ISSN 2278 – 4799

International Journal of Drug Discovery and Medical Research

URL:ijddmr.org Email : editor@ijddmr.com

ARTIFICIAL INTELLIGENCE - BASED DRUG DISCOVERY, VIRTUAL SURGERY SOFTWARE AND MACHINE LEARNING IN PHARMACOLOGICAL TARGETS DISCOVERY

MD FAIYAZ[1], YOGESH CHAND YADAV[2], ALISHA SINGH[3], MD GULJAR AHMAD[4], MILTON SINGH[5] NISHA CHAUDHARY[6]

[1,] PG student Department of Pharmacology, Faculty of Pharmacy, Uttar Pradesh University of Medical Sciences, Saifai, Etawah UP India, [2]Head, Department of Pharmacology, Faculty of Pharmacy, Uttar Pradesh University of Medical Sciences, Saifai, Etawah UP India, [3]Assistant professor SRLT Institute of Pharmaceutical Sciences and Technology Ekdil Etwaha 206126 India , [4]Assistant Professor, Petrol Pump, New, Bailey Rd, near Gola Road, Ramjaipal Nagar, Patna, Bihar 801503, India, [5] Pharmacist Company - Apollo Pharmacies Limited , New Delhi, India, [6] Lecturer SRLT Institute of Pharmacy and Management Ekdil Etawah(U.P.) India

Corresponding author: Dr Yogesh Chand Yadav, drycy31@gmail.com, +916308394804,

DOI: https://doi.org/10.63665/ijddmr.v15.i01.01

Received 15. December .2025; accepted 24 Jan 2026; Published 17 March 2026

ABSTRACT

The pharmaceutical industry is currently experiencing a transition crisis that has been called the “analog” to digital biology transition. The traditional drug development process is plagued by a high failure rate and a clinical success rate of about 10 - 12 per cent. Preclinical development expenses are usually more than $ 2.5 billion / new molecular entity. The technologies of Artificial Intelligence (AI) and Machine Learning (ML) have become groundbreaking, which can reverse this trend and bring a chance to reduce the development time-frames and expenditures to a significant margin. This is a review article on the critical and high-resolution integration of AI in the upstream of the drug discovery pipeline. The development of classical Quantitative Structure-Activity Relationship (QSAR) models via the state of the art deep Graph Neural Networks (GNNs) up to state of the art Transformer based models of pharmacological targets identification is theorized on and selectively analyzed. The revolutionary impact of Alpha-fold to the growth of the so-called druggable genome and applications of generative AI (such as Diffusion Models and Reinforcement Learning) to de novo molecular design are given special consideration. We also discuss the ease of the acceleration of virtual screening (VS) thanks to active learning schemes, which are performed with billions of chemical structures in chemical libraries. But lastly, we also openly address critical bottlenecks for widespread adoption (like bias of the data, model interpretability (Explainable AI) and the so-called hallucination of synthetically invalid structures) and outline the future of entirely autonomous and closed loop discovery laboratories.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Drug Discovery, Virtual Screening, Target Identification, Alpha Fold, Generative Artificial Intelligence, Graph Neural Networks.

1. INTRODUCTION

The search of new medicines is one of the most complex intellectual problems of modern science. It is a venture, and requires the manoeuvrance of two unimaginable spaces, biological space of the mechanisms that give rise to the disease and the chemical space of possible therapeutic agents. The pharmaceutical industry is traditionally based upon the reductionist approach i.e. the hypothesis driven approach which could be summed up as one gene, one drug, one disease. Although this paradigm has seen the rise of the blockbuster drugs in the latter half of the 20th century the industry is now experiencing a law of diminishing returns known as Erooms Law (Moore Law backwards) which indicates that the cost of creating a new drug doubles every 9 years with despite advancements in technology. (1)

The piped style of conducting business are inefficient. The idea of finding a valid biological target, protein, or gene driver of a disease, is also prone to failure when animal models are used because there is a failure to recapitulate in humans the pathology. Having a target which has been validated, physical compound libraries, high-throughput screening (HTS) are now required to find a chemical hit, and it is expensive and time consuming and also still has a limitation due to the size of the compound library (typically a few million compounds). Since after identifying a hit, it needs to be subject to years of iterations of medicinal chemistry optimizing it to an effective, selective and safe lead candidate (ADMET profiling), thus takes years of time. The end result overall is a 12-15 years process, which costs over 2.5 billion dollars and has been a failure in 90 % of the clinical trials. (2,3)

