Original Article

Deep Learning Driven Drug Discovery and Use of Machine Learning Strategies: A Review

Taha Shoaib[1*], Dr. Suraiya Parveen[2]

[1*]Department of Computer Science and Engineering, Jamia Hamdard University, New Delhi, India. E-mail: tahashoaib7@gmail.com

[2]Department of Computer Science and Engineering, Jamia Hamdard University, New Delhi, India. E-mail: husainsuraiya@gmail.com

Abstract

Computational strategies have transformed the entire drug design and discovery process. However, the primary concerns associated are time consumption and production costs. Other hurdles include incompetency, inexact target delivery, and insuitable dosage. Such challenges can be eliminated with technological excellence by integrating computer-assisted drug design and AI algorithms. The computational technologies like deep learning and single-cell methods which conquer extensive biological facts from images, can speed-up the discovery process. ML approaches have facilitated the improvement at many steps in drug discovery for analysis of high-dimensional profiling data and technological advances to generate huge data sets. Using Image-based profiling, information in biological images is scaled-down to multidimensional profile, which reveals unanticipated biological activity that is applicable in drug discovery at many steps like recognizing disease-associated phenotypes, perceiving disease mechanisms and predicting drug's activity, toxicity or MOA.

In this review, we have discussed how the Machine Learning & Deep Learning approaches have been applied in functional profiling workflows by recent studies, the use of advanced techniques to optimize the challenges and the potential of emerging techniques in drug discovery which are anticipated to amplify the applicability of ML in drug discovery. Further, we focus on image-based profiling applications to the drug discovery process.

Keywords: Deep Learning (DL), Drug Discovery, Drug Perturbation, Drug-dug Interaction (DDI), Drug - Target Interaction (DTI), High-throughput Screening (HTS), Machine Learning (ML), Artificial Intelligence (AI), Image-based Profiling.

DOI: 10.47750/pnr.2022.13.S03.112

INTRODUCTION

Drug designing and development is an essential domain of research for the chemical scientists and the pharmaceutical companies. Though some hurdles and challenges are imposed that impact drug discovery process which include off-target delivery, low efficacy, high cost and time consumption. The other obstacles in the drug discovery pipeline are convoluted and big data from genomics, proteomics, micro-array and clinical trials. Such challenges and hurdles can be eliminated with technological excellence in Computer-assisted drug design integrating AI algorithms. [6] Methodological aspects are currently being concentrated on the AI context in drug discovery such as deep learning. AI makes faster decisions and more cheaply i.e. predictions are faster and cost less than experiments respectively. AI also leads to better decisions where appropriate data or simulations allow, and where decisions are better if braced by facts. Collaborated with the above analysis, better decisions will, by far, have the biggest impact on the success of drug-discovery programmes, compared with both faster and cheaper decisions. [11]

New pathways for data-rich phenotypic profiling of small molecules in drug discovery are unlocked by the escalation in imaging throughput, new analytical frameworks and high-performance computational resources. Subsequent studies require to develop methodologies that curtail the entry barriers to high-throughput profiling experiments by streamlining image-based profiling assays and providing applications for advanced learning technologies such as easy to deploy deep neural networks, to enable discoveries.[5] The aim of Image-based profiling is to use suitable computational models to recognize biologically relevant similarities and differences between samples on the basis of profiles.[14] Profiling can sort hits into clusters with biologically same effects at minimum and at the best, it can imply a compound's MOA and formerly unsuspected off-target activities, Profile-based phenotype discovery and screening, and lead generation. [13]

Machine learning which is a subset of AI, comprises supervised learning, unsupervised learning, and reinforcement learning. Further, a subset of machine learning called deep learning has been extensively implemented in drug design and development. The

examples of the algorithms used for drug design and discovery process includes ANN, deep neural network, classification and regression, SVMs, GANs, meta-learning and symbolic learning. AI has been applied in different fields of drug design and development process, such as from peptide synthesis to molecule design, virtual screening to molecular docking, QSAR to repositioning of drug, misfolding of protein to protein–protein interactions, and identification of molecular pathway to poly-pharmacology. The principles of AI have been used in the classification of active and inactive, monitoring drug release, pre-clinical and clinical development, primary and secondary drug screening, pharmaceutical manufacturing, biomarker development, bioactivity identification and physiochemical properties, prediction of toxicity, and identification of MOA. [6]

