Deep Learning, Genomics, and Precision Medicine

For preliminary authorship information, see the contributors on GitHub.

Abstract

Abstract goes here.

Introduction

Biology and medicine are rapidly becoming data-intensive with respect to both research and practice. A recent comparison of genomics with social media, online videos and other data-intensive scientific disciplines suggested that the field of genomics alone would equal or surpass other fields in data generation and analysis within the next decade [1]. These data present new opportunities, but also new challenges. The data volume and complexity both indicate that automated algorithms will be needed to extract meaningful patterns and provide actionable knowledge allowing us to better treat, categorize, or study disease, all within data constrained and privacy critical environments.

Concurrent with this explosive growth in biomedical data, a new enthusiasm for a class of machine learning algorithms, known as deep learning, is revolutionizing domains from image search to the game of Go [2]. As recently applied to image analysis problems, these architectures readily surpass previous best-in-class results, and computer scientists are now building many-layered neural networks from collections of millions of images. In a famous and early example, scientists from Google demonstrated that a neural network could learn to identify cats simply by watching online videos [3].

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? Deep learning has transformed image analysis, but what about biomedicine more broadly? In this review, we examine whether this transformation is simply a matter of time or if there are unique challenges posed by biomedical data that render deep learning methods more challenging or less fruitful to apply.

What is deep learning?

Deep learning is built on a biologically-inspired approach from machine learning termed neural networks. Each neuron in a computational neural network, termed a node, has inputs, an activation function, and outputs. Each value from the inputs is usually multiplied by some weight and combined and summarized by the activation function. The value of the activation function is then multiplied by another set of weights to produce the output **TODO: we probably need a figure here - I see no way that we don't include this type of description in our paper, despite the fact that it's been done tons of times before. - I'm really partial to this nature review's explanation about making non-linear problems linear - figure 1 [4]** These neural networks are trained by identifying weights that produce a desired output from some specific input.

Neural networks can also be stacked. The outputs from one network can be used as inputs to another. This process produces a stacked, also known as a multi-layer, neural network. The multi-layer neural network techniques that underlie deep learning have a long history. Multi-layer methods have been discussed in the literature for more than five decades [5]. Given this context, it's challenging to consider "deep learning" as a new advance, though the term has only become widespread to describe analysis methods in the last decade. Much of the early history of neural networks has been extensively covered in a recent review [6]. For the purposes of this review, we identify deep learning approaches as those that use multi-layer neural networks to construct complex features.

We also discuss a class of algorithms that we term "shallow learning" approaches. We do not use this as a pejorative term, but instead to denote algorithms which have all of the hallmarks of deep approaches except that they employ networks of limited depth. We found it valuable to include these as we sought to identify the current contributions of deep learning and to predict its future impact. Researchers may employ these shallow learning methods for a number of reasons including: 1) shallow networks provide a degree of interpretability that better matches their use case; 2) the available data are insufficient to support deeper architectures, however new datasets that will support deep methods are expected; 3) or as building blocks to be combined with other non-neural-network-based approaches at subsequent stages.

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 chose 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 [7]. Here, we ought to identify whether deep learning was an innovation that would induce a strategic inflection point on the practice of biology or medicine. We

considered this with an eye towards the concept of precision medicine.

There are numerous examples where deep learning has been applied to biological problems and produced somewhat improved results, and there are numerous reviews that have focused on general applications of deep learning in biology [8–13]. We sought cases where deep learning was enabling researchers to solve challenges that were previously considered infeasible, or if it made difficult, tedious, and non-routine analyses routine.

Based on our guiding question, we focused on the application of deep learning to topics of biomedical importance. We divided the large range of topics into three broad classes: Disease and Patient Categorization, Fundamental Biological Study, and Patient Treatment. 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 involve histology, requiring manual human expertise for quantification, or small panel of molecular markers, such as cell surface receptors or genes' 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 [14,15]. 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. For example, Tan et al. applied denoising autoencoders (DA), an unsupervised approach, in order to cluster breast cancer patients [16]. Ciresan et al. utilized convolutional neural networks (CNN) to count mitotic divisions in histological images; a feature that is highly correlated with disease outcome [17]. Despite these recent advances, a number of challenges exist in this area of research, such as the integration of disparate types of data, including electronic health records (EHR), imaging and histology data, and molecular omics data.

Fundamental Biological Study

Deep learning can be applied to answer more fundamental biological questions, and is especially suited to leveraging large amounts of data from high throughput omics studies. One classic biological problem where machine learning has been extensively applied is the prediction of molecular targets. Recent advances using deep learning have shown higher accuracy in determining molecular targets. For example, Lee et al. used deep recurrent neural networks (RNN) to predict gene targets of micro-RNAs [18]. Wang et al. used a residual CNN to predict protein-protein contact on a genome-wide scale [19]. Other biological questions that have been investigated include the prediction of protein secondary structure based on sequence data [20,21], recognition of functional genomic elements such as enhancers and promoters [22–24], predicting the deleterious effects of nucleotide polymorphisms[25], etc.

