A Review: Drug Discovery Using Deep Learning

[1]Ms. Sneha Khaire, [2]Dr. Pawan Bhaladhare
[1]*Research Scholar, School of Computer Science*
[2]*Professor, School of Computer Science*
[1, 2]*Sandip University, Nashik, Maharashtra, India*

Abstract: In a wide range of applications, deep learning algorithms have achieved the best possible results. Images can be classified, objects can be detected, and semantically segmented using a Convolutional Neural Network (CNN). An RNN and its descendants like the LSTM, the GRU, and the transformers come to mind first when trying to solve problems like neural language translation and speech recognition as well as the creation of new texts or music. The pharmaceutical industry stands to gain greatly from the success of deep learning. Deep learning can be utilized to expedite the discovery and development of new drugs while also lowering the associated costs. Deep learning can be used in this area, as demonstrated in this paper, which provides an overview of recent research on the subject and shows how it can be used.

Keywords: Deep learning, drug, neural networks, drug discovery

1. Introduction

Developing medicines that are both safe and effective for human use is the primary goal of drug discovery efforts. Time and money are required for the entire drug development process, from identifying the target to conducting step-by-step clinical trials. As the costs rise with each milestone, selecting the right drug candidates for the next phase is critical. The "hit-to-lead" process is a critical step in identifying promising lead compounds from hits and evaluating their therapeutic potential. The concept of polypharmacology, which states that a single or multiple drugs often interact with multiple targets, is a common cause of side effects and lack of in vivo efficacy in clinical trials. While conducting in vivo tests for each disease model would be ideal, it would take a tremendous amount of time and effort to do so. Since the 1980s, computer-aided drug discovery or design methods have played a significant role in modern pharmaceutical R & D (R & D) [1–3]. Even with this in-silico approach, pharmaceutical R&D productivity has been declining since the mid-1990s.

Researchers and pharmaceutical companies have recently invested a lot of money in artificial intelligence (AI) because of its ability to help them find new drugs. The advent of high-performance processors like graphics processing units and decades of chemical and biological data have allowed AI to be used in drug development. With a variety of approaches, artificial intelligence (AI) has become an integral part of the drug development process [4–6]. A better understanding of biological space's complex contexts can be gained through the use of deep neural networks. By using hidden nonlinear models, it is possible to extract complex patterns from multi-level representations. Automated data preprocessing and feature selection also save time and effort. Accurate drug-target interactions (DTIs) predictions and novel molecules with desired properties have been achieved thanks to the development of deep learning (DL)-based methods. Even though drug development datasets are different from those used in traditional AI data, such as images and texts in terms of the types and distributions of the data, there is still a need for new DL techniques to be applied.

2. Applications

There are three main ways in which deep learning is used in drug discovery:

2.1 Drug properties prediction

supervised learning, self-supervised learning, and reinforcement learning are all examples of machine learning problems. Using a supervised learning problem, it is possible to predict the properties of a drug in advance. By using algorithms, a drug's (compound's) properties are determined (e.g., drug toxicity or solubility).

- Input: A drug
- Output: If the drug has certain properties or not, it will be labeled as such on its packaging. It can also be framed as a multi-label classification or regression task.

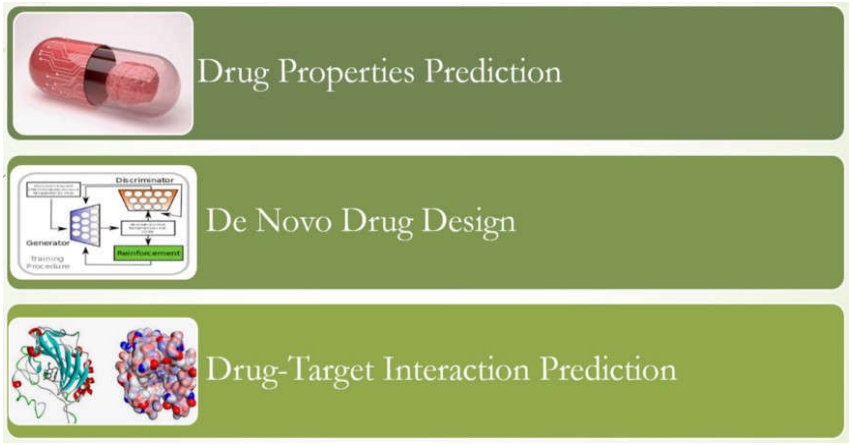


Figure 1: There are three areas where deep learning can assist in the search for new drugs.