The researchers in drug discovery use simpler model systems as ethically and practically, the evaluation of the safety and efficacy of all candidate compounds in humans is not practicable. Screening assays are often used for testing thousands to millions of small molecules to identify target hits. [15] Before screening, up to thousands of compounds are profiled in exploratory assays in discovery stages, which combine imaging, primary or induced pluripotent stem cell-derived cells and/or genetic editing. Thus, key biological readouts which can be used for the subsequent phase of screening at high throughput profiling, may be revealed in the early stages of drug discovery. [13]

Latest Literature Review

1. Easy-to-Collect Information (EI) & Hard-to-Collect Information (HI)

Many experiments face a trade-off between gathering easy-to-collect information (EI) on many samples or hard-to-collect information (HI) on a smaller number of samples due to costs in terms of both money and time. With conditional GAN model known as feature mapping GAN (FMGAN), the results of expensive experiments can be predicted, saving on the costs of actually performing the experiment. This could have an impact on pharmaceutical drug discovery, where early phase experiments need casting a wide net to determine just a few potential leads to follow. [1]

2. Drug-Target Interaction (DTI)

In aspects of computational methodologies, Chen et al. classified existing models for DTI prediction into Network-based, Machine Learning based, and other models. **Network-based methods** apply graph-theoretic algorithms to approach the DTI prediction task, where the nodes represent drugs and targets, and the edges model interactions between the nodes. Therefore, the DTI prediction task becomes a link prediction problem. **ML methods** train a parametric or a non-parametric model recursively to approach the DTI prediction problem using supervised, unsupervised, or semi-supervised algorithms, with a training set which is independent and identically distributed and is made up of drug–target pairs. [3]

Mainly, the ML methods in the literature are Similarity-based and feature-based. **Similarity-based approaches** leverage the drug–target, target–target and drug–drug similarities to forecast new interactions. **Feature-based methods** use a numerical vector to represent each drug or target, which reflects the entity's properties like physicochemical features and then the numerical vectors are applied to train a Machine Learning model to anticipate unknown interactions. Furthermore, several DL methods have been proposed to learn the features of compounds and targets for DTI prediction, whereas others have proposed DL models that take predefined features as inputs. [3]

3. Drug-Drug Interaction (DDI)

Drug–drug interaction contributes to 30% of the unpredicted clinical adverse drug events and is a crucial public health problem. Machine learning framework "SMDIP", using Drug-Bank as one of the most reliable pharmaceutical knowledge bases, is a promising framework for discovering DDIs, which can be multifariously feasible in drug development, post marketing surveillance, and public health fields. Research efforts have been oriented towards predictive modeling, including machine learning, to reveal undisclosed DDIs during the drug discovery process prior to the commercial availability or market launch date. [2]

A meta-learning framework with the node2vec for drug representation and the bagging SVM as the classification algorithm is designed to predict DDIs, provided with the lack of negative interaction data for the DDI prediction task and the availability of ample data sources publishing new features related to drugs. [4]

4. High-Throughput Screening (HTS)

Automated methods such as large-scale High-Throughput Screening (HTS) have enhanced traditional in vivo and in vitro methods for analyzing bio-activities. The automation is influenced by the quest to curtail the cost and time-to-market challenges that are related with the drug development process. [3]

The availability of large-scale chemo-genomic and pharmacological data (such as Drug-Bank, KEGG, STITCH, and Chem BL, Davis, KIBA, Pub Chem.), coupled with advances in computational resources and algorithms have given rise to the growth of the in-silico Virtual Screening (VS) domain. In-silico methods have the potential to tackle the above mentioned challenges that plague HTS due to their ability of analyzing assay data, reveal inherent relationships, and exploit such latent information for drug discovery tasks. [3]

