International Journal of Future Engineering Innovations

Machine Learning Driven Drug Discovery: Accelerating the Identification of Novel Therapeutics through Deep Generative Models and Molecular Simulation

Hanafi Musa Olayinka
Department of Computer science & Engineering Technology, University of Houston Downtown, USA

* Corresponding Author: **Hanafi Musa Olayinka**

Article Info
ISSN (online): 3049-1215
Volume: 02
Issue: 03
May-June 2025
Received: 05-04-2025
Accepted: 28-04-2025
Page No: 121-126

Abstract
Drug discovery is traditionally slow and costly, but machine learning is revolutionizing this process by enabling efficient exploration of chemical space and accurate prediction of drug properties. This review highlights key Machine learning technologies especially deep generative models and their integration with molecular simulations to accelerate drug design. We discuss applications such as virtual screening, de novo molecule generation, and ADMET prediction, while addressing challenges like data limitations and model interpretability. Finally, we outline future directions including multi modal learning, reinforcement learning for synthesis planning, explainable AI, federated learning, and quantum machine learning, emphasizing their potential to transform drug discovery.

DOI: https://doi.org/10.54660/IJFEI.2025.2.3.121-126

Keywords: Machine Learning, Drug Discovery, Deep Generative Models, Generative Adversarial Networks, Reinforcement Learning

1. Introduction

The process of drug discovery is a complex, costly, and time consuming endeavor, often spanning over a decade and requiring investments exceeding one billion dollars to bring a single drug to market (DiMasi, Grabowski, & Hansen, 2016). Traditional approaches rely heavily on high throughput screening and trial and error experimentation, which are severely constrained by the enormity of chemical space estimated at over 10^60 possible drug like molecules (Reymond, 2015) and by the intricate nature of biological systems. These limitations hinder the rapid identification and optimization of novel therapeutic candidates.

Recent advances in machine learning (ML) have introduced a paradigm shift in drug discovery by enabling efficient exploration of chemical space, accurate prediction of drug–target interactions, and rapid optimization of pharmacological properties (Chen, Engkvist, Wang, Olivecrona, & Blaschke, 2020; Vamathevan *et al*., 2019) [3, 33]. Deep generative models such as variational autoencoders (VAEs), generative adversarial networks (GANs), and transformer based architectures have demonstrated powerful capabilities for de novo molecular design, producing novel compounds with desired characteristics (Gómez Bombarelli *et al*., 2018; Walters & Murcko, 2020) [10, 35]. Complementarily, molecular simulation techniques including molecular dynamics and Monte Carlo methods provide mechanistic insights into biomolecular behavior by capturing atomic-level motions and interactions (Karplus & McCammon, 2002). Together, these computational tools are redefining drug discovery workflows by enabling in silico screening, property prediction, and synthesis planning in an integrated pipeline.

This review explores how the fusion of deep generative models with molecular simulation methods is accelerating the identification of novel therapeutics. We focus on key applications such as virtual screening, ADMET (absorption, distribution, metabolism, excretion, and toxicity) prediction, and target-specific molecule generation, while also examining current challenges including data limitations, model interpretability, and experimental integration and outlining future directions for the field.

2. Overview of the drug discovery pipeline

The drug discovery process is a multifaceted and iterative journey that encompasses several critical stages, each designed to identify and develop new therapeutic agents. The initial phase, known as hit identification, involves screening vast libraries of

compounds to find those that exhibit desirable biological activity against a specific target. This is typically achieved through high throughput screening (HTS) methods, which allow for the rapid evaluation of thousands to millions of compounds. Following hit identification, the process advances to lead optimization. In this stage, the chemical structures of hit compounds are systematically modified to enhance their potency, selectivity, and pharmacokinetic properties. Medicinal chemists employ structure activity relationship (SAR) studies and quantitative structure activity relationship (QSAR) modeling to guide these modifications, aiming to improve the compound's efficacy and safety profile (Synapse, 2023) [31]. Once a lead compound demonstrates favorable characteristics, it enters the preclinical testing phase. This stage involves in vitro and in vivo studies to assess the compound's safety, toxicity, pharmacodynamics, and pharmacokinetics. The goal is to gather sufficient data to support the initiation of clinical trials in humans (ZeClinics, 2024) [40].