Big Data has become a new variable introduced with the fourth industrial revolution for biology. And now we are generating petabytes of biological data of high resolution sequence data to cryo-electron microscopy structures. But this data is so complex and there is a human being who is unable to comprehend this data. Artificial Intelligence (AI) and presumably more specifically Machine Learning (ML) is what provides computational support to the understanding of this complexity (4). In contrast to classical computational chemistry (e.g. molecular mechanisms) that model the behaviour of molecules by explicit physical formulae, ML-algorithms are based on stochastic functions i.e. they are learnt by data and without any physical laws about how they should be implemented. (5)

The review tries to develop a thorough state-of-the art review of the AI in the early-stage drug discovery. We are singularly focused on two most critical upstream stages:

1.1 Pharmacological Target Identification: How does AI Put Together Multi-Omics Data to Find the "Biological Lock.

1.2 Virtual Screening and Lead Discovery: AI scours the immense space of the so called chemical key in search of these molecules that will bind to these targets.

Intrigued by the progress that AI designs have achieved, we say that we have entered the era of AI reality after the first wave of AI-designed molecules has found its way into clinical studies. (6, 7) Nevertheless, we also provide a sound critique of the limitations that are currently in place and estimate that AI is not magic, but a multifaceted tool which requires good quality data and human expertise control.

2. Computational Engine Learning based Chemical Biology Solutions.

To understand the revolution that AI is having on the discovery of drugs, we must first understand what the algorithms are doing, and probably equally important to our understanding, how chemical and biological objects are being modeled to a computer. A molecule does not matter to a machine, that said, it works in numbers of vectors. The first step of any pipeline in AI is the conversion of the molecular structure to something that the machine could understand.

2.1. Molecular Representations: From String to Graph.

The quality of input representation is highly dependent on the performance of an ML model.

1D Representations (SMILES): The Simplified Molecular Input Line Entry System (SMILES), This is the textual representation of a molecular structure (i.e. “CCO” ethanol). It implies that Natural Language Processing (NLP) has the potential to offer robust architectures to the researchers. Similarly, as ChatGPT can charge you on predicting the next word in a sentence, the chemical language models (CLMs) such as the Transformers such as ChemBERTa can also be trained to predict the next atom in a molecule, or the translation of a sequence of reactants in a molecule into a sequence of products.(8,9)

2D Representations (Molecular Fingerprints): Conventionally, molecules were modelled as binary vectors (fingerprints) of the existence of some substructures (e.g., ECFP4). Fingerprints, though computationally efficient, have the disadvantage of loss of information in the layout of the atoms in space. (10)

Graph Representations (The Rise of GNNs): The most significant trend about the reality that has appeared in the last few years is that molecules are presented as mathematical graphs (atoms are nodes and chemical bonds are edges). These graph structures are dealt directly by the GNNs, such as the Message Passing Neural Networks (MPNNs). At this point of the message passing stage, each atom will replace its feature vector with the information about its neighbours. In this way, the representation of the local chemical environment (e.g., electronegativity, hybridization) of the model can be obtained and is independent of the orientation of a given molecule. GNNs have achieved state-of-the-art in predicting the properties on molecules (11, 12).

3D Representations (Voxel Grids): The 3D representations are of high importance in the case of protein-ligand binding. Arguably, developed initially to identify images, the convolutional neural networks (CNNs) can be converted to 3D image, with the binding pocket being visualized as a grid of 3D pixels (voxels). This grid is scanned by the network to find spatial structures which has a hydrophobic patch or a hydrogen bond donor which is correlated with binding affinity.(13) These representations of molecules are used to represent molecules as in Figure 1 (this is a more detailed explanation of these representations.