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 outcome, and guide future development of new therapies. Specifically, effort in this area aims to identify drug targets, identify drug interactions or predict drug response. One recent approach for predicting drug response is the use of protein structure to predict drug interactions and drug bioactivity through CNN [26]. Since CNNs leverage spatial relationships within the data, this particular deep learning framework is well suited to the problem. Drug discovery and drug "repurposing" are two other hot topics. Aliper et al. used transcriptomic data to predict which drugs might be repurposed for other diseases through deep fully connected neural networks [27]. In a similar vein, Wang et al. used restricted boltzman machines (RBM) to predict drug molecular targets [28].

Deep learning and patient categorization

In a health care 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 health care. 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.[29].

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 autoantibodies, 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. [30]. However, this represents a potential hope of 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 isn't, 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 health care

One of the general areas 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. Mammography has been one area with numerous contributions [31–34]. In all of this work, the researchers must work around a specific challenge - the limited number of well annotated training images. To expand the number and diversity of images, the researchers have employed approaches where they employ adversarial examples [34] or first train towards human-created features before subsequent fine tuning [31]. The presence of a large bank of well-annotated mammography images would aid in the application of deep neural networks to this area. Though this strategy has not yet been employed in this domain, large collections of unlabeled images might first be used in an unsupervised context to construct high-quality feature detectors. Then 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 [35].

In addition to radiographic images, histology slides are also being analyzed with deep learning approaches. Ciresan et al. [17] 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.[36] 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. Their algorithm had about twice the error rate of a pathologist. However, their algorithms errors were not strongly correlated with the pathologist. Theoretically, combining both could reduce 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 record. Recently Lee et a[37] 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 [38]. 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.

TODO: Potential remaining topics: #122 & #151 looked interesting from an early glance. - Do we want to make the point that most of the imaging exar

Electronic health records

EHR data include substantial amounts of free text, which remains challenging to approach[39]. Often, researchers developing algorithms that perform well on specific tasks must design and implement domain- specific features [40]. 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 [41]. 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 [41]. 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[42] 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 indicates a potential strength of deep methods. It may be possible to repurpose features from task to task, improving overall predictions as the field tackles new challenges.

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. [43], 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 [44]. In addition, it pointed towards learned features being useful for subtyping within a single disease. A concurrent large- scale analysis of an electronic health records system found that a deep denoising autoencoder architecture applied to the number and co-occurrence of clinical test events, though not the results of those tests, constructed features that were more useful for disease prediction than other existing feature construction methods [45]. Taken together, these results support the potential of unsupervised feature construction in this domain. 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 [46]. Early approaches, such as the Faraggi-Simon feed-forward network, aimed to relax the linearity assumption, but performance gains were lacking [47]. 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 [48]. 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 [49]. 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 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 [50], 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 [51].

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 [52]. 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 [53], 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've discussed reduce the need for labeled examples [44]. The adversarial training example strategies that we've mentioned can reduce overfitting, if transformations are available that preserve the meaningful content of the data while transforming irrelevant features [34]. While adversarial training examples can be easily imagined for certain methods that operate on images, it's more challenging to figure out what an equivalent transformation would be for a patient's clinical test results. Consequently, it may be hard to employ adversarial training examples, not to be confused with generative adversarial neural networks, with other applications. Finally, approaches that transfer features can also help use valuable training data most efficiently. Rajkomar et al. trained a deep neural network using generic images before tuning using only radiology images [54]. Datasets that require many of the same types of features might be used for initial training, before fine tuning takes place with the more sparse biomedical examples. Though the analysis hasn't yet been attempted, it's 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 [55]. In data programming, noisy automated labeling functions are integrated. Numerous writers have described data as the new oil [56,57]. 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" [58]. In this framing, data are abundant and not a scarce resource. Instead, new approaches to solve 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.

In addition to methodological improvements, a robust culture of data sharing - and in particular the sharing of high-quality labeled datasets - would do much to speed advances in this domain. The cultural barriers are perhaps best captured by the implications of using the term "research parasite" to describe scientists who use data from other researchers [59]. 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. Though not a methodological consideration, efforts are underway to recognize those who promote an ecosystem of rigorous sharing and analysis [60].

We expect both improved methods and an improved culture of sharing to play an important role if deep learning is going to transform how we analyze data to 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

EHRs are designed and optimized primarily for patient care and billing purposes, meaning research is at most a tertiary priority. This presents significant challenges to EHR based research in general, and particularly to data intensive deep learning research. EHRs are used differently even within the same health care system [61,62]. Individual users have unique usage patterns, and different departments have different priorities which introduce missing data in a non-random fashion. Just et al. demonstrated that even the most basic task of matching patients can be challenging due to data entry issues [63]. This is before considering challenges caused by system migrations and health care system expansions through acquisitions. Replication between hospital systems requires controlling for both these systematic biases as well as for population and demographic effects. Historically, 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 [64]. Wiley et al. [65] showed that warfarin dosing algorithms often under perform in African Americans, illustrating that some of these issues are unsolved even at a treatment best practices level. Lack of standardization 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.

Even if data were perfectly standardized, attempts to share data in this domain would still encounter technological and legal barriers. A responsibility to protect patient privacy limits the ability to openly share large patient datasets. As described above, labeled data are already expensive to obtain. Even after they are generated restrictions on sharing can hamper their broad distribution. All of these factors combine to result in small samples sizes that restrict the number of parameters that can be trained in a model. Though the lack of sharing may also hamper reproducibility and physician confidence in results, recently described techniques such as Continuous Analysis [66] can allow reproducible computing on private data. Using such techniques intermediate results can be automatically tracked and shared without sharing

the original data, which may help to address concerns around physician confidence.