5. Image-Based Profiling in Drug Discovery

The highly expansive the profile, the higher biological information it can conceivably encode, however difficult it is to extract the information for any provided individual task from among irrelevant information. Such a problem is called curse of dimensionality which can be tackled by combining the features by weighted aggregation and/or more powerful representations generated by machine learning, having various levels of supervision. In a way, an assay developer applies feature weighting when exploring and optimizing biological models, reagents and readouts for a screening assay. [13]

Deep learning approaches have great potential to automate analysis at various stages such as segmentation, feature extraction, classification to reduce manual tuning of analysis pipelines. A novelty detection framework to identify unexpected phenotypes is recently explored application of supervised learning in image-based profiling, particularly deep neural networks. Label-free profiling and the prediction of targeted drug screening assays are also future approaches exploiting image-based profiling data. [5]

At the least, profiling can organize hits into clusters with same biological effects and at the best, it can hint at a compound's mechanism of action and prior unsuspected off-target activities. [13] The screening of extensive collections of small molecules to identify research probes and therapeutic leads with useful biological properties (called high-content screening) is the most popular application of high-throughput imaging. High-throughput image-based screens involve the development of assays that measure particular morphological properties of single cells. [14]

THE TECHNOLOGY USED

1. Deep Learning in Drug Profiling

Deep learning, a subset of machine learning, is essentially a neural network with multiple layers which attempts to imitate the behavior of the human brain, allowing it to learn from large datasets. Deep learning Neural Networks or ANNs attempt to simulate the human brain through a combination of data inputs, weights, and bias. These elements work collectively to precisely recognize, classify, and describe objects within the data. [12]

A computer model in deep learning learns to perform classification tasks from images, text, or sound directly. The DL models are trained by using a huge labeled datasets and multiple layered neural network architectures which learn features directly from the data without requiring manual feature extraction and can therefore acquire state-of-the-art accuracy, sometimes beyond human-level performance. [17]

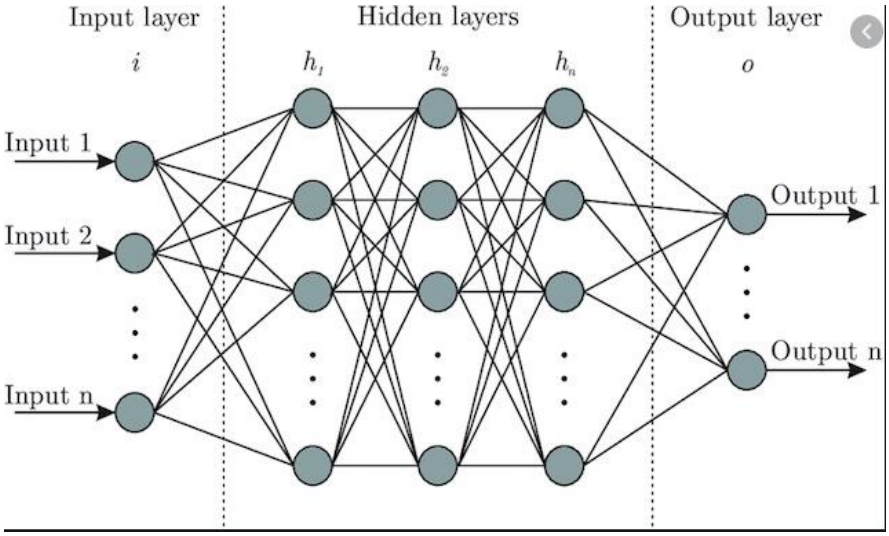

Fig.1. A neural network comprising set of interconnected nodes within organized layers and consists of the input layer, the hidden layers and the output layer. [17]

The Deep learning algorithms are incredibly complex, and specific problems or datasets can be addressed by different types of neural networks. For instance, Convolution neural networks can identify features and patterns within an image, enabling object detection or recognition tasks. Recurrent neural network leverage sequential or times series data and are therefore used in NLP and speech recognition applications. [12]

Profiling, an alternative strategy to screening, focuses to capture a wide variety of features that might have relevance to potential treatment or a disease and therefore represents model systems with a more comprehensive set of features. [15] Few or none of these features may have previously validated relevance to a disease or potential treatment and thus it reveals unexpected biology.[13] Profiling enables high-throughput experimentation and multiplexed readouts

to produce massive amounts of mineable data and is a powerful approach for systems biology and drug discovery applications. [14]

