APPLICATIONS OF DEEP LEARNING IN BIOINFORMATICS AND DRUG DISCOVERY

Roshan Kotkondawar¹, Sanjay Sutar¹, and Arvind Kiwelekar¹

¹Department of Information Technology, Dr. Babasaheb, Ambedkar Technological University

January 4, 2024
APPLICATIONS OF DEEP LEARNING IN BIOINFORMATICS AND DRUG DISCOVERY

1ROSHAN KOTKONDAWAR, 2SANJAY SUTAR, 3ARVIND KIWELEKAR

1,2,3 Department of Information Technology, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad 402103, Maharashtra, India.
E-mail: 1kotkondawarroshan@gmail.com, 2srsutar@dbatu.ac.in, 3awk@dbatu.ac.in

Abstract - The remarkable progress and growth in the fields of Artificial Intelligence (AI) and Bio-Chemical sciences in recent times have been continuously solving many complex problems. AI has shown a measurable impact in almost all areas of life including healthcare. The exponential growth of data, technologies for efficient storage, retrieval of this huge data, and ever-increasing computational power have given rise to data-driven AI-based approaches like Deep Learning (DL). Productive use of Big Data has been recognized in many phases of drug discovery. The availability of massive biochemical data catalyzed by the fast processing abilities of GPUs is pushing up the huge success of deep learning approaches for some key activities in drug discovery like drug-target interaction or virtual screening. DL techniques have shown the potential to strategically reduce the time and huge cost spent in the drug discovery pipeline. This paper discusses the significance of DL in the context of biochemical Big Data and the drug discovery process.

Keywords - Bioinformatics, Healthcare, Artificial Intelligence, Big Data, Drug Discovery, Deep Learning

I. INTRODUCTION

Artificial Intelligence (AI) and Big Data are two technologies which are driving research and innovation in almost all disciplines. Numerous research areas including Bioinformatics and Healthcare have been revolutionized with the emergence of AI and Big Data. Several nations have emphasized the need for collective efforts to improve global health and AI has a big role to play in achieving this goal. Drug Development process holds a major share in global Healthcare. Providing affordable healthcare to all is a challenge in front of the developing countries. The safety of human life directly depends on the outcome of the drug discovery process. Hence, accuracy and precision are of immense importance in the drug discovery pipeline. Recent advances in areas of AI and Big data along with some impressive interdisciplinary research have built the pillars for adopting the In-silico trend of problem-solving. Researchers and Pharmaceutical companies across the globe are working together on complex tasks in the drug discovery pipeline by applying AI-based techniques. The continuous advancements in the AI domain have the potential to facilitate the process of drug discovery and lower the expenses which will surely benefit healthcare systems. This paper aims to highlight Drug Discovery process in a nutshell along with the usefulness of AI-driven approaches like Machine Learning (ML) and Deep Learning (DL) through different drug discovery activities to deliver health care services. Our objective is to seek attention to the applications of ML and DL in the fields of Bio-Chem-informatics and Drug Discovery. We review existing works in this area and suggest a few promising future works. The paper is organized in a total of 8 sections. Section 2 describes the basics of DL and gives an account of various DL architectures. Section 3 defines frequently used basic terms and concepts in Drug Discovery. Section 4 briefly explains the technologies behind Compute raided Drug Discovery. Section 5 deals with the actual applications of DL algorithms in the process of Drug Discovery. Section 6 interprets the influence of Big Data on Drug Discovery and describes some standard data sets available in this area. Section 7 discusses the details of some of the most widely used DL frameworks. Section 8 concludes the paper specifying some prominent works in near future.