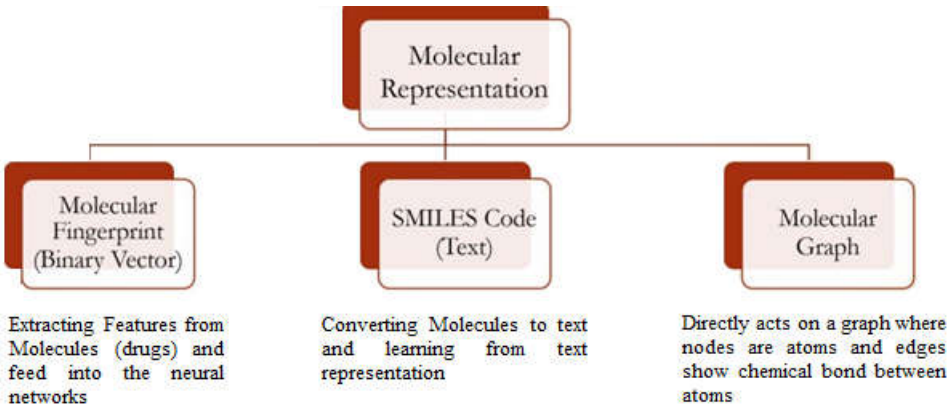


Figure 2: Numerous representations of a molecule.

A drug (compound) can be depicted in a variety of ways.

- Molecular swab
- Text-based representation
- Graph design is an example of this (2d or 3d graph)

2.2 Molecular fingerprint

It is possible to represent a drug's molecular fingerprint in the machine learning framework's input pipeline A binary digit (bit) fingerprint can indicate whether or not a molecule contains a particular structure. This is used by many scientists. Because of this, it is possible to view a drug as an array of 0s and 1s.

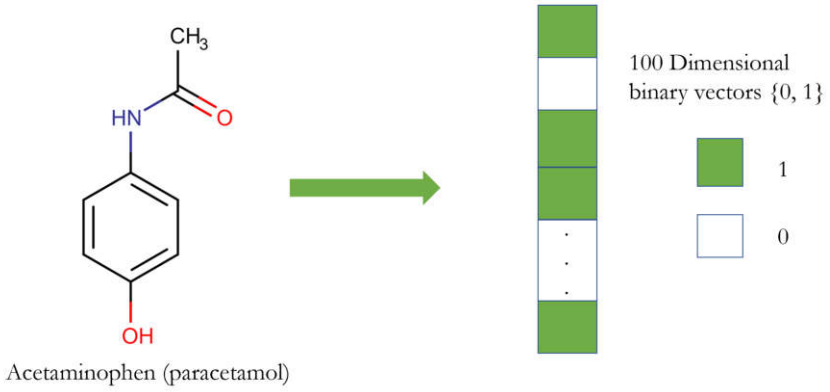
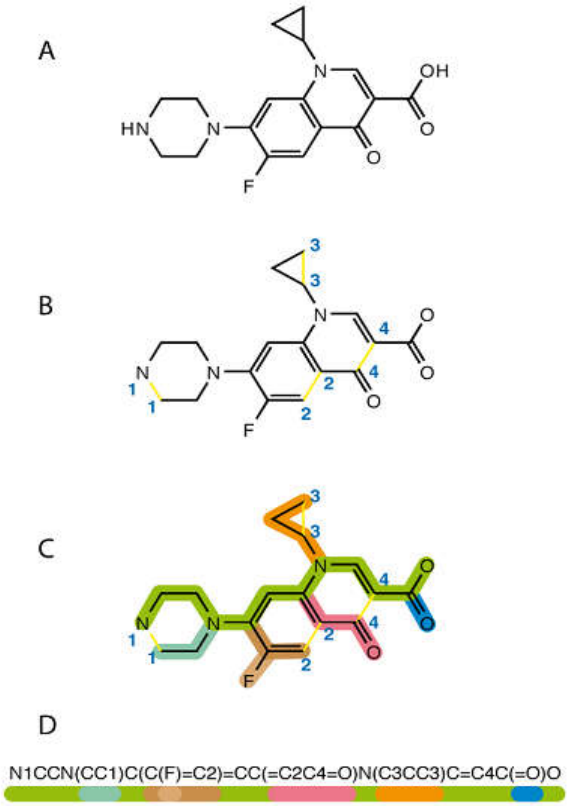