The Process of Drug Discovery

The contemporary drug discovery involves the recognition of screening hits, medicinal chemistry and optimization of those hits to improve the affinity, selectivity (to lower the potential of side effects), efficacy or potency, metabolic stability (to increase the half-life), and oral bioavailability. When a compound attaining all of these requirements is recognized, the process of drug development can continue. If successful, the clinical trials are developed. [9]

The use of deep learning in drug discovery is primarily categorized into: Drug properties prediction, Drug-target interaction prediction and De Novo drug design. [7]

1. Drug Properties Prediction

It can be outlined as a supervised learning problem or a multi-label classification or regression task. The Input is a drug (small molecule) and the Output is 0–1 label to specify whether a drug has certain properties or not (like drug toxicity or solubility). There are distinct ways to denote a drug such as Molecular Fingerprint, SMILES Codes, Graph-Structured Data. [7]

Molecular fingerprint. It is a sequence of bits (binary digits) that indicate the presence or absence of particular substructures in the molecule. Therefore, a small compound (drug) is described as a vector (array) of zeros and ones. [7]

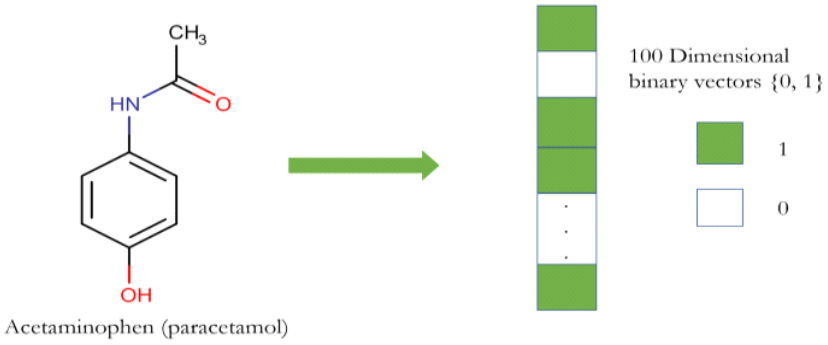

Fig. 2. A molecule of Paracetamol is represented into binary vectors {0,1} where 0 implies the absence of a specific substructure and 1 implies presence of a specific substructure in the molecule. [7]

SMILES code. An approach of transforming graph-structured data into textual content and applying the encoded string or text in the machine learning pipeline is called Simplified Molecular-Input Line-Entry System (SMILES). After transformation, we can use powerful algorithms from natural language processing literature to process the drug and for instance, forecast the properties, side effects, or even chemical-chemical interaction. [7]

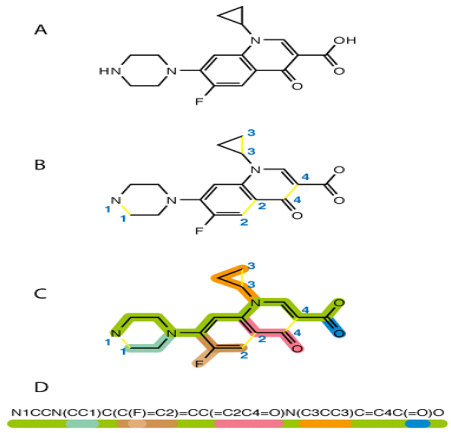

Fig. 3. (A).Chemical Structure of molecule to be represented into SMILES code. (B).The chemical graph is transformed to eliminate hydrogen atoms and the cycles are split to turn it into a spanning tree. (C)At each point where cycle has been broken, numeric suffix labels are incorporated to specify the connected nodes. (D) Parentheses are used to show points of branching on the tree and the color coding is done as well. SMILES string is attained by printing the symbol nodes found in a depth-first tree traversal of a chemical graph of the molecule. [7]

Graph-structured data. The universality of deep learning on graph-structured data, such as graph convolution network, has enabled the use of graph data directly as an input to the deep learning pipeline. E.g., a compound can be contemplated as a graph, where vertices are atoms, and edges are chemical bonds between atoms. [7]

