of an analysis. For example, Chen et al. [145] aimed to infer the expression level of all genes with a single neural network, but due to memory restrictions they randomly partitioned genes into two halves and analyzed each separately. In other cases, researchers limited the size of their neural network [27,308]. Some have also chosen to use slower CPU implementations rather than sacrifice network size or performance [370].

While steady improvements in GPU hardware may alleviate this issue, it is unclear whether advances can occur quickly enough to keep up with the growing amount of available biological data or increasing network sizes. Much has been done to minimize the memory requirements of neural networks [317,358–361,371,372], but there is also growing interest in specialized hardware, such as field-programmable gate arrays (FPGAs) [367,373] and application-specific integrated circuits (ASICs). Specialized hardware promises improvements in deep learning at reduced time, energy, and memory [367]. Obviously, there is as yet less software available for such highly specialized hardware [373], and it could be a difficult investment for those not solely interested in deep learning. However, it is likely that such options will find increased support as they become a more popular platform for deep learning and general computation.

Distributed computing is a general solution to intense computational requirements, and has enabled many large-scale deep learning efforts. Early approaches to distributed computation [374,375] were not suitable for deep learning [376],

but much progress has been made. There now exist a number of algorithms [360,376,377], tools [378–380], and high-level libraries [381,382] for deep learning in a distributed environment, and it is possible to train very complex networks with limited infrastructure [383]. Besides handling very large networks, distributed or parallelized approaches offer other advantages, such as improved ensembling [384] or accelerated hyperparameter optimization [385,386].

Cloud computing, which has already seen wide adoption in genomics[387], could facilitate easier sharing of the large datasets common to biology [388,389], and may be key to scaling deep learning. Cloud computing affords researchers flexibility, and enables the use of specialized hardware (e.g., FPGAs, ASICs, GPUs) without major investment. As such, it could be easier to address the different challenges associated with the multitudinous layers and architectures available [390]. Though many are reluctant to store sensitive data (e.g. patient electronic health records) in the cloud, secure/regulation-compliant cloud services do exist [391].

## Data, code, and model sharing

A robust culture of data, code, and model sharing would do much to speed advances in this domain. The cultural barriers of data sharing in particular are perhaps best captured by the implications of using the term "research parasite" to describe scientists who use data from other researchers [392]. In short, a field that honors only discoveries and not the hard work of generating useful data will have difficulty encouraging scientists to share their hard-won data. Unfortunately, it's precisely those data that would help to power deep learning in the domain. Efforts are underway to recognize those who promote an ecosystem of rigorous sharing and analysis [393].

The sharing of high-quality, labeled datasets will be especially valuable. In addition, researchers who invest time to preprocess datasets to be suitable for deep learning can make the preprocessing code (e.g. Basset [173] and variationanalysis [252]) and cleaned data (e.g. MoleculeNet [306]) publicly available to catalyze further research. However, there are complex privacy and legal issues involved in sharing patient data that cannot be ignored. Furthermore, in some domains, some of the best training data has been generated privately, for example, 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 sufficient data for improved methods to be developed.

Code sharing and open source licensing is essential for continued progress in this domain. We strongly advocate following established best practices for sharing source code, archiving code in repositories that generate digital object identifiers, and open licensing [394] regardless of the minimal requirements, or lack thereof, set by journals, conferences, or preprint servers. In addition, it is important for authors to share not only code for their core models but also scripts and code used for data cleaning (see above) and hyperparameter optimization. These improve reproducibility and serve as documentation of the detailed decisions that impact model performance but may not be exhaustively captured in a manuscript's methods text.

Because many deep learning models are often built using one of several popular software frameworks, it is also possible to directly share trained predictive models. The availability of pre-trained models can accelerate research, with image classifiers as an apt example. A pre-trained neural network can be quickly fine-tuned on new data and used in transfer learning, as discussed below. Taking this idea to the extreme, genomic data has been artificially encoded as images in order to benefit from pre-trained image classifiers [251]. "Model zoos" -- collections of pre-trained models -- are not yet common in biomedical domains but have started to appear in genomics applications [210,395]. Sharing models for patient data requires great care because deep learning models can be attacked to identify examples used in training. We discuss this issue as well as recent techniques to mitigate these concerns in the patient categorization section.

DeepChem [305–307] and DragoNN [395] exemplify the benefits of sharing pre-trained models and code under an open source license. DeepChem, which targets drug discovery and quantum chemistry, has actively encouraged and received community contributions of learning algorithms and benchmarking datasets. As a consequence, it now supports of a large suite of machine learning approaches, both deep learning and competing strategies that can be run on diverse test cases. This realistic, continual evaluation will play a critical role in assessing which techniques are most promising for chemical screening and drug discovery. Like formal, organized challenges such as the ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge[396], TODO: placeholder URL until the preprint is available DeepChem provides a forum for the fair, critical evaluations that are not always conducted in individual methodological papers, which can be biased toward favoring a new proposed algorithm. Likewise DragoNN (Deep RegulAtory GenOmic Neural Networks), offers not only code and a model zoo but also a detailed tutorial and partner package for simulating training data. These resources, especially the ability to simulate datasets that are sufficiently complex to demonstrate the challenges of training neural networks but small enough to train quickly on a CPU, are important for (human) training and attracting machine learning researchers to problems in genomics and healthcare. We have even included DragoNN and hands-on model training into the curriculum of a graduate student course.