Figure 3: The binary vector representation of a molecule.

Literature [7] is replete with references to it. We can't fully reconstruct the fingerprint from the molecular structure, so it's clear that the process of encoding molecules as vectors is not reversible (it's a lossy transformation). A small molecule can be represented by a plethora of different fingerprints.

2.3 SMILES code

It's also possible to use molecules in the form of textual representations. Data from a graph structure is encoded into a string and used in the machine learning pipeline. SMILES is a well-known standard representation of molecular-input (SMILES). Following conversion, we can utilize powerful NLP algorithms to process the medicine, such as predicting its qualities, adverse effects, or even chemical-chemical interactions. For more information, please contact



Sunyoung Kwon [9].

Figure 4: How to represent a molecule in SMILES code.

Even though SMILES is a popular text-based representation of medicine, it is not the only one. If you're looking for an academic representation, InChIKey is a good option. Selfies (SELF-referenced Embedded Strings) were proposed by Mario Krenn and his colleagues [11] using a Chomsky type-2 grammar. They (advantages) are discussed in greater detail in the section on De Novo Drug Design.

2.4 Graph-structured data

Since graph convolution networks [Thomas Kipf] [12] have become so common in deep learning, it is now possible to feed graph data directly into the pipeline.

A compound, for example, can be viewed as a graph with atoms as the vertices and chemical bonds as the edges. Deep Graph Library, PyTorch Geometric, and PyTorch-BigGraph are some of the libraries dedicated to this work that we've seen a lot of success with.

2.5 Drug-Target Interaction Prediction

As the basic building blocks of living organisms, proteomes play an essential role in many of the processes that occur both inside and outside of cells. Apoptosis, cell differentiation, and other critical processes are all aided by specific proteins. A protein's ability to carry out a specific function is directly correlated with the three-dimensional structure of the protein. Changes in protein structure can have significant effects on protein function, which is an important consideration in the search for new drugs. Many drugs (small molecules) target specific proteins in order to bind to and alter their structure in order to change their function. Even a small change in the function of a single protein can have a significant impact on the performance of the cell. There are proteins in cells that can directly affect other proteins, repressing or activating their production, known as transcription factors (e.g., you can see protein-protein networks). Consequently, even a slight change in a protein's function can have a tremendous influence on cells and develop an altogether new cellular pathway.

