Review

Applications of deep learning in biomedicine

Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov

Mol. Pharmaceutics, **Just Accepted Manuscript** • DOI: 10.1021/acs.molpharmaceut.5b00982 • Publication Date (Web): 23 Mar 2016

Applications of deep learning in biomedicine

Polina Mamoshina[1], Armando Vieira[2], Evgeny Putin[1], Alex Zhavoronkov[1]

1 Artificial Intelligence Research, Insilico Medicine, Inc, ETC, Johns Hopkins University, Baltimore, MD, 21218
2 RedZebra Analytics, 1 Quality Court, London, WC2A 1HR

Abstract
Increases in throughput and installed base of biomedical research equipment led to a massive accumulation of -omics data known to be highly variable, high-dimensional, and sourced from multiple often incompatible data platforms. While this data may be useful for biomarker identification and drug discovery, the bulk of it remains underutilized. Deep neural networks (DNNs) are efficient algorithms based on the use of compositional layers of neurons, with advantages well matched to the challenges -omics data presents. While achieving state-of-the-art results and even surpassing human accuracy in many challenging tasks, the adoption of deep learning in biomedicine has been comparatively slow. Here, we discuss key features of deep learning that may give this approach an edge over other machine learning methods. We then consider limitations and review a number of applications of deep learning in biomedical studies demonstrating proof of concept and practical utility.

Keywords: deep learning, deep neural networks, RBM, genomics, transcriptomics, artificial intelligence, biomarker development

Introduction

The amount of biomedical data in public repositories is rapidly increasing[1], but the heterogeneous nature of this data renders integrative analysis increasingly difficult.[2] Computational biology methods are essential and routinely used in various fields of biomedicine from biomarker development to drug discovery[3], and machine learning methods are extensively and increasingly applied[4]. Deep learning is a broad class of machine learning techniques that shows particular promise in extracting high level abstractions from the raw data of very large, heterogeneous, high-dimensional datasets. This is precisely the type of data biology now has to offer.

Here, we review deep learning as a versatile biomedical research tool with many potential applications, including the resurrection of cold repository datasets for new use in drug discovery and biomarker development. We first define basic concepts and the rationale for applying these techniques to biological data. We then discuss various implementation considerations and follow with a review of recent biomedical studies that have used deep learning, while highlighting areas of biomedicine that could benefit from this approach, in particular those important for biomarker development and drug discovery. Finally, we discuss limitations and future directions.

1. Basic concepts of deep learning

Machine learning, or learning that occurs without explicit programming, can take place in one of two forms: conventional, "shallow" learning (neural networks with a single hidden layer or support vector machines), or deep learning (neural networks with many hierarchical layers of non-linear information processing). Deep learning was recently reviewed in detail by LeCun et al[5]. While deep and shallow learning differ in more than one way and both approaches have value in specific applications, the takeaway difference is that shallow learning does not deal well with raw data, requiring extensive human input to set up and maintain, whereas deep learning can be largely unsupervised once set in motion, learning intricate patterns from even high-dimensional raw data with little guidance[5]. Bengio and LeCun referred to this as optimizing the breadth/depth tradeoff[2]; that is, only a deep circuit can perform exponentially complex computational tasks without requiring an infinite number of elements[7].

The importance of this is most readily apparent in the areas where deep learning has shown to be useful: image and language recognition[8] and video games[9] are two common examples, or, perhaps more interestingly, replication of painting styles or even composition of classical music[10]. The type of learning required in these tasks is representation learning; that is, detecting or classifying patterns, or representations, from raw data[5], particularly when this data is heirarchical in structure. Image recognition, for example, begins with learning a progressive heirarchy of sub-images from pixels, starting with edges, then motifs, until the final output is a whole object[5]. Representations are formed through simple associations using, for example, pixels as raw data, not by human labeling or pre-programmed logic. Being essentially unsupervised algorithms, deep neural networks can act as feature detector units at each layer (level) that gradually extract more sophisticated and invariant features from the original raw input signals.

One can imagine the impossible effort of annotating the millions of images that machines can now accurately identify. That machines can now distinguish images of two nearly identical objects or complete a sentence is all possible increasingly with help from deep learning. These and other recent developments in DNN architectures have boosted enthusiasm within the machine learning community, with unprecedented performance in many challenging tasks[11,12]. They have also raised important questions about whether deep learning could also automate tasks like annotation, image recognition, prediction, and classification in similar biological applications, where the sheer amount and complexity of data has surpassed human analytical capabilities.

2. Why deep learning may benefit biomedical research

With some imagination, parallels can be drawn between biological data and the types of data deep learning has shown the most success with—namely image and voice data. A gene expression profile, for instance, is essentially a "snapshot," or image, of what is going on in a given cell or tissue under given conditions, with patterns of gene expression representative of physical states in a cell or tissue in the same way that patterns of pixelation are representative of objects in a picture.

In the same way that two similar but categorically different images must be discerned by deep learning algorithms regardless of background or position, two similar but categorically different disease pathologies may be difficult to distinguish if certain unimportant background conditions happen to match (e.g. tissues, time-points, individual, species, platform), thus selectivity of key differences is essential. Alternatively, one

pathology may appear to differ from itelf when imposed on a variety of different experimental "backgrounds" and in several different states of progression, so invariance to non-target-related differences is also key. These features, selectivity and invariance, are requirements for both image recognition and gene expression analysis and are also two hallmarks of CNNs, the powerhouses of modern visual image processing.[5]

The same type of analogies can be drawn with other applications of deep learning—language prediction, for example, requires sequential learning with recurrent neural networks[5] and can be paralleled with signaling cascades in biology, where one event can be predicted from previous upstream events in the same way that one word in a sentence can be predicted from the previous set of words. Structural prediction would be another example. The possibilities are endless; with enough interest in the topic, any number of other parallels can be drawn and new applications conceived.

These parallels, while illustrative and hypothetical in nature, are also backed up by several practical advantages of DNNs that strengthen the case for biological application. First, DNNs require very large datasets, which biology is teeming with at this time. Secondly, DNNs are well-equipped to handle high dimensional, sparse, noisy data with non-linear relationships, all of which are common to transcriptomics and other –omics data in biology. Third, DNN have high generalization ability; once trained on a dataset they can be applied to other, new datasets; this is a requirement for binding and interpretation of heterogeneous multiplatform data, such as gene expression data.

Finally, these considerations are further supported by the fact that the small number of deep learning studies in biomedicine that now exist have shown success with this method. These are to be discussed below.

Importantly, despite the suitability of DNN for biological data and the potential applications, the adoption of deep learning methods in biology has been slow. This may have several explanations. While deep architectures can be exponentially more efficient than conventional models, capturing fine subtleties in the structure of the data[13], DNN, especially recurrent networks, are very complex machines containing hundreds of millions of weights, which makes training and regularization difficult. Deep models are still not optimized, still lack an adequate formulation, require more research, and rely heavily on computational experimentation. It should also be emphasized that despite being able to extract latent features from the data, DNNs are black boxes that learn by simple associations and co-occurrences. They lack the transparency and interpretability of other methods and may be unable to uncover complex causal and structural relationships common in biology without some human interpretation. Nevertheless, their many benefits may outweigh these obstacles, some of which may be overcome with time.

3. Important Considerations for Deep Learning Implementation

Deep learning represents a broad class of techniques, and one of the challenges in applying deep learning is selecting the appropriate DNN type for the task at hand. Here we briefly summarize some of the considerations in developing DNN for a particular application.

Although new deep learning approaches and architectures are increasingly being proposed, most DNNs can be classified into three major categories [14]:

1. *Networks for unsupervised learning:* designed to capture high-order data correlation by identifying jointly statistical distributions with the associated classes when available. Bayes rule can later be used to create a discriminative learning machine[15].