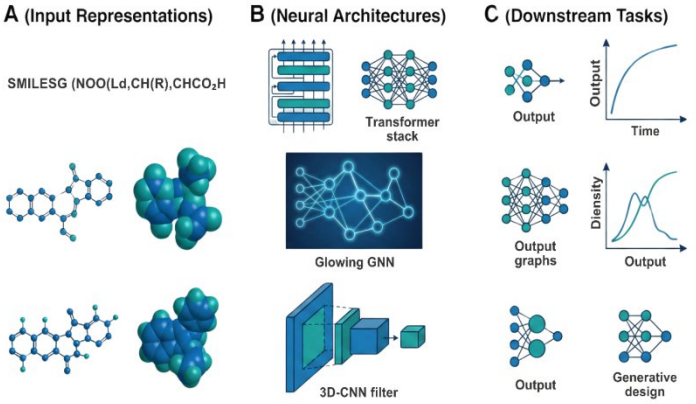

Figure 1. Overview of deep learning workflows in chemical science, illustrating the mapping from various input **molecular representations (A) to specific neural architectures (B) and their subsequent downstream tasks (C).**

Abbreviations: SMILES: Simplified Molecular Input Line Entry system, GNN-Graph Neural Network, 3D-CNN- Three-Dimensional Convolutional Neural Network.

2.2. Key Algorithms

Depending on the scope of data and task at hand, the type of the algorithm adopted depends on level of data.

Random Forests (RF) and Support Vector Machines (SVM): Both the algorithms are of relevance with small volume of data which is the case with ADMET prediction. They are also robust and less likely to overfit than the deep neural networks and provide some level of interpretability. (14, 15)

Deep Neural Networks (DNN's): The deep learning is good in cases where the quantity of data available are high. Multi-task DNNs can be of a huge use in drug-discovery. Rather than training a model which predicts a single endpoint (i.e. solubility of a particular drug, how toxic the drug is, etc), a single model is trained which predicts multiple endpoints at once. This can let the model to be able to study shared representations, i.e. chemical properties behind diverse chemical properties that drive different tasks (16), which, in most cases, leads to better performance than different tasks in models concerning. (16)

Generative Models (VAEs, GANs, Diffusion): Unlike generative models, which are used for the classification of known molecules, generative models can be used to generate new molecules. Variational Autoencoders are used for the mapping of the discrete molecular frameworks to a continuous latent space. The capability of moving about this mathematical space helps researchers decode new points back into new molecules, which are similar to known drugs, however, with a different chemical make-up. (17, 18)

3.Pharmacological Target Identification: Artificial intelligence in use.

The garbage in, garbage out notion results in a conclusion that even the best inhibitor will not work if it makes a contribution of a protein that is not fat burning the cause of the disease. The most dangerous steps involved in the drug discovery process are target finding and validation. AI is changing this step and giving rise to so-called Network Medicine - i.e. the

understanding of disease as a distortion of the complex biological networks and not only one genetic mutation.

3.1. Mining Multi-Omics Data

Contemporary biology is creating torrents of heterogeneous information genomics (DNA), transcript, proteomics (proteins) and metabolomics (metabolites) information. AI is capable of perfect implementation of these layers.

Network-Based Approaches: Failure of AI algorithms to build large Protein-Protein Interaction (PPI) networks. Using algorithms such as the Random Walk with Restart (RWR) through overallizing disease-related genes on this network, it would be possible to identify the topologically central proteins in the disease path even if they were not surfactant in the first place when dealing with genetic screens in the first place. (19, 20)

Synthetic Lethality Prediction: A prediction in oncology, which states that when two genes are both perturbed and a cell dies, but when a single gene is perturbed, a cell survives is called synthetic lethality. It is a wonderful method to arm the cancer cells at which bear a mutation. It is possible to predict new synthetic lethals using deep learning models that have been trained using a big dataset of CRISPR (i.e., the DepMap project) screens and create new drug targets in some cancer subtypes. (21, 22)

Natural Language Processing (NLP) in Literature Mining: The publication speed of science is much higher than the human reading speed. One can mine millions of abstracts in PubMed one can use NLP models (e.g., BioBert) to extract gene-disease relationships to construct knowledge graphs. e.g., links genes with drugs with phenotypes. Such automated hypothesis generation gives the researcher a way to connect dots that are otherwise scattered in the literature. (23)

3.2. The expansion of the Druggable genome and AlphaFold.

The most widely promoted development in AI, applied to biology, is protein structure prediction. The extent to which a target is druggable - its ability to interact with a small molecule - usually has been determined using X-ray crystallography or NMR spectroscopy, which is both slow and expensive. This has led to the pharmaceutical industry always being concentrated on a small cross-section of the proteome (primarily on the kinases, GPCRs, and ion channels) where structures could be obtained.