## Multimodal, multi-task, and transfer learning

As discussed above, the fact that biomedical datasets often contain a limited number of instances or labels can be a

cause of poor performance of machine learning algorithms. When trained on such datasets, deep learning 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 in hand. The first step allows a model to take advantage of a larger amount of data and/or labels to extract better feature representations. Transferring learnt 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 [397].

In image analysis, previous examples of deep transfer learning applications proved large scale natural image sets[63] to be useful for pre-training models that can then serve as generic feature extractors applied to various types of biological images [14,203,398,399]. More recently, deep learning models trained to predict protein sub-cellular localization successfully performed predictions for proteins not originally present in the training set [400]. Moreover, in this type of task, learnt features performed reasonably well even when applied to images obtained using different fluorescent labels, imaging techniques, and different cell types [401]. However, there are no established theoretical guarantees for feature transferability between distant domains such as natural images and various modalities of biological imaging. Because learnt 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 multiple networks' top layers on the specific dataset in order to re-learn domain-specific high level concepts (e.g. fine-tuning for radiology image classification [61]). Fine-tuning on specific biological datasets enables more focused predictions. The Basset package [173] for prediction of functional activities from DNA sequences 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. [176] 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 [309,311].

Related to transfer learning, multimodal learning assumes simultaneous learning from various types of inputs, such as images and text. It allows capture of features that describe common concepts across input modalities. Generative graphical models like restricted Boltzmann machines (RBM) and their stacked versions, deep Boltzmann machines (DBM), and deep belief networks (DBN), demonstrate successful extraction of more informative features for one modality (images or video) when jointly learnt with other modalities (audio or text) [402]. Deep graphical models such as DBNs are considered to be well suited for multimodal learning tasks since they learn a joint probability distribution from inputs. They can be pre-trained in an unsupervised fashion on large unlabeled data and then fine-tuned on a smaller number of labeled examples. When labels are available, convolutional neural networks (CNN) are ubiquitously used since they can be trained end-to-end with backpropagation and demonstrate state-of-the-art performance in many discriminative tasks [14].

Jha et al. [153] showed that an integrated training approach delivers better performance compared to 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. Results showed that the integrative deep model generalized well for combined data, offering large performance improvement for alternative splicing event estimation. Chaudhary et al. [403] 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 omics as different modalities outperformed both traditional methods (PCA) 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. [404] used deep belief networks to learn phosphorylation states of a common set of signaling proteins in primary cultured bronchial cells collected from rats and humans treated with distinct stimuli. By interpreting species as different modalities representing similar high-level concepts, they showed that DBNs were able to capture cross-species representation of signaling mechanisms in response to a common stimuli. Another application used DBNs for joint unsupervised feature learning from cancer datasets containing gene expression, DNA methylation, and miRNA expression data [147]. This approach allowed for the capture of intrinsic relationships in different modalities and for better clustering performance over conventional k-means based methods.

Multimodal learning with CNNs is usually implemented as a collection of individual networks in which each learns representations from single data type. These individual representations are further concatenated before or within fully-connected layers. FIDDLE [405] is an example of multimodal CNNs that represents an ensemble of individual networks that take as inputs a number of genomic datasets, including NET-seq, MNase-seq, ChIP-seq, RNA-seq, and raw DNA sequence to predict Transcription Start Site-seq (TSS-seq) outputs. The combined model radically improves performance over separately trained datatype-specific networks, suggesting that it learns the synergistic relationship between datasets.

Multi-task learning (MTL) is an approach related to transfer learning. In an MTL framework a model co-learns a number of tasks simultaneously such that features are shared across them. DeepSEA framework [162] implemented a multi-task joint learning of diverse chromatin factors sharing predictive features from raw DNA sequence. This allowed, for example, a sequence feature that is effective in recognizing binding of a specific TF to be simultaneously used by another predictor for a physically interacting TF. Similarly, TFImpute [154], a CNN-RNN architecture learned information shared across transcription factors and cell lines to predict cell-specific TF binding for TF-cell line combinations based on only a small fraction (4%) of the combinations using available ChIP-seq data. On multiple test sets that excluded specific TFs and cell lines, TFImpute showed comparable or superior performance compared to the state-of-the-art. Yoon et al.[77], previously discussed in the section on Electronic Health Records, demonstrated that predicting the primary cancer site from the cancer pathology reports together with its laterality substantially improved the performance for the latter task, suggesting that MTL can effectively leverage the commonality between two tasks using a shared representation. A number of studies previously mentioned in the section on developing new treatments employed multi-task learning approach to predict a large number of compound and target interactions for drug discovery [293,296] and drug toxicity prediction [297,406]. Kearnes et al. [290] did a systematic comparison of single-task and multi-task deep models on a set of industrial ADMET datasets. They confirmed that multi-task learning can improve performance over single-task models. They further showed that smaller datasets tend to benefit more from multitask learning than larger datasets. Results emphasized that multi-task effects are highly dataset-dependent, suggesting the use of dataset-specific models to maximize overall performance.

MTL approach is complementary to multimodal and transfer learning. All three techniques can be used together in the same model. For example, Zhang et al. [398] combined deep model-based transfer and multi-task learning for cross-domain image annotation. One could imagine extending that approach to multimodal inputs as well. Common characteristic of these methods lies in better generalization of extracted features by leveraging relationships between information in provided in inputs and task objectives, represented at various hierarchical levels of abstraction in a deep learning model structure.