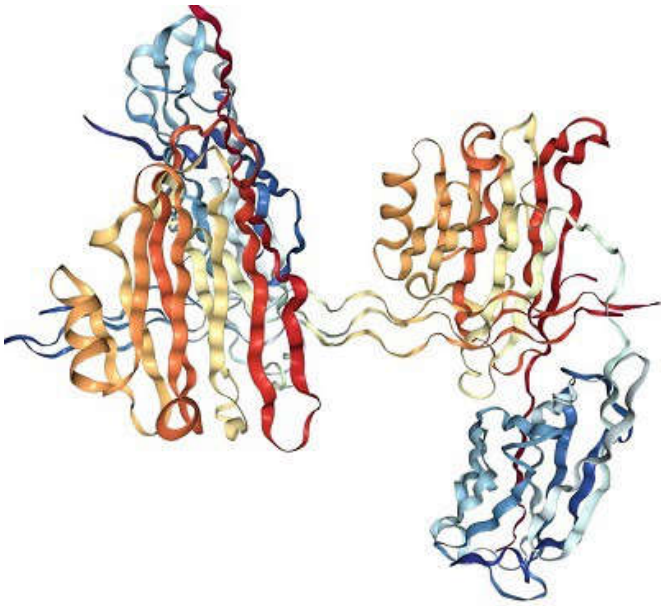


Figure 5: Complex Structure of Collagen Adhesin.

Computational drug development relies heavily on the ability to anticipate whether certain medications can bind to a specific protein.
Predicting the interaction between a drug and its intended target is known as drug-target interaction (DTI) prediction.
The task of DTI prediction can be conceptualized as follows:

- Compound and protein binding affinity can be predicted using this binary classification (it can be formalized as a regression task or binary classification)
- Compounds and proteins are represented as inputs.
  The output can be either [0–1] or a real number.

Feng and colleagues [13] created a deep learning-based framework for predicting drug-target interactions. When using deep learning frameworks to predict DTI, the representations used to feed neural networks often include both chemical and protein information. A binary fingerprint, SMILES code, or features generated from graph convolution networks can all be used to represent compounds. DTI prediction can be done using a variety of architectures, depending on the input data. When it comes to text-based representations of both compounds and proteins, RNN-based architectures spring to mind first (SMILES code for compounds and Amino acids or other sequence-based descriptors for proteins).
Using a Convolutional Neural Network, Matthew Ragoza and his colleagues developed a method for scoring proteins and ligands [14]. A protein-ligand complex has been depicted in 3D, rather than in text, to make it easier to understand. Because of this 3D structure, convolutional neural networks, which are capable of predicting Protein-Ligand binding affinity, have been selected.
Molecular Transformer Drug Target Interaction has recently been proposed by researchers as a method for identifying commercially available drugs that may target viral proteins of 2019-nCoV. (MT-DTI).
DTI prediction algorithms based on deep learning have become a popular trend, but the papers are all very similar and there is only one area of innovation: the choice of input representation and the architecture used to act on it. Consequently, we can sum up this task thusly:

- A database of compounds and their targets, as well as whether or not they interact with each other, is available (E.g., the STITCH database).
- DTI prediction networks typically use a pair of compounds and proteins as input.
- Make sure the representation of compounds and proteins is clear to you. However, I haven't come across any other depictions.
- Choose a neural network architecture that is compatible with the representation that you've chosen. Convolutional neural networks can be used when dealing with images or 3D structures, while RNNs can be used for text-based inputs (GRU, LSTM,...) and transformers.
- As a framework for solving the problem, binary classification (whether the compound binds or not) or regression can be used (prediction of the strength of affinity between compound and proteins).

That concludes our look ahead to DTI. Initially, it may appear to be a difficult and challenging task, but the papers we've reviewed are employing simple techniques and strategies to address the issue.

2.6 De Novo Drug Design

Algorithms for predicting side effects and other properties of a drug have so far been limited to discriminative algorithms, such as algorithms that can predict the likelihood of binding between two compounds. For example, what if we want to create a compound with specific properties?

In other words, we'd like to create a compound that has the ability to bind to a specific protein, alter certain pathways without interfering with others, and have some specific physical properties like a narrow range of solubility. This problem cannot be addressed with the toolkits we discussed in the previous sections. This problem is best understood in the context of generative models. Automatic regression algorithms, normalizing flows, and Variational autoencoders (VAEs) are generative models; as are Generative adversarial networks (GANs). Recent attempts have been made to incorporate them into the design of new drugs.