As a matter of fact, not only the target identification methods involved in the given task can be compared, but also the collective characteristics of the outcomes the related techniques provide. Human| In fact, the comparative aspect of the traditional and AI-driven target identification methods can be applied not only to the target identification methods involved in the process of the given task, but also to the overall nature of the results that the corresponding techniques can offer.

Table 01 Description: Comparative table depicting the change in the process between the traditional methods to the application of the AI methods.

Table 1: Evolution of Target Identification Methods.

Feature	Traditional Target ID	AI-Powered Target ID
Primary Data Source	Literature, Genetic Screens, Animal Models	Multi-omics integration (Genomics, Proteomics), Clinical Data
Methodology	Reductionist (Single gene analysis)	Systems Biology (Network-based, Hub detection)
Structural Insight	Limited to available PDB structures (X-ray/NMR)	**AlphaFold** predictions for the entire proteome
Key Limitation	High failure rate, limited druggable genome	Data quality dependence, biological interpretability
Speed	Years	Months

With the release of AlphaFold 2 (at time called AlphaFold 3) by Google DeepMind, this has changed the situation fundamentally. (24) AlphaFold consists of an attention-based neural network, which is able to create the 3D structure of proteins provided its amino acid sequence with an accuracy that is inside the range of near-experimental accuracy. The AlphaFold 3 takes account of this ability to predict the arrangement of protein-ligand and protein-DNA and protein-RNA based complexes.(25)

Impact: It is due to this that the researchers are able to screen the entire human proteome with druggable pockets. These desired structures can be scanned, e.g., using tools such as DeepSite and P2Rank, for cryptic binding sites, i.e., transient binding pockets that open when needed on proteins that have been considered undruggable, such as Intrinsically Disordered proteins (IDPs) or transcription factors.(26,27) This implicitly increases the universe of potential targets to tens of thousands, rather than the several thousand.

3.Virtual Screening 2.0: Docking To Deep Learning.

When a target has been identified to be valid the hunt towards a hit is begun. Physical screening of compounds using robotics is called High-Throughput Screening (HTS). It is the gold standard and is capital intensive and restricted to the size of accumulation of the compound of the company (usually 1-5 million molecules). Virtual Screening (VS) of libraries with the assistance of computers to actual assess libraries of order of magnitude larger.

4.1. The conversion to the use of Data Driven scoring as compared to the Physics based scoring.

Traditional VS Uses Molecular Docking: - binding energy of the poses of the ligand in a protein pocket are determined from the calculation of the position of the ligand and binders using a physics-based force field (e.g. CHArom and AMBeR respectively). These techniques are physical in nature and are not compatible with the complexity of the effects of entropy and solvent (water), resulting in high rates of false positives. (28)

The difference between AI and the way is fundamentally different. The AI models do not compute the energy but make predictions of binding, like the way it is done in a pattern recognition problem.

Structure-Based AI Scoring: RF-Score or DeepAtom, which is trained on extreme scale data of experimental protein-ligand structures (i.e. PDBbind). They are trained to correlate some characteristics (e. g. 3 Angstroms distal to an aspartic acid residue) with high affinity to bind. Such data-driven

functionality in many cases is much better than classical force fields in ranking active compounds in that they explicitly consider solvent effects and entropy as is observed in training data. (29, 30)

Ligand-Based Screening: In case of the protein structure was not available, the ai can compute the candidate molecules vs. the known active ligands. Individually, Siamese neural networks can be trained upon a much more sophisticated similarity measure than simple Tanimoto similarity and will be able to search to structurally distinct molecules (scaffold hopping), and which will behave similarly. (31, 32)

4.2. Filtering of Ultra Large Library

The space of chemicals that can be made is getting too big. Over 30 billion compounds are available today in the so-called make-on-demand libraries, including the Enamine REAL Space. This is computationally infeasible to most organizations to do 30 billion molecule docks.