II. ESSENTIALS OF DEEP LEARNING

Deep Learning (DL) is a highly established branch in the arena of AI and ML concerning the complex computational aspects of problem-solving. DL comprises a variety of algorithms that essentially work to establish a mapping in between the input variables and the output variable(s) by exploring a pattern in the feature variables. Numerous tasks like classification or prediction can be suitably formulated adopting one of the DL frameworks. The selection of a specific DL framework is governed by the class of problem being addressed along with the kind of data. The relationship between input features or feature variables and the output variable(s) is mathematically formulated as equation 1:

\[ Y = F(a_1, a_2, \ldots, a_n) \] (1)

Here, \( a_1, a_2, \ldots, a_n \) represents input features which are also called as independent variables and variable \( Y \) represents the output. The function \( F \) in the equation is also called hypothesis or model and it relates input with output following the association among variables [1]. Based on the availability of the labels or output values in the preferred data-set also called as training data-set, there are two broad classes of DL algorithms namely Supervised and Unsupervised algorithms. In the supervised DL approach, the model
is trained to learn the mapping using both inputs and labels whereas in the unsupervised DL approach the model learns the mapping only from input features. The ultimate objective of a DL algorithm is to minimize the cost or error in the prediction of the result. Different cost functions are utilized to reduce the error and to improve the skill of the learning algorithm. The gradient-based policy of error reduction is the most popular which employs the stochastic gradient descent optimization algorithm during training. The classical ML algorithms are well suited for the limited size of the data where a linear or direct relationship exists between the input features and the labels. However, the ML-based approaches do not suffice in the case of the huge data-set, especially for unsupervised learning problems. The algorithms fall short to accurately grasp the complex and mostly non-linear association among the huge input variables in training data [1]. DL-based approaches have always been found efficient and productive as compared to the classical ML-based methods in the case of the huge datasets. The data handling capacity of DL algorithms is superior to the ML-based methods and can address the most challenging problems today. This section discusses various DL frameworks briefly as these frameworks appeared as promising alternatives for various drug discovery activities.

2.1 Deep Learning Architectures

Deep Learning approaches are influenced by the dynamic operation of the human brain and thus attempt to mimic the same but in a simulated manner. The functioning of the DL method is centered around a fundamental unit called a neuron similar to the human brain. A long chain of such neurons in close association makes an Artificial Neural Network (ANN). ANN is the building block of all the DL architectures.

2.2 Deep Neural Network (DNN)

The ANN is altogether a layered-architecture built over the three core layers namely the input layer, output layer, and hidden layer. Each of these layers is responsible for the processing of the data using a special unit called a neuron. A neuron holds the value during data crunching in the network and these values are only visible in the case of the input and output layer but not in the hidden layer.

In a Deep Neural Network (DNN), the information in the form of the inputs and weights is processed at each layer starting at the input layer with m-inputs to the final output layer with n-outputs. DNNs are sometimes referred to as feed-forward networks or Deep feed-forward networks since the information runs through these three layers while estimating a function [2]. This information processing is carried out in two steps.

In the first step, all the inputs are summed up using the weighted multiplication of these inputs. The output $Z$ for the first step can be formulated as equation 2:

$$Z = (W_{ij} X_j)$$  \hspace{1cm} (2) 

Where $W_{ij}$ denotes the weight corresponding to $i$th node with $j$th input and $X_j$ represents the value for $j$th input [1].

In next (second) step, the output $Z$ is put through the typical non-linear regulative function named as the Activation Function. The output $Y_i$ at the end of this second step using an activation function $g$ can be formulated as equation 3:

$$Y_i = g(Z), \quad (3)$$

Several activation functions exist in the literature among which Sigmoid function, Hyperbolic tangent (tanh) function, and Rectified Linear Unit (ReLU) function are to mention a few. A skilful selection of activation function is done depending on the problem being addressed. The current DNN comprise of few hidden layers having numerous neurons or nodes at each layer using ReLU as an activation function, as depicted in Figure 1.

![Figure 1: Structure of the Deep Neural Network.](image)

The DNNs are trained iteratively using a suitable data-set and a special technique called Back-propagation until an optimized version of the model is found. The weights of parameters are adjusted at random initially and are continuously monitored during training. The gradient-based model optimization technique is preferred in most tasks which proceeds by finding the prediction error at the final or output layer.

2.3 Convolutional Neural Network (CNN)

The DNNs are very effective to handle two-dimensional (2-D) data-sets where the data is characterized by a 2-D matrix simply. In practice, most of the multi-model data-sets are usually unstructured and multidimensional such as images, audio, video, or textual data. This increases the complexity of calculations discussed in the previous section since the dimension of matrices increases exponentially. The situation can even be chaotic for high-resolution image data (1024 X 1024). The Convolutional Neural Network (CNN) denotes the
of DL architectures which are skilled to handle the multidimensional especially 3-D data. CNNs are adopted to solve the complex problems in Computer Vision and Image Processing. The data-sets here can be a collection of high-quality images with three channels (Red, Green, Blue). Here, the data is 3-D which makes the size of the matrix used to represent this data sufficiently large.