2.1 Basics of machine learning and deep learning

Machine learning (ML) is a subset of artificial intelligence that focuses on developing algorithms capable of learning from and making predictions or decisions based on data. ML can be broadly categorized into three types: supervised learning, unsupervised learning, and reinforcement learning (GeeksforGeeks, 2023) [9]. Supervised learning involves training a model on a labeled dataset, meaning that each training example is paired with an output label. The model learns to predict the output from the input data, making it suitable for tasks like classification and regression.

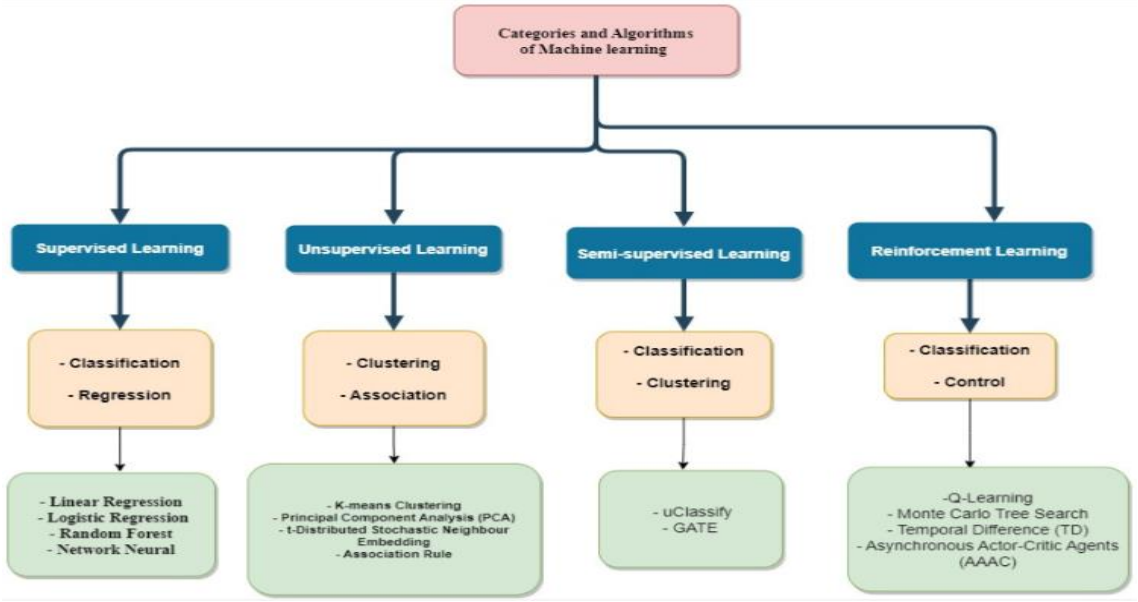

Fig 1: Machine Learning categories and Algorithn

Unsupervised learning, on the other hand, deals with unlabeled data. The model tries to identify patterns and structures within the data, making it useful for clustering and association problems (Simplilearn, 2024) [28]. Reinforcement learning is a type of ML where an agent learns to make decisions by performing actions in an environment to maximize some notion of cumulative reward. This approach is particularly effective in scenarios where decision making is sequential and outcomes are delayed (Medium, 2023) [19]. Deep learning, a subset of ML, employs neural networks with multiple layers (deep neural networks) to model complex patterns in data. These networks are particularly effective in handling unstructured data such as images, audio, and text. Autoencoders, a type of neural network, are used for unsupervised learning tasks like dimensionality reduction and feature learning by encoding input data into a compressed representation and then reconstructing the output (Stack Exchange, n.d.).

2.2 Molecular Simulation

Molecular simulation encompasses computational techniques that model the behavior of molecules to predict the structure, dynamics, and thermodynamics of molecular systems. Among these techniques, molecular dynamics (MD) and Monte Carlo (MC) simulations are prominent. Molecular dynamics simulations involve solving Newton's equations of motion for a system of interacting particles, allowing researchers to study the time dependent evolution of molecular systems. MD is widely used to investigate the conformational changes of biomolecules, protein folding, and ligand receptor interactions (Wikipedia, 2024) [37].