2. *Networks for supervised learning:* designed to provide maximum discriminative power in classification problems and are trained only with labeled data; all outputs must be tagged.[15]

3. *Hybrid or semi-supervised networks:* where the objective is to classify data using the outputs of a generative (unsupervised) model. Normally, data is used to pre-train the network weights to speed up the learning process prior to the supervision stage.

DNN can also be built with a variety of architectures. The most commonly used architectures-- CNNs, Stacked Autoencoders, Deep Belief Networks, and Restricted Boltzmann Machines, are illustrated in Figure 1, with more detailed summaries in the supplementary material (Supp A).

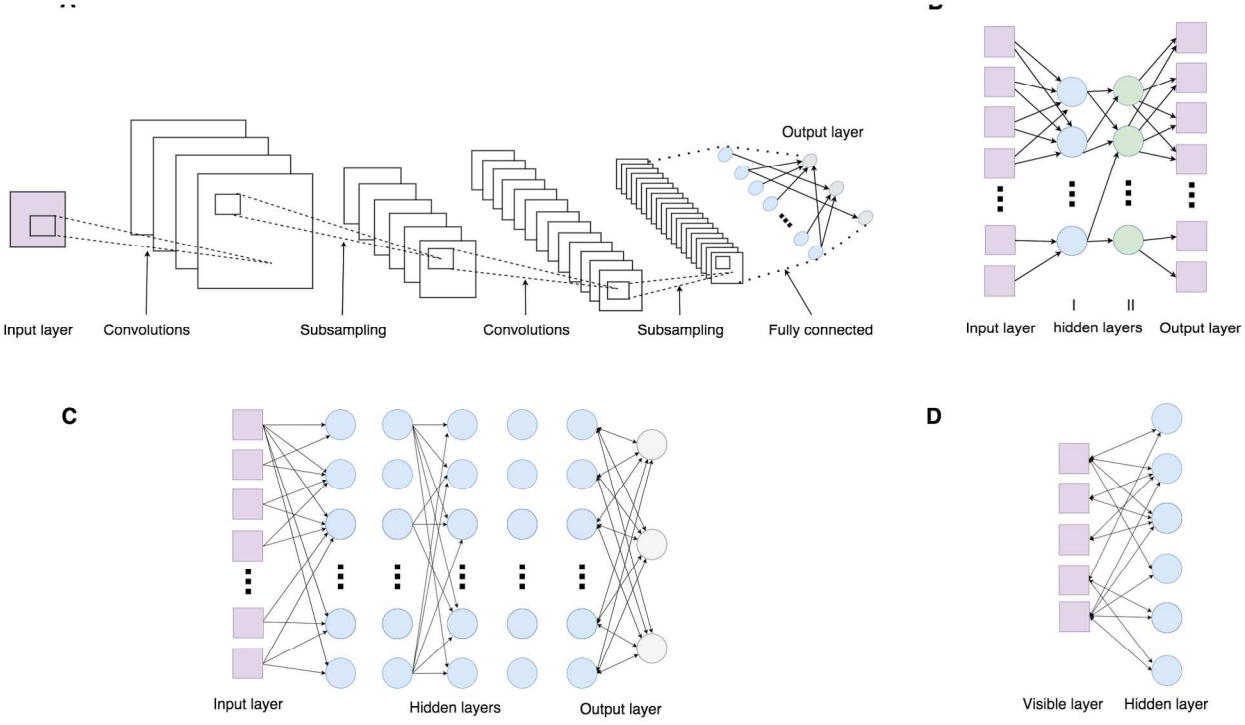


Figure 1. Four of the most popular classes of Deep Learning architectures in biological data analysis **A.** Convolutional neural network (CNN): has several levels of convolutional and subsampling layers optionally followed by fully connected layers with deep architecture. **B.** Stacked autoencoder: consists of multiple sparse autoencoders **C.** Deep Belief Network (DBN): trained layerwise by freezing previous layers weights and feeding the output to the next layer **D.** Restricted Bolzmann Machine architecture: includes one visible layer and one layer of hidden units.

The importance of hyperparameter optimization

Another important consideration of DNNs is their many hyperparameters; these are either architectural, such as layer sizes and transfer functions, optimization types, such as learning rates and momentum values, or regularization, such as the dropout probabilities for each layer or the noise level injected on inputs. Optimization of DNNs is in general extremely challenging due the large number of parameters and non-linearities in the models. Careful fine-tuning of the numerous hyperparameters is one of the most difficult and time consuming tasks in implementing these solutions.

A common approach to optimizing neural networks hyperparameters is using Bayesian optimization to maximize the validation AUC (Area Under the Roc curve) or minimize the loss function. Bayesian optimization is ideally suited for globally optimizing noisy functions while being parsimonious in the number of function evaluations. The method proposed by Snoek et al is very useful as they used Spearmint with warping[16]. Labeled training runs that diverged were considered as constraint violations.

Bayesian optimization assumes that the unknown function is sampled from a Gaussian Process updating the posterior distribution as observations are gathered—in this case, the AUC or the loss function. The hyperparameters of the next iteration are selected through the optimization of the expected improvement, as suggested by Mockus[17], considering the current best result.

4. Deep learning studies and potential applications in biomedicine

Deep learning has already shown success in a variety of biological applications. In this section, we review challenges and opportunities for deep learning in various areas of research and, where possible, review studies that apply deep learning to these problems (Table 1). We first review areas important for biomarker development, including genomics, transcriptomics, proteomics, structural biology and chemistry. We then review the prospects for drug discovery and repurposing, including the use of multi-platform data.

Biomarkers

One important task in biomedicine is the translation of biological data to valid biomarkers that reflect phenotype and physical states, such as disease. Biomarkers are critical in assessing clinical trial outcomes[18] and detecting and monitoring disease, particularly for diseases as heterogeneous as cancer[19,20]. Identification of sensitive specific biomarkers is a great challenge for modern translational medicine[21,22]. Computational biology is an essential tool for biomarker development. Virtually any source of data could be used, from genomics to proteomics; these are discussed in the following section.

Genomics

Next Generation Sequencing (NGS) technology has allowed production of a massive amount of genomics data. Much of the analysis of this data can be performed *in silico* with modern computational approaches. This includes structural annotation of genomes (including non-coding regulatory sequences, protein binding site prediction, and splicing sites).

One important division of genomics is metagenomics, also known as environmental, ecogenomics or community genomics. NGS technology has shed light on the natural diversity of microorganisms that are not cultivated and previously not well studied.

There are several bioinformatic challenges in metagenomics. One major challenge is functional analysis of sequence data and analysis of species diversity. The use of deep belief networks and recurrent neural networks have allowed classification of both metagenomics pH data and human microbiome data by phenotype. These did not improve classification accuracy compared to baseline methods as reinforcement learning, but did provide the ability to learn hierarchical representations of a data set[23]. However, Ditzler et al highlighted that DNN could improve existing algorithms of metagenomics classification especially on large datasets and with proper selection of networks parameters.

Transcriptomics

Transcriptomics analysis exploits variation in the abundance of various types of transcripts (messenger RNA (mRNA), long non-coding RNA (lncRNA), microRNA (miRNA), etc.) to gather a range of functional information, from splicing code to biomarkers of various diseases.

Transcriptomics data are often obtained from different types of platforms (various microarray platforms, sequencing platforms) that differ by the gene set measured and method of signal detection. Many factors contribute to variability of gene expression data. Thus normalization is needed even for single platform analysis. Cross-platform analysis requires normalization techniques, which can be a major challenge[24]. DNNs are particularly well suited for cross-platform analysis because of their high generalization ability. They are also well equipped to handle some of the other major issues with gene expression data, such as the size of the datasets and the need for dimension reduction and selectivity/invariance, and in the following section we review several DNNs that have been used with different types of gene expression data with varying levels of success.

