WORLD JOURNAL OF PHARMACEUTICAL RESEARCH
SJIF Impact Factor 8.453
Volume 13, Issue 23, 672-692. Review Article ISSN 2277–7105

ROLE OF ARTIFICIAL INTELLIGENCE (AI) IN DRUG DISCOVERY PROCESS

Dhanraj Pundlik Jadhav[1*], Poonam Khade[2] and Megha T. Salve[3]

Department of B. Pharmacy.

Article Received on
14 October 2024,

Revised on 03 Nov. 2024,
Accepted on 24 Nov. 2024

DOI:10.20959/wjpr202423-34783

***Corresponding Author**
Dhanraj Pundlik Jadhav
Department of B. Pharmacy.

ABSTRACT

Artificial intelligence (AI) has been transforming the practice of drug discovery in the past decade. Various AI techniques have been used in many drug discovery applications, such as virtual screening and drug design. In this survey, we first give an overview on drug discovery and discuss related applications, which can be reduced to two major tasks, i.e., molecular property prediction and molecule generation. We then present common data resources, molecule representations and benchmark platforms. As a major part of the survey, AI techniques are dissected into model architectures and learning paradigms. The use of data augmentation, explainable AI, and the integration of AI with traditional experimental methods, as well as the potential advantages of AI in pharmaceutical research, are also discussed. Overall, this review highlights the potential of AI in drug discovery and provides insights into the challenges and opportunities for realizing its potential in this field. To reflect the technical development of AI in drug discovery over the years, the surveyed works are organized chronologically. We expect that this survey provides a comprehensive review on AI in drug discovery. We also provide a GitHub repository with a collection of papers (and codes, if applicable) as a learning resource, which is regularly updated.

KEYWORDS: Artificial intelligence; drug discovery; AI-assisted content generation; AI-limitations.

INTRODUCTION

Artificial intelligence (AI) is the simulation of human intelligence in machines that are programmed to analyze and process data in a human-like manner. The term may be applied to any machines or algorithms that exhibit traits normally associated with human behaviors such

as pattern recognition, extrapolation, and problemsolving. The ideal characteristic of AI is its ability to rationalize and identify solutions that have the best chance of achieving a specific goal. AI can be categorized into different subsets such as machine learning and deep learning (Fig. 1). Machine learning refers to computer programs that can automatically learn from and adapt to new data and can be supervised by a teacher or unsupervised and learns independently. One form of machine learning involves use of artificial neural networks (ANNs). These are statistical models directly inspired by and modeled on biological neural networks. ANNs have unique capabilities that help to solve tasks that rote learning models could never solve. Deep learning is a subset of machine learning enabling automatic learning using multilayered ANNs trained on vast amounts of data. Within each layer of the neural network, deep learning algorithms perform calculations and make predictions repeatedly, progressively learning and progressively improving the accuracy of the outcome over time as the size of the data increases, all without further human intervention.[1] AI is integrated in our lives in many ways, through personal assistants, computers, smartphones, smartwatches, social media algorithms, and gaming. In 1955, John McCarthy coined the term "artificial intelligence" to describe "the science and engineering of making intelligent machines, especially intelligent computer programs".[2]

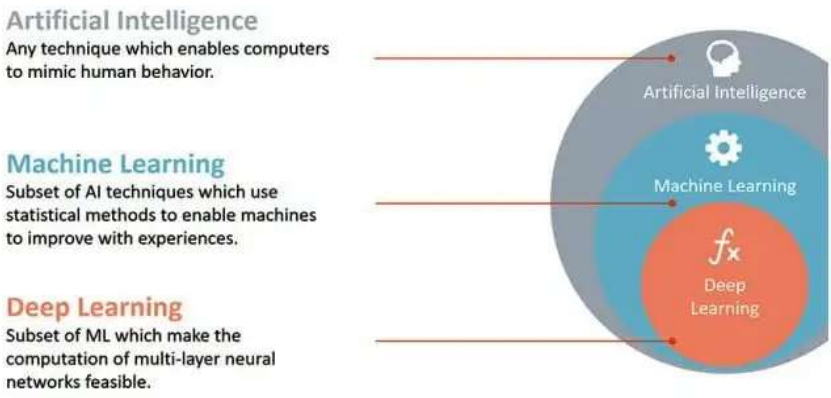


Fig. 1: Overview of the different types of technologies used to simulate human-like learning behaviors in machines, known broadly as artificial intelligence (AI). Machine learning is a subfield of AI in which computer algorithms process and learn from inputs in order to improve accuracy of future predictions. These can be supervised by a teacher or unsupervised and learnt independently. Deep learning consists of multiple layers of artificial neural networks (ANNs) or collections of connected nodes that can progressively learn and improve their predictions independent of human intervention.

Computers perform tasks that replicate human intelligence but can ultimately expand even beyond biological Fig. 1 Overview of the different types of technologies used to simulate human-like learning behaviors in machines, known broadly as artificial intelligence (AI). Machine learning is a subfield of AI in which computer algorithms process and learn from inputs in order to improve accuracy of future predictions. These can be supervised by a teacher or unsupervised and learnt independently. Deep learning consists of multiple layers of artificial neural networks (ANNs) or collections of connected nodes that can progressively learn and improve their predictions independent of human intervention 132 Sunil Thomas et al. systems. AI programming benefits from large datasets to expand its capability to learn; therefore, resulting in higher accuracy.[3] Though AI was first introduced in the 1950s, it was only in the last decade that computer scientists aided by growth in computational power have managed to unlock many new applications. Advantages of AI-based approaches to medicine and health care are that AI can automatically learn to recognize intricate patterns in the input data and create predictive models even when human understanding of the underlying biological processes is limited.[4]