Raw data isn't the only point of concern in the domain with regards to privacy. 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. [67] showed the ability to steal trained models via public APIs and Dwork and Roth[68] 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 [69] - in this case, whether a person's record was present. There are solutions to this challenge. Training algorithms in a differentially private information from accurate answers in a machine learning model. There are solutions to this challenge. Training algorithms in a differentially private manner provides a limited guarantee that the algorithms 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. [70] present the ability to perform GWASs in a differentially private manner and Abadi et al. [71] show the ability to train deep learning classifiers under the differential privacy framework. Federated learning [72] and secure aggregations [73,74] 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, training process in these approaches places constraints on the algorithms used and can make fitting a model substantially more challenging.

Applying deep learning algorithms to this domain provides considerable opportunity as well as challenges - such as patient privacy - that cannot be ignored. Techniques that enable training on data without sharing the raw data may have a part to play. Those applying deep learning to the domain should also consider the potential of deep neural networks to inadvertently leak the training data of participants. 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) [75]. 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 algorithm interpretability alongside accuracy. For example, if we train an algorithm to predict which drugs would be prescribed during a patient's visit to the doctor and there's an existing pattern of racial differences in prescription behavior
(TODO: @traversc - you can pick a different example but I think we need one - whichever one you think is most fully supported by the literature.), this pattern could become baked into the predictions made by the algorithm. Machine learning practitioners, and particularly those who use deep neural networks, which are challenging to interpret, must remain cognizant of this possibility 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 and quality of life [76–79]. 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 [80] but require exceedingly large patient sizes due to discrete combinatorial bucketing. Lasko et al. [43] 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) [81] and the incorporation of past and current state (Recurrent Neural Networks, Long Short Term Memory Networks)[82]. This may be a particular area of opportunity for deep neural networks. The ability to discover relevant sequences of events from large number of trajectories requires powerful and flexible feature construction methods - an area at which deep neural networks tend to excel.

Has deep learning already induced a strategic inflection point for one or more aspects?

I have looked through the papers that we have. I don't see a case in our collection where I felt that we'd be justified to say that deep learning has transformed how we categorize individuals with respect to health and disease. There are definitely interesting applications, but I don't see anything that we couldn't do similarly with some other method.

Will deep learning induce a strategic inflection point for categorization?

This section attempts to get at whether or not we think that deep learning will be transformational. Since we have some room to provide our perspective, I'd suggest that we take a relatively tough look at this once we review where we are in the parts above.

What unique potential does deep learning bring to this?

Are there areas that we expect deep learning to transform how we categorize disease that we haven't seen yet? Let's get fun with speculation/dreaming on this one.

Where would you point your deep learning efforts if you had the time?

This can be fun. We might eventually merge this with the section immediately above on deep learning's unique potential here.

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.

Gene expression

Gene expression measurements 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 such next generation sequencing technologies continue 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 [83] and also to stratify yeast strains based on chemical and mutational perturbations [84]. Shallow (one hidden layer) denoising autoencoders have also been fruitful in extracting biological insight from thousands of *Pseudomonas aeruginosa* experiments [85,86] and in aggregating features relevant to specific breast cancer subtypes [16]. 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 agregated into features. Precisely attributing node activations to specific biological functions risks overinterpreting models and can lead to incorrect conclusions.

Alternatively, deep learning approaches are also being applied for 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 [87]. However, while the deep learning model outperformed already 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. Furthermore, a convolutional neural network applied to histone modifications, termed DeepChrome, [88] was shown to predict gene expression output. DeepChrome greatly improved high or low expression prediction accuracy over existing methods. Deep learning applied to epigenetic data for gene expression inference is a promising approach to study gene regulation. Deep learning approaches have also been applied to study cancer gene expression data with goals of identifying subtypes of patients with different molecular features and clinical manifestations [89]. In the study, the authors combine RBMs to integrate gene expression, DNA methylation, and miRNA data and use the constructed features in search of ovarian cancer subtypes. 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 in its infancy but the future is bright. Many hypotheses can now be interrogated because of increasing amounts of data and new developing technologies. For example, there is a growing appreciation for the large impact of disease heterogeneity on research and treatment strategies for disease. New technologies are being developed, such as single cell RNA-seq and high throughput fluorescence based imaging that are good matches for deep learning. Concurrently, deep learning methods are being developed to address novel problems such as adjusting for batch effects in single-cell RNA-seq data [90]. Moreover, deep learning is already well established in the image processing community, so the marriage of fluorescence based imaging techniques and deep learning is natural. These technologies are growing in popularity and will provide increasingly novel perspectives with respect to how cellular heterogeneity impacts gene expression coordination within a sample. In general, as the flow of gene expression data increases, and techniques to integrate heterogeneous genomic measurements made on the same samples are enhanced, the quality and types of questions deep learning can address is poised to improve.

Splicing

Pre-mRNA transcripts can be spliced into different isoforms by retaining or skipping subsets of exons, or including parts of introns. This alternative splicing provides cells with enormous spatiotemporal flexibility to generate multiple distinct proteins from a single gene. Splicing is catalyzed by small nuclear RNAs (snRNAs) and spliceosomal proteins, which detect sequence motifs such as splice sites and exon sequence enhancers and silencers (ESE and ESS). Various RNA-binding proteins and noncoding RNAs can bias these reactions by altering binding affinities, blocking splice sites, or sequestering splicing factors. This remarkable complexity unfortunately lends itself to defects that underlie many diseases [91]. For instance, in Becker muscular dystrophy, a point mutation in dystrophin creates an ESS 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 [92].