In Active Learning, it has been possible to screen such ultra-large libraries by AI. (33)

I. **Selection:** Molecules (library): Reduced fraction of library (e.g., 1 million molecules) is selected, which is reduced in biodiversity (docked).

II. **Training:** These results are used in the training of an AI model which makes a difference in high and low scoring compounds.

III. **Inference:** The remaining billions of molecules then is predicted to have their score rapidly (in the range of a few seconds per mole as compared to a few seconds/minutes to dock).

IV. **Iteration:** Most prediced precautions are constraint on the estanrhoes confirm and sameoreticam re- trained.

A study in one of the landmark studies for this philosophy was performed by Lyu et al. (Nature, 2019), which was a screening of 138 million compounds against the D4 dopamine receptor on the basis of the identified potential compounds yielding hits evaluation include potent compounds with novel scaffolds that would have been missed in a traditional HTS. (34) Academic clusters such as Deep Docking have since been democratized and nowadays billions of compounds can be screened by a common academic cluster.(35)

This can be believed to be an artificial intelligence augmented screening funnel, and better understood by looking at Figure 2.

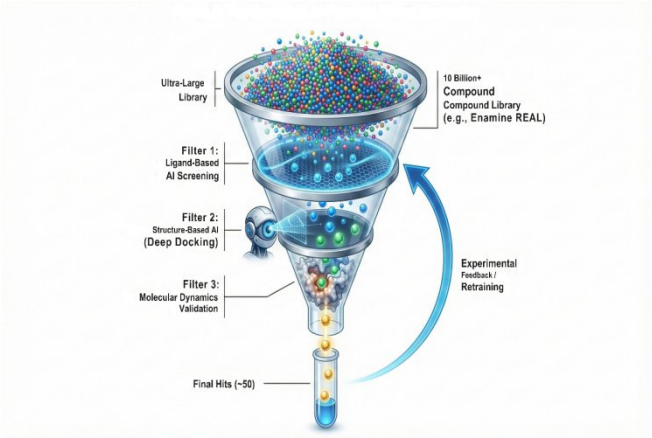

Figure 2: The Funnel of AI-Powered Virtual Screening.

However, with the current state of affairs, things are little bit different (To see its comparison with HTS, Classical Docking and AI-Powered Screening, see Table 2).

Table 2: Comparison of Virtual Screening Paradigms.

Paradigm	Throughput (Compounds)	Cost per Compound	Accuracy / Main Issue	Primary Technology
High-Throughput Screening (HTS)	10^5 - 10^6	High	High Accuracy / False Positives (PAINS)	Robotics, Wet-lab assays
Classical Virtual Screening	10^6 - 10^7	Medium	Medium / Force field inaccuracies, Slow	Molecular Docking (e.g., Glide, AutoDock)
AI-Powered Screening	10^9 - 10^{10}	Low	Variable / Domain Applicability, Very Fast	Deep Learning, Active Learning, GNNs

Description: Contrasts HTS, Classical Docking, and AI Screening

3. Generative Artificial Intelligence and De Novo Drug Designing.

Whereas Virtual Screening as used to find the best needle in a haystack, Generative AI is used to make a needle. This is where the process of the discovery of drugs comes to a halt and production of drugs begins.

5.1. The Generative Toolkit

Variational Autoencoders (VAEs): As it was stated, the Variational Autoencoders (VAEs) learn a continuous numerical representation of the chemical space. This hovers the latent space navigable in a direction pointing towards a risqué, or solubility and or increase in potency, and then decode a new point that points to a new molecule. (36,37)

Generative Adversarial Networks (GANs): There are two types of networks which plays its part in GANs: (Competing networks) The Generator is the one that creates the fake molecules, and the Discriminator is the one that attempt to detect them, and differentiate the faked drugs and the real drugs. Gradually the Generator learns how to manufacture molecules which are indistinguishable to real pharmaceutical agents. This is supplemented by the ORGAN (Objective-Reinforced GAN) where an awarding mechanism is introduced whereby the Generator is forced to generate molecules which will not only appear natural but also fulfil desired qualities (when they bind to a kinase, etc.). (38, 39)