Predicting the three-dimensional structures of potential target proteins, solely from their amino acid sequence, is often necessary for drug discovery, and AI systems had a major recent success in this, with AlphaFold2[5] winning the Critical Assessment of Structure Prediction CASP14.[6] Existing deep learning-based libraries, such as DeepChem and DeepAffinity, and databases, including PubChem, PDB, and ChEMBL, that could help drug discovery are discussed, along with AlphaFold2. As drug discovery applications focus on the three-dimensional structures of molecules (proteins, DNA, RNA, and drugs/medicines) and their interactions, the atom is the fundamental unit of these structures and can be considered as a "machine learning datatype". Molecular systems contain poorly described higher-level patterns, which could be learned from their data. Interrelations among biomedical data are attributes that could be represented in the form of graphs in the design of data-driven systems. Graph machine learning allows modeling of unstructured multimodal datasets[7] and so could model more complex relationships between drugs and disease, protein-protein interactions, side effects of drugs, prediction of responses to a drug and drug re-purposing.[8]

When coupled with an attention mechanism, graph machine learning may identify drug binding sites[9], highly communicating residues/atoms, and provide more interpretable models.[10] A detailed discussion of molecular representation, GNNs, and their application in

the context of drug discovery processes is presented. Experimental high-throughput screening, combinatorial chemistry, and other technical methods have been the main choices to create new chemical entities with specific desired features[11] but AI applications now have the potential to be better than a human expert.[12] The application of GNN, generative models and RL for de novo molecule generation and optimization is presented. Simulation of bio-molecular structures by detailed, physics-based atomic methods, such as molecular dynamics (MD),[13] is central to drug discovery and biotechnology. The 3D structures of proteins and drugs from the Protein Data Bank (PDB) and DrugBank (or structures predicted by AlphaFold2) can be docked for MD simulations, to investigate the stability, dynamics, geometry, and binding efficacy of a protein-drug complex, giving a time-trajectory of atomic movements. Deep learning or advanced data analysis methods can be applied to analyze these trajectories of biological systems, hopefully leading to new hypotheses about the structural changes and interactions in complex biological systems, that may answer questions about diseases, pathways, and drug response / resistance mechanisms. Structure-based drug design, with the application of MD simulations, for the analysis of drug response and resistance, is discussed.[14]

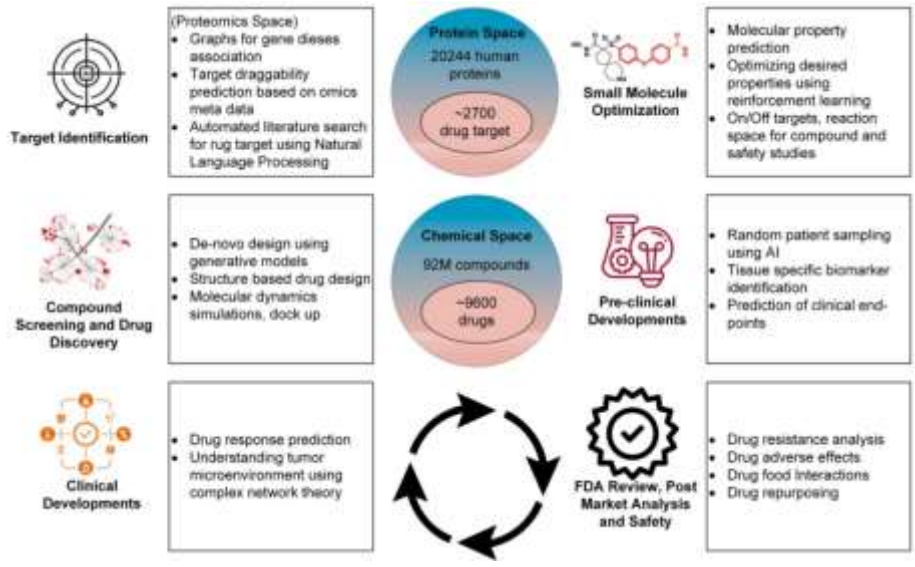


Fig. 2: Applications of AI-based methods at different stages of a drug discovery pipeline. There are about 2700 known potential drug target proteins in the human body and about 9600 FDA-approved small molecule drugs. Machine learning can be used to identify the targeted protein, GNNs can be used for predicting drug-target interactions and binding affinity, and reinforcement learning can be used to optimize the properties of a molecule. Computer vision can determine the spatial state of the tumor

microenvironment. Generative models can be employed to design new molecules, simulation-based studies can suggest properties of protein drug complexes, such as stability and dynamics, and NLP can be used to mine the existing scientific literature for drug re-purposing, FDA review, and post-market analysis.

Application of data science in the drug discovery process

The emergence of epidemics and pandemics, such as influenza and COVID-19[15], and the prevalence of severe diseases, such as cancer and heart disease, demonstrate the ongoing need to discover new drugs. A multi-stage process (Fig. 2), requiring target identification, validation, high throughput screening, animal studies, safety and efficacy protocols, clinical trials, and regulatory approval, is usually followed.[16] Development of a new drug takes approximately 14.6 years and costs about US$ 2.6 billion[17] on average. AI-based methods could be utilized at several stages in this process: identifying novel targets[18], evaluating drug-target interactions[19,20], examining disease mechanisms[21], and improving small molecule compound design and optimization. These methods can also be used to identify and develop prognostic bio-markers, and study drug efficacy, response, and resistance.[22]

Target identification in drug discovery

Target identification during drug discovery aims to identify molecules, usually proteins, that could alter a disease state if their activity was modulated. Machine learning algorithms can analyze various types of data, including gene expression profiles, protein-protein interaction networks, and genomic and proteomic data, to identify potential targets that are likely to be involved in disease pathways.[23] Of the approximately 20,000 proteins in the human proteome, only about 3,000 have been identified as potential therapeutic targets.[24] Future knowledge might expand our understanding of which proteins could become drug targets. The first step in identifying a target is to establish a causal relationship between the target and the disease.[25] Causal relationships between genes and diseases can be identified using graphs, GNNs, or tree-based methods. A decision tree-based meta-classifier trained on a network topology involving protein-protein, metabolic, and transcriptions interactions, and tissue expression and subcellular localization of proteins was proposed in[26] to predict morbidity-associated genes that are also druggable. Regulation by multiple transcription factors (TFs), centrality in metabolic pathways, and extracellular location were identified as key parameters from the decision tree. Machine learning-based methods classified proteins as drug targets or non-targets for specific diseases, such as lung, pancreatic, and ovarian cancer,

based on features such as protein-protein interaction, gene expression, DNA copy number, and occurrence of mutations.[27] The primary source of information on target association with disease is the literature. Text mining and Natural Language Processing (NLP) approaches can also be used to identify relevant target-disease pairs from literature and develop databases for target identification.[28] BeFree[29], PKDE4J[30] and other deep learning-based tools[31] can be used to mine articles to identify drug–disease, gene–disease, and target–drug associations. Drug–target interactions may also be inferred, based on descriptor similarity to reference ligands, in the same cell without explicitly addressing the target identity of those reference ligands. A software tool (SPiDER)[32] discretizes the input feature similarity vector onto a so-called feature map using a neural network-inspired approach.