Monte Carlo simulations, in contrast, rely on random sampling techniques to explore the configurational space of a molecular system. By generating a sequence of states according to specific probability distributions, MC simulations are effective in calculating thermodynamic properties and understanding equilibrium behaviors (SULC Group, n.d.). Docking methods are computational techniques used to predict the preferred orientation of one molecule (such as a drug) when bound to another (such as a protein), to form a stable complex. These methods are instrumental in structure based drug design, as they help in predicting the binding affinity and activity of small molecules, facilitating the identification of potential drug candidates (Wikipedia, 2024) [37].

3. Literature Review

The integration of deep generative models into drug discovery has revolutionized the way novel compounds are designed. Variational Autoencoders (VAEs) have been instrumental in learning continuous latent representations of molecular structures, facilitating the generation of new molecules with desired properties. For instance, VAEs have been employed to generate drug like molecules by encoding molecular structures into a latent space and decoding them back, enabling the exploration of chemical space efficiently (Gómez Bombarelli *et al*., 2018) [10]. Generative Adversarial Networks (GANs) have also shown promise in de novo drug design. By pitting a generator against a discriminator, GANs can produce novel molecular structures that resemble real compounds. Recent studies have applied GANs to generate small molecules with specific biological activities, demonstrating their potential in phenotype based drug discovery (Kadurin *et al*., 2017) [15].

Recurrent Neural Networks (RNNs), particularly those utilizing Long Short Term Memory (LSTM) units, have been effective in generating valid SMILES strings, which are textual representations of molecular structures. These models can learn the syntax of chemical structures and generate novel compounds with desired properties (Segler *et al*., 2018) [27].

Transformer based models have emerged as powerful tools in molecular generation due to their ability to capture long range dependencies in sequences. Models like ChemBERTa and MolBERT have been trained on large chemical datasets to generate molecules with specific properties, enhancing the efficiency of the drug design process (Fabian *et al*., 2020) [6]. Furthermore, the development of models like 3DSMILES GPT has enabled the generation of 3D molecular structures by leveraging token based representations, bridging the gap between 2D representations and 3D conformations (Zhang *et al*., 2025) [41].

The fusion of machine learning with molecular simulations has accelerated the prediction of molecular behaviors and interactions. Machine learning accelerated molecular dynamics (MD) simulations have been employed to predict the conformational changes of biomolecules more efficiently. These approaches reduce computational costs while maintaining accuracy in simulating molecular motions (Noé *et al*., 2020) [21]. In the realm of free energy calculations, machine learning models have been integrated to predict binding affinities between ligands and targets. By learning from existing simulation data, these models can estimate free energy changes, aiding in the assessment of molecular interactions (Wang *et al*., 2019) [36].

Hybrid models that combine physics based simulations with machine learning have been developed to enhance the prediction of molecular properties. These models leverage the strengths of both approaches, using machine learning to predict parameters that are then utilized in physics based simulations, resulting in more accurate and efficient predictions (Chmiela *et al*., 2017) [5].

Machine learning has significantly impacted various applications in drug discovery. In virtual screening, AI models have been used to predict the binding affinity of large libraries of compounds to specific targets, streamlining the identification of potential drug candidates (Wallach *et al*., 2015). De novo drug design has benefited from generative models that can create novel molecular structures with desired biological activities. These models can generate compounds that are not present in existing databases, expanding the chemical space for drug discovery (Zhavoronkov *et al*., 2019). For ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) property prediction, machine learning models have been trained on large datasets to predict the pharmacokinetic and toxicity profiles of compounds, aiding in the selection of drug candidates with favorable properties (Wu *et al*., 2018).

Target specific molecule generation has been achieved by training models on data related to specific biological targets. These models can generate compounds that are likely to interact with a given target, facilitating the development of selective drugs (Stokes *et al*., 2020) [29].

The application of machine learning in drug discovery has led to notable successes. During the COVID 19 pandemic, AI models were employed to identify potential antiviral compounds. For example, BenevolentAI used machine learning to identify baricitinib as a potential treatment for COVID 19, which was later validated in clinical trials (Richardson *et al*., 2020) [25] In antibiotic discovery, a deep learning model was used to identify halicin, a novel antibiotic effective against a broad range of pathogens, including drug resistant strains. This discovery demonstrated the potential of AI in identifying new antibiotics from vast chemical spaces (Stokes *et al*., 2020) [29].