Despite demonstrated improvements, transfer learning approaches also pose a number of challenges. As mentioned above, there are no theoretically sound principles for pre-training and fine-tuning. Most best practice recommendations are heuristic and have to take into account additional hyper-parameters that depend on specific deep architectures, sizes of pre-training and target datasets, and similarity of domains. However, similarity of datasets and domains in transfer learning and relatedness of tasks in MTL are difficult to access. Most current studies address these limitations by empirical evaluation of the model using established best practices or heuristics and cross-validation. Unfortunately, negative results are typically left out and not presented in the final study publications. Results by Rajkomar et al. [61] showed that a deep CNN trained on natural images can boost radiology image classification performance. However, due to differences in imaging domains, target task required either re-training the initial model from scratch with special pre-processing or fine-tuning of the whole network on radiographs with heavy data augmentation to avoid overfitting. Exclusively fine-tuning top layers led to much lower validation accuracy (81.4 vs 99.5). Fine-tuning procedure for the discussed Basset model pre-trained on data from different cell types required no more than one training pass. Otherwise, the model started overfitting new data [173]. DeepChem successfully improved results for low-data drug discovery with one-shot learning for related tasks. However, it demonstrated clear limitations to cross-task generalization across unrelated tasks in one-shot models, specifically nuclear receptor assays and patient adverse reactions [305].

Overall, multimodal, multi-task and transfer learning strategies demonstrate high potential for many biomedical applications that are otherwise limited by data volume and presence of labels. However, these methods not only inherit most methodological issues from natural image, text, and audio domains, but also pose new challenges, specific to biological data. Making negative results, source code, and pre-trained models publicly available helps to accelerate progress in this direction. However, there are privacy considerations for models trained on sensitive data, such as patient-related information (see Data sharing section). Thus, there is a compelling need for the development of privacy-preserving transfer learning algorithms, such as Private Aggregation of Teacher Ensembles (PATE-G) [123], that can leverage publicly available data. We suggest that these types of models deserve deeper investigation to establish sound theoretical guarantees and best practices and determine limits for the transferability of features between various closely related and distant learning tasks.

## Conclusions

Deep learning-based methods now match or surpass the previous state of the art in a diverse array of tasks in patient and disease categorization, fundamental biological study, genomics, and treatment development. Returning to our central question: given this rapid progress, has deep learning transformed the study of human disease? Though the answer is highly dependent on the specific domain and problem being addressed, we conclude that deep learning has not *yet* realized its transformative potential or induced a strategic inflection point. Despite its dominance over competing machine learning approaches in many of the areas reviewed here and quantitative improvements in predictive performance, deep learning has not yet definitively "solved" those 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% [407] and finally approaching

or exceeding human performance in the past year [408,409] TODO: need better source for this error trajectory. The phenomenal improvements on benchmark datasets are undeniable, but halving the error rate on these benchmarks did not fundamentally transform the domain. Widespread adoption of conversational speech technologies will require not only improvements over baseline methods but truly solving the problem, in this case exceeding human-level performance, as well as convincing users to embrace the technology [407]. We see parallels to the healthcare domain, where achieving the full potential of deep learning will require outstanding predictive performance as well as acceptance and adoption by biologists and clinicians.

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 [48], diabetic macular edema [48], tuberculosis [62], and skin lesion [4] classifiers are highly accurate and comparable to clinician performance in the latter case.

In other domains, perfect accuracy will not be required because deep learning will be used primarily to 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 [62], though the risk of false negatives must be addressed. In protein structure prediction, errors in individual residue-residue contacts can be tolerated when using the contacts jointly for 3D structure modeling. Improved contact map predictions [27] have led to notable improvements in fold and 3D structure prediction for some of the most challenging proteins, such as membrane proteins [198].

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

We are witnessing deep learning models achieving human-level performance across a number of biomedical domains, and yet do not believe that biologists and clinicians will be out of a job anytime soon. While deep learning methods will soon be (or already are) better than scientists at specific tasks, they may not fully grasp the bigger picture. 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 [410,411], a reminder that these algorithms do not understand the semantics of the objects presented. Despite progress in addressing some of these limitations [412,413], until true and reliable artificial intelligence becomes standard in the laboratory and in the clinic, the human factor still has a critical role to play. In some cases, cooperation between human experts and deep learning algorithms can achieve better performance than either individually [71]. Especially for sample and patient classification tasks, we expect deep learning methods to complement and assist biomedical researchers rather than compete with or even replace them.

Even if deep learning in biology and healthcare is not yet transformative today, we are extremely optimistic about its future. Given how rapidly deep learning is evolving, its full potential in biomedicine has not been explored. We have highlighted numerous challenges beyond improving training and predictive accuracy, such as preserving patient privacy and interpreting models. Ongoing research has begun to address these problems and shown they are not insurmountable. Deep learning offers the flexibility to model data in its most natural form, for example, longer DNA sequences instead of k-mers for transcription factor binding prediction and molecular graphs instead of pre-computed bit vectors for drug discovery. These flexible input feature representations have spurred creative modeling approaches that would be infeasible with other machine learning techniques. Unsupervised methods are currently less developed than their supervised counterparts, but they may have the most potential because of how expensive and time-consuming it is to label large amounts of biomedical data. When deep learning algorithms can summarize very large collections of input data into interpretable models that spur scientists to ask questions that they didn't know how to ask, it will be clear that deep learning has transformed biology and medicine.