Virtual screening and optimization of compounds

AI can be used to virtually screen and optimize compounds, estimate their bio-activities, and predict protein-drug interactions.[33] One way AI can help in virtual screening is through the development of predictive models, that can identify compounds with a high probability of binding to a target protein. These models can be trained using various types of data, such as known protein-ligand complexes, structural information, and molecular descriptors. Physico-chemical properties of the drug, such as solubility, partition coefficient (logP), degree of ionization, and intrinsic permeability, may have an indirect effect on a drug's interaction with a target receptor family and must be considered when designing a new drug.[34] AI can also be used to plan efficient routes for chemical synthesis and develop insights into the reaction mechanisms of drugs to identify potentially unwanted interactions with other molecules. Candidate structures of drugs are refined and modified to improve target specificity and selectivity, and their pharmacodynamics, pharmacokinetics, and toxicological properties. A virtual chemical space with structure and ligand information may provide profile analysis, faster elimination of non-lead structures, and speed up the drug discovery process by avoiding costly time-consuming laboratory work. Multi-objective optimization methods can tune molecules in a desired direction. MD simulation and docking methods can be used to model the orientation, stability, and dynamics of the compounds.

Pre-clinical and clinical development

Predicting possible responses to a drug is a critical step in a drug design pipeline. Similarity or feature-based machine learning methods can be used to predict the response of a drug on individual cells and the efficacy of a drug-target interaction by binding affinity or free energy

of binding. Similarity methods assume that similar drugs act on similar targets[35], while feature-based methods find individual features of drugs and targets and feed the drug-target feature vector to the classifier. Deep learning-based methods, such as DeepConv-DTI[36] and DeepAffinity[37] are examples methods, where the embedding of drugs and targets are learned using convolution and attention mechanism. AI-based techniques can assist in selecting potential patients for pre-clinical trials by identifying relevant human-disease biomarkers and anticipating potential toxic or unnecessary side effects[38] and by filtering a high dimensional set of clinical variables to select a cohort of patients. AI can also help in predicting the outcome of clinical trials well ahead of the actual trial minimizing the chance of any harmful effect on patients.[39]

FDA approval and post-market analysis

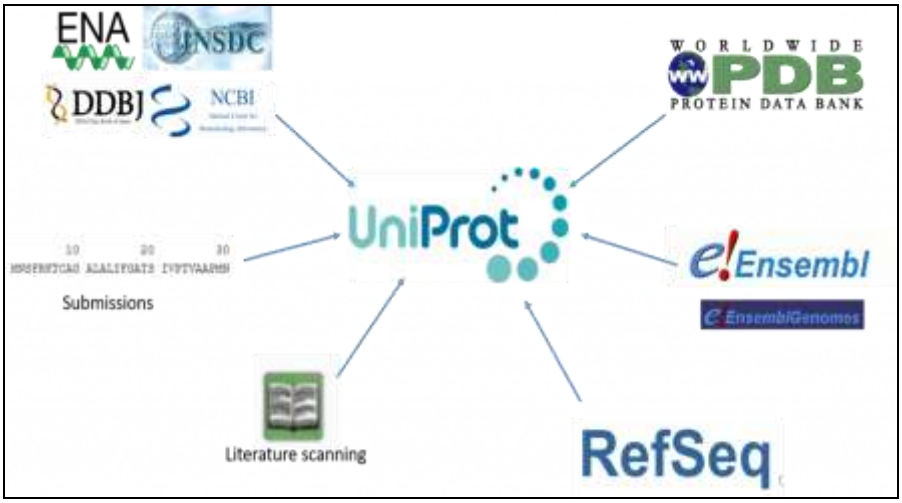
Natural Language Processing (NLP) can be used to mine scientific literature to report adverse effects, such as toxicity, of a drug or resistance to it and prepare automated evaluations for regulatory (FDA) approval or a patent application.[40] NLP-based sentiment analysis methods can be used to recommend drugs.[41] Prediction of likely sales of a product by machine learning-based systems could help pharmaceutical companies optimize their business resources.[42]

❖ **Existing databases and tools for drug development**

1. **PubChem:** PubChem[43] is the largest free database of chemical information, with about 111 Million compounds, 279 Million substances, 295 Million bio-activities, and 34 Million articles, organized into three inter-linked web data pages; substance, compound, and bio assay.[44] The descriptions of, and test results from, bio-assays are stored in the bio-assay database. Data mining methods can be used to identify compounds for a particular target or protein.