![Convolution Layer](image1)

**Figure 2: Structure of the Convolutional Neural Network.**

The CNN is functionally divided into three basic sub-networks or layers as depicted in Figure 2. The first two layers of CNN namely, convolution layer and pooling layer deal with data pre-processing especially feature engineering. The feature engineering activities can be feature identification, data compression, or reduction of features. The last layer of fully-connected DNN performs the primary task of prediction or classification based on the problem. The convolution operation is executed at the convolution layer using a specific filter size depending on input data and a feature map is created. The filter is simply a weight matrix that represents a pattern or mapping. The convolution operation is essentially performed to recognize the low-level or front-line features from the data for example lines or edges in the image. The convolution layer works on an assumption that the low-level or front-line features are associated with the pixel configuration. The pooling layer accomplishes an important task of limiting the spatial size of data representation by implementing the aggregation activity to achieve weight sharing across the network. This functionality of the pooling layer helps to reduce the count of features used during computations in the network. The most famous pooling strategies are max pooling and average pooling. CNNs have been actively used with some outstanding results in the areas like Computer Vision, Image Processing, and Speech Recognition since CNNs work around the principle that the front-line or low-level features constitute the high-level features.

### 2.4 Recurrent Neural Network (RNN)

The third and important class of DL architectures is Recurrent Neural Network (RNN) which are intended to handle sequential information such as text data, time-series data, and DNA sequence. RNNs have been employed in numerous fields like sentiment analysis, time-series forecasting, music generation, machine translation, and DNA sequence analysis. The sequence data exists in almost all forms which make handling the data difficult. Many issues arise while dealing with the sequence data such as the dependency among sequence or sequences, varying length of sequences, and the heterogeneous nature of data. The RNNs are skilled to learn the association in a given sequence(s) of data. RNNs are smart architectures which process the sequence data at ease tackling the above-mentioned challenges. RNNs master the association in sequence(s) of input data at different time steps or instances. RNNs are the only architectures among DL frameworks which have an inbuilt memory or a state vector used to remember the long sequences. This state vector is actualized in the network by chaining the hidden states or units of neighbouring layers in a feed-forward manner as depicted in Figure 3. Therefore the output is calculated over all the data accessed by the network thus far and is a combination of current input as well as the input traced so far. Consequently, the output of RNN may vary for the identical input sequence over two instances.

One difficulty in using RNNs being the problem of vanishing-gradients. A more admired form of RNNs is the RNN implemented using reasonable units of LSTM (Long Short Term Memory). These networks are very effective to work over the most complex sequences in the case of machine translation, question-answering systems or chat-bots, and even DNA or drug sequence.

### 2.5 Autoencoders (AE)

The fourth class of neural network architectures belongs to the unsupervised DNN known as Autoencoder. In contrast to the other DL architectures like DNN, CNN, or RNN, the auto-encoder transforms the input data after learning the mapping in the input. One can imagine the spatiality reduction task for a high-dimensional array as a specific transformation activity example. Here, the auto-encoders take an input vector of dimension m and translate it to the corresponding output vector with dimension n where n < m.

Two basic operations carried in an auto-encoder network are encoding and decoding using two separate functions Encode and Decode. The input representation x is turned to a different form y by implementing the Encode function as defined by equation 4:
y = Encode(x) \ (4)

Where, x and y are vectors having distinct representations. Likewise, the vector y is then put back to its initial form by implementing Decode function as defined by equation 5:

x = Decode(y) \ (5)

Figure 4 shows an Auto-encoder architecture with Encoder and Decoder sections performing encoding and decoding activities during input transformation.

Deep auto-encoders with a long chain of hidden layers are usually employed for the highly unstructured and non-linear kind of data. Auto-encoders are preferably adopted to reduce the dimensions of input data as well as for pre-training a DL model. The pre-training of DL model with auto-encoders has been proved highly advantageous in improving the performance of a CNN or RNN [4].