## Author contributions

TODO: not sure if it should go here, but somewhere we should talk about how we wrote this thing, since it is still somewhat unconventional to have a
We recognized that writing a review on a rapidly developing area in a manner that allowed us to provide a forward-looking perspective on diverse approaches and biological problems would require expertise from across computational biology and medicine. We created an open repository on the GitHub version control system and engaged with numerous authors from papers within and outside of the area. Paper review was conducted in the open by # individuals, and the manuscript was drafted in a series of commits from # authors. Individuals who met the ICJME standards of authorship are included as authors. 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 one or more of these ways, but who did participate, are acknowledged at the end of the manuscript.

TODO: update after finalizing discussion in #369

1. Stephens ZD *et al.* 2015 Big Data: Astronomical or Genomical? *PLOS Biology* **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. *Nature Communications* **5**. (doi:10.1038/ncomms5308)

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

5. Wu Y *et al.* 2016 Google's neural machine translation system: Bridging the gap between human and machine translation.

6. In press. See http://research.google.com/archive/unsupervised_icml2012.html.

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

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

9. Niu F, Recht B, Re C, Wright SJ. 2011 HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent.

10. In press. See http://www.deeplearningbook.org/.

11. In press. See http://www.intel.com/pressroom/archive/speeches/ag080998.htm.

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

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

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

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

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

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

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

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

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

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

22. Parker JS *et al.* 2009 Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes.*Journal of Clinical Oncology* **27**, 1160–1167. (doi:10.1200/jco.2008.18.1370)

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

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

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

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

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

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

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

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

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

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

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

34. Wallach I, Dzamba M, Heifets A. 2015 AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery.

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

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

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

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

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

40. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Laak JAWM van der, Ginneken B van, Sánchez CI. 2017 A survey on deep learning in medical image analysis.

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

42. 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.

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

44. Jafari MH, Nasr-Esfahani E, Karimi N, Soroushmehr SMR, Samavi S, Najarian K. 2016 Extraction of skin lesions from non-dermoscopic images using deep learning. (doi:10.1007/s11548-017-1567-8)

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

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

47. Leibig C, Allken V, Berens P, Wahl S. 2016 Leveraging uncertainty information from deep neural networks for disease detection. (doi:10.1101/084210)

48. 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)

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

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

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

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

53. Zhu W, Xie X. 2016 Adversarial Deep Structural Networks for Mammographic Mass Segmentation. (doi:10.1101/095786)

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

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

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

57. Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. 2017 Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical Physics* **44**, 1017–1027. (doi:10.1002/mp.12110)

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

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

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

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

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

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

64. Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, Kim L, Summers RM. 2016 Improving Computer-Aided Detection Using_newlineConvolutional Neural Networks and Random View Aggregation. *IEEE Transactions on Medical Imaging* **35**, 1170–1181. (doi:10.1109/tmi.2015.2482920)

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

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

67. Wang X, Lu L, Shin H-c, Kim L, Bagheri M, Nogues I, Yao J, Summers RM. 2017 Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition.

68. 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.

69. In press. See https://console.cloud.google.com/storage/browser/gcs-public-data--nih/radiology_2017/Chest_X-Ray_CVPR17.

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

71. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. 2016 Deep learning for identifying metastatic breast cancer.

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

73. In press. See https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

74. Ohno-Machado L. 2011 Realizing the full potential of electronic health records: the role of natural language processing. *Journal of the American Medical Informatics Association* **18**, 539–539. (doi:10.1136/amiajnl-2011-000501)

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

76. Chalapathy R, Borzeshi EZ, Piccardi M. 2016 Bidirectional lstm-crf for clinical concept extraction.

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

78. Mikolov T, Chen K, Corrado G, Dean J. 2013 Efficient estimation of word representations in vector space.

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

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

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

82. 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)

83. Beaulieu-Jones BK, Greene CS. 2016 Semi-Supervised Learning of the Electronic Health Record for Phenotype Stratification. (doi:10.1101/039800)

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

85. Razavian N, Marcus J, Sontag D. 2016 Multi-task prediction of disease onsets from longitudinal lab tests.

86. Ranganath R, Perotte A, Elhadad N, Blei D. 2016 Deep survival analysis.

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

88. Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, Kluger Y. 2016 Deep survival: A deep cox proportional hazards network.

89. Ranganath R, Tang L, Charlin L, Blei DM. 2014 Deep exponential families.

90. Hoffman M, Blei DM, Wang C, Paisley J. 2012 Stochastic variational inference.

91. Ranganath R, Tran D, Blei DM. 2015 Hierarchical variational models.

92. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. 2017 A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics* **97**, 120–127.

(doi:10.1016/j.ijmedinf.2016.09.014)

93. In press. See https://phekb.org/implementations.

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

95. Ratner A, Sa CD, Wu S, Selsam D, Ré C. 2016 Data programming: Creating large training sets, quickly.

96. In press. See http://ana.blogs.com/maestros/2006/11/data_is_the_new.html.

97. In press. See https://medium.com/twenty-one-hundred/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294.

98. In press. See http://hazyresearch.github.io/snorkel/blog/weak_supervision.html.

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

100. Bowman S. 2013 Impact of Electronic Health Record Systems on Information Integrity: Quality and Safety Implications. *Perspect Health Inf Manag* **10**, 1c.

101. Botsis T, Hartvigsen G, Chen F, Weng C. 2010 Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translat Bioinforma* **2010**, 1–5.

102. 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)

103. 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.