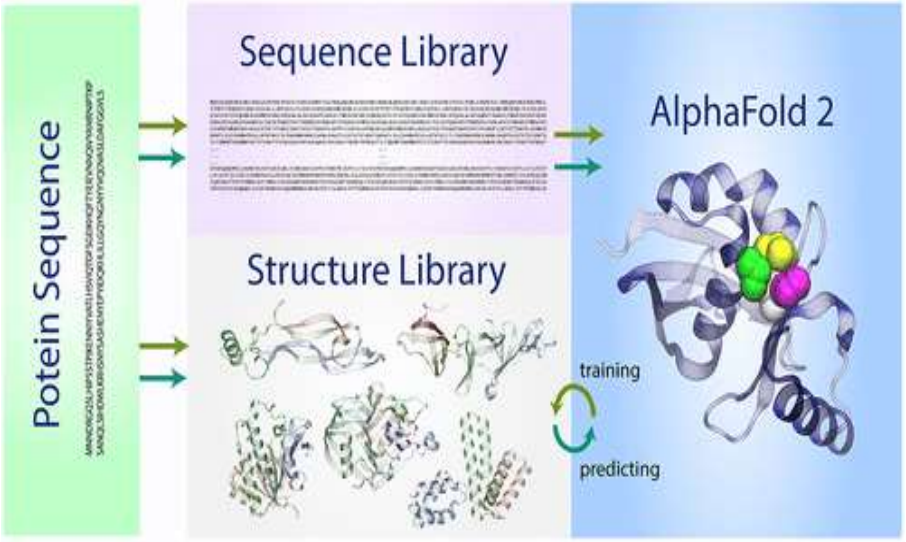
2. ChEMBL: ChEMBL[45] is an open-access drug discovery database, developed by the European Molecular Biology Laboratory (EMBL). Data on authorized and candidate medications, such as the mechanism of action and therapeutic indications, are gathered from full-text papers in high-impact publications and combined with data on small, compounds and their biological activity. The bio-activity data is exchanged with another database; such as BindingDB[46] and PubChem Bioassay. The ChEMBL database has been used to identify chemical tools for a target of interest, to predict drug-target interactions, to re-purpose a drug, to determine target tractability, and to integrate with existing drug discovery tools.[47]

3. UniProt database: UniProt [48]is a public database of protein sequences annotated with taxonomic data and information on biological functions. There are four components; UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), UniProt Archive (UniParc), and UniProt Metagenomic and Environmental Sequences (UniMES). Uniprot contains more than 189 million records; more than half were curated by human experts.



❖ **List of AI-based software for drug discovery, development, and analysis.**

1. AlphaFold2: Predicting the 3D structures of proteins from their amino acid sequence is a very complex and challenging problem. AlphaFold2, developed by DeepMind, has achieved a breakthrough level of accuracy[49]) and is openly available via Google Colab.



2. DeepChem: The DeepChem[50] library is a Tensorflow wrapper that understands and streamlines the analysis of chemical datasets. It has been used for algorithmic research into one-shot deep-learning algorithms for drug discovery and application projects such as modeling inhibitors for BACE-1).[51,52] DeepChem can be used to analyze protein structures, predict the solubility of small molecule drugs and their binding affinity to targets, and count the number of cells in a microscopic image. MoleculeNet[53], which contains the properties of 700,000 compounds has been integrated into the DeepChem package.

3. DeeperBind: DeeperBind[54] is a long short-term recurrent convolutional network that predicts protein binding specificity in relation to DNA probes, which can model the interaction between transcription factors (TF) and their corresponding (DNA/RNA) binding sites. DeeperBind can effectively predict the dynamics of probe sequences. It can also be trained and tested on datasets with sequences of variable lengths.

4. DeepAffnity: DeepAffinity[55] is a semi-supervised model that unifies recurrent and convolutional neural networks to predict the binding affinity between a drug and target sequences. The model uses both labeled and unlabeled data to jointly encode molecular representations under unique structurally annotated protein sequence representations. DeepAffinity outperformed random forest, ensemble methods, and RNN-CNN models.

❖ Applications of artificial intelligence in drug development process

The most arduous and desponding step in the drug discovery and development process is identifying suitable and bioactive drug molecules present in the vast size of chemical space, which is in the order of 1060 molecules. Further, the drug discovery and development process are considered a time- and cost-consuming process. The most infuriating point is that nine out

of ten drug molecules usually fail to pass phase II clinical trials and other regulatory approvals.[56-58] The above-said limitations of drug discovery and development can be addressed by implementing AI-based tools and techniques. AI is involved in every stage of the drug development process such as small molecules design, identification of drug dosage and associated effectiveness, prediction of bioactive agents, protein–protein interactions, identification of protein folding and misfolding, structure and ligand-based VS, QSAR modeling, drug repurposing, prediction of toxicity and bioactive properties, and identification of mode of action of drug compounds as discussed below.

- **Peptide synthesis and small molecule design**

Peptides are a biologically active small chain of around 2–50 amino acids, which are increasingly being explored for therapeutic purposes as they have the ability to cross the cellular barrier and can reach the desired target site.[59] In recent years, researchers have taken advantage of AI and used it to discover novel peptides. For instance, Yan et al. 2020 developed Deep-AmPEP30, a DL-based platform for the identifi cation of short anti-microbial peptides (AMPs).[60] Deep AmPEP30 (https:// cbbio. online/ AxPEP/) is a CNN-driven tool that predicts short AMPs from DNA sequence data. Using Deep-AmPEP30, Yan et al. identified novel AMPs from the genome sequence of C. glabrate, a fungal pathogen present in the GI tract. Likewise, Plisson et al. 2020 combined the ML algorithm with an outlier detection technique to discover AMPs with non-hemolytic profiles.[61] In addition, Kavousi et al. developed IAMPE (http:// cbb1. ut. ac. ir/), a web server for the identification of anti-microbial peptides, which integrates 13CNMR-based features and phys icochemical features of peptides as input to ML algorithms, in order to identify novel AMPs.[62] Similarly, Yi et al. 2019 devised ACP-DL (https:// github. com/ haich engyi/ ACP- DL), a DL-based tool for the discovery of novel anti-cancer peptides.[63] ACP-DL uses the LSTM algorithm, which is an improved version of the recursive neural network (RNN), for differentiating anti-cancer peptides from non-anti-cancer peptides. Moreover, Yu et al.[64] proposed DeepACP, a deep recurrent neural network-based model for identifying anti-cancer peptides. Likewise, Tyagi et al. 2013 developed an SVM-based platform for identifying new anti-cancer pep tides.[65] In addition, Rao et al. 2020 combined a graphical convolutional network and one-hot encoding to design ACP-GCN for the discovery of anti-cancer peptides.[66] Moreover, Grisoni et al. used an ensemble of four counter propagation ANN for identifying new anti-cancer peptides. Likewise, Wu et al.[67] proposed PTPD, a tool based on CNN and word2vec, for the discovery of novel peptides for therapeutics.