2. Drug-Target Interaction Prediction

The functionality of a protein explicitly depends on its 3D structure which means changing the protein's structure can remarkably modify the functionality of the proteins, and it is one of the essential facts for drug discovery process. However, to anticipate whether a particular drug can bind to particular proteins or not, is critical problem in computational drug discovery. This concept is called drug-target interaction (DTI) prediction. [7]

We can frame the Drug-Target Interaction prediction as binary classification that anticipates the binding affinity of compound and protein which can be characterized as a regression task or binary classification. The Inputs are Compounds and proteins representation, and the Output is 0–1 or a real number in [0–1]. [7]

3. De Novo Drug Design

Generative models, from the autoregressive algorithm, Normalizing flows, Variational auto-encoders, and generative adversarial networks, have become pervasive and are applied for the task of De Novo drug design. The obstacle is, generating a compound, given certain desirable properties. The space of feasible chemical molecules is extremely large and searching in such a space to find a proper drug is very gradual and futile task. [7] Rafael Gomez-Bombarelli et al. recommended a technique for an automatic chemical design using a data-driven continuous representation of molecules that produces the SMILES as the output, and at the end, transforms the SMILES into the chemical space. [8]

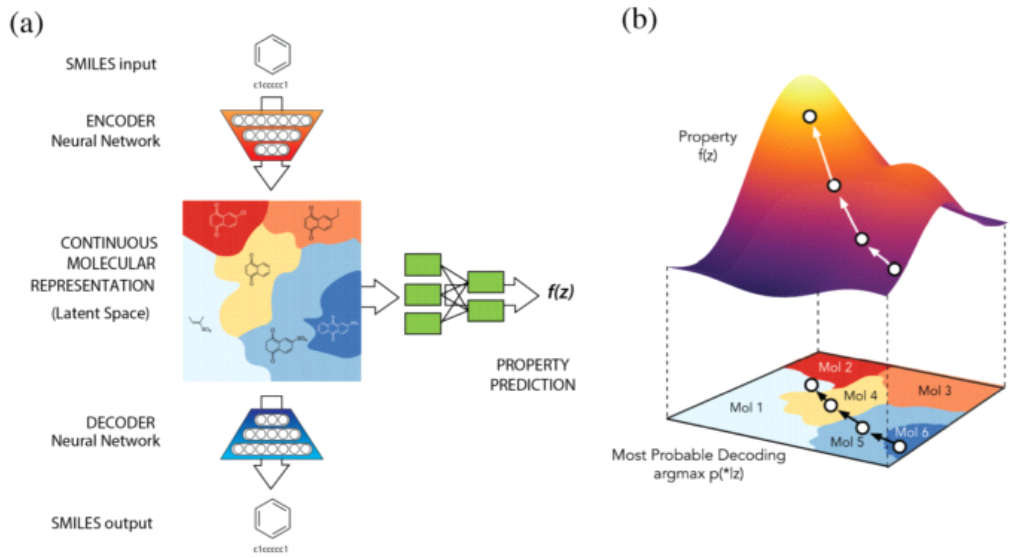

Fig. 4. (a) Auto-encoder and joint property prediction model are used for the molecular design where the encoder network transforms each molecule of SMILES string into a vector in the latent space which is efficiently a continuous molecular representation, and on opposite the decoder network generates corresponding SMILES string for a given point in the latent space. A muti-layer perceptron network evaluates the value of target properties associated with each molecule. (b) The gradient-based optimization in continuous latent space after training a model $f(z)$ to forecast molecule's properties based on their latent representation z, can optimize $f(z)$ with respect to z to discover new latent representations predicted to have high values of desired properties. Those can then be decoded into SMILES strings, at which point their properties can be tested empirically. [8]

Challenges and Future Work

The analysis and production of data needs to be promoted from what can be done to what should be done to arrive at safe and efficacious drugs faster and at lower cost. The potential of AI in drug discovery cannot be fully utilized by the current proxy measures and available data, specifically when it comes to drug efficacy and safety in vivo. Thus, the key to advance clinically relevant decision-making in the future will be addressing the questions of which data to generate and which end points to model. [11]