Sequencing studies routinely return thousands of unannotated variants. 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 [93,94]. With the advent of deep learning, more complex models were built that provided better predictive accuracy [95,96]. Importantly, these new approaches can take in not only genomic features, but also tissue identity and CLIP-seq measurements of interactions between splicing factors and RNA, which all improve predictive accuracy.

The massive improvement seen with deep learning seems to stem from hidden layers being able to create new higher-order “features”, whereas earlier approaches often assumed independence of features and were unable to generalize. Higher-order understanding is especially important in splicing, which depends not only on the primary sequence, but also local RNA structure, tissue identity, splicing factor binding, and other currently unknown factors — all of which interact in complex, incompletely characterized ways. With new tools to interpret these meta-features, a major focus of current deep learning research, we will soon have the ability to extract a more nuanced biological understanding of splicing — perhaps by interrogating informative hidden nodes within neural networks that take in tissue type as part of the input, or by building separate networks for each tissue type and looking for common versus distinctive nodes [97].

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 was able to describe motifs that influence the strength of alternative splice sites [98]. 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 framework. This case is instructive in that clever sources of data, not just more powerful models, can lead to 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 alternative splicing genome-wide gives rise to or is involved in complex, non-Mendelian diseases such as autism, schizophrenia, Type 1 diabetes, and multiple sclerosis [99]. As a proof of concept, Xiong et al. [95] sequenced five ASD and 12 control samples, each with an average of 42,000 rare variants, and identified 19 genes with neural functions that are mis-spliced. 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 more comprehensive and diverse data sources, will likely provide the path forward to a better understanding of 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 [100] 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 TFBSs on a DNA sequence, researchers initially used consensus sequences and position weight matrices to match against a test sequence [101]. 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 [102]. Later, SVM techniques outperformed the generative methods by using k-mer features [103,104], but string kernel based SVM systems are limited by expensive computational cost proportional to the number of training and testing sequences. More recently, [105] 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. [106] introduced several new convolutional and recurrent neural network models for predicting TFBSs, but it remains unclear which neural architectures work best for all samples and TFs. 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.

While accurately predicting transcription factors computationally is useful, it is important to understand how these computational models make their predictions. To handle this, several papers have focused on understanding machine learning models [105–107]. [105] 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. [106] introduced a suite of state-of-the-art deep learning models and new visualizations techniques for a more in-depth analysis of TFBSs. Furthermore, [107] introduced an advanced visualization method and toolbox for analyzing possible TFBS sequences. [105] also introduced mutation maps, where they could easily mutate, add, or delete basepairs 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: Add discussion about the large number of deep learning works in this area since the DeepBind paper. In particular, add [#43](https://github.co

Promoters, enhancers, and related epigenomic tasks

We may want to be selective about what we discuss and not list every application in this area.

Micro-RNA binding

miRNAs are important biologically, but have neural networks produced anything particularly notable in this area?

Protein secondary and tertiary structure

Proteins play fundamental roles in all biological processes including the maintenance of cellular integrity, metabolism, transcription/translation, and cell-cell communication. Complete description of protein structures and functions is a fundamental step towards understanding biological life and also highly relevant in the development of therapeutics and drugs. UnitProt currently has about 94 millions of protein sequences. Even if we remove redundancy at 50% sequence identity level, UnitProt still has about 20 millions of protein sequences. However, fewer than 100,000 proteins have experimentally-solved structures in Protein Data Bank (PDB). As a result, computational structure prediction is essential for a majority number of protein sequences. However, predicting protein 3D structures from sequence alone is very challenging, especially when similar solved structures (called templates) are not available in PDB. In the past decades, various computational methods have been developed to predict protein structure from different aspects, including prediction of secondary structure, torsion angles, solvent accessibility, inter-residue contact map, disorder regions and side-chain packing.

Machine learning is extensively applied to predict protein structures and some success has been achieved. For example, secondary structure can be predicted with about 80% of 3-state (i.e., Q3) accuracy by a neural network method PSIPRED [108]. Starting from 2012, deep learning has been gradually introduced to protein structure prediction. The adopted deep learning models include deep belief network, LSTM(long short-term memory), deep convolutional neural networks (DCNN) and deep convolutional neural fields[21,109]. Here we focus on deep learning methods for two representative subproblems: secondary structure prediction and contact map prediction. Secondary structure refers to local conformation of a sequence segment while a contact map contains information of global conformation. Secondary structure prediction is a basic problem and almost an essential module of any protein structure prediction package. It has also been used as sequence labeling benchmark in the machine learning community. Contact prediction is much more challenging than secondary structure prediction, but it has a much larger impact on tertiary structure prediction. In recent years, contact prediction has made good progress and its accuracy has been significantly improved [110–113].