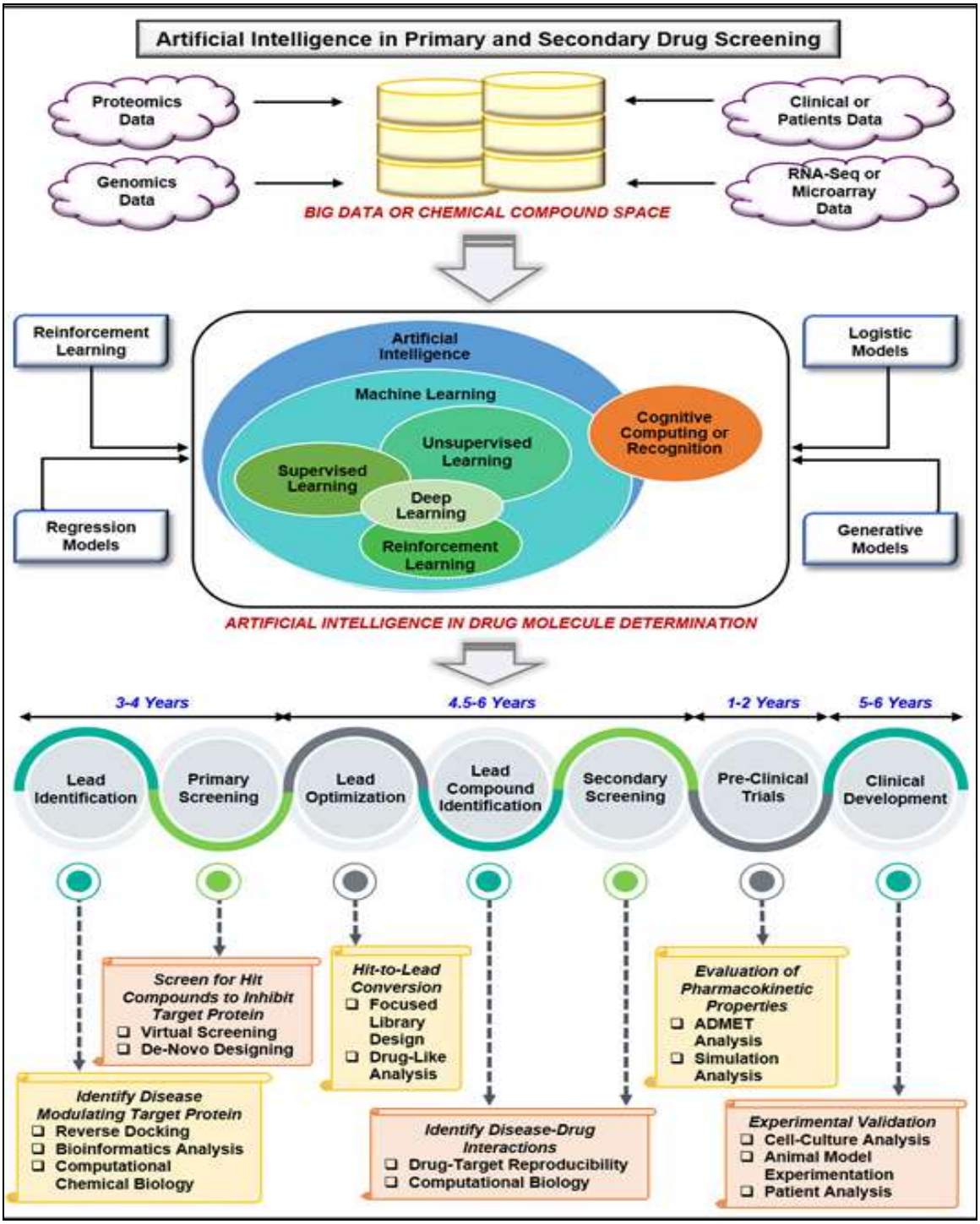


Fig: Artificial intelligence in primary and secondary drug screening: in drug discovery and designing pipeline, screening of potential lead is crucial, and artificial intelligence plays a great role in identifying novel and potential lead compounds. There are approximately 106 million chemical structure presents in chemical space from different studies such as OMIC studies, clinical and pre-clinical studies, in vivo assays, and microarray analysis. With machine learning models such as reinforcement models, logistic models, regression models, and generative models, these chemical structures are screened out based on active sites,

structure, and target binding ability. The complete drug discovery process through artificial intelligence will take about 14–18 years, which is comparatively less than the traditional drug discovery process. The first step in the drug discovery pro cess is lead identification, in which disease-modifying target protein is identified through reverse docking, bioinformatics analysis, and computational chemical biology. In the second step, primary screening of compounds is done to select potential lead compounds, which can inhibit target protein. This can be done through virtual screening and de novo designing. The next step in the drug discovery process includes lead optimization and lead compound identification through focused library design, drug-like analysis, drug-target reproducibility, and computational biology. Afterward, secondary screening of com pounds is performed, followed by pre-clinical trials. The drug discovery process's final step is clinical development through cell-culture analysis, animal model experimentation, and patient analysis.

- **Identification of drug dosage and drug delivery effectiveness**

Administering an improper dose of any drug to a patient can lead to undesirable and lethal side effects; hence, it is crucial to determine a safe drug dose for treatment purposes. Over the years, it has been challenging to ascertain the optimum dose of a drug that can achieve the desired efficacy with minimum toxic side effects.[68] With the emergence of AI, lots of researchers are taking the help of ML and DL algorithms to determine appropriate drug dosage. For instance, Shen et al.[69] developed an AI-based platform, referred to as AI-PRS, to determine the optimum dose and combinations of drugs to be used for HIV treatment through antiretroviral therapy. AI-PRS is a neural network-driven approach, which relates drug combinations and dosage to efficacy through a parabolic response curve (PRS). In their study, Shen et al. administered a combination of tenofovir, efavirenz, and lamivudine to 10 HIV patients, and in due course, using the PRS method, they found out the dose of tenofovir could be reduced by 33% of the starting dose without causing virus relapse. Hence, using AI-PRS optimum drug dosage can be found out for other diseases as well.