III. BASIC TERMINOLOGY USED IN DRUG DISCOVERY

Drug discovery process can be best understood if some fundamental terms in overall development are introduced before diving deep into it. This section gives a glimpse of these terms.

1. **Drug**
   A drug is a central object in any form of drug development activity. The drug is an agent potentially utilized for diagnosis, treatment, and inhibition (prevention) of a disease. A completely developed drug ready to be taken in a dosage form is known as medicine.

2. **Rational drug design**
   Rational drug design is a process that involves the discovery of new medications based on the prior facts about the targets obtained by studying the structure and functions of the target molecules. Rational drug design is based on some hypothesis and a methodology to generate new drug molecules in contrast to the traditional drug design that relies on the trial and error-based experimentation.

3. **Lead compound**
   A lead compound in the simplest terms is a substance or a molecule or a chemical compound that can be converted into a candidate drug having therapeutically useful activity by minimizing the adverse effects and improving the desired effects. The chemical structure of a lead compound initiates its development to a potential drug by improving its safety and efficacy.

4. **Ligand**
   Ligand is a molecule that strongly binds to a biomolecule or a target compound in the drug discovery process. In pharmacological terms, a ligand is a chemical entity capable of forming a complex with bio-molecule to achieve a biological target [5]. Drug design can sometimes be referred as the ligand design since it is a ligand molecule which is repeatedly optimized for the occurrence of all the desired properties and therapeutic effects in it before it finally becomes a safe and effective drug.

5. **Assay**
   An assay is a procedural technique designed to test the biological activity of the drug target. In the drug development pipeline, assays are designed for many purposes considering the drug target and receptors. The process of drug design sets strict regulations on the amount of drug intake.

6. **Molecular docking**
   In drug discovery, docking is a method to predict the alignment of a complex formed from the interaction between two molecules. The knowledge about orientation of the newly formed complex is useful in determining the binding affinity of the complex.

7. **Protein Engineering**:
   Protein engineering is an emerging area of study that deals with the design of useful proteins. In this new discipline, research is being carried out to understand protein folding and to recognize design principles for proteins [6].

8. **Drug Re-positioning**:
   Drug Re-positioning is also known as Drug Re-purposing or Drug Recycling. It is a process of investigating and reprocessing an already licensed and approved drugs for discovering new medications. It involves using an existing drug for a different therapeutic activity. This process is beneficial in terms of saving time for clinical trials and also the huge cost required for the same.

IV. COMPUTER AIDED DRUG DISCOVERY

The drug development process is an interconnection between businesses and biological sciences. Pharmaceutical companies are involved in the development and marketing of new drugs. Drug development is a long process that involves small processes like the creation of a bioactive drug molecule, testing of such drugs for their desired or side effects, and marketing these drugs to customers. Drug design and discovery form a part of this
complex drug development process. The drug design
process is sometimes also termed as rational drug
design and is dedicated to detecting novel candidate
drug molecules based on their biological properties
and targets. Drug discovery is a way to develop new
candidate drugs by the application of various drug
design techniques. Drug discovery is a patient-
oriented process where the new drug is more potent
than the already existing drug of the category.

Traditional approaches to this rely on laboratory-
based experimentation done by an expert scientist or
a chemist with a piece of strong domain knowledge.
It is a long step-wise activity involving various stages
namely: (i) identification of biological targets, (ii)
validation of biological targets, (iii) lead structure
search, (iv) lead structure optimization, (v) pre-
clinical studies, (vi) clinical trials, and (vii)
formulations for clinical studies [7]. The field of
computer science and statistics attempt to automate
this complex drug development process to reduce
development time and cost. The adoption of emerging
techniques and tools of Artificial Intelligence (AI)
during drug discovery process usually referred to as
Computer Aided Drug Discovery (CADD). The
CADD utilizes Computers, Statistics, Information
Sciences, AI tools like ML, and DL most finely to
discover the candidate drugs. Many ML and DL-
based tools and softwares have been found effective
for drug discovery activities like drug-target
interaction (DTI) prediction [8,9,10,11] compound
and bioactivity prediction [12,13] target prediction
[14], molecular property prediction [15], protein-
ligand or drug-protein interaction [16,17],
identification of essential proteins [18], compound
property interaction [19], virtual screening (target
prediction) [20].