There is a difficulty in creating a compound with desired properties. Although it should be self-evident, this problem is significantly more difficult than the two that we discussed earlier. The number of possible chemical molecules is enormous, making it difficult, if not impossible, to search through all of them in search of a useful drug. I've noticed a rise in the use of generative models to create new chemical compounds. More mature methods are still needed to further progress in this field despite promising results that have been published. In this section, we'll take a look at some of the best papers in the field. At the end of the paper, the SMILES are generated as an output and then converted into chemical space. Rafael Gomez-Bombarelli et al. proposed a data-driven continuous representation of molecules for the purpose of automatic chemical design [16].

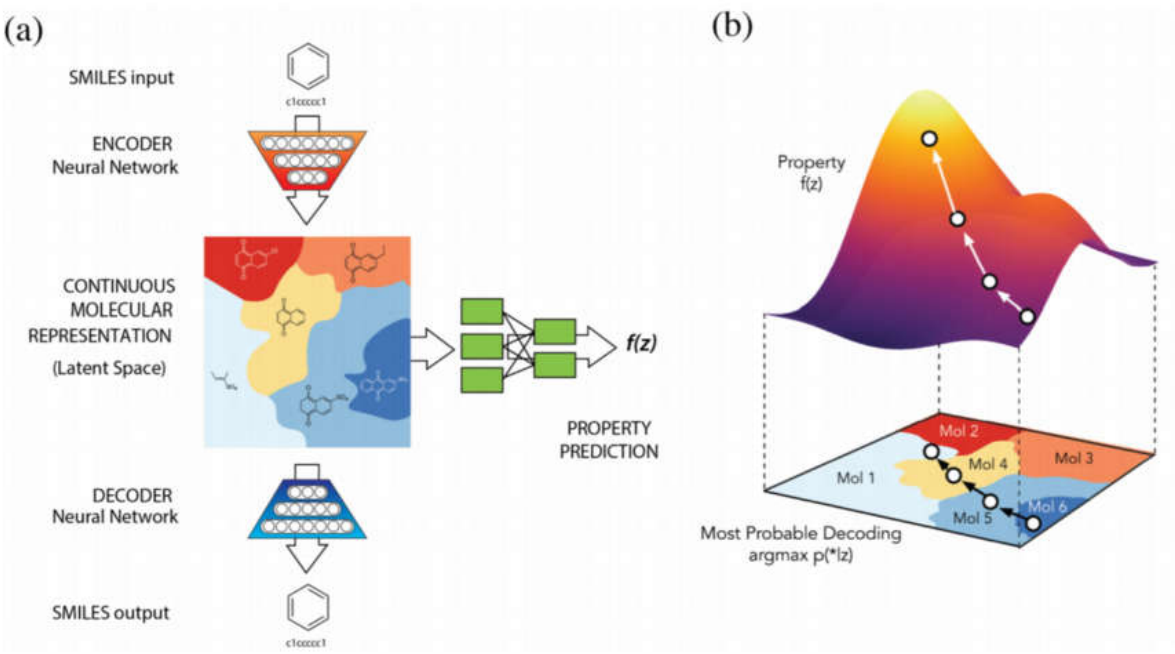


Figure 5: Compounds can be created using the variational autoencoder. In the work of Rafael Gomez-Bombarelli and colleagues [15]

Molecules have been created using VAEs. Using SMILES code, the input and output can be depicted. By using the Gaussian process within latent space (a continuous space), this paper is able to arrive at its intended destination. Create the SMILES code from the latent space point using the decoder. I think the author did an amazing job, so I'd say go ahead and check it out. The SMILES code and molecules do not have a one-to-one connection, which creates a problem. As a result, not all of the created code can be transferred back to the chemical space in which it was initially formed, resulting in SMILES code that does not always match genuine molecules.

SMILES, for all their popularity, have one major flaw: they can't hold up under close examination. For example, a single character change in the SMILES can invalidate the validity

of a molecule. For this problem, Matt J. Kusner et al. proposed Grammar VAE, which was designed specifically to deal with it [16].