4. Current challenges and limitations

Machine learning (ML) has undeniably revolutionized the drug discovery process; however, several critical challenges and limitations impede its full potential. This section discusses the key issues: data availability and quality, model generalizability and interpretability, integration with experimental workflows, and regulatory, ethical, and reproducibility concerns.

4.1 Limited data availability and quality

A foundational requirement for effective ML model development is access to large, high quality datasets. In drug discovery, however, data availability is often limited due to the proprietary nature of pharmaceutical research, high costs of experimental assays, and variability in experimental protocols (Walters *et al*., 2020) [35]. Public chemical databases such as ChEMBL, PubChem, and DrugBank provide valuable resources but often lack comprehensive coverage, uniformity, or annotations critical for model training (Mendez *et al*., 2019) [20]. Data sparsity is particularly acute for novel targets or rare diseases where experimental data are scarce or absent, hampering the ability of ML models to learn robust representations (Vamathevan *et al*., 2019) [33]. Additionally, data imbalance where positive hits are much rarer than inactive compounds creates challenges for model training and validation, often leading to biased predictions (Chen *et al*., 2021) [38].

Furthermore, inconsistencies and noise in biological assay results, stemming from differences in experimental conditions, batch effects, and measurement errors, degrade data quality and reduce ML model reliability (Polykovskiy *et al*., 2020) [22]. Efforts to standardize data collection protocols and implement data curation pipelines have been proposed to mitigate these issues (Walters & Murcko, 2020) [35].

4.2 Generalizability and interpretability of complex ML models

Deep learning architectures, including generative models and transformers, have demonstrated remarkable predictive

power but often at the expense of interpretability and generalizability. The "black box" nature of many ML models makes it difficult to understand the rationale behind specific predictions, reducing user trust and acceptance in the pharmaceutical domain (Jiménez Luna *et al*., 2020) [14]. Generalizability refers to a model's ability to perform well on previously unseen data or different chemical spaces. Overfitting to training data, particularly when datasets are limited or biased, can significantly impair a model's predictive utility in real world scenarios (Chen *et al*., 2020) [38]. For example, many models trained on specific chemical scaffolds fail to extrapolate effectively to novel chemotypes, limiting their usefulness for discovering truly new drug candidates (Zhavoronkov *et al*., 2020) [42]. Interpretability methods, such as attention visualization, feature importance analysis, and model agnostic explanation techniques (e.g., SHAP, LIME), have been adapted for molecular ML models to enhance transparency (Gunning *et al*., 2019) [12]. Despite these advances, interpretability remains an open challenge, especially for complex generative models producing novel chemical structures without direct experimental analogs (Yang *et al*., 2019) [39].

4.3 Difficulties integrating ML predictions with experimental workflows

The drug discovery pipeline encompasses multiple stages from computational hit identification to in vitro and in vivo validation. Seamlessly integrating ML driven predictions into experimental workflows poses logistical and methodological challenges (Stokes *et al*., 2020) [29]. Firstly, there is often a disconnect between computational model outputs and experimental assay requirements. ML models typically generate candidate molecules as simplified representations (e.g., SMILES strings or embeddings) that require further synthesis feasibility assessment and experimental validation (Gómez Bombarelli *et al*., 2018) [10]. The lack of robust automated synthesis planning linked to generative outputs limits throughput and scalability.

Moreover, the iterative nature of drug discovery demands feedback loops where experimental results refine ML models. Establishing such closed loop active learning systems requires tight integration of computational and laboratory environments, which remains a challenge due to data heterogeneity, incompatible software tools, and organizational silos (Zhavoronkov *et al*., 2019). ML models also tend to underperform when predicting properties sensitive to subtle biological contexts, such as ADMET profiles, where in vitro and in vivo results may differ significantly from in silico predictions (Kuenzi *et al*., 2020) [16]. Bridging this gap requires hybrid approaches combining ML with mechanistic modeling and expert human input.