Reinforcement Learning (RL): In RL, one has an "agent" which is constructing a molecule, one atom at a time (adding atoms or bonds). Whenever it makes any step it to be rewarded had it expected that the molecule is active requires synthesize-able, and nontoxic. The agent gets to know about certain policy and maximises this reward. A famous example of such a strategy was performed by Insilico Medicine where they managed it in only 21 days to discover a potentali DDR1 kinase inhibitor which was validated in mice.(40)

5.2. The Next Generation Diffusion Models.

The latest generation of generative AI as made famous to the world through the image generators such as DALL-E, is based upon Diffusion Models. These models are trained using the process of undoing which has brought noises to the data. In the discovery of drugs they are being utilised in the Structure Based Design. Models such as DiffDock or TargetDiff are given

the 3D structure of a protein pocket and are supposed to produce a ligand that fits that pocket, placing atoms in 3D space to ensure that as many interactions as possible occur. It is similar to the intuition of a medicinal chemist, except it is done in three-dimensional coordinate space instead of the one-dimensional strings. (41)

To accomplish this, the different methodologies used by Generative AI are explained in Figure 3. (A better explanation of each is provided there).

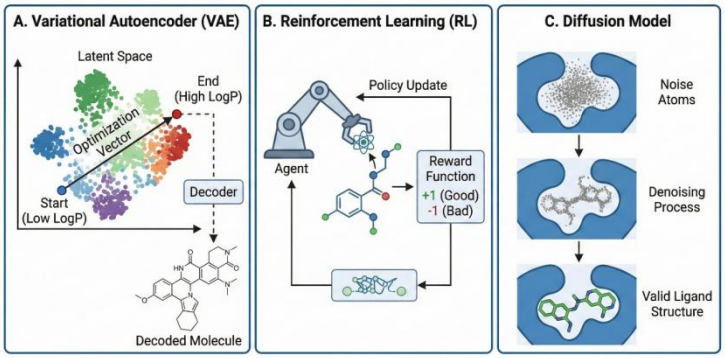

Figure 3: Generative AI Methodologies – From Latent Space to Structure.

4. Challenges and Limitations

Notwithstanding the presence of such a good possibilities, AI application to pharmaceutical R&D are not devoid of its share of complications. The technical and systemic barriers should be mentioned and the hype should be looked through.

6.1. The Data Quality Bottleneck

Data sucking mega size artificial intelligence. However, with biological situations, data tends to be small, sparse, and noisy.

Bias: Public databases such as ChEMBL or PubChem are very biased. They contain millions of data points for popular targets (e.g., kinases, GPCRs) and few data points for novel targets (e.g., RNA-binding proteins). A kinase inhibitor-trained artificial intelligence program will be finding it hard to design an ion channel blocker - this is referred to as domain shift. (42)

The lack of Negative Data: Positive publication bias in the sciences: Scientists publish positive data. He is missing Negative data. Nonetheless, samples of inactive compounds (negative data) are needed to ensure that a classifier learns efficiently. This absence of such data leads to over-optimistic models and a false positive rate. (43, 44)

Experimental Noise: Not only is it biological noisy, but so is noisy biological assays. The IC50 of one lab against another one may differ by an order of magnitude. One of the issues that need to be trained is high precision model on noisy labels. (45)

6.2. The Interpretability of the Black Box

Deep learning models are very obfuscated. An example is that a Graph Neural Network could theorize that a given molecule is toxic with a 99% confidence but not at all tell why. Such impossibility of interpretation is a potent impediment to the medicinal chemists trying to adapt the field as mechanistic knowledge is the precondition to the optimization of the molecules (SAR).

Solution: This is something Solution Explainable AI (XAI) is doing. Specific methods can be used to find out specific atoms, or substructure that induced the model to predict something (e.g. a nitro group that is making it toxic). These methods are known as feature attribution or attention mapping. This has the merits to make the trust and make the rational adjustments of the molecule. (46, 47)

6.3. Synthesizability (Makeability Issue)

A generative model would take into consideration a molecule which would be a perfect fit (in silico) but that cannot be formed in the laboratory (e.g. determinants strained rings, unstable bonds). This hysterical illusion of invalid chemistry is to waste of resources.