104. In press. See http://www.ncrr.nih.gov/publications/informatics/ehr.pdf.

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

106. Oemig F, Snelick R. 2016 *Healthcare Interoperability Standards Compliance Handbook* Springer International Publishing. (doi:10.1007/978-3-319-44839-8)

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

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

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

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

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

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

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

114. Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. 2016 Stealing machine learning models via prediction apis.

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

116. Shokri R, Stronati M, Song C, Shmatikov V. 2016 Membership inference attacks against machine learning models.

117. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. 2017 Generating multi-label discrete electronic health

records using generative adversarial networks.

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

119. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. 2016 Deep learning with differential privacy. (doi:10.1145/2976749.2978318)

120. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. 2016 Communication-efficient learning of deep networks from decentralized data.

121. In press. See http://proceedings.mlr.press/v54/mcmahan17a.html.

122. In press. See https://eprint.iacr.org/2017/281.pdf.

123. In press. See https://openreview.net/forum?id=HkwoSDPgg.

124. Goodman B, Flaxman S. 2016 European union regulations on algorithmic decision-making and a 'right to explanation'.

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

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

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

128. Price AL, Zaitlen NA, Reich D, Patterson N. 2010 New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463. (doi:10.1038/nrg2813)

129. Sebastiani P *et al.* 2011 Retraction. *Science* **333**, 404–404. (doi:10.1126/science.333.6041.404-a)

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

131. Lum K, Isaac W. 2016 To predict and serve? *Significance* **13**, 14–19. (doi:10.1111/j.1740-9713.2016.00960.x)

132. Hardt M, Price E, Srebro N. 2016 Equality of opportunity in supervised learning.

133. Joseph M, Kearns M, Morgenstern J, Neel S, Roth A. 2016 Rawlsian fairness for machine learning.

134. Jagsi R, Surender R. 2004 Regulation of junior doctors' work hours: an analysis of British and American doctors' experiences and attitudes. *Social Science & Medicine* **58**, 2181–2191. (doi:10.1016/j.socscimed.2003.08.016)

135. Liapis CD. 2003 Effects of limited work hours on surgical training.*Journal of the American College of Surgeons* **196**, 662–663. (doi:10.1016/s1072-7515(03)00097-8)

136. Gravenstein JS, Cooper JB, Orkin FK. 1990 Work and Rest Cycles in Anesthesia Practice *Anesthesiology* **72**, 737–742. (doi:10.1097/00000542-199004000-00024)

137. Firth-Cozens J, Greenhalgh J. 1997 Doctors' perceptions of the links between stress and lowered clinical care. *Social Science & Medicine* **44**, 1017–1022. (doi:10.1016/s0277-9536(96)00227-4)

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

139. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. 2016 Deepr: A convolutional net for medical records.

140. Pham T, Tran T, Phung D, Venkatesh S. 2016 DeepCare: A deep dynamic memory model for predictive medicine.

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

142. 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**. (doi:10.1186/s12859-015-0852-1)

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

144. Tan J *et al.* 2016 Unsupervised extraction of stable expression signatures from public compendia with eADAGE. (doi:10.1101/078659)

145. 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)

146. Singh R, Lanchantin J, Robins G, Qi Y. 2016 DeepChrome: Deep-learning for predicting gene expression from histone modifications.

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

148. Scotti MM, Swanson MS. 2015 RNA mis-splicing in disease. *Nature Reviews Genetics* **17**, 19–32. (doi:10.1038/nrg.2015.3)

149. 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)

150. 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)

151. Xiong HY, Barash Y, Frey BJ. 2011 Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* (doi:10.1093/bioinformatics/btr444)

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

153. Jha A, Gazzara MR, Barash Y. 2017 Integrative Deep Models for Alternative Splicing. (doi10.1101/104869)

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

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

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

157. 2004 The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640. (doi:10.1126/science.1105136)

158. Stormo GD. 2000 DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23. (doi:10.1093/bioinformatics/16.1.16)

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

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

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

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

163. 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)

164. Lanchantin J, Singh R, Wang B, Qi Y. 2016 Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks.

165. Shrikumar A, Greenside P, Kundaje A. 2017 Reverse-complement parameter sharing improves deep learning models for genomics. (doi:10.1101/103663)

166. Shrikumar A, Greenside P, Shcherbina A, Kundaje A. 2016 Not just a black box: Learning important features through propagating activation differences.

167. Werner T. 2003 The state of the art of mammalian promoter recognition. *Briefings in Bioinformatics* **4**, 22–30. (doi:10.1093/bib/4.1.22)

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

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

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

171. Fickett JW, Hatzigeorgiou AG. 1997 Eukaryotic Promoter Recognition. *Genome Research* **7**, 861–878. (doi:10.1101/gr.7.9.861)

172. Matis S, Xu Y, Shah M, Guan X, Einstein J, Mural R, Uberbacher E. 1996 Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Computers & Chemistry* **20**, 135–140. (doi:10.1016/s0097-8485(96)80015-5)

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

174. Bromberg Y, Rost B. 2007 SNAP: predict effect of non-synonymous polymorphisms on function.*Nucleic Acids Research* **35**, 3823–3835. (doi:10.1093/nar/gkm238)

175. 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)

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

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

178. Li Y, Shi W, Wasserman WW. 2016 Genome-Wide Prediction of cis-Regulatory Regions Using Supervised Deep Learning Methods. (doi:10.1101/041616)

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

180. Berezikov E. 2011 Evolution of microRNA diversity and regulation in animals.*Nature Reviews Genetics* **12**, 846–860. (doi:10.1038/nrg3079)

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

182. Lee B, Baek J, Park S, Yoon S. 2016 DeepTarget: End-to-end learning framework for microRNA target prediction using deep recurrent neural networks.