Protein secondary structure can exhibit three different states (alpha helix, beta strand and loop regions) or eight finer-grained states. More methods are developed to predict 3-state secondary structure than 8-state. A predictor is typically evaluated by 3-state (i.e., Q3) and 8-state (i.e., Q8) accuracy, respectively. Qi et al. developed a multi-task deep learning method to simultaneously predict several local structure properties including secondary structures [114]. Spencer, Eickholt and Cheng predicted secondary structure using deep belief networks [20]. Heffernan and Zhou et al. developed an iterative deep learning framework to simultaneously predict secondary structure, backbone torsion angles and solvent accessibility [115]. However, none of these deep learning methods achieved significant improvement over PSIPRED [116] in terms of Q3 accuracy. In 2014, Zhou and Troyanskaya demonstrated that they could improve Q8 accuracy over a shallow learning architecture conditional neural fields [117] by using a deep supervised and convolutional generative stochastic network[118], but did not report any results in terms of Q3 accuracy. In 2016 Wang and Xu et al. developed a deep convolutional neural fields (DeepCNF) model that can significantly improve secondary structure prediction in terms of both Q3 and Q8 accuracy[21]. DeepCNF possibly is the first that reports Q3 accuracy of 84-85%, much higher than the 80% accuracy maintained by PSIPRED for more than 10 years. It is also reported that DeepCNF can improve prediction of solvent accessibility and disorder regions [109]. This improvement may be mainly due to the introduction 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 similar amount of 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 an effective contact prediction method for some proteins with a very large number (>1000) of sequence homologs [113], but ECA fares poorly for proteins without many sequence homologs. Since (soluble) proteins with many sequence homologs are likely to have a good template in PDB, to make contact-assisted folding practically useful for ab initio folding, it is essential to predict accurate contacts for proteins without many sequence homologs. By combining ECA with a few other protein features, shallow neural network-based methods such as MetaPSICOV [111] and CoinDCA-NN [119] have shown some advantage over ECA for proteins with a small number of sequence homologs, but their accuracy is still not very good. In recent years, deep learning methods 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[120]. Eickholt and Cheng combined deep belief networks and boosting techniques to predict protein contacts [121] 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 [122]. However, blindly tested in the well-known CASP competitions, these methods did not show any advantage over MetaPSICOV [111], a method using two cascaded neural networks. Very recently, Wang and Xu et al. proposed a novel deep learning method RaptorX-Contact [110] that can significantly improve contact prediction over MetaPSICOV especially for proteins without

many sequence homologs. RaptorX-Contact 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 [123]), RaptorX-Contact is ranked first in terms of the total F1 score (a widely-used performance metric) on free-modeling targets as well as the whole set of targets. In the CASP12 test, the group ranked second also employed a deep learning method. Even MetaPSICOV, which ranked third in CASP12, employed more and wider hidden layers than its old version. Wang and Xu et al. have also demonstrated in another blind test CAMEO (which can be interpreted as a fully-automated CASP) [124] that their predicted contacts can help fold quite a few proteins with a novel fold and only 65-330 sequence homologs and that their method also works well on membrane protein contact prediction even if trained mostly by non-membrane proteins. In fact, most of the top 10 contact prediction groups in CASP12 employed some kind of deep learning techniques. The RaptorX-Contact method performed better mainly due to introduction of residual neural networks and exploiting contact occurrence patterns by simultaneous prediction of all the contacts in a single protein. It is still possible to further improve contact prediction by studying new deep network architectures. However, current methods fail when proteins in question have almost no sequence homologs. It is unclear if there is an effective way to deal with this type of proteins or not except waiting for more sequence homologs. Finally, the deep learning methods summarized above also apply to interfacial contact prediction of a protein complex, but may be less effective since on average protein complexes have fewer sequence homologs.

Signaling

There is not much content here. Can[125] be covered elsewhere?

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 [126]. 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 [10,126,127]. 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 [128]. 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. But images of cells contain a wide variety of quantitative information, and deep learning may just be the tool to extract it. 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. Already, there is evidence deep learning can surpass the efficacy of classical methods [129], even using generic deep convolutional networks trained on natural images[130], 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 [131–133]. 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.

TODO: Make sure that at the end we clearly emphasize our excitement around unsupervised uses.

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 [134]. For instance, tumor cells and neurons can both harbor extensive somatic variation [135]. 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 [136,137].

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. [138] 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 [97,139].

Examining populations of single cells can reveal biologically meaningful subsets of cells as well as their underlying gene regulatory networks [140]. 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. [141] 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 [142] 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 [143], or by feeding cell morphology and movement into neural networks [128]. Reinforcement learning algorithms [2] 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 [144] and then delved into other clustering algorithms, such as self-organizing maps (SOM)[145]. Then came the taxonomic classification problem, with researchers naturally using BLAST [146], followed by other machine learning techniques such as SVMs [147], naive Bayesian classifiers [148], 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 [149], (WGS)Quikr [150], and some configurations of tools like OneCodex[151] and LMAT [152]. 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 [153] 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 [154,155]. 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 [156]. One of the first attempts was a survey of supervised classification methods for microbes->phenotype classification [157], followed by similar studies that are more massive in scale[158,159]. There have been techniques that bypass the taxonomic classification step altogether [160], (sequence composition to phenotype classification). Also, researchers have looked into how feature selection can improve classification [159,161], and techniques have been proposed that are classifier-independent [162,163].

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 [164] and FragGeneScan [165]), which usually has plenty of training examples. Representations (similar to Word2Vec [166] in natural language processing) for protein family classification has been introduced and classified with a skip-gram neural network [167]. Recurrent neural networks show good performance for homology and protein family identification [168,169]. 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[145]. 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

[170] 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 [171] 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 [172]. 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 [173]. 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) [174]. 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 [175].