Over the last decade, drug discovery has been greatly
influenced by the fields of Chemoinformatics and
Bioinformatics. Both these fields run hand in hand
concerning drug discovery. The term Informatics
itself justifies the use of computers in these fields.

Cheminformatics is a multidisciplinary science that
combines Chemistry, Mathematics, Biochemistry,
Statistics, and Informatics. In a similar way,
Bioinformatics deals with biological data to extract
essential information. Bioinformatics is also a
multidisciplinary field that combines Biology,
Mathematics, Statistics, and Informatics. The major
difference between these two fields is Chemo
informatics deals with small molecules, like drug
molecules, or lead compounds, whereas,
Bioinformatics deals with large molecules like
proteins or genes [21]. Both these streams have laid
an essential foundation in the drug discovery
phenomenon and are collectively referred to as Bio-
Chem-Informatics. One major applicative area of
Chemo informatics is drug discovery. Chem
informatics approaches extract the meaningful data,
and this technique is used for target identification
stage of drug discovery to identify the drug target,
which can either be a gene or protein [21]. Many chemo
informatics tools along with the chemical data
available in data repositories speed up the drug
discovery process.

V. APPLYING DEEP LEARNING FOR DRUG
DISCOVERY

The drug discovery process is centered on the
creation of novel drug molecules adopting an
efficient design technique. It is one of the most
complex activities which needs an efficient and
prompt approach to be adopted. With the rapid
enhancement in computational power and advances in
fields of Bio-Chem Informatics, CADD has gained
attention in recent times. ML and DL algorithms
accompanied by large chemical and pharmacological
data have re-coined the drug discovery process as
data-driven drug discovery [22]. Besides, Drug
Discovery is one of the major applications of AI in
healthcare and medicine. The strength of deep
learning lies in the fact that it is capable of learning
the complex features automatically without any prior
training. This has made the application of deep
learning approaches like dense neural networks for
making better predictions in less time and cost.

5.1 Challenges in adopting DL for drug discovery

Drug discovery is a challenging task. The first
challenge comes from the stringent requirement to
meet standard procedures for drug development. It is
an area where the accuracy and safety of human life
are the biggest concerns. Hence, all the
experimentation and modeling need to follow the
standard protocols in the drug discovery process.
These procedures guarantee accurate measurement
of drug dosage, identification of therapeutic effects,
and ensures drug safety. The second challenge concerns
about data. Biomedical data is complex and comes in
different forms like images or genetic data.
The data needs careful prepossessing and analysis
before the development of ML models. The third
challenge is the availability of large data sets. DL
models work best with high volumes of data. Hence
the appropriate data should be obtainable for a
particular task to train the model effectively. The
fourth challenge is legal constraints. An automated or
AI-based system predict the outcomes based on the
training data set, and it lacks human touch in
predictions. It may raise some legal issues.

5.2 Important activities in drug discovery

Drug Discovery is a highly sophisticated procedure
consisting of many intermediate stages. Deep
learning techniques were effectively utilized to
perform some crucial activities in the drug discovery
pipeline over the last few years. In this section, we
review and illustrate the application of DL for
performing these activities.

Proceedings of ISETE International Conference, Nagpur, India, 06th June, 2022
5.3 Drug-Target Interaction

Drug-Target interaction (DTI) stands among the pivotal activities in the initial stages of drug discovery. A biological target refers to an element in the human body to which the other substance binds during a reaction. This binding introduces a change in the nature and functionality of the target element. Most of the biological targets are proteins such as enzymes and receptors. The drug discovery process aims to generate novel drug-like molecules by monitoring the therapeutic effects of the molecules on a target through prolonged experimentation and pilot studies.

A drug having desired therapeutic properties on the target is retained in the process. A biological target is acknowledged in drug design if the molecules which are active to a specific target can be generated with positive therapeutic effects [23]. Hence, the successful detection of DTI is crucial for both drug discovery and drug repositioning. Three classes of statistical approaches exist in the literature to estimate DTI: (i) ligand-based, (ii) docking simulation, and (iii) chemogenomic approaches. Among these three methods, chemogenomic approaches are widely used today for both drug discovery and drug repositioning. Targets frequently thought of in predicting DTI are protein, disease, gene, and side effects [24].