183. Park S, Min S, Choi H, Yoon S. 2016 DeepMiRGene: Deep neural network based precursor microRNA prediction.

184. Wang S, Sun S, Xu J. 2016 AUC-Maximized Deep Convolutional Neural Fields for Protein Sequence Labeling. In *Machine Learning and Knowledge Discovery in Databases*, pp. 1–16. Springer International Publishing. (doi:10.1007/978-3-319-46227-1_1)

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

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

187. 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)

188. 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)

189. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y. 2015 Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports* **5**, 11476. (doi:10.1038/srep11476)

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

191. Zhou J, Troyanskaya OG. 2014 Deep supervised and convolutional generative stochastic network for protein secondary structure prediction.

192. 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)

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

194. Eickholt J, Cheng J. 2012 Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics* **28**, 3066–3072. (doi:10.1093/bioinformatics/bts598)

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

196. In press. See http://www.predictioncenter.org/casp12/rrc_avrg_results.cgi.

197. In press. See http://www.cameo3d.org/.

198. Li Z, Wang S, Yu Y, Xu J. 2017 Predicting membrane protein contacts from non-membrane proteins by deep transfer learning.

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

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

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

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

203. Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. 2016 Automating Morphological Profiling with Generic Deep Convolutional Networks. (doi:10.1101/085118)

204. Caicedo JC, Singh S, Carpenter AE. 2016 Applications in image-based profiling of perturbations. *Current Opinion in Biotechnology* **39**, 134–142. (doi:10.1016/j.copbio.2016.04.003)

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

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

207. 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)

208. Liu S, Trapnell C. 2016 Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* (doi:10.12688/f1000research.7223.1)

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

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

211. Koh PW, Pierson E, Kundaje A. 2016 Denoising genome-wide histone ChIP-seq with convolutional neural networks. (doi:10.1101/052118)

212. Gaublomme J *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)

213. Arvaniti E, Claassen M. 2016 Sensitive detection of rare disease-associated cell subsets via representation learning. (doi:10.1101/046508)

214. He K, Zhang X, Ren S, Sun J. 2015 Deep residual learning for image recognition.

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

216. 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)

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

218. Abe T. 2003 Informatics for Unveiling Hidden Genome Signatures.*Genome Research* **13**, 693–702. (doi:10.1101/gr.634603)

219. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool.*Journal of Molecular Biology* **215**, 403–410. (doi:10.1016/s0022-2836(05)80360-2)

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

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

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

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

224. In press. See https://www.onecodex.com/.

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

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

227. 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)

228. Soueidan H, Nikolski M. 2017 Machine learning for metagenomics: methods and tools.*Metagenomics* **1**. (doi:10.1515/metgen-2016-0001)

229. In press. See http://www.fasebj.org/content/30/1_Supplement/406.3.

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

231. 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)

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

233. 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**. (doi:10.1186/s12859-015-0753-3)

234. 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.0097958)

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

236. Ditzler G, Polikar R, Rosen G. 2015 A Bootstrap Based Neyman-Pearson Test for Identifying Variable Importance. *IEEE Transactions on Neural Networks and Learning Systems***26**, 880–886. (doi:10.1109/tnnls.2014.2320415)

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

238. Rho M, Tang H, Ye Y. 2010 FragGeneScan: predicting genes in short and error-prone reads.*Nucleic Acids*

*Research* **38**, e191–e191. (doi:10.1093/nar/gkq747)

239. 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)

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

241. Sønderby SK, Sønderby CK, Nielsen H, Winther O. 2015 Convolutional lstm networks for subcellular localization of proteins. (doi:10.1007/978-3-319-21233-3_6)

242. Kelley DR, Salzberg SL. 2010 Clustering metagenomic sequences with interpolated Markov models.*BMC Bioinformatics* **11**, 544. (doi:10.1186/1471-2105-11-544)

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

244. In press. See https://repozitorij.uni-lj.si/IzpisGradiva.php?id=85515.

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

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

247. In press. See http://alifar76.github.io/sklearn-metrics/.

248. Bengio Y, Boulanger-Lewandowski N, Pascanu R. 2012 Advances in optimizing recurrent networks.

249. Boža V, Brejová B, Vinař T. 2016 DeepNano: Deep recurrent neural networks for base calling in minion nanopore reads.

250. Sutskever I, Vinyals O, Le QV. 2014 Sequence to sequence learning with neural networks.

251. 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. (doi:10.1101/092890)

252. Torracinta R, Campagne F. 2016 Training Genotype Callers with Neural Networks. (doi10.1101/097469)

253. Chollet F. 2016 Xception: Deep learning with depthwise separable convolutions.

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

255. Hamburg MA, Collins FS. 2010 The Path to Personalized Medicine.*New England Journal of Medicine***363**, 301–304. (doi:10.1056/nejmp1006304)

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

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

258. Baxt WG. 1991 Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction*Annals of Internal Medicine* **115**, 843. (doi:10.7326/0003-4819-115-11-843)

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

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

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

262. Johansson FD, Shalit U, Sontag D. 2016 Learning representations for counterfactual inference.

263. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. 2015 Causal Phenotype Discovery via Deep Networks. *AMIA Annu Symp Proc***2015**, 677–686.

264. Lipton ZC, Kale DC, Wetzel R. 2016 Modeling missing data in clinical time series with rnns.

265. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. 2016 Recurrent neural networks for multivariate time series with missing values.