Tabular data applications

One way in which gene expression data can be represented is in tabular form as matrices, which contain quantitative information about transcript expression. These data are high-dimensional, making statistical analysis problematic due to loss of signal to noise in the data[25].

High dimensional data can be handled in two ways:

I. dimensionality reduction:
    A. feature extraction, for instance with SVM or Random Forest algorithms;
    B. feature subset selection;
    C. pathway analysis;
II. use of methods less sensitive to high-dimensionality, such as Random Forest or deep belief networks.

Methods such as Principal Component Analysis (PCA), Singular Value Decomposition, Independent Component Analysis or Non-negative Matrix Factorization are common first front approaches. However, the above-mentioned methods transform the data into a number of components that can be difficult to interpret

biologically. Also, such dimensionality reduction methods extract features based on gene expression profiles regardless of interactions between genes. Pathway analysis allows reduction of the number of variables, reducing error rate and retaining more biologically relevant information[25,26].

Deep learning has also showed some success in handling high-dimensional matrix transcriptomics data. In one alternative approach, features from gene expression were extracted together with regions of non-coding transcripts such as miRNA; this was implemented using deep belief networks and active learning, where deep learning feature extractors were used to reduce the dimensionality of six cancer datasets and outperformed basic feature selection methods[27]. Application of active learning with classification improved accuracy and allowed selection of features related to cancer (improved cancer classification) not solely based on gene expression profile. Feature selection with miRNA data was implemented using relationships with target genes from previously selected subsets of features.

In another deep learning application, Fakoor et al. took advantage of an auto-encoders network for generalization and applied this to cancer classification using microarray gene expression data obtained from a different type of microarray platform (Affimetrix family) with different set of genes[28]. They used a combination of dimensionality reduction through PCA and unsupervised non-linear sparse feature learning, through auto-encoders, to build features for general classification of microarray data. Results for classification of cancer and non-cancer cells showed important improvements, especially using supervised fine tuning, which makes the features less generic but achieves higher classification accuracy even for data without cross-platform normalization. The global generalization ability of auto-encoders facilitates the use of data collected using different microarray technologies and thus may be promising in massive integrative analysis of data from the public domain.

Image processing applications

Gene expression can also be stored in visual forms as images, such as image fluorescence signal from microarray or RNA *in situ* hybridisation fluorescence or radioactive signal. In several applications, CNNs, known for superior performance in image processing, have shown potential in improving the analysis of these images.

In microrarray analysis, detection of a signal and recognition of fluorescence spots can be challenging because of variation in spot size, shape, location or signal intensity, and fluorescence signal intensity often corresponds poorly to gene or sequence expression level. In one application of deep learning techniques to this problem, a CNN was used for microarray image segmentation and demonstrated results in accuracy that resembled baseline approaches in accuracy, but with easier training and fewer requirements of computational sources[29].

Another opportunity for the application of CNNs to image-based gene expression data has been RNA *in situ* hybridization, a tedious technique that enables localization and visualization of gene expression in a group of cells, tissue slice, or whole organism when such manipulations are allowed. This method facilitates powerful longitudinal studies that illustrate changes in expression patterns during development. It was implemented for construction of the detailed Allen Developing Mouse Brain Atlas, which contains expression maps for more than 2000 genes, each illustrated in multiple brain sections. In the past, these were annotated manually, which was time-consuming, expensive and at times inaccurate. Recently, however, Zeng et al. performed automatics

annotation using a deep pre-trained CNN[30]. To do this, neural network models were trained on raw natural *in situ* hybridization images of different levels of developing brain without exact information about coordinate (spatial information); this technique achieved superior accuracy at multiple brain levels throughout four stages of development.

Splicing

Yet another area of application of deep learning is splicing. Splicing is one of the major factors that provides biological diversity of proteins in eukaryotic organisms; moreover, recent studies show a connection between "splicing code" and various diseases[31]. However, modern science is still not able to provide a comprehensive understanding of the mechanisms controlling splicing regulation. The modern concept of splicing regulation includes transcript level, presence of specific signaling regulatory sequence elements (splicing enhancers or silencers), structure of splice site, and state of splicing factors (e.g. phosphorylation of specific sites could change splicing factor activity). All of these factors complicate the analysis because of the huge number of elements and complex non-linear interactions between them. Existing splicing prediction software requires as input high-throughput sequencing data and is faced with the problem of raw reads, which are shorter than regular genes, along with high duplication level and presence of pseudogenes in genomes. Thus algorithms for analysis of splicing mechanisms are slow, requiring high computational sources for combinatorics, and deep learning may offer improvement in this respect. In one deep learning application using five tissue-specific RNA-seq datasets, a DNN was developed using hidden variables for features in both genomic sequences and tissue types and was shown to outperform Bayesian methods in predicting tissue splicing within individuals and across tissues, specifically the change of the percentage of transcripts with an exon spliced (PSI), a metric for splicing code[32].

Non-coding RNA

Non-coding RNA is another problem in biology that will require sophisticated computational methods like deep learning. Non-coding RNAs are surprisingly important, involved in the regulation of transcription, translation, and epigenetics[33], but they remain difficult to differentiate from protein-coding RNAs. This task is well resolved for short non-coding RNAs, but is quite challenging for lncRNA. lncRNAs make up a heterogeneous class that may contain putative origins of replication (ORF), short protein-like sequences. A novel deep learning approach, called IncRNA-MFDL, was developed to identify lnc-RNAs, using a combination of multiple features such as ORF, k neighboring bases, secondary structure, and predicted coding domain sequences[34]. This approach used as input five individual features extracted from sequences data from Gencode (IncRNA) and Refseq (protein coding mRNA data) and resulted in 97.1% prediction accuracy on the human dataset.

Expression Quantitative Trait Loci analysis

Finally, there is potential for deep learning in quantitative trait loci (QTL) analysis. QTL analysis identifies genetic loci containing polymorphisms that contribute to phenotypic variation of complex, polygenic traits (e.g. body weight, drug response, immune response). One such "trait" showing genetic variation is expression, or transcript abundance for any given gene in a given tissue and/or condition. Expression QTL (eQTL) are loci that influence genetic variation in transcript abundance. eQTL analysis has led to insights into the regulation of human gene expression, but is faced with a number of challenges. eQTL that regulate

expression locally (cis-eQTL) are relatively easy to identify with a limited number of statistical tests, but loci that regulate expression of genes elsewhere in the genome (trans-eQTL) are more difficult to detect. A deep learning approach, MASSQTL, was recently implemented to solve the problem of trans-eQTL prediction using various encoded biological features, such as physical Protein Interaction Networks, gene annotation, evolutionary conservation, local sequence information and different functional elements from the ENCODE project[35]. The DNN utilized 9 DNN models from their respective cross-validation folds, outperformed other machine learning models, and provided new insight into the mechanisms underlying the regulatory architecture of gene expression. Deep Decoding systems were also employed to cluster trans-eQTL feature vectors and then visualised through t-SNE dimensionality reduction technique.

Proteomics

Compared to transcriptomics, proteomics is a much less developed area of research, with data still scarce and fewer computational approaches available for analysis. The lack of human proteomics data and difficulty of translation of results from model organisms to humans also complicates analysis, even if similar signal-encoding and transmitting mechanisms are in place.

Deep learning can benefit proteomics in several ways, as some approaches do not require a large number of training cases as other machine learning algorithms. Other strengths of deep learning methods are that they build hierarchical representations of data and learn general features from complex interactions, thus benefiting proteomic and network analysis of proteins. For example, Bimodal deep belief networks have been used to predict the human cellular response to stimuli from rats' cellular response to the same stimuli, using phosphorylation data[36]. The developed algorithm achieved considerable accuracy compared to classical pipelines.