Though there exist many Chemogenomic approaches, ML-based approaches are effectively utilized to estimate DTI accurately. ML-based approaches adopt both supervised and unsupervised learning techniques for precise DTI prediction. Some of the most widely ML-based approaches include Support Vector Machines (SVM), K-Nearest Neighbour, Random Forest (RF), and Decision Trees.

Due to the rapid explosion of the biochemical data, researchers have started adopting DL-based methods. The Merck Molecular Activity challenge hosted by Kaggle in 2012 has laid a strong foundation for DL-based approaches in drug discovery, especially in DTI phase [25]. The DL-based methods have been efficient than the ML-based methods for predicting DTI. Figure 5 depicts the utility of DL models for prediction where the relationships in between compounds and molecules are interpreted using a typical DL framework.

5.4 Quantitative Structure Activity Relationship (QSAR)

According to the vocabulary of key terms used in the computational drug discovery (IUPAC recommendations 1997), Quantitative Structure-Activity Relationship (QSAR) determine the analytical association amongst chemical structures and pharmacological activities quantitatively for a set of compounds. Various pattern recognition and regression techniques are usually adopted to find these relationships. In a broad sense, QSAR is similar to the development of a mathematical model relating a set of chemical compounds to their biological activity. Consequently, the QSAR prediction method can essentially detect the biological actions of a new compound. Therefore QSAR modeling is a critical stage in Rational Drug Design. This stage also saves time and dependency on preclinical trials for testing whether the drug is giving an expected response. As mentioned in previous sections, ML and DL algorithms establish relationships between variables by recognizing the patterns inside data. QSAR can hence be thought of as an ML-based model that tests the activity of a molecule. Moreover, predicting physicochemical, pharmacokinetic and toxicological properties of the lead compounds during initial phases is crucial because it reduces the number of expensive and late development failures [26]. The relationships between molecular descriptors and the activities are mostly non-linear. Hence, ML-based techniques like Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) are useful for QSAR modeling, especially in virtual screening. Deep Neural Nets have effectively used for better predictions in QSAR13. An ensemble-based ML model has been adopted by S. Kwon et al. [27] for QSAR prediction with outstanding results. Recently, Hu et al. [28] designed a model combining Autoencoder and CNN in QSAR prediction to identify an active molecule.

5.5 Molecular property prediction

An essential and tricky task in the drug discovery pipeline is to track the exact activity of a unique molecule or compound. A drug with high efficiency and the lowest side effect is the ultimate goal of drug discovery. The general protocol is to test the biological activity of compound for both the intended target molecule as well as off the target molecule. Finally, a drug that is highly active for target molecule and least or not active for off the target molecule is preferred as a result. This complete task is tedious and time-consuming. Hence the prediction of such an activity saves both time and cost. We have already discussed such predictive methods under an
umbrella of QSAR. There are many activities in QSAR one of which is molecular property prediction. Machine learning approaches have been employed for this task successfully. Also, the DL approaches are preferred when benchmark data-set is available. In 2017 Gilmer et al.15 proposed a DL based Message Passing Neural Network which predicts molecular quantum properties using standard data-set consisting of molecular fingerprint images and deep CNN. Simm et al. [29] in 2018 developed a method to predict the activities of compounds using DNNs and image-based molecular fingerprints.