266. 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)

267. Lipton ZC, Kale DC, Wetzel RC. 2015 Phenotyping of clinical time series with lstm recurrent neural networks.

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

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

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

271. 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. (doi:10.1101/095653)

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

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

274. 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)

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

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

277. 2016 OUP accepted manuscript. *Briefings In Bioinformatics* (doi:10.1093/bib/bbw110)

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

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

280. 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**. (doi:10.1186/s12859-015-0845-0)

281. 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)

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

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

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

285. Wang C, Liu J, Luo F, Tan Y, Deng Z, Hu Q-N. 2014 Pairwise input neural network for target-ligand interaction prediction. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE.

(doi:10.1109/bibm.2014.6999129)

286. Duan Q *et al.* 2016 L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Systems Biology and Applications* **2**. (doi:10.1038/npjsba.2016.15)

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

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

289. Swamidass SJ, Azencott C-A, Lin T-W, Gramajo H, Tsai S-C, Baldi P. 2009 Influence Relevance Voting: An Accurate And Interpretable Virtual High Throughput Screening Method. *Journal of Chemical Information and Modeling* **49**, 756–766. (doi:10.1021/ci8004379)

290. Kearnes S, Goldman B, Pande V. 2016 Modeling industrial admet data with multitask networks.

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

292. Todeschini R, Consonni V, editors. 2009 *Molecular Descriptors for Chemoinformatics*. Wiley-VCH Verlag GmbH & Co. KGaA. (doi:10.1002/9783527628766)

293. Dahl GE, Jaitly N, Salakhutdinov R. 2014 Multi-task neural networks for qsar predictions.

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

295. In press. See http://www.bioinf.at/publications/2014/NIPS2014a.pdf.

296. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. 2015 Massively multitask networks for drug discovery.

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

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

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

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

301. Rogers D, Hahn M. 2010 Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754. (doi:10.1021/ci100050t)

302. In press. See http://papers.nips.cc/paper/5954-convolutional-networks-on-graphs-for-learning-molecular-fingerprints.

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

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

305. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. 2017 Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* (doi:10.1021/acscentsci.6b00367)

306. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. 2017 MoleculeNet: A benchmark for molecular machine learning.

307. In press. See https://github.com/deepchem/deepchem.

308. 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.

309. Segler MHS, Kogej T, Tyrchan C, Waller MP. 2017 Generating focussed molecule libraries for drug discovery with

recurrent neural networks.

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

311. Olivecrona M, Blaschke T, Engkvist O, Chen H. 2017 Molecular de novo design through deep reinforcement learning.

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

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

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

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

316. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. 2016 Protein-ligand scoring with convolutional neural networks.

317. Ba LJ, Caruana R. 2013 Do deep nets really need to be deep?

318. Nguyen A, Yosinski J, Clune J. 2014 Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.

319. Ribeiro MT, Singh S, Guestrin C. 2016 'Why should i trust you?': Explaining the predictions of any classifier.

320. Zeiler MD, Fergus R. 2013 Visualizing and understanding convolutional networks.

321. Zintgraf LM, Cohen TS, Adel T, Welling M. 2017 Visualizing deep neural network decisions: Prediction difference analysis.

322. Shrikumar A, Greenside P, Kundaje A. 2017 Learning important features through propagating activation differences.

323. Fong R, Vedaldi A. 2017 Interpretable explanations of black boxes by meaningful perturbation.

324. Simonyan K, Vedaldi A, Zisserman A. 2013 Deep inside convolutional networks: Visualising image classification models and saliency maps.

325. 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)

326. Kindermans P-J, Schütt K, Müller K-R, Dähne S. 2016 Investigating the influence of noise and distractors on the interpretation of neural networks.

327. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. 2014 Striving for simplicity: The all convolutional net.

328. In press. See https://www.robots.ox.ac.uk/~vedaldi/assets/pubs/mahendran16salient.pdf.

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

330. Sundararajan M, Taly A, Yan Q. 2017 Axiomatic attribution for deep networks.

331. Lundberg S, Lee S-I. 2016 An unexpected unity among methods for interpreting model predictions.

332. Mahendran A, Vedaldi A. 2014 Understanding deep image representations by inverting them.

333. Finnegan AI, Song JS. 2017 Maximum Entropy Methods for Extracting the Learned Features of Deep Neural Networks. (doi:10.1101/105957)

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

335. In press. See http://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html.

336. In press. See http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/247.

337. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. 2015 Understanding neural networks through deep visualization.

338. Bahdanau D, Cho K, Bengio Y. 2014 Neural machine translation by jointly learning to align and translate.

339. 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.

340. Deming L, Targ S, Sauder N, Almeida D, Ye CJ. 2016 Genetic architect: Discovering genomic structure with learned neural architectures.

341. 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.

342. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. 2016 GRAM: Graph-based attention model for healthcare representation learning.

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

344. Karpathy A, Johnson J, Fei-Fei L. 2015 Visualizing and understanding recurrent networks.

345. Strobelt H, Gehrmann S, Huber B, Pfister H, Rush AM. 2016 Visual analysis of hidden state dynamics in recurrent neural networks.

346. Murdoch WJ, Szlam A. 2017 Automatic rule extraction from long short term memory networks.

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

348. Gal Y, Ghahramani Z. 2015 Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

349. Koh PW, Liang P. 2017 Understanding black-box predictions via influence functions.

350. Kahng M, Andrews P, Kalro A, Chau DH. 2017 ActiVis: Visual exploration of industry-scale deep neural network models.