Solution: It is now common practice that tools to Retrosynthesis Prediction (e.g. AiZynthFinder, ASKCOS) are nowadays directly part of the generative loop of the pipeline. They are artificial intelligence-powered tools used to predict the actions needed to generate a molecule. In case there is no potential viable synthetic path to a molecule generated, then that molecule is destroyed or punished before a chemist can even lay his or her eyes on it. (48, 49)

These challenges can be overcome by using the key tools and databases that are more appropriate, as shown in **Table 3.**

Table 3: Key AI Tools and Databases for Drug Discovery.

Category	Tool / Database	Developer / Institution	Function
Database	ChEMBL	EMBL-EBI	Massive open bioactivity database for training models.
Database	Binding DB	UC San Diego	Collection of measured binding affinities (K_d, K_i).
Structure Prediction	AlphaFold 3	Google DeepMind	Predicts protein and complex structures from sequence.
Generative AI	REINVENT	AstraZeneca	Framework for *de novo* design using Reinforcement Learning.
Synthesis	AiZynthFinder	AstraZeneca / Univ. Manchester	Retrosynthesis prediction to verify synthetic routes.
Screening	DeepDocking	Univ. British Columbia	Protocol for screening billion-scale libraries.

Description: A resource guide for researchers.

5. Autonomous Discovery Loop: The future of the Autonomous Discovery Loop.

The ultimate goal of this discipline would be to have the human removed on the design- make- test loop. We are going to Self-Driving Laboratories. (50)

In these closed-loop systems:

AI Design: An ML model is used in generating an hypothetical (a set of compounds).

Robotic Synthesis Automated synthesis systems (based on flow chemistry or modular synthesis) synthesize the molecules physically.

Automated Testing: Automated testing of the molecules on its behave and toxicity is done.

Active Learning: The information is sent back to the AI that giveth an update to the internal model and produces the next round of better compounds.

This circular process works 24/7, every day, without any human bias, fatigue, or reliance on intuition. Even the first versions of these systems have already shown themselves to be able to optimise chemical reactions and to find new materials.(51)

More so, the next horizon is the merger of quantum computing. Quantum Machine Learning (QML) can solve the Schrödinger equation for drug-target interactions with the exact precision required by current state-of-the-art methods to replace the approximations of existing force fields and enhance the accuracy of AI predictions. (52, 53)

6. Conclusion

Discovery of drugs using AI has stopped being an exciting yet somewhat hypothetical notion and become a background to contemporary pharmacology. It is not just an efficiency tool, but an exploration tool to help to navigate the vastness of the 'druggable genome', 'chemical space' with nothing but a compass as compared to a map.

Since, indeed, the multi-omics data mining of various targets of previously unknown diseases, the screening of billion-compound libraries in days and the design of novel classes of therapeutics the ML algorithms are systematically eliminating the bottlenecks of preclinical research. Concept Proof: Adequate success with the intervention, such as the discovery of the antibiotic Halicin. (54) and the clinical development of anti-fibrotic drugs developed with the assistance of AI are some of the examples of success with concept proof.

Nevertheless, the future requires both symbiotic relationship calculation and experimentation. The biological validation of a target or a clinical trial can never be replaced with AI. Rather, it is the tool to help scientists to start making better, fewer decisions. The better the quality of the data, the easier it is to explain by algorithms and automations, which, as this area of research starts to scale, means the revolution in the field of digital biotech could well be in safer and more effective medicines getting into patients' hands cheaper and faster than to date.