5.6 High Throughput Screening
Screening concerning drug discovery is an art of selecting compounds also known as leads having the desired biological activity to make the admissible drug. The utmost popular compound or molecule screening technique used today is high throughput screening (HTS). HTS is a fully mechanized system designed for testing a large number of compounds together concerning certain activities or properties. These properties can be either the inhibitors or activators of a specific target, a cell-surface receptor, or a metabolic enzyme [30]. The vast expansion in Chemical Big Data holding enormous chemical databases like chEMBL [31] has taken HTS a step further. HTS is one of the best application of Internet of Things (IoT) and Robotics. HTS techniques let an investigator carry out millions of bio-chemical or pharmaceutical tests all at once. This type of automated screening greatly reduces the time required for the process as compared to manual screening in laboratories by chemists. In recent years, the concept of virtual screening has also fueled up this process a step away. The virtual screening method is a highly adopted technique in computational drug discovery and centered at detecting the leads which are proficient in binding a target molecule [32]. The virtual screening technique employed can either be ligand-based or structure-based. The use of ML and DL methods has boosted virtual screening since an optimized ML or DL algorithm when worked on a preferred data-set can detect the novel molecules much faster. DNNs have been already employed for the ligand-based approach [13] and structure-based approaches [19]. In both these approaches, DL based methods have done reasonably well and can be used as a promising option while dealing with a huge data-set.

5.7 Identification of essential proteins
Rational drug discovery is a target-oriented procedure where the activity of a drug molecule is checked on the target molecule in many stages. In addition to this, a new drug molecule is checked for safety and efficacy several times during preclinical and clinical trials. Targets for most of the drugs are generally proteins in our body, hence the essential proteins must be identified before the application of a respective drug. Proteins are the polymers in which amino acids are strongly linked together forming a chain. One of the techniques to identify essential proteins is to predict them through protein-protein interaction (PPI). Several methods to discover the necessary proteins exist in the literature but it is still a challenging assignment. PPI network topologies are mostly used for finding the essential proteins. DL-based methods have been proved fruitful for this task too. M. Zeng et al.18 in 2018 designed a CNN-based DL model to recognize essential proteins using the gene expression data. In 2019 M. Zeng et al. [33] proposed one more method for the same by combining several versions of biological data. In this approach, a DL network is used to automatically learn the biological features. In a very recent work, L Yu et al. [34] employed a novel DL-based approach to accurately identify the druggable proteins by visualizing the trained data-set of proteins.

VI. BIG DATA FOR DRUG DISCOVERY
The transition from the conventional drug discovery practice to AI-based computer-aided drug discovery is a big success involving multiple disciplines like Bio-chem-informatics, AI, Statistics, and Big Data. This section discusses the significance of Big Data in decision-making concerning drug discovery and also presents a few biochemical databases.

6.1 Significance of Big Data in drug discovery
Proficient data usage has always proved fruitful in almost all fields. This act of skilful data processing is managed in practice by AI methods. ML algorithms are always suitable for small or medium-sized data sets whereas a huge data-set can be handled by adopting a DL-based technique. The volume of experimental data utilized by a DL model alters the decisive skill of the model since a large data-set can efficiently stock all the possible patterns or sequences in biochemical data. Hence, Big Data has appeared as an impressive strategy in most activities while applying DL algorithms to the drug discovery pipeline. Drug discovery is a patient-centered process and hence there is a need for real-world data. Some huge biochemical data repositories are coming to rescue in this regard. Handling such a huge amount of complex data is itself a challenge. AI-based strategies should hence be applied for data mining and data integration for drug discovery.

6.2 Data Sets for Drug Discovery
We list some benchmark data repositories along with data-sets used for carrying out different activities in drug discovery.

6.3 chEMBL Database
chEMBL is one of the most favoured data-sets for performing drug discovery activities. It is a manually developed database that contains data sets of different
Applications of Deep Learning In Bioinformatics and Drug Discovery

categories like target, assay, compounds etc. chEMBL database [31] is the best example of Chemical Big Data and contains huge data sets for biochemical research. These data-sets are available in CSV, tsv, and SDF formats for download. These data-sets are successfully utilized for drug discovery activities like virtual screening [20], target prediction [14], bioactivity prediction [12] and other activities.

6.4 PubChem Database
PubChem is the freely available chemical database. It is one of the largest and ever-growing databases for activities related to molecular chemistry like drug discovery. Some of the most crucial activities like drug target interaction prediction [10], compound property prediction [19] are done using these data sets.

6.5 DrugBank Database
DrugBank is an open-source data repository in the fields of Bio-cheminformatics. It is a freely available database containing detailed information related to drugs and drug targets. Researchers have used this drug-target paired database for drug-target interaction prediction [11] activity of drug discovery.