351. Liu M, Shi J, Li Z, Li C, Zhu J, Liu S. 2016 Towards better analysis of deep convolutional neural networks.

352. Che Z, Purushotham S, Khemani R, Liu Y. 2015 Distilling knowledge from deep networks with applications to healthcare domain.

353. Lei T, Barzilay R, Jaakkola T. 2016 Rationalizing neural predictions.

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

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

356. In press. See https://openreview.net/pdf?id=Sk-oDY9ge.

357. Schmidhuber J. 2015 Deep learning in neural networks: An overview.*Neural Networks* **61**, 85–117. (doi:10.1016/j.neunet.2014.09.003)

358. Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. 2015 Deep learning with limited numerical precision.

359. Courbariaux M, Bengio Y, David J-P. 2014 Training deep neural networks with low precision multiplications.

360. Sa CD, Zhang C, Olukotun K, Ré C. 2015 Taming the wild: A unified analysis of hogwild!-style algorithms.

361. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. 2016 Quantized neural networks: Training neural networks with low precision weights and activations.

362. Hinton G, Vinyals O, Dean J. 2015 Distilling the knowledge in a neural network.

363. Raina R, Madhavan A, Ng AY. 2009 Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, ACM Press. (doi:10.1145/1553374.1553486)

364. In press. See https://research.google.com/pubs/pub37631.html.

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

366. Hadjis S, Abuzaid F, Zhang C, Ré C. 2015 Caffe con troll: Shallow ideas to speed up deep learning.

367. Edwards C. 2015 Growing pains for deep learning. *Communications of the ACM* **58**, 14–16. (doi:10.1145/2771283)

368. Su H, Chen H. 2015 Experiments on parallel training of deep neural network using model averaging.

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

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

371. Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, Shelhamer E. 2014 CuDNN: Efficient primitives for deep learning.

372. Chen W, Wilson JT, Tyree S, Weinberger KQ, Chen Y. 2015 Compressing neural networks with the hashing trick.

373. Lacey G, Taylor GW, Areibi S. 2016 Deep learning on fpgas: Past, present, and future.

374. Dean J, Ghemawat S. 2008 MapReduce. *Communications of the ACM* **51**, 107. (doi:10.1145/1327452.1327492)

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

376. In press. See http://research.google.com/archive/large_deep_networks_nips2012.html.

377. In press. See https://papers.nips.cc/paper/5717-taming-the-wild-a-unified-analysis-of-hogwild-style-algorithms.pdf.

378. Moritz P, Nishihara R, Stoica I, Jordan MI. 2015 SparkNet: Training deep networks in spark.

379. Meng X *et al.* 2015 MLlib: Machine learning in apache spark.

380. In press. See http://download.tensorflow.org/paper/whitepaper2015.pdf.

381. In press. See https://github.com/fchollet/keras.

382. In press. See https://github.com/maxpumperla/elephas.

383. In press. See http://www.jmlr.org/proceedings/papers/v28/coates13.html.

384. Sun S, Chen W, Liu T-Y. 2016 Ensemble-compression: A new method for parallel training of deep neural networks.

385. In press. See https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf.

386. In press. See http://dl.acm.org/citation.cfm?id=2188395.

387. Schatz MC, Langmead B, Salzberg SL. 2010 Cloud computing and the DNA data race. *Nature Biotechnology* **28**, 691–693. (doi:10.1038/nbt0710-691)

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

389. Stein LD. 2010 The case for cloud computing in genome informatics. *Genome Biology* **11**, 207. (doi:10.1186/gb-2010-11-5-207)

390. Krizhevsky A. 2014 One weird trick for parallelizing convolutional neural networks.

391. Armbrust M *et al.* 2010 A view of cloud computing. *Communications of the ACM* **53**, 50. (doi:10.1145/1721654.1721672)

392. Longo DL, Drazen JM. 2016 Data Sharing. *New England Journal of Medicine* **374**, 276–277. (doi:10.1056/nejme1516564)

393. Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. 2017 Celebrating parasites. *Nature Genetics* **49**, 483–

484. (doi:10.1038/ng.3830)

394. 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)

395. In press. See http://kundajelab.github.io/dragonn/.

396. In press. See https://www.synapse.org/#!Synapse:syn6131484/wiki/402026.

397. In press. See https://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.

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

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

400. Pärnamaa T, Parts L. 2017 Accurate Classification of Protein Subcellular Localization from High Throughput Microscopy Images Using Deep Learning. *G3: Genes|Genomes|Genetics*, g3.116.033654. (doi:10.1534/g3.116.033654)

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

402. In press. See https://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf.

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

404. 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)

405. Eser U, Churchman LS. 2016 FIDDLE: An integrative deep learning framework for functional genomic data inference. (doi:10.1101/081380)

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

407. In press. See http://www.businessinsider.com/ibm-edges-closer-to-human-speech-recognition-2017-3.

408. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G. 2016 Achieving human parity in conversational speech recognition.

409. Saon G *et al.* 2017 English conversational telephone speech recognition by humans and machines.

410. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. 2013 Intriguing properties of neural networks.

411. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and harnessing adversarial examples.

412. Papernot N, McDaniel P, Sinha A, Wellman M. 2016 Towards the science of security and privacy in machine learning.

413. Xu W, Evans D, Qi Y. 2017 Feature squeezing: Detecting adversarial examples in deep neural networks.