SMILES code is being converted to a parse tree instead of being sent directly to the network (by utilizing SMILES context-free grammar). In order to create molecules that are more syntactically correct, they use grammar. While our model delivers more accurate results, it also demonstrates the ability to learn a more coherent latent space, in which close points decode into comparable discrete outputs, as explained in the work by the authors Recently, Mario Krenn et al [11] came up with an entirely new method for analyzing VAEs. The main selling point of the SELFIES is their robustness. The examples given here are just a few of the many attempts to use deep learning to synthesize small molecules that have been made. Numerous approaches are available in the literature, including normalizing flow and genetic algorithms for the compound's representation and generation algorithm (next table).

Table 1: Normalising Flow and Genetic Algorithms for the Compound's Representation and Generation Algorithm

Method	Input (Output Presentation)	Algorithm	Comment
Character [15]	SMILES	VAE	Bayesian optimization on the latent space
Grammar VAE [16]	Grammar rules form SMILES	VAE	Producing more grammatically valid SMILES
ORGAN [17]	SMILES	GAN + RL	1. Reconciling with RL 2. Based on SeqGAN
JT-VAE [18]	Fragments	VAE	Packing molecules into the meaningful
GENTRL [19]	SMILES	VAE +RL	They used GENTRL to discover potent inhibitors of (DDR1)
Aksjat Kumar Nigan et al. [20]	SELFIES	GA + DNN	1. Genetic algorithm 2. Robustness of SELFIES
Synthesizable Molecules [21]	Fragments	VAE	1. Focus is on the synthesizability 2. Provide synthesis route

3. Conclusion

Because of rapid advancements in computing power and massive amounts of chemical and biological data, artificial intelligence has had a positive impact on drug discovery projects. Deep learning studies have grown in number even in the last few decades. We wrote this paper to better understand how deep learning can be applied to drug discovery. This paper provides a good starting point for newcomers to deep learning-based drug discovery.

References

[1] Sachdev, K.; Gupta, M.K. A comprehensive review of feature-based methods for drug-target interaction prediction. J. Biomed. Inform. 2019, 93, 103159.

[2] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 2019, 18, 463–477.

[3] Kimber, T.B.; Chen, Y.; Volkamer, A. Deep learning in virtual screening: Recent applications and developments. Int. J. Mol. Sci. 2021, 22, 4435.

[4] Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov, M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. 2019, 37, 1038–1040.

[5] Feng, Q.; Dueva, E.; Cherkasov, A.; Ester, M. PADME: A Deep Learning-Based Framework for Drug-Target Interaction Prediction. arXiv 2018, arXiv:1807.09741.

[6] Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: Inpainting binding pockets using 3Dconvolutional neural networks. Bioinformatics 2019, 35, 243–250.

[7] Database fingerprint (DFP): an approach to represent molecular databases, Eli Fernández-de Gortari et al.

[8] Fingerprints in the RDKit

[9] DeepCCI: End-to-end Deep Learning for Chemical-Chemical Interaction Prediction, Sunyoung Kwon

[10] OpenSMILES specification. link

[11] SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry, Mario Krenn et al.

[12] GRAPH CONVOLUTIONAL NETWORKS, Thomas Kipf

[13] PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction, Qingyuan Feng et al.

[14] Protein-Ligand Scoring with Convolutional Neural Networks, Matthew Ragoza, et al.

[15] Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, Rafael Gomez-Bombarelli, et al.

[16] Grammar Variational Autoencoder, Matt J. Kusner et al.

[17] Objective-Reinforced Generative Adversarial Networks (ORGAN) for SequenceGeneration Models, Gabriel Guimaraes, et al.

[18] Junction Tree Variational Autoencoder for Molecular Graph Generation, Wengong Jin et al.

[19] Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Alex Zhavoronkov et al.

[20] Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space, AkshatKumar Nigam et al.

[21] A Model to Search for Synthesizable Molecules, John Bradshaw et al.