Structural Biology and Chemistry

Structural biology includes analysis of protein folding, protein dynamics, molecular modeling, and drug design. Secondary and tertiary structures are important features of proteins and RNA molecules. For proteins, proper structure determination is important for enzymatic function prediction, formation of catalytic centers and substrate binding, immune function (antigen binding), transcriptional factors (DNA binding), and post-transcriptional modifications (RNA binding). Loss of proper structure leads to loss of function and, in some cases, aggregation of abnormal proteins which can lead to neurodegenerative diseases such as Alzheimer's or Parkinson's[37].

Comparative modeling, based on compound homology, is one possible way to predict protein secondary structure, but is limited by the amount of existing well-annotated compounds. Machine learning *de novo* prediction, on the other hand, is based on recognized patterns of compounds with well-known structure, but has not been accurate enough to be of practical use. Employing deep learning methods *de novo* has improved structure prediction by using protein sequencing data[38]. Similarly, deep learning has been applied to predict contacts and orientations between secondary structure elements and amino acid residues using ASTRAL database data and a complex, three stage approach[39]. The methods used were effective tools in analyzing biased and highly variable data.

The constancy of three-dimensional structure is also functionally important. However, there are several proteins with no unique structure that are involved in fundamental biological processes, such as control of cell cycle, regulation of gene expression, and molecular signal transmission. Moreover, recent studies show markedness of some disordered proteins[37]; many oncogene proteins have unstructured domains, and abnormal aggregations of misfolded proteins lead to disease development[40]. Such proteins, without fixed three-dimensional structure, are called Intrinsically Disordered Proteins (IDP), and domains without constant structure are called Intrinsically Disordered Regions (IDR).

Many parameters distinguish IDP/IDR from structured proteins, thus making the prediction process challenging. This issue can be resolved using deep learning algorithms that are able to consider a wide variety of features. In 2013, Eickholt and Cheng published a sequence-based deep learning predictor, DNdisorder, that improved prediction of disordered proteins compared to state-of-the-art predictors[41]. Later, in 2015, Wang et al. presented a new method, DeepCNF, that enables accurate prediction of multiple parameters such as IDPs or proteins with IDR, using experimental data from Critical Assessment of protein Structure Prediction (CAPS9 and CASP10). Through utilization of numerous features, the DeepCNF algorithm performed better than baseline single *de novo (ab intio)* predictors[42].

Another important class of proteins is that of RNA-binding proteins, which bind single or double stranded RNA. These proteins participate in all kinds of post-transcriptional modifications of RNA: splicing, editing, regulation of translation (protein synthesis), and polyadenylation. RNA molecules form different types of arms and loops, secondary and tertiary structures required for recognition and formation of the connection between RNA and protein. Secondary and tertiary structures of RNA are predictable and have been used for modeling structural binding preferences and predicting binding sites of RBPs by applying deep belief networks[43]. The deep learning framework was validated on the real CLIP-seq (cross-linking immunoprecipitation high-throughput sequencing) datasets to show the ability to extract hidden features from both raw sequence and structure profiles and to accurately predict RBP's sites.

Drug Discovery and Repurposing

Computational drug biology and biochemistry are broadly applied on almost every stage in drug discovery, development and repurposing. A huge number of computational approaches for *in silico* drug discovery and target extension have been developed worldwide by different research groups and companies in past decades to reduce time and resource consumption. While many methods exist[44], none are yet optimal (inability to perform throughput screening or limitation by class of proteins, for example) and several studies now show that deep learning is an important approach to consider (Table 1).

One of the important tasks in drug discovery is prediction of drug-target interaction. Targets (proteins) often have one or more binding sites with substrates or regulatory molecules; these can be used for building prediction models. However, including other protein sites could bring bias into the analysis. The ability of pairwise input neural network (PINN) to accept two vectors with features obtained both from protein sequences and target profiles was used by Wang et al. to compute target-ligand interaction[45]. This advantage of NNs resulted in a better accuracy than other representative target-ligand interaction prediction methods.

Drug discovery and evaluation is expensive, time-consuming, and risky; computational approaches and various prediction algorithms can help reduce risks and save resources. One potential risk is toxicity; for

example, liver toxicity (hepatotoxicity) is a frequent cause of removal of a drug from production. Prediction of hepatotoxicity with computational approaches could help to avoid likely hepatotoxic drugs. Using deep learning, it is possible to effectively determine compound toxicity with raw chemical structure without requiring a complex encoding process[46]. Using CNNs, it is also possible to predict properties such as epoxidation, which means high reactivity and possible toxicity; this was first implemented by Hughes et al. by using Simplified Molecular Input Line Entry Specification (SMILES) format data of epoxidized molecules and hydroxides molecules, as a negative control[47].

Multi-platform data (multi-omics)

The ability to work with multi-platform data is a major advantage of deep learning algorithms. Since biological systems are complex, with multiple interrelated elements, the systems level integration of genomics, epigenomics, and transcriptomics data is key to extracting the most valid, biologically meaningful results. The integration process is not computationally trivial, but the payoff is a gain in biomarker specificity and sensitivity over single-source approaches.

One of the major fields in computational biology that requires analysis of combined data is computational epigenetics. Only joint analysis of genome, transcriptome, methylome characteristics and histone modifications provides accurate epigenome predictions.

Several investigators have developed deep learning approaches useful in analyzing data from multiple sources (Table 1). Alipanahi et al. developed the deep learning-based approach DeepBind, (tools.genes.toronto.edu/deepbind/), to calculate the ability of nucleotide sequences to bind transcription factors and RNA-binding proteins and characterize the effects of single point mutations on binding properties in various diseases. DeepBind software was inspired by CNN, so is not sensitive to technology; rather, it is compatible with qualitatively different forms of data, from microarray to sequences. Implementation of CPU also allows users to parallelize the computational processes[48]. In another CNN-based application, Zhou and Troyanskaya designed the DeepSEA framework for prediction of chromatin features and evaluation of disease-associated sequence variants. Unlike other computational approaches, their algorithm is capable of capturing large-scale context sequence information for each binding site, for annotation of *de novo* sequence variants[49] Kelley et al developed analogous CNN pipelines, which reveal effects of sequence variants on chromatin regulation, with training and testing on DNase-seq (DNAse I sequencing) data[50]. A deep learning software, called Bassed, outperformed baseline approaches and acheived a mean AUC 0.892 over all datasets. Finally, deep learning has been used to identify active enhancers and promotors with the development of a model called Deep Feature Selection, which harnesses the ability of DNN to model complex nonlinear interactions and learn high-level generalized features[51]. A model selects features from multi-platform data and ranks them by importance. In each of these applications, deep learning methods were more sensitive and powerful predictors of chromatin properties and key to the development of complex biomarkers.

Cancer is a broad name for a group of heterogeneous diseases, some of which are caused by genetic mutations, and as such cancer classification using multi-platform data could shed light on underlying pathology. Liang et al. developed a deep belief network model with multi-platform data for clustering cancer patients[52]. Restricted Boltzmann machines were applied to encode features defined by each input modality.

One advantage of this approach is that deep belief networks do not require data with a normal distribution, as other clustering algorithms, and genetic (biological) data is not normally distributed

Finally, from the point of view of Natural Language Processing, deep learning could be very useful in navigating through the immense unstructured (research publications and patents) and structured data (knowledge annotated graphs, like Gene Ontology[53] or Chembl[54]) by testing the plausibility of hypotheses. These databases together form a massive, multiplatform masterset of data that would be much more information-rich and comprehensive if combined.