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 [176]), 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 [177], 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 [178], 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

We have one nanopore paper in the issues and very recent work on variant calling that looks worthy of inclusion.

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

There will be some overlap with the Categorize section, and we may have to determine which methods categorize individuals and which more directly match patients with treatments. The sub-section titles are merely placeholders.

Categorizing patients for clinical decision making

There has been sustained interest in applying deep learning to clinical decision making for over two decades. In 1996, Tu [179] 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 [180] or mortality [181]. He posited that while neural networks have several advantages in representational power, the difficulties in interpretation may limit clinical applications. In 2006, Lisboa and Taktak [182] examined the use of artificial neural networks in medical journals, concluding that neural networks provided an increase in benefit to healthcare relative to traditional screening methods in 21 of 27 studies.

While significant progress has been made in developing deep learning methods for diagnosis, it is not clear that these methods have yet transformed clinical decision making. The difficulty in applying deep learning to clinical decision making represents 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. 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.

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 [183–186] have used deep recurrent networks to categorize patients but most stop short of suggesting clinical decisions. Nemati et al [187] 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 [2] have relied on knowledge of the underlying processes (e.g. the rules of the game). Some models have been developed for targeted medical problems [188], 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 [189] used a randomized denoising autoenconder 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 [190] 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 proability, 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. Because deep learning models are data-intensive, directly applying current deep learning models to diagnose patients with rare diseases would require prohibitively large datasets. 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.

Effects of drugs on transcriptomic responses

We discussed a few papers that operate on Library of Network-Based Cellular Signatures (LINCS) gene expression data. We could briefly introduce the goals of that resource and comment on the deep learning applications. In the Issues, we had reservations about whether the improvements in expression prediction are good enough to make a practical difference in the domain and feature selection and construction.

Ligand-Based Prediction of Bioactivity

In the biomedical domain, high-throughput chemical screening aims to improve therapeutic options over a long term horizon [191] TODO: add another general screening reference. 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.
TODO: clarify desired outputs, what will be done with the top-ranked hits, one vs multiple hits, abandoning lead compounds This screening process can serve as the first step in the long drug discovery pipeline, where novel chemicals are pursued for their ability to inhibit or enhance disease-relevant biological mechanisms. 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.
TODO: is the sufficient number target dependent? This task has been treated as a classification, regression, and 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 to other inactives is less important [192]. TODO: can improve this first attempt at an intro by reviewing more existing literature on the topic
TODO: check which other existing reviews should be referenced

We primarily focus on ligand-based approaches that train on chemicals' features without requiring knowledge of the target, as opposed to alternative strategies that use target features such as the protein structure TODO: add examples. 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 [193]. 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 [194]
TODO: can add additional references besides this review, 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 [195], and the Merck 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" [196]. Subsequent work explored the effects of jointly modeling far more targets than the Merck challenge [197,198], with [198] showing that the benefits of multitask networks had not yet saturated even with 259 targets. Although a deep learning approach, DeepTox [199], 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. A reliance on AUC ROC TODO: define here? for the evaluation (see Discussion) further hinders the ability to declare Tox21 as an outright success for deep learning.

In retrospect, 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 absorption distribution, metabolism, excretion, and toxicity (ADMET) tasks demonstrated that there are limitations to multitask transfer learning that are in part a consequence of the degree to which tasks are related [200]. Some of the ADMET datasets showed far superior performance in multitask models of only the 22 ADMET tasks relative to more expansive multitask networks that included over 500 less-similar tasks.
TODO: also has a good discussion of information leakage in cross validation, include that in the Discussion section 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 [201]. 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% or 1% of all tested chemicals for a typical target TODO: verify those estimates. This is consistent with the strong performance of low-parameter neural networks that emphasize compound-compound similarity, such as influence-relevance voter, [192,202] instead of predicting compound activity directly from chemical features.
TODO: include recent DeepChem IRV benchmarks?

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 [203]. An overly simplistic but approximately correct view of these 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 [204] 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 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 [205]. 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 [206,207].

Advances in chemical representation learning have not been limited to graph-based neural networks. The simplified molecular-input line-entry system (SMILES) is a standard way to transform a chemical into string-based representation. A SMILES-to-SMILES autoencoder learns a continuous latent feature space for chemicals [208].
TODO: connect to related EHR paper TODO: autoencoder doesn't fit cleanly with supervised methods, revisit organization after adding GANs In this learned continuous space is it possible to train some types of supervised learning algorithms or interpolate between continuous representations of chemicals in a manner that is not possible with discrete (e.g. bit vector or string) features. A drawback is that not all SMILES strings produced by the autoencoder's decoder correspond to valid chemical structures. TODO: could mention GANs here Altae-Tran et al developed a one-shot learning network[207] 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 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 [204] were narrowly beaten by regression using traditional circular fingerprints on a drug efficacy prediction task (but were superior for predicting solubility or photovoltaic efficiency). Graph convolutions [206] performed comparably to a multitask network using standard fingerprints and slightly better than the neural fingerprints [204] on the drug efficacy task but were slightly worse than the influence-relevance voter method on an HIV dataset. [192].
TODO: there are also problems with some papers using ROC primarily for benchmarking, which skews results and makes it hard to assess absolute p

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 and expanded public datasets to advance the state of the art and convincingly demonstrate superiority over traditional cheminformatics techniques. Fortunately, there has recently been significant progress in this direction. The DeepChem software [207,209] and MoleculeNet benchmarking suite [210] 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 [197,198]. Other studies have shown the limitations of transfer learning when the prediction tasks are insufficiently related [200,207]. This has important implications for representation learning. The typical approach of improving 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 [208]. Future work will establish whether it is better to train on massive collections of diverse compounds, drug-like small molecules, or specialized subsets.