9. References

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ. 2016;47:20–33.
2. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 2020;323(9):844–53.
3. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov. 2015;14(7):475–86.
4. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
5. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97–113.
6. Insilico Medicine. Insilico Medicine announces positive Phase IIa results for ISM001-055. Nat Biotechnol. 2024;42:1–2.
7. Mullard A. AI-designed drugs in the clinic. Nat Rev Drug Discov. 2024;23:1–3.
8. Chithrananda S, Grand G, Ramsundar B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885. 2020.
9. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. ACS Cent Sci. 2019;5(9):1572–83.
10. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
11. Wieder O, Kohlbacher S, Kuenemann M, Garon A, Seidel P, Langer T, et al. A compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol. 2020;37:1–12.
12. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. arXiv preprint arXiv:1704.01212. 2017.
13. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv preprint arXiv:1510.02855. 2015.
14. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947–58.
15. Cortes-Ciriano I, Bender A. Deep Confidence: A Deep Learning-Based Confidence Estimation Method for Molecular Property Prediction. J Chem Inf Model. 2019;59(3):1269–81.
16. Ramsundar B, Kearnes B, Riley P, Webster D, Konerding D, Pande V. Massively Multitask Networks for Drug Discovery. arXiv preprint arXiv:1502.02072. 2015.
17. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Continuous Representation of Molecules. ACS Cent Sci. 2018;4(2):268–76.
18. Jin W, Barzilay R, Jaakkola T. Junction Tree Variational Autoencoder for Molecular Graph Generation. arXiv preprint arXiv:1802.04364. 2018.
19. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
20. Ferrero E, Dunham I, Sanseau P. In silico prediction of novel therapeutic targets using gene-disease association data. J Transl Med. 2017;15(1):182.
21. Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, Cable D, Cohen W, et al. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell. 2014;158(5):1199–209.
22. Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature. 2019;568(7753):511–6.
23. Kim S, Thiessen PA, Bolton EE, Bryant SH. PUG-SOAP and PUG-REST: web services for programmatic access to PubChem information. Nucleic Acids Res. 2015;43(W1):W605–11.
24. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
25. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
26. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33(19):3036–42.

27. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018;10(1):39.
28. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;432(7019):862–5.
29. Li H, Sze KH, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug discovery: Principles, progress and implications. Drug Discov Today. 2020;25(1):27–35.
30. Ain QU, Aleksandrova A, Roessler HI, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci. 2015;5(6):405–24.
31. Hadsell S, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE; 2006. p. 1735–42.
32. Preuer K, Lewis R, Hochreiter S, Bender A, Lewis R, Hochreiter S, et al. Deep learning in drug discovery. Comput Struct Biotechnol J. 2018;16:285–97.
33. Reker D, Schneider G. Active-learning strategies in computer-aided drug discovery. Drug Discov Today. 2015;20(4):458–65.
34. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019;566(7743):224–9.
35. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep Docking: A Deep Learning Platform for Virtual Screening of Big Data. ACS Cent Sci. 2020;6(6):939–49.
36. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. Front Pharmacol. 2020;11:565644.
37. Winter R, Montanari F, Noé F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci. 2019;10(6):1692–701.
38. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. arXiv preprint arXiv:1705.10843. 2017.
39. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360–5.
40. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038–40.
41. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv preprint arXiv:2210.01776. 2022.
42. Sieg J, Flachsenberg F, Rarey M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. J Chem Inf Model. 2019;59(2):647–61.
43. Chen L, Cruz A, Ramsey S, Dikicioglu CJ. The hidden uncertainty in machine learning for drug discovery. Drug Discov Today. 2021;26(5):1128–34.
44. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. J Chem Inf Model. 2020;60(8):3770–80.
45. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The experimental uncertainty of heterogeneous public Ki data. J Med Chem. 2012;55(11):5165–73.
46. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573–84.
47. Preuer K, Klambauer G, Rippmann F, Hochreiter S, Unterthiner T. Interpretable Deep Learning in Drug Discovery. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer; 2019. p. 331–45.
48. Thakkar A, Reymond JL, Bjerrum EJ, Engkvist O. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. Chem Sci. 2021;12(9):3339–49.
49. Genheden S, Thakkar A, Chadimova V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform. 2020;12(1):35.
50. Coley CW, Eyke NS, Jensen KF. Autonomous discovery in the chemical sciences part I: Progress. Angew Chem Int Ed. 2020;59(51):22858–93.
51. Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LP, Hein JE, et al. ChemOS: An Orchestration Software for Autonomous Discovery in Chemistry. PLoS One. 2018;13(4):e0192750.
52. Cao Y, Romero J, Olson JP, Degroote M, Johnson PD, Kieferová M, et al. Quantum Chemistry in the Age of Quantum Computing. Chem Rev. 2019;119(19):10856–915.
53. Batra R, Zhai H, Loeffler TD, Chen L, Sankaran B, Chan H, et al. Quantum machine learning for molecular properties. arXiv preprint arXiv:2104.03222. 2021.
54. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020;180(4):688–702.