5. Challenges and Limitations of Deep Learning Systems

While deep learning algorithms have demonstrated advantages in recognition, classification and feature extraction from complex and noisy data, these methods have also some limitations that should be considered compared with traditional machine learning methods. These include:

1. The "black box" problem
2. Overfitting and the need for large training datasets
3. The selection problem in choosing a type of DNN
4. The high computational costs of training

1. The "black box." One of the major limitations of deep learning in biological context relates to quality control and interpretation. Most DNNs are "black boxes" that learn by simple asociations and co-occurrences. They have limited means with which to interpret the representations, although some, like CNNs, are also very powerful in creating high level representations. When working with image, voice or textual data, developers can rapidly test classification performance and evaluate the quality of the output data. High dimensional biological data, on the other hand, is not easy for humans to interpret and requires additional quality control and interpretation pipelines. DNNs thus lack the transparency and interpretability of other methods and are unable to uncover complex causal and structural relationships common in biology without human input.

2. The need for large datasets. Another limitation is the requirement of large training data sets that may not be readily available. Many commonly-used machine learning methods outperform DNNs where experimental data is scarce. When datasets are not sufficiently large, one of the major challenges with training DNNs is dealing with the risk of overfitting, i.e., when training error is low but the test error is high, thus the model fails to learn a proper generalization of knowledge contained in data. There are ways to regularize the DNN, such as dropout, i.e., temporal removal of a random subset of units with their connections, which reduces conspiracy between units, but overfitting is often still a threat within small biological data sets, especially with unbiased and noisy data.

3. The selection problem. With many types of DNNs available, task-appropriate selection is not always straightforward. While there are some tools to aid selection, such as hyperparameter optimization techniques[15], the pipeline architecture is significantly more complex than other machine learning methods, and new architectures are increasingly being proposed.

4. The computation costs. Finally, while DNNs require few computational resources when trained, the training process is usually computationally-intensive, time consuming and often requires access to and programming knowledge for graphics processing units (GPU). Tensorflow (http://tensorflow.org), a recent framework open sourced by Google and inspired in Theano, greatly simplifies the implementation and debugging of deep architectures. This and other frameworks are summarized in Supp. B.

Thus despite the current enthusiasm for deep learning, traditional models still play an important role in –omics, especially when the amount of data is not very large or when the number of variables is large and non-numeric support vector machines or ensemble methods, like Random Forest, may be a better option.

Discussion and future perspectives

DNNs have the potential to benefit a wide range of biological research applications including annotation, semantic linking and even interpretation of complex biological data in areas including biomarker development, drug discovery, drug repurposing and clinical recommendations. One area where DNNs can have major impact is transcriptomic data analysis. Several million samples of human transcriptomic data are available from almost two decades of experiments, residing in multiple repositories, including GEO, ArrayExpress, ENCODE and TCGA, as well as cell line data available from the Broad Institute Connectivity Map and LINCS projects (Figure 2).

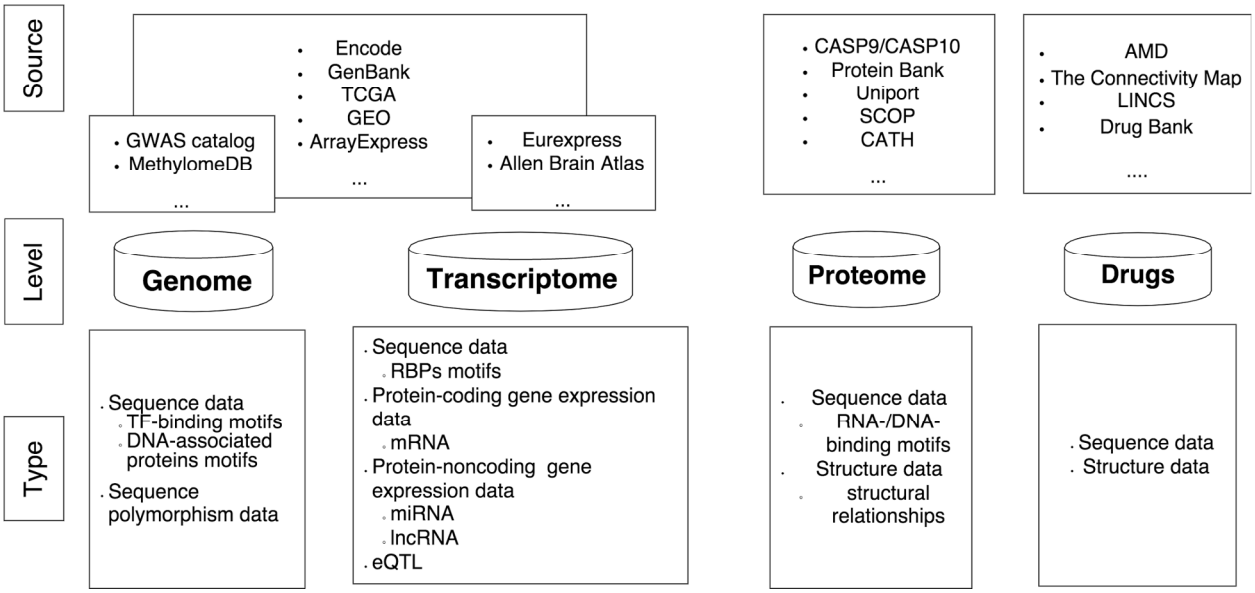


Figure 2. Potential sources of biological data that could be fed into DNN. Biological data abundant with millions of human samples available from almost two decades of experiments residing in multiple repositories, containing various types of data, from genetic variants (GWAS catalog) to transcriptome profiles of the response to drugs (LINCS and The Connectivity Map projects).

Deep learning algorithms may also benefit drug discovery, not only from the structural side (e.g. by increasing the accuracy in the challenging drug ligand target problem), but also, and perhaps more importantly, from the abstract representations learned by upper layers of CNNs in guiding formulation of new hypotheses.

Here, we covered several applications of deep learning methods, including CNN, deep belief networks, auto-encoders, and recurrent neural networks, involving a variety of analytical tasks in biomedicine. A combination of these approaches using supervised, unsupervised and reinforcement learning applied to disparate types of biomedical data already play a key role in understanding fundamental biomedical processes and have helped lead to the development of personalized, or stratified, medicine. CNNs outperformed baseline approaches not only in classical deep learning applications, such as image recognition (microarray segmentation and gene expression annotation), but also in annotation of sequence polymorphism data in tools such as DeepSEA, DeepBind and Basset. The above mentioned tools are successfully used as frameworks in the public domain (DeepSEA- http://deepsea.princeton.edu/job/analysis/create/ or Basset- https://github.com/davek44/Basset). Deep belief networks, as a universal technique that can also be configured to avoid overfitting, can be applied to various types of biomedical data (from structural to gene expression). Auto-encoders, with their ability to learn flexible and rich representations of data, can be successfully used for feature extraction and dimensionality reduction and as independent classifiers of gene expression data. Restricted boltzmann machines, perhaps due to relative ease in training, have been applied primarily in early works with structural data.

DNN can deliver substantial improvements compared to traditional machine learning methods, e.g. nearest neighbours, boosted trees, or support vector machines, but usually require much larger data sets—sometimes surpassing tens of thousands of samples—and availability of high-performance GPU-based computational resources. They also require a level of experience in computer science and mathematics not commonly available in organizations working with biological data. Another factor impeding mainstream use of DNNs in biomedicine is communication of research results.

One of the aims of this review was to introduce the concept of deep learning, a new topic for many biologists, and plant a seed of interest for those working with this data as to how deep learning may offer solutions to some of their data analysis problems. We see a lot of value in increased collaboration between biologists and the computational biology community or larger deep learning community. For readers from the machine-learning community, we aimed to review existing problems with biological data and open a discussion about the many opportunities biological data offers for application of this approach, including obstacles to overcome and many possible directions moving forward.