TODO: other papers to add such as generative models

TOOD: relationship to traditional docking (some networks include docking scores), deep learning with structure (e.g. [@Z7fd0BYf; @bNBiliTt; @17Ya

TODO: analogies to other domains where deep learning can capture the behavior of complex physics (e.g. quantum physics example)?

Modeling Metabolism and Chemical Reactivity

Add a review here of metabolism and chemical reactivity.

Discussion

This section provides meta-commentary that spans the Categorize, Study, and Treat subject areas. The candidate sub-sections below are initial ideas that can be further pruned.

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.

Interpretation

As the challenge of interpretability is common across many domains, there is significant interest in developing generic procedures for knowledge extraction from deep models. Ribeiro et al [211] focus on interpreting individual predictions rather than interpreting the model. By fitting simple linear models to mimic the predictions of the deep learning model in a small neighborhood of a data sample, they generated an interpretable model for each prediction. While this procedure can provide interpretable models for each sample, it is unclear whether these interpretable models are reliable. Theoretical guarantees on the curvature of the predictions of deep learning models are not known, and it is unclear whether predictions from deep learning models are robust to sample noise. 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 [212] provided confidence intervals under the assumption of normally distributed error. However, Nguyen et al [213] showed that the confidence of convolutional neural networks is not reliable; they can output confidence scores over 99.99% even for samples that are purely noise. Recently, Fong and Vedaldi [214] provided a framework for understanding black box algorithms by perturbing input data.

For domain-specific models, we previously described approaches for the interpretation and visualization of neural networks that prediction transcription factor binding [105–107]. Other studies have primarily focused on integrating attention mechanisms with the neural networks. Attention mechanisms dynamically weight the importance the neural network gives to each feature. By inspecting the attention weights for a particular sample, a practitioner can identify the important features for a particular prediction. Choi et al [215] inverted the typical architecture of recurrent neural networks to improve interpretability. In particular, they only used recurrent connections in the attention generating procedure, leaving the hidden state directly connected to the input variables. In the clinical domain, this model was able to produce accurate diagnoses in which the contribution of previous hospital visits could be directly interpreted. Choi et al [216] 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. Che et al [217] introduced a knowledge-distillation approach which used gradient boosted trees to learn interpretable healthcare features from trained deep models.

Data limitations

Related to evaluation, are there data quality issues in genomic, clinical, and other data that make this domain particularly challenging? Are these worse than what is faced in other non-biomedical domains?

Many applications have used relatively small training datasets. We might discuss workarounds (e.g. semi-synthetic data, splitting instances, etc.) and how this could impact future progress. Might this be why some studies have resorted to feature engineering instead of learning representations from low- level features? Is there still work to be done in finding the right low-level features in some problems?

Biomedical data is often "Wide"

Biomedical studies typically deal with relatively small sample sizes but each sample may have millions of measurements (genotypes and other omics data, lab tests etc).

Classical machine learning recommendations were to have 10x samples per number of parameters in the model.

Number of parameters in an MLP. Convolutions and similar strategies help but do not solve

Bengio diet networks paper

Hardware limitations and scaling

Several papers have stated that memory or other hardware limitations artificially restricted the number of training instances, model inputs/outputs, hidden layers, etc. Is this a general problem worth discussing or will it be solved naturally as hardware improves and/or groups move to distributed deep learning frameworks? Does hardware limit what types of problems are accessible to the average computational group, and if so, will that limit future progress? For instance, some hyperparameter search strategies are not feasible for a lab with only a couple GPUs.

Some of this is also outlined in the Categorize section. We can decide where it best fits.

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. As such, neural networks have only recently found widespread use [6].

Many have sought to curb the costs of deep learning, with methods ranging from the very applied (e.g., reduced numerical precision [218–221]) to the exotic and theoretic (e.g., training small networks to mimic large networks and ensembles [222,223]). The largest gains in efficiency have come from computation with graphics processing units (GPUs) [6,224–228], 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 [225,226,229,230]. However, GPUs also have a limited quantity of memory, making it difficult to implement networks of significant size and complexity on a single GPU or machine [224,231]. This restriction has sometimes forced computational biologists to use workarounds or limit the size of an analysis. For example, Chen et al. [87] 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 [19,208]. Some have also chosen to use slower CPU implementations rather than sacrifice network size or performance [232].

Steady improvements in GPU hardware may alleviate this issue somewhat, but it is not clear whether they 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 [218–222,233,234], but there is also growing interest in specialized hardware, such as field-programmable gate arrays (FPGAs) [228,235] and application-specific integrated circuits (ASICs). Specialized hardware promises improvements in deep learning at reduced time, energy, and memory [228]. Logically, there is less software for highly specialized hardware[235], 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 [236,237] were not suitable for deep learning [238], but significant progress has been made. There now exist a number of algorithms [220,238,239], tools [240–242], and high-level libraries [243,244] for deep learning in a distributed environment, and it is possible to train very complex networks with limited infrastructure [245]. Besides handling very large networks, distributed or parallelized approaches offer other advantages, such as improved ensembling [246] or accelerated hyperparameter optimization [247,248].