- **Structure-based and ligand-based virtual screening**

In drug designing and drug discovery, VS is one of the crucial methods of CADD. VS refers to the identification of a small chemical compound that binds to a drug target. VS is an efficient method to screen out the promising therapeutic compound from a pool of compounds.[70] Thus, it becomes an important tool in high-throughput screening, which incurred the problem of high-cost and low-accuracy rate. In general, there are two important

types of VS that are structure-based VS (SBVS) and ligand-based VS (LBVS).[71,72] The LBVS depends on the chemical structure and empirical data of both active and inactive ligands, which uses the chemical and physiochemical similarities of active ligands to predict the other active ligand from a pool of compounds with high bioactivity. However, the LBVS does not depend on the 3-D structure of the target protein, and thus, this method is implemented where target structure or information is missing, and the obtained structural accuracy is low.[73] On the other hand, SBVS has been implemented in such cases where 3-D structural information of protein or target has been elucidated either through in vitro or in vivo experiments or through computational modelling.[74, 75] In general, this method is used to predict the interaction between the active ligand or its associated target and to predict the amino acid residues, which are involved in drug-target binding. In comparison with LBVS, SBVS possesses high accuracy and precision. How ever, SBVS is associated with the problem of an increasing number of disease-causing proteins and their complicated conformations.[76] To use ML for VS, there should be a filtered training set comprising of known active and inactive compounds. These training data are used to train a model using supervised learning techniques. The trained model is then validated, and if it is accurate enough, the model is used on new data sets to screen compounds with desired activity against a target.[77] After that, the shortlisted compounds can go for ADMET analysis, followed by various bioassays before entering clinical trials. Hence, ML has the power to speed up VS, make it more robust, and can even reduce false positives in VS. Docking is the main principle applied in SBVS, where several AI and ML-based scoring algorithms have been developed such as NNScore, CScore, SVR-Score, and ID-Score. Similarly, ML and DL methods such as RFs, SVMs, CNNs, and shallow neural networks have been constructed to predict protein–ligand affinity in SBVS. Moreover, AI-based algorithms have been developed for molecular dynamic simulation assays in SBVS. On the other hand, LBVS consists of several steps, and each step comes up with novel AI- and ML-based algorithms to speed up the process and increase reliability. For example, several ML- and DL-based algorithms have been constructed for the preparation of useful decoy sets such as Gaussian mixture models (GMMs), isolation forests, and artificial neural networks (ANNs).[78]

CONCLUSION

AI-based methods are being adopted in the health care industry where low-cost, intelligent, and flexible methods are affecting areas such as drug design, support for clinical decision making, diagnosis, prevention, and making clinical recommendations. AI applications were

previously thought to be inferior to experimental high-throughput screening, combinatorial chemistry, and other technical drivers. It was difficult to create new chemical entities using computer programs, with desired features from the ground up, potentially even better than a human expert. The long and costly process of drug design can be accelerated by employing data science methods for target identification, De novo molecular design, drug repurposing, retrosynthesis and prediction of reactivity and bio-activity, FDA approval, and post-market analysis. AI has been implemented by some pharmaceutical organizations, with revenue from AI-based solutions in the pharmaceutical sector estimated to reach US $2.199 billion by 2022. Deep neural networks (DNNs) can be used to boost prediction power when inferring the properties of small molecules, and one-shot learning can be used if a large amount of experimental data is not available. Understanding technical and human errors, labeling constraints, and biological variability associated with the underlying data is crucial to create useful predictive models. It is difficult to represent the experimental data in numerical or computer-assisted form. AI is now being utilized to create representations of trials that allow for data categorization and, ultimately, the development of predictive models. Great things happen in minds and are never done alone, AI is delivering only a platform to execute the plans. We need to develop novel hypotheses for drug discovery by employing the knowledge from different domain experts. After that, we can design a data analysis algorithm, and then we can learn from the data to modulate the hypothesis or modify the algorithms. In short, both mind and machine need to work in synergy. We hope that the use of machine learning, especially deep learning, will increase in the future and help us understand complex biological systems, generate particles with the desired properties, and lead to semi-automated smart healthcare systems. We also expect that AI would be a valuable tool in understanding human biology, a catalyst in combating human diseases and will accelerate drug design. In terms of drug discovery, quality, and safety are more important than speed and cost, devising an AI system that can meet this multi-objective optimization in a multi-dimensional complex space is a huge challenge, which needs collaborative efforts from multiple disciplines in academia and industry.

AKNOWLEDGEMENT

The corresponding author would like to express sincere gratitude to Mrs. Poonam Khade Mam for their aspiring guidance, valuable contributions and support throughout the preparation of this review article. Her dynamic vision, genuine sincerity, and unwavering motivation have been a profound source of inspiration for me. It was a tremendous privilege

and a source of great honour to have the opportunity to learn and collaborate under their exceptional guidance.

REFERENCES

1. Frankenfield J (2021) Artificial intelligence. Retrieved from: https: //www.investopedia. com/terms/a/artificial-intelligence-ai.asp#: ~: text¼Artificial%20intelligence%20(AI)% 20refers%20to,as%20learning%20and%20prob lem%2Dsolving
2. McCarthy J (2004) What is Artificial Intelligence? Retrieved from: http: //www-formal. stanford.edu/jmc/whatisai.pdf
3. Panesar A (2020) What is artificial intelligence? In: Machine learning and AI for healthcare. pp 1–18.
4. Bishop CM (2013) Model-based machine learning. Philos Trans A Math Phys Eng Sci, 371: 20120222
5. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature, 2021; 596(7873): 583–589.
6. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (casp)—round x, Proteins, Struct. Funct. Bioinform, 2014; 82: 1–6.
7. T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell, 2018; 41(2): 423–443.
8. T. Gaudelet, B. Day, A.R. Jamasb, J. Soman, C. Regep, G. Liu, J.B. Hayter, R. Vickers, C. Roberts, J. Tang, et al., Utilizing graph machine learning within drug discovery and development, Brief. Bioinform, 2021; 22(6): bbab159
9. M. Karimi, D. Wu, Z. Wang, Y. Shen, Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, Bioinformatics, 2019; 35(18): 3329–3338.
10. C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol, 2016; 12(7): 878.
11. G.D. Geromichalos, C.E. Alifieris, E.G. Geromichalou, D.T. Trafalis, Overview on the current status of virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; part I, J. Buon, 2016; 21(4): 764–779.