Conclusions

In conclusion, the massive scale of modern biological data is simply too large and complex for human-led analysis. Machine learning and particularly deep learning, combined with human expertise, is the only approach that stands a chance at fully integrating the multiple, huge repositories of multi-platform data. Deep learning has enabled humans to do what was previously unimagined—image recognition with millions of inputs, voice recognition and speech automation that approaches human abilities. While deep learning and particularly unsupervised deep learning is still in its infancy, particularly in biological applications, initial studies support it as a promising approach that, while not free of limitations and challenges in implementation, may overcome some of the problems with biological data and lead to new insights into the millions of indirect and interconnected mechanisms and pathways that underlie disease.

Acknowledgements

We would like to thank Dr. Leslie C. Jellen, Dr. Quentin Vanhaelen, Dr. Qingsong Zhu and Dr. Andrew Kazennov of Artificial Intelligence Research department at Insilico Medicine for reviewing and editing the manuscript.

References

(1) EMBL-European Bioinformatic Institute. EMBL-EBI Annual Scientific Report 2014. **2014**, 142.

(2) Greene, C. S.; Troyanskaya, O. G. Chapter 2: Data-Driven View of Disease Biology. *PLoS Comput. Biol.* **2012**, *8* (12), e1002816.

(3) Nussinov, R. Advancements and Challenges in Computational Biology. *PLoS Comput. Biol.* **2015**, *11* (1), e1004053.

(4) Libbrecht, M. W.; Noble, W. S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16* (6), 321–332

(5) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444.

(6) Bengio, Y.; LeCun, Y. Scaling Learning Algorithms towards AI. *Large Scale Kernel Mach.* **2007**, No. 1, 321–360.

(7) Bengio, Y.; Delalleau, O.; Simard, C. Decision Trees Do Not Generalize To New Variations. *Comput. Intell.* **2010**, *26* (4), 449–467.

(8) Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks. *Cvpr* **2014**, 1717–1724.

(9) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Graves, A.; Riedmiller, M.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518* (7540).

(10) Gatys, L. A.; Ecker, A. S.; Bethge, M.; Sep, C. V. A Neural Algorithm of Artistic Style. 3–7.

(11) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. **2014**, 1–88.

(12) Solovyeva, K. P.; Karandashev, I. M.; Zhavoronkov, A.; Dunin-Barkowski, W. L. Models of Innate Neural Attractors and Their Applications for Neural Information Processing. **2015**.

(13) (1) Baralis, E.; Fiori, A. Exploring Heterogeneous Biological Data Sources. In *2008 19th International Conference on Database and Expert Systems Applications*; IEEE, 2008; pp 647–651.

(14) Bengio, Y.; Goodfellow, I. J.; Courville, A. *Deep Learning*; 2015.

(15) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, 1–9.

(16) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. **2012**, 1–12.

(17) Mockus, J.; Tiesis, V.; Zilinskas, A. The application of Bayesian methods for seekeng the extremum. Towards Global Optimization. **1978,** *2*: 117-129.

(18) Bakhtiar, R. Biomarkers in Drug Discovery and Development. *J. Pharmacol. Toxicol. Methods* **2008**, *57* (2), 85–91.

(19) Lezhnina, K.; Kovalchuk, O.; Zhavoronkov, A. A.; Korzinkin, M. B.; Zabolotneva, A. A.; Shegay, P. V; Sokov, D. G.; Gaifullin, N. M.; Rusakov, I. G.; Aliper, A. M.; Roumiantsev, S. A.; Alekseev, B. Y.; Borisov, N. M.; Buzdin, A. A. Novel Robust Biomarkers for Human Bladder Cancer Based on Activation of Intracellular Signaling Pathways. *Oncotarget* **2014**, *5* (19), 9022–9032.

(20) Shepelin, D.; Korzinkin, M.; Vanyushina, A.; Aliper, A. Molecular Pathway Activation Features Linked with Transition from Normal Skin to Primary and Metastatic Melanomas in Human. **2015**.

(21) Borisov, N. M.; Terekhanova, N. V; Aliper, A. M.; Venkova, L. S.; Smirnov, P. Y.; Roumiantsev, S.; Korzinkin, M. B.; Zhavoronkov, A. A.; Buzdin, A. A. Signaling Pathways Activation Profiles Make Better Markers of Cancer than Expression of Individual Genes. *Oncotarget* **2014**, *5* (20), 10198–10205.

(22) Brooks, J. D. Translational Genomics: The Challenge of Developing Cancer Biomarkers. *Genome Res.* **2012**, *22* (2), 183–187.

(23) Ditzler, G.; Polikar, R.; Member, S.; Rosen, G.; Member, S. Multi-Layer and Recursive Neural Networks for Metagenomic Classification. **2015**, *14* (6), 608–616.

(24) Järvinen, A.-K.; Hautaniemi, S.; Edgren, H.; Auvinen, P.; Saarela, J.; Kallioniemi, O.-P.; Monni, O. Are Data from Different Gene Expression Microarray Platforms Comparable? *Genomics* **2004**, *83* (6), 1164–1168.

(25) Hira, Z. M.; Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv. Bioinformatics* **2015**, *2015* (1), 198363.

(26) Buzdin, A. A.; Zhavoronkov, A. A.; Korzinkin, M. B.; Roumiantsev, S. A.; Aliper, A. M.; Venkova, L. S.; Smirnov, P. Y.; Borisov, N. M. The OncoFinder Algorithm for Minimizing the Errors Introduced by the High-Throughput Methods of Transcriptome Analysis. *Front. Mol. Biosci.* **2014**, *1* (August), 8.

(27) Ibrahim, R.; Yousri, N. A.; Ismail, M. A.; El-Makky, N. M. Multi-Level gene/MiRNA Feature Selection Using Deep Belief Nets and Active Learning. *Eng. Med. Biol. Soc. (EMBC), 2014 36th Annu. Int. Conf. IEEE* **2014**, 3957–3960.

(28) Fakoor, R.; Huber, M. Using Deep Learning to Enhance Cancer Diagnosis and Classification. *Proceeding 30th Int. Conf. Mach. Learn. atlanta, Georg.* **2013**, *28*.

(29) Jones, A. L. *Segmenting Microarrays with Deep Neural Networks*; 2015.

(30) Zeng, T.; Li, R.; Mukkamala, R.; Ye, J.; Ji, S. Deep Convolutional Neural Networks for Annotating Gene Expression Patterns in the Mouse Brain. *BMC Bioinformatics* **2015**, *16* (1), 147.

(31) Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, Brendan J. Frey, "The human splicing code reveals new insights into the genetic determinants of disease", Vol. 347, No. 6218, Science (2015).

(32) Leung, M. K. K.; Xiong, H. Y.; Lee, L. J.; Frey, B. J. Deep Learning of the Tissue-Regulated Splicing Code. *Bioinformatics* **2014**, *30* (12), i121–i129.

(33) Cech, T. R.; Steitz, J. A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones.pdf. *Cell* **2014**, *157* (1), 77–94.

(34) Fan, X.-N.; Zhang, S.-W. lncRNA-MFDL: Identification of Human Long Non-Coding RNAs by Fusing Multiple Features and Using Deep Learning. *Mol. BioSyst.* **2015**, *11* (3), 892–897.

(35) Witteveen, M. J. Identification and Elucidation of Expression Quantitative Trait Loci (eQTL) and Their Regulating Mechanisms Using Decodive Deep Learning. **2014**, 1–17.

(36) Chen, L.; Cai, C.; Chen, V.; Lu, X. Trans-Species Learning of Cellular Signaling Systems with Bimodal Deep Belief Networks. **2015**, 1–8.

(37) Ross, C. A.; Poirier, M. A. Opinion: What Is the Role of Protein Aggregation in Neurodegeneration? *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (11), 891–898.