Cloud computing, which has already seen adoption in genomics[249], could facilitate easier sharing of the large datasets common to biology [250,251], and may be key to scaling deep learning. Cloud computing affords researchers significant flexibility, and enables the use of specialized hardware (e.g., FPGAs, ASICs, GPUs) without significant investment. With such flexibility, it could be easier to address the different challenges associated with the multitudinous layers and architectures available [252]. Though many are reluctant to store sensitive data (e.g., patient electronic health records) in the cloud, secure/regulation-compliant cloud services do exist [253].

TODO: Write the transition once more of the Discussion section has been fleshed out.

Code, data, and model sharing

Reproducibiliy is important for science to progress. In the context of deep learning applied to advance human healthcare, does reproducibility have different requirements or alternative connotations? With vast hyperparameter spaces, massively heterogeneous and noisy biological data sets, and black box interpretability problems, how can we best ensure reproducible models? What might a clinician, or policy maker, need to see in a deep model in order to influence healthcare decisions? Or, is deep learning a hypothesis generation machine that requires manual validation? DeepChem and DragoNN are worth discussing here.

Transfer learning/transferability of features

- https://github.com/greenelab/deep-review/issues/139#issuecomment-268901804

Conclusions

Final thoughts and future outlook here. The Discussion will give an overview and the Conclusion will provide a short, punchy take home message.

Points to mention based on discussion thus far that may make the bar for conclusions:

- Limitations of data & workarounds (availability impacts on amount, etc)
- Transferability of features
- Strong enthusiasm for unsupervised approaches.
- Right to an explanation (possibly, depends if raised in multiple areas)

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.

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

2. 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)

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

4. LeCun Y, Bengio Y, Hinton G. 2015 Deep learning. *Nature* **521**, 436–444. (doi:10.1038/nature14539)

5. 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)

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

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

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

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

10. 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)

11. 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)

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

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

14. 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)

15. 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)

16. 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)

17. 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)

18. 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)

19. Wang S, Sun S, Li Z, Zhang R, Xu J. 2016 Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. (doi:10.1101/073239)

20. 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)

21. 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)

22. 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)

23. 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)

24. 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)

25. 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)

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

27. 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)

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

29. 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)

30. 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)

31. 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)

32. 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)

33. 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)

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

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

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

37. 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)

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

39. 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)

40. 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)

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

42. 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)

43. 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)

44. 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)

45. 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)

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

47. 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)

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

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

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

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

52. 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)

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

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

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

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

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

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

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

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

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

62. 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.

63. 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.

64. 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)

65. 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)

66. 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)

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

68. 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)

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

70. 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)

71. 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)

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

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

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

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

76. 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)

77. 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)

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

79. 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)

80. 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)

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

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

83. 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)

84. 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)

85. 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)

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

87. 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)

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

89. 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)

90. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. 2016 Removal of batch effects using distribution-matching residual networks.

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

92. 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)

93. 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)

94. 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)

95. 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)

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

97. 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)

98. 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)

99. 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)

100. 2004 The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640. (doi:10.1126/science.1105136)

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

102. 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)

103. 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)

104. 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)

105. 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)

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

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

108. McGuffin LJ, Bryson K, Jones DT. 2000 The PSIPRED protein structure prediction server.*Bioinformatics* **16**, 404–405. (doi:10.1093/bioinformatics/16.4.404)

109. 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)

110. 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)

111. 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)

112. 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)

113. 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)

114. 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)

115. 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)

116. 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)

117. Wang Z, Zhao F, Peng J, Xu J. 2011 Protein 8-class secondary structure prediction using conditional neural fields. *PROTEOMICS* **11**, 3786–3792. (doi:10.1002/pmic.201100196)

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

119. 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)

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

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

122. 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)

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

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

125. 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)

126. 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)

127. 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)

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

129. 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)

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

131. 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)

132. 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)

133. 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)

134. 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)

135. 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)

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

137. 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)

138. 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)

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

140. 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)

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

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

143. 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)

144. 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)

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

146. 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)

147. 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)

148. 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)

149. 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)

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

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

152. 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)

153. 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)

154. 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)

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

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

157. 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)

158. 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)

159. 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)

160. 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)

161. 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)

162. 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)

163. 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)

164. 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)

165. 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)

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

167. 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)

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

169. 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)

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

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

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

173. 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)

174. 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)

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

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

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

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

179. 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)

180. 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)

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

182. 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)

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

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

185. 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)

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

187. 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)

188. 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)

189. 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)

190. 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)

191. 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)

192. 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)

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

194. 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)

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

196. 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)

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

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

199. 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)

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

201. 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)

202. 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)

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

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

205. 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)

206. 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)

207. 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)

208. 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.

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

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

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

212. 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)

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

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

215. 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.

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

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

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

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

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

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

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

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

224. 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)

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

226. 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)

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

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

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

230. 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)

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

232. 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)

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

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

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

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

237. 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)

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

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

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

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

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

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

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

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

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

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

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

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

250. 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)

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

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

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