6.6 Kaggle Database
These datasets are available on the Kaggle website and are part of the Merck Molecular Activity Challenge hosted by Kaggle. These datasets are successfully used for QSAR [13], molecular activity prediction and many other activities.

6.7 STITCH Database
STITCH is a vast protein-related database used for activities like target protein interactions. This database covers a huge number of proteins from different organisms. Some of the recent works include Compound Protein Interaction (CPI) [19] using STITCH protein database.

6.8 MoleculeNet Database
MoleculeNet is a database specially designed for ML research using molecules. It is a public domain database based on the DeepChem python package. It contains 17 benchmark data sets [35]. Some benchmark datasets like Quantum Machine 9 (QM9) data-set and PDBind core set are utilized for activities like molecular property prediction [15] and Protein-ligand scoring [17].

VII. DEEP LEARNING FRAMEWORKS FOR DRUG DISCOVERY

Drug discovery activities require a combination of hardware, software, and a large database for accurate results. Since the accuracy of prediction is of extreme importance in drug discovery, specialized deep learning tools along with the software tools, hardware tools are necessary for fast execution of the processes. Hence, the latest technologies like Graphical Processing Unit (GPU), GPU enabled workstations, fast computing and processing models like CUDA are being utilized. In this section, we list some of the widely used DL frameworks in drug discovery activities.

7.1 Pytorch
Pytorch is a powerful open-source framework for DL developed by Facebook in 2016. It is built over Python, and it makes use of GPUs [36,37]. It is a tensor library useful for implementing the DL applications. It supports all DL platforms and have been widely applied for drug discovery.

7.2 Tensor Flow
TensorFlow (TF) is among the most favoured open-source APIs designed by Google in 2015 for implementing the DL algorithms. It is accessible for Python, Java, and Go programming languages. All the DL architectures like ANN, CNN, DNN, RNN, and Auto-encoders can be implemented using TF. Deep learning models developed in TF can be easily deployed over hardware platforms like CPU, GPU and TPU etc. TF has been already employed in drug discovery by researchers [38].

7.3 Caffe /Caffe2
Convolutional architecture for fast feature embedding (Caffe) is a fast-growing open-source DL framework designed by Berkeley AI Research (BAIR) in conjunction with community contributors in 2013. The framework has been designed by Yangqing Jia at Berkeley, the University of California [39]. As developed by community contributors, researchers can use the projects stored over there. Features that make Caffe powerful are its speed of operation, modularity, and the extensible code. Researchers have used this framework for drug discovery activities [12,17].

7.4 DeepChem
DeepChem is a Python package utilized for the design of DL models in life sciences like chemistry, biology, and drug discovery. DeepChem is most preferred tool adopting machine learning for molecular data [40]. It has numerous applications in molecular chemistry and drug discovery. It has been actively used along with the chemical datasets for DL. DeepChem is a Python package utilized for the design of DL models in life sciences like chemistry, biology, and drug discovery.

7.5 Theano
Theano is also an open-source DL framework designed over Python in 2007. Theano has been used for the design of computationally complex models using DL and has found superior performance [41]. It is fast, efficient and generates dynamic C code that makes expression evaluation easy.
VIII. CONCLUDING REMARKS AND FUTURE SCOPE

This work is focused on exploring the applications of DL in drug discovery along with the recent works in this domain. This study presents an introduction to drug discovery as a pipe-lined procedure enclosing all the necessary terms and tasks. In this work, we reviewed literature scrutinizing all the phases of the drug discovery process, and DL architectures turned out to be highly advantageous. Applications of DL can boost up many complex tasks in the drug discovery process. DL architectures accompanied by suitable data-sets and specialized hardware can be well organized to build the best prototype addressing drug design activity. Following the application of DL in drug design activities, many of the complex activities like DTI prediction, virtual screening or molecular property prediction are carried out in much less time with a measurable saving of cost. AI and Big Data have made a big transition in the way the drug discovery process has been carried out to date with some outshining results. This will surely benefit healthcare systems. DL applications can lead to a crucial change towards affordable drug discovery for all, especially in developing countries. A huge opportunity exists for researchers working in this area in the coming decade since mechanization of processes like screening has laid the foundation for the effective use of chemical Big Data and AI-based techniques.

REFERENCES