(38) Spencer, M.; Eickholt, J.; Cheng, J. A Deep Learning Network Approach to *ab Initio* Protein Secondary Structure Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2015**, *12* (1), 103–112.

(39) Di Lena, P.; Nagata, K.; Baldi, P. Deep Architectures for Protein Contact Map Prediction. *Bioinformatics* **2012**, *28* (19), 2449–2457.

(40) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically Disordered Proteins: Regulation and Disease. *Curr. Opin. Struct. Biol.* **2011**, *21* (3), 432–440.

(41) Eickholt, J.; Cheng, J. DNdisorder: Predicting Protein Disorder Using Boosting and Deep Networks. *BMC Bioinformatics* **2013**, *14* (1), 88.

(42) Wang, S.; Weng, S.; Ma, J.; Tang, Q. DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields. *Int. J. Mol. Sci.* **2015**, *16* (8), 17315–17330.

(43) Zhang, S.; Zhou, J.; Hu, H.; Gong, H.; Chen, L.; Cheng, C.; Zeng, J. A Deep Learning Framework for Modeling Structural Features of RNA-Binding Protein Targets. *Nucleic Acids Res.* **2015**, gkv1025.

(44) Schirle, M.; Jenkins, J. L. Identifying Compound Efficacy Targets in Phenotypic Drug Discovery. *Drug Discov. Today* **2015**, *00* (00)

(45) Wang, C.; Liu, J.; Luo, F.; Tan, Y. Pairwise Input Neural Network for Target-Ligand Interaction Prediction. **2014**, 67–70.

(46) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, 151013124508007.

(47) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci.* **2015**, *1* (4), 168–180.

(48) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **2015**, No. November 2014, 1–9.

(49) Zhou, J.; Troyanskaya, O. G. Predicting Effects of Noncoding Variants with Deep Learning–based Sequence Model. *Nat. Methods* **2015**, *12* (10), 931–934.

(50) Kelley, D. R.; Snoek, J.; Kelley, D. R. Basset : Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks . **2015**.

(51) Li, Y.; Chen, C.-Y.; Wasserman, W. Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. In *Research in Computational Molecular Biology SE - 20*; Przytycka, T. M., Ed.; Lecture Notes in Computer Science; Springer International Publishing, 2015; Vol. 9029, pp 205–217.

(52) Liang, M.; Li, Z.; Chen, T.; Zeng, J. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2015**, *12* (4), 928–937.

(53) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25* (1), 25–29.

(54) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2015**, gkv1253 – .

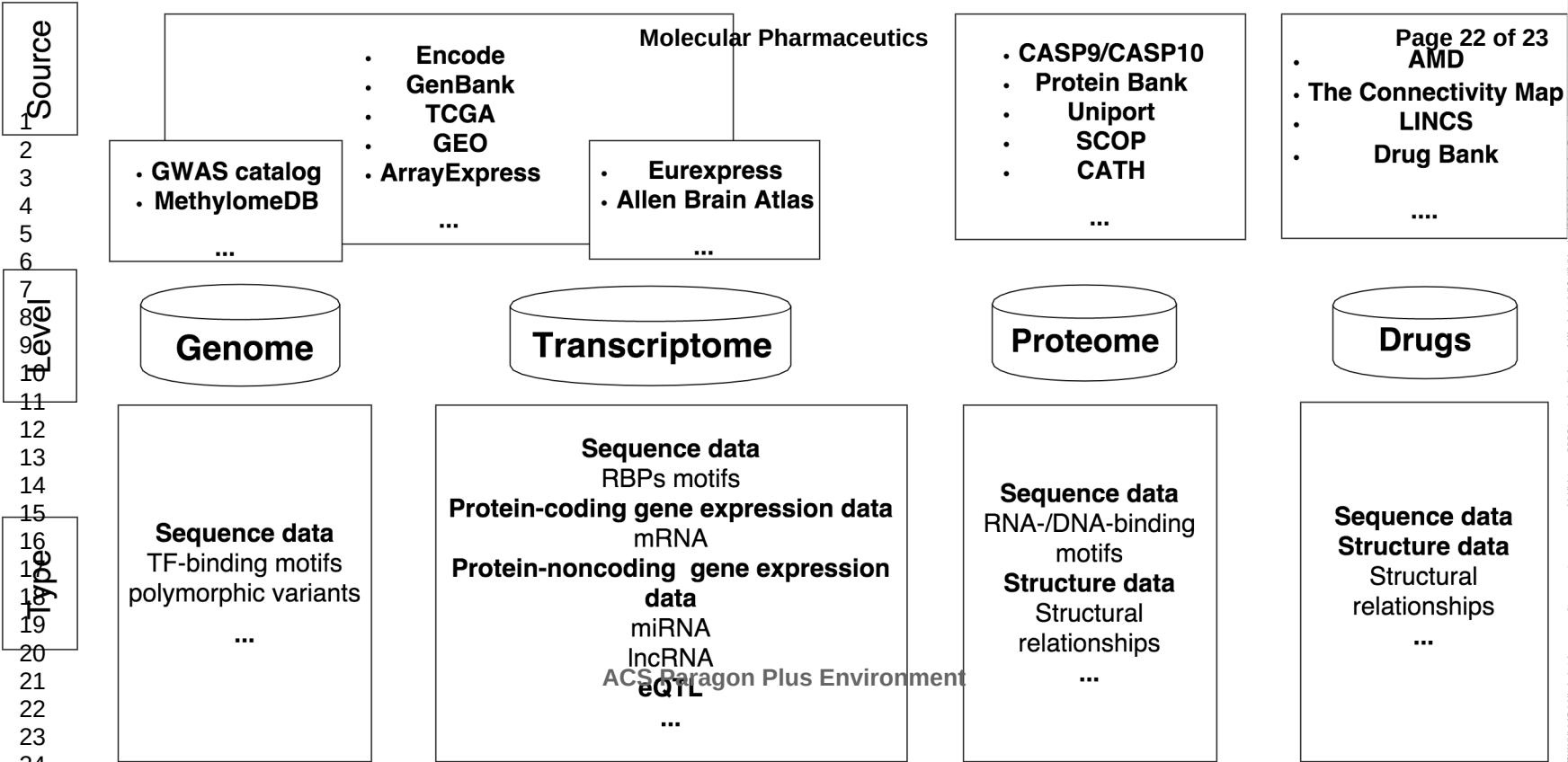
Table 1. Summary of deep learning techniques applied to different types of biomedical data