12. G. Schneider, An insight into artificial intelligence in drug discovery: an interview with professor gisbert Schneider, Expert Opin. Drug Discov, 2021; 16(9): 933–935.
13. A. Ganesan, M.L. Coote, K. Barakat, Molecular dynamics-driven drug discovery: leaping forward with confidence, Drug Discov. Today, 2017; 22(2): 249–269.
14. Y. Wang, J.M.L. Ribeiro, P. Tiwary, Machine learning approaches for analyzing and enhancing molecular dynamics simulations, Curr. Opin. Struct. Biol, 2020; 61: 139–145.
15. Y. Zheng, Y. Ma, J. Zhang, X. Xie, Covid-19 and the cardiovascular system, nature reviews cardiology, 2020.
16. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov, 2019; 18(6): 463–477.
17. O.J. Wouters, M. McKee, J. Luyten, Estimated research and development investment needed to bring a new medicine to market, 2009-2018, JAMA, 2020; 323(9): 844–853.
18. J. Jeon, S. Nim, J. Teyra, A. Datti, J.L. Wrana, S.S. Sidhu, J. Moffat, P.M. Kim, A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening, Gen. Med, 2014; 6(7): 1–18.
19. I. Lee, J. Keum, H. Nam, Deepconv-dti: prediction of drug-target interactions via deep learning with convolution on protein sequences, PLoS Comput. Biol, 2019; 15(6): e1007129.
20. T. Katsila, G.A. Spyroulias, G.P. Patrinos, M.-T. Matsoukas, Computational approaches in target identification and drug discovery, Comput. Struct. Biotechnol. J., 2016; 14: 177–184.
21. C.A. Nicolaou, N. Brown, Multi-objective optimization methods in drug design, Drug Discovery Today. Technologies, 2013; 10(3): e427–e435.
22. R. Qureshi, B. Zou, T. Alam, J. Wu, V. Lee, H. Yan, Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: recent advances in drug design, challenges and future prospects, IEEE/ACM Trans. Comput. Biol. Bioinform, 2022.
23. G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe, Computational methods in drug discovery, Pharmacol. Rev, 2014; 66(1): 334–395.
24. T.M. Bakheet, A.J. Doig, Properties and identification of human protein drug targets, Bioinformatics, 2009; 25(4): 451–457.
25. B.-M. Lv, Y. Quan, H.-Y. Zhang, Causal inference in microbiome medicine: principles and applications, Trends Microbiol, 2021; 29(8): 736–746.

26. P.R. Costa, M.L. Acencio, N. Lemke, A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data, in: BMC Genomics, vol. 11, Springer, 2010; 1–15.
27. J. Jeon, S. Nim, J. Teyra, A. Datti, J.L. Wrana, S.S. Sidhu, J. Moffat, P.M. Kim, A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening, Gen. Med, 2014; 6(7): 1–18.
28. J.Y. Khan, M.T.I. Khondaker, I.T. Hoque, H.R. Al-Absi, M.S. Rahman, R. Guler, T. Alam, M.S. Rahman, Toward preparing a knowledge base to explore potential drugs and biomedical entities related to Covid-19: automated computational approach, JMIR Med. Inform, 2020; 8(11): e21648.
29. À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L.I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC Bioinform, 2015; 16(1): 1–17.
30. M. Song, M. Kim, K. Kang, Y.H. Kim, S. Jeon, Application of public knowledge discovery tool (pkde4j) to represent biomedical scientific knowledge, Front. Res. Metr. Anal, 2018; 3: 7.
31. T. Alam, S. Schmeier, Deep learning in biomedical text mining: contributions and challenges, in: Multiple Perspectives on Artificial Intelligence in Healthcare, Springer, 2021; pp. 169–184.
32. D. Reker, T. Rodrigues, P. Schneider, G. Schneider, Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus, Proc. Natl. Acad. Sci, 2014; 111(11): 4067–4072.
33. Ó. Álvarez-Machancoses, J.L. Fernández-Martínez, Using artificial intelligence methods to speed up drug discovery, Expert Opin. Drug Discov, 2019; 14(8): 769–777.
34. Q. Zang, K. Mansouri, A.J. Williams, R.S. Judson, D.G. Allen, W.M. Casey, N.C. Kleinstreuer, In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning, J. Chem. Inf. Model, 2017; 57(1): 36–49.
35. K. Sachdev, M.K. Gupta, A comprehensive review of feature based methods for drug target interaction prediction, J. Biomed. Inform, 2019; 93: 103159.
36. I. Lee, J. Keum, H. Nam, Deepconv-dti: prediction of drug-target interactions via deep learning with tion on protein sequences, PLoS Comput. Biol, 2019; 15(6): e1007129. Convolu.

37. M. Karimi, D. Wu, Z. Wang, Y. Shen, Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, Bioinformatics, 2019; 35(18): 3329–3338.
38. M. Woo, An ai boost for clinical trials, Nature, 2019; 573(7775): S100.
39. S. Harrer, P. Shah, B. Antony, J. Hu, Artificial intelligence for clinical trial design, Trends Pharmacol. Sci, 2019; 40(8): 577–591.
40. Y.Shi, P. Ren, Y. Zhang, X. Gong, M. Hu, H. Liang, Information extraction from fda drug labeling to enhance product-specific guidance assessment using natural language processing, Front. Res. Metr. Anal, 2021; 6.
41. S. Garg, Drug recommendation system based on sentiment analysis of drug reviews using machine learning, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2021; pp. 175–181.
42. N. Khalil Zadeh, M.M. Sepehri, H. Farvaresh, Intelligent sales prediction for pharmaceutical distribution companies: a data mining based approach, Math. Probl. Eng, 2014 (2014).
43. S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, et al., Pubchem substance and compound databases, Nucleic Acids Res, 2016; 44(D1): D1202–D1213.
44. Y. Wang, S.H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B.A. Shoemaker, P.A. Thiessen, S. He, J. Zhang, Pubchem bioassay: 2017 update, Nucleic Acids Res, 2017; 45(D1): D955–D963.
45. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al., Chembl: a large-scale bioactivity database for drug discovery, Nucleic Acids Res, 2012; 40(D1): D1100–D1107.
46. U. Consortium, Uniprot: a hub for protein information, Nucleic Acids Res, 2015; 43(D1): D204–D212.
47. D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, et al., Chembl: towards direct deposition of bioassay data, Nucleic Acids Res, 2019; 47(D1): D930–D940.
48. U. Consortium, Uniprot: a hub for protein information, Nucleic Acids Res, 2015; 43(D1): D204–D212.
49. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature, 2021; 596(7873): 583–589.