1. The Quality is more crutial than Speed and Cost in Drug Discovery

This categorically implies to the constant utilization of proxy metrics, such as correlation coefficients or root mean

squared error (RMSE) of models instead of impact on project success and proxy measures, for instance activity on a target instead of readouts related to efficacy or safety. Beyond Pearson correlation coefficient (R2) or RMSE values, how could quality be defined in the context of an AI model? [11]

2. Associated Chemical and Biological Properties for Drug Discovery and the range to which they are Conquered in Current Data

High quality decisions are required and the data produced by high-throughput systems in recent decades can only be used to make this type of decision in some cases, as an outcome of proxy nature of the data. To improve drug discovery, acknowledging the appropriateness of a given end point to answer a given question is at least as important as modeling a particular end point, as judged by a numerical performance measure. [11]

3. AI in Drug Discovery or AI in Ligand Discovery? What is needed to be Advanced?

Better-quality decisions need to be made on which compound to advance, but currently much of the proxy data provides only limited value for making such decisions in reality. So, on what basis those decisions need to be made on? And will AI enhance the decision making on the data currently available and to what extent? [11]

4. Competitions, Data Sharing and Validation of AI Models in Drug Discovery

Mostly the present efforts of AI in drug discovery are directed towards ligand discovery, and this can certainly support in validating a target with respect to its ability to recover the diseased phenotype. However, a ligand is not yet a drug in the context of drug discovery. We need to shift to more complex biological systems earlier, and more often to validate AI systems in drug discovery. This implies including more predictive end points in models, associated to both efficacy and safety, at the computational level. [11]

5. Improving the present strategies of using AI in drug discovery

Better compounds are needed for clinical trials, including the right dosing/PK for acquiring a safe therapeutic index, which involves selection through efficacy and safety-relevant end points. Better patient selection through biomarkers probably increases the chance of clinical success in the future, in relation to both efficacy and safety end points. For quality-based decision making, only comprehending relevant biological end points will allow us to produce the data required which is needed to make drug discovery overall more successful. [11]

Deep learning-based methods have only begun to decode some elementary problems in drug discovery. Definite methodological advances, such as message-passing models, spatial-symmetry-preserving networks, hybrid de novo design, and other creative machine learning models, will probably help address some of the most challenging questions. In advancing the drug discovery with AI, the open data sharing and model development will play a vital role. [10] Above technological innovations like AI and machine learning for drug discovery and early development, innovations will appear in clinical trial models to reduce costs and put potential therapies on a faster track to marketing authorization. [16]

CONCLUSION

The domain of Artificial Intelligence in drug discovery has got recently much attraction, however using the current approaches of producing and utilizing data, we are not likely to acquire the remarkably better decisions that are needed to make drug discovery prosperous. The key reason for this is the use of proxy data at many stages of decision making i.e., core type of data available on a large scale for computational models. To really improve the domain, we require to infer the biology better and produce data that contains a signal of interest in a hypothesis-driven manner, associated both to efficacy and safety end points. We require to infer what to measure, and how to measure it, to be able to anticipate drug efficacy and safety with the desired quality, and then with reduced cost and increased speed. Also, the application of ML in drug discovery will assist from a strategic and unified database as the quality of the predictions made by the models will depend on the quality of the data.

Further, Image-based profiling studies detailed the capability to enhance the pre-clinical development of small molecules at almost any step of the pipeline from target identification over mechanism of action prediction to toxicity profiling. Expanding the throughput and broadening more complex analysis methods of image based phenotypic screens and profiling approaches will help to extend the methodological portfolio of cellular screens to reinforce the drug development process.