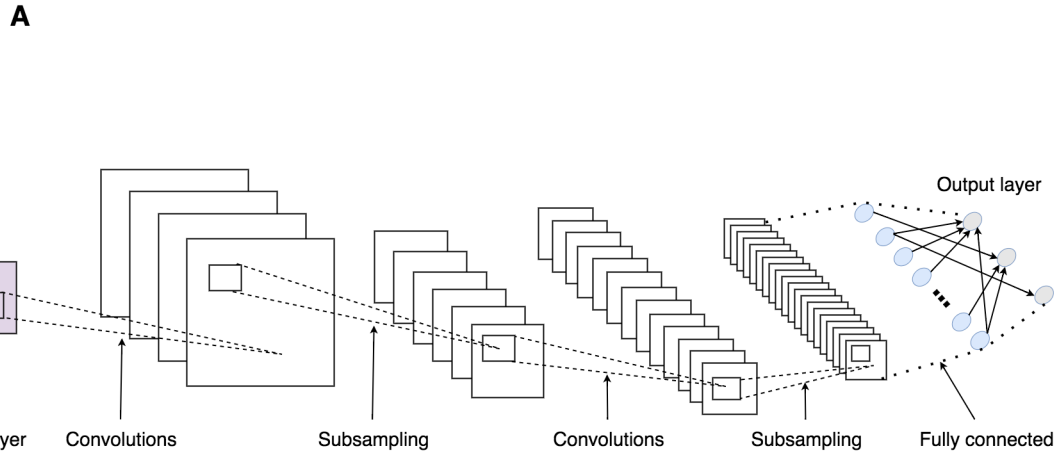
Application	Source of data	Research aim	DL techniques	Accuracy
Using deep learning to enhance cancer diagnosis and classification[23]	13 different gene expression datasets of cancers	cancer detection, cancer type classification	Sparse & Stacked Autoencoders + Softmax Regression	For each dataset accuracy better than baseline
Deep learning of the tissue-regulated splicing code[30]	11 019 mouse alternative exons profiled from RNA-Seq data	splicing patterns recognition	Autoencoders + DNN (3 layers) + spearmint (hyperparameter selection)	AUC better than baseline accuracy
Deep CNNs for annotating gene expression patterns in the mouse brain[28]	*ISH* images of four developing stages of mouse brain by Allen Institute for Brain Science	gene expression annotation	CNN (OverFeat)	AUC=0.894
Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach[50]	ovarian and breast cancer datasets	clustering cancer patients	DBNs	-
lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning[32]	protein-coding and non-coding sequences from Gencode and RefSeq	identification of long non-coding RNAs	IncRNA-MFDL (Deep Stacked network, each unit DNN)	ACC=97.1%
Multi-Layer and Recursive Neural Networks for Metagenomic Classification[21]	pH microbiome sequencing dataset and human microbiome sequencing dataset	metagenomic classification	MLP, DBN, RNN	Comparison
Multi-Level Gene/MiRNA Feature Selection using Deep Belief Nets and Active Learning[25]	MiRNA expression data from 6 type of cancers	Gene/MiRNA Feature Selection (gene expression)	MLFS (DBN + Feature Selection + Unsupervised Active Learning)	F1=84.7%
Pairwise Input Neural Network for Target-Ligand Interaction Prediction[43]	sc-PDB database (sc-pdb: a database for identifying variations and multiplicity of 'druggable' binding	protein-ligand prediction	PINN (SVD + Autoencoder/RBM)	AUC=0.959

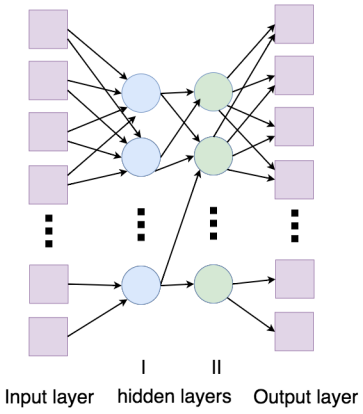
	sites in proteins)			
Predicting effects of noncoding variants with deep learning–based sequence model[47]	690 TF binding profiles for 160 different TFs, 125 DHS profiles and 104 histone-mark profiles from ENCODE and Roadmap Epigenomics projects	predict the noncoding-variant affects de novo from sequence	DeepSEA (CNN)	AUC=0.923 (histone)
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning[46]	506 ChIP-seq experiments DREAM5 TF-DNA Motif Recognition Challenge	classification of specificities of DNA- and RNA-binding proteins	DeepBind (CNN)	train: AUC=0.85 validation: AUC>0.7
Trans-species learning of cellular signaling systems with bimodal DBNs[34]	phosphoproteomic data from SBV IMPROVER challenge	Trans-species learning (simulate cellular signaling systems)	bDBN (Bimodal DBN) and sbDBN (Semi-Restricted Bimodal DBN)	AUC=0.93
Identification and Elucidation of eQTL and their regulating mechanisms using Decodive Deep Learning[33]	GEUVADIS (combination of RNA-Seq and Whole Genome Wide SNP-Array data from a selection of 337 lymphoblastoid cell-lines extracted from individuals participating in the 1000 Genomes Project)	identification of eQTL	MASSQTL (DNN)	AUC=0.85
A deep learning framework for modeling structural features of RNA-binding protein targets[41]	24 datasets derived from doRiNA (database of RNA interactions in post-transcriptional regulation)	predicting binding sites of RNA-binding proteins (RBP target recognition)	DBN (Multimodal DBNs)	AUC=0.983 on PTB HITS-CL
DeepCNF-D: Predicting Protein Order/Disorder Regions by	CASP9, CASP10 datasets from CASP (Critical	Predicting protein order/disorder Regions	DeepCNF (CRF+CNN)	AUC=0.855 on CASP9

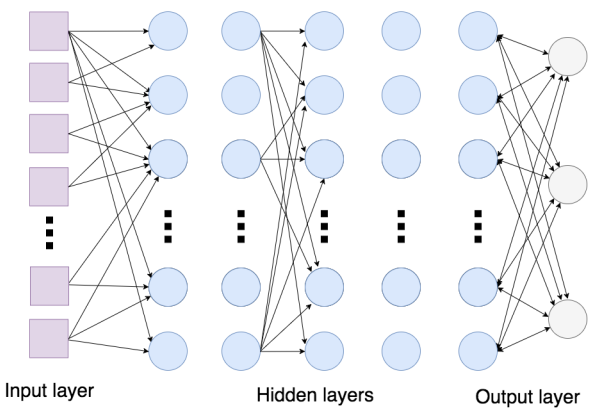
Weighted Deep Convolutional Neural Fields[40]	Assessment of protein Structure Prediction)			AUC=0.898 on CASP10
Segmenting Microarrays with Deep Neural Networks[27]	Two datasets microarray images from Lehmussola et al. 2006	microarray segmentation	CNN	MAE=0.25
Deep Learning for Drug-Induced Liver Injury[44]	Four datasetsща compounds chemical structure annotated DILI-positive or DILI-negative properties	drug-induced liver Injury prediction	RNN (Recursive Neural Network)	AUC=0.955
A Deep Learning Network Approach to *ab initio* Protein Secondary Structure Prediction[36]	train: Protein Data Bank val: CASP9, CASP10 (Critical Assessment of protein Structure Prediction)	*ab initio* protein secondary structure predictions	DNSS (Multimodal RBMs)	Q3=90.7% Sov=74.2%
Deep architectures for protein contact map prediction[37]	ASTRAL database	protein contact map prediction	RNN + DNN	ACC~30%
Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network[45]	Accelrys Metabolite Database (AMD): 389 epoxidized molecules 811 non-epoxidized molecules	modeling epoxidation properties of molecules	CNN	AUC better than baseline accuracy
DNdisorder: predicting protein disorder using boosting and deep networks[39]	DISORDER723, CASP9, CASP10	predicting protein order/disorder Regions	RBM	AUC better than baseline accuracy
Basset: Learning the regulatory code of the accessible genome with deep CNNs[48]	DNAsel-seq data of 164 cell types from ENCODE and Epigenomics Roadmap projects	learn functional activities of DNA sequences	CNN	AUC=0.892

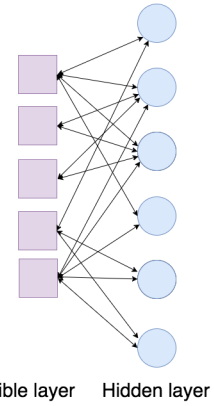
Glossary of Acronyms: CNN=convolutional neural networks, DNN=deep neural network, RNN=recursive neural network, DBN=deep belief network, RBM=restrictive boltzmann machine, MLP=multilayer perceptron, MLFS=multi-level feature selection, PINN=pairwise input neural network, CRF= Conditional Random Fields




A
Output layer
Input layer
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected


B
I
II
Input layer
hidden layers
Output layer


C
Input layer
Hidden layers
Output layer


D
Visible layer
Hidden layer