50. B. Ramsundar, Molecular machine learning with DeepChem, PhD thesis, Stanford University, 2018.
51. G. Subramanian, B. Ramsundar, V. Pande, R.A. Denny, Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches, J. Chem.
52. Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, Moleculenet: a benchmark for molecular machine learning, Chem. Sci, 2018; 9(2): 513–530.
53. B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of dna- and rna-binding proteins by deep learning, Nat. Biotechnol, 2015; 33(8): 831–838.
54. M. Karimi, D. Wu, Z. Wang, Y. Shen, Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, Bioinformatics, 2019; 35(18): 3329–3338.
55. Álvarez-Machancoses Ó, Fernández-Martínez JL (2019) Using artificial intelligence methods to speed up drug discovery. Expert Opin Drug Discov, 14(8): 769–777. https: //doi.org/10.1080/17460 441.2019.16212 84.
56. Fleming N (2018) How artificial intelligence is changing drug discovery. Nature. https: //doi. org/ 10. 1038/ d41586- 018- 05267-x 108. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neu ral networks. ACS Cent Sci. https://doi.org/10.1021/acscentsci.7b00512
57. Bruno BJ, Miller GD, Lim CS (2013) Basics and recent advances in peptide and protein drug delivery. Ther. Deliv, 4(11): 1443–67. https: //doi.org/10.4155/tde.13.104.
58. Yan J, Bhadra P, Li A et al (2020) Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. Mol Ther-Nucleic Acids, 20: 882–894. https: //doi. org/ 10. 1016/j. omtn. 2020.05.006
59. Plisson F, Ramírez-Sánchez O, Martínez-Hernández C (2020) Machine learning-guided discovery and design of non-hemo lytic peptides. Sci Rep, 10: 1–19. https: //doi. org/ 10. 1038/ s41598-020-73644-6
60. Kavousi K, Bagheri M, Behrouzi S et al (2020) IAMPE: NMR assisted computational prediction of antimicrobial peptides. J Chem Inf Model, 60: 4691–4701. https: //doi. org/ 10. 1021/ acs. jcim.0c00841
61. Yi HC, You ZH, Zhou X et al (2019) ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Ther-Nucleic Acids, 17: 1–9. https: //doi.org/10.1016/j.omtn.2019.04.025

62. Yu L, Jing R, Liu F et al (2020) DeepACP: a novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. Mol Ther-Nucleic Acids, 22: 862–870. https: //doi.org/10.1016/j.omtn.2020.10.005
63. Tyagi A, Kapoor P, Kumar R et al (2013) In silico models for designing and discovering novel anticancer peptides. Sci Rep, 3: 1–8. https: //doi.org/10.1038/srep02984
64. Rao B, Zhang L, Zhang G (2020) ACP-GCN: the identification of anticancer peptides based on graph convolution networks. IEEE Access, 8: 176005–176011. https: //doi.org/10.1109/access.2020. 3023800
65. Wu C, Gao R, Zhang Y, De Marinis Y (2019) PTPD: predicting therapeutic peptides by deep learning and word2vec. BMC Bio informatics, 20: 1–8. https: //doi.org/10.1186/s12859-019-3006-z
66. Dimmitt S, Stampfer H, Martin JH (2017) When less is more efficacy with less toxicity at the ED50. Br J Clin Pharmacol, 83(7): 1365–1368. https: //doi.org/10.1111/bcp.13281
67. Shen Y, Liu T, Chen J et al (2020) Harnessing artificial intel ligence to optimize long-term maintenance dosing for antiretro viral-naive adults with HIV-1 Infection. Adv Ther, 3: 1900114. https: //doi.org/10.1002/adtp.201900114
68. Lavecchia A, Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. Curr Med Chem. https: //doi.org/10.2174/09298673113209990001
69. Gonczarek A, Tomczak JM, Zaręba S et al (2018) Interaction prediction in structure-based virtual screening using deep learn ing. Comput Biol Med. https://doi.org/10. 1016/j. compb iomed. 2017.09.007
70. Goh GB, Hodas NO, Vishnu A (2017) Deep learning for compu tational chemistry. J Comput Chem, 38(16): 1291–1307. https: //doi.org/10.1002/jcc.24764
71. Yang X, Wang Y, Byrne R et al (2019) Concepts of artificial intelligence for computer-assisted drug discovery. Chem. Rev, 119(18): 10520–10594. https: //doi. org/ 10. 1021/ acs. chemr ev. 8b00728
72. Arciniega M, Lange OF (2014) Improvement of virtual screen ing results by docking data feature analysis. J Chem Inf Model. https: //doi.org/10.1021/ci500028u
73. Feinstein WP, Brylinski M (2015) Calculating an optimal box size for ligand docking and virtual screening against experimen tal and predicted binding pockets. J Cheminform, https://doi.org/10.1186/s13321-015-0067-5
74. Gazgalis D, Zaka M, Zaka M et al (2020) Protein binding pocket optimization for virtual high-throughput screening (vHTS) drug discovery. ACS Omega, https://doi.org/10. 1021/ acsom ega. 0c005 22

75. Carpenter KA, Huang X (2018) Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a review. Curr Pharm Des. https: //doi.org/10.2174/1381612824 666180607124038
76. Serafim MSM, Kronenberger T, Oliveira PR et al (2020) The application of machine learning techniques to innovative anti bacterial discovery and development. Expert Opin Drug Discov. https: //doi.org/10.1080/17460441.2020.1776696
77. Melville J, Burke E, Hirst J (2009) Machine learning in virtual screening. Comb Chem High Throughput Screen. https: //doi. org/ 10.2174/138620709788167980
78. Wójcikowski M, Ballester PJ, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep. https: //doi.org/10.1038/srep46710