REFERENCES

"Matthew Amodio, Dennis Shung, Daniel B. Burkhardt, Patrick Wong, Michael Simonov, Yu Yamamoto, Davidvan Dijk, Francis Perry Wilson, Akiko Iwasaki, Smita Krishnaswamy: Generating Hard-to-Obtain Information from Easy-to-Obtain Information: Applications in Drug Discovery and Clinical Inference. Patterns 2(7) (9-JULY-2021) https://www.sciencedirect.com/science/article/pii/S2666389921001215

"Heba Ibrahimab, Ahmed M. El Kerdawycd, A. Abdoae, A. Sharaf Eldinaf": Similarity-Based Machine Learning Framework for Predicting Safety Signals of Adverse Drug–Drug Interactions. Informatics in Medicine Unlocked 26 (2021) https://www.sciencedirect.com/science/article/pii/S2352914821001830

"Brighter Agyemangab, Wei-Ping Wuab, Michael Yelpengne Kpiebaarehab, Zhihua Leiab, Ebenezer Nanorab, Lei Chenb": Multi-View Self-Attention for Interpretable Drug–Target Interaction

Prediction. Journal of Biomedical Informatics 110 (October 2020) https://www.sciencedirect.com/science/article/pii/S1532046420301751

"S.S. Deepika, T.V. Geetha": A Meta-Learning Framework Using Representation Learning to Predict Drug-Drug Interaction. Journal of Biomedical Informatics 84 (August 2018) https://www.sciencedirect.com/science/article/pii/S1532046418301217

"Christian Scheeder, Florian Heigwer, Michael Boutros": Machine Learning and Image-Based Profiling in Drug Discovery. Current Opinion in Systems Biology 10 (August 2018) https://www.sciencedirect.com/science/article/pii/S2452310018300027

"Rohan Gupta, Devesh Srivastava, Mehar Sahu, Swati Tiwari, Rashmi K. Ambasta & Pravir Kumar": Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. Molecular Diversity 25 (2021) https://link.springer.com/article/10.1007/s11030-021-10217-3

"Hosein Fooladi": Review: Deep Learning In Drug Discovery. (26-Feb-2020) https://towardsdatascience.com/review-deep-learning-in-drug-discovery-f4c89e3321e1

"Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik": Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science 8(1) (12 January 2018) https://pubs.acs.org/doi/full/10.1021/acscentsci.7b00572

Wikipedia: Drug Discovery.

https://en.wikipedia.org/wiki/Drug_discovery#:~:text=In%20the%20fields%20of%20medicine,new%20candidate%20medications%20are%20discovered.&text=Once%20a%20compound%20that%20fulfills,successful%2C%20clinical%20trials%20are%20developed.

"José Jiménez-Luna, Francesca Grisoni, Nils Weskamp & Gisbert Schneider": Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. Expert Opinion on Drug Discovery 16(9) (Mar 2021) https://www.tandfonline.com/doi/full/10.1080/17460441.2021.1909567

"Andreas Bender, Isidro Cortés-Ciriano": Artificial Intelligence in Drug Discovery: What is Realistic, What are Illusions? Part 1: Ways to Make an Impact, and why we are not There Yet. Drug Discovery Today 26(2) (feb-2021) https://www.sciencedirect.com/science/article/pii/S1359644620305274

IBM cloud learn Hub: What is Deep Learning?, https://www.ibm.com/cloud/learn/deep-learning

"Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd & Anne E. Carpenter": Image-Based Profiling for Drug Discovery: due for a Machine-Learning Upgrade?. Nature Reviews Drug Discovery 20 (22-December-2020) https://www.nature.com/articles/s41573-020-00117-w

"Juan CCaicedo, Shantanu Singh, Anne E Carpenter": Applications In Image-Based Profiling of Perturbations. Current Opinion in Biotechnology 39 (June-2016) https://www.sciencedirect.com/science/article/pii/S0958166916301112

"Shannon Gunn": Image-Based Profiling for Drug Discovery. (22-January-2021) https://d4-pharma.com/image-based-profiling-for-drug-discovery/

"Belén Garijo": What to Expect from the Next Decade of Drug Development. World Economic Forum (28 Feb 2020) https://www.weforum.org/agenda/2020/02/technology-in-drug-discovery-and-development/

What is Deep Learning? Things You Need to Know, https://www.mathworks.com/discovery/deep-learning.html#:~:text=Deep%20learning%20is%20a%20machine,to%20humans%3A%20learn%20by%20example.&text=In%20deep%20learning%2C%20a%20computer,images%2C%20text%2C%20or%20sound.