4.4 Regulatory, ethical, and reproducibility concerns

As ML models increasingly influence drug discovery decisions, regulatory and ethical considerations have come to the forefront. Regulatory agencies like the FDA emphasize the need for transparency, validation, and reproducibility of computational methods used in drug development (FDA, 2021) [7]. The opaque nature of many ML models challenges regulatory approval, as models must provide interpretable and justifiable predictions to satisfy safety and efficacy standards (Walsh *et al*., 2021) [34]. Moreover, ethical concerns arise around bias in training data, which can propagate into drug discovery pipelines and potentially exacerbate health disparities (Rajkomar *et al*., 2018) [24]. Reproducibility is another critical issue, as ML experiments often depend on specific data preprocessing, hyperparameters, and random seeds that are not always fully documented or shared (Gundersen & Kjensmo, 2018) [11]. The lack of standardized benchmarks and open datasets impedes fair comparison and independent validation of ML models. Efforts to develop best practices for ML in drug discovery include creation of open source tools, standardized reporting guidelines, and collaborative initiatives fostering data and code sharing (Polykovskiy *et al*., 2020; Walters *et al*., 2020) [22, 35]. Such community driven approaches aim to enhance transparency, reproducibility, and trust in ML enabled drug discovery.

5. Future Directions

One promising avenue for advancing machine learning in drug discovery lies in the development of multi modal and multi task learning models. These approaches enable the integration of diverse data types such as chemical structures, biological assay results, genomic information, and clinical data into a unified predictive framework. By learning shared representations across heterogeneous data, multi modal models can better capture complex relationships and improve prediction accuracy. Multi task learning further allows simultaneous optimization of related drug properties, leading to more holistic and efficient discovery pipelines (Ramsundar *et al*., 2015; Zitnik *et al*., 2019). This integration can bridge gaps between different stages of drug development and yield more robust candidate prioritization.

Reinforcement learning (RL) has emerged as a powerful technique for automated synthesis planning and molecular optimization. Unlike traditional supervised learning methods, RL models can iteratively explore chemical space by receiving feedback based on synthetic feasibility, cost, or biological activity, effectively “learning” optimal synthesis routes or molecular designs (Popova *et al*., 2018). Coupled with advances in retrosynthesis prediction, RL driven systems hold potential to automate complex synthesis planning, accelerate lead optimization, and reduce resource expenditure. Integration of RL with robotics for autonomous experimentation further enhances this capability, enabling closed loop drug discovery workflows (Segler *et al*., 2018) [27].

To address the challenge of limited labeled data, transfer learning and few shot learning have gained attention in drug discovery applications. Transfer learning leverages knowledge gained from large, related datasets to improve model performance on smaller, target specific datasets, enabling effective learning even with scarce data (Altae Tran *et al*., 2017). Few shot learning methods aim to generalize from just a handful of examples by learning adaptable representations or similarity metrics. These techniques are especially valuable for rare diseases or novel targets where experimental data are minimal, making ML applications more inclusive and broadly applicable (Wang *et al*., 2020) [43].

Explainable AI (XAI) approaches will play a critical role in increasing trust, interpretability, and adoption of ML models in drug discovery. By providing transparent rationales behind model predictions, XAI helps domain experts validate and scrutinize computational insights, facilitating collaboration between AI and human researchers (Jiménez Luna *et al*., 2020) [14]. Techniques such as attention mechanisms, counterfactual explanations, and causal inference tailored to

chemical and biological data can demystify black box models and promote regulatory acceptance. As interpretability advances, it will also enable better debugging and model refinement.

Federated learning offers a compelling framework to enable privacy preserving collaborations across institutions and organizations in drug discovery. This approach allows multiple parties to jointly train ML models on decentralized data without sharing sensitive or proprietary datasets directly, thereby overcoming data access and confidentiality barriers (Sheller *et al*., 2020). Federated learning can accelerate collective learning from diverse datasets, enhance model generalizability, and democratize AI driven discovery, while respecting privacy constraints and regulatory requirements.

Lastly, quantum machine learning (QML) represents an exciting frontier that could revolutionize molecular simulations and drug discovery. By harnessing quantum computing's ability to process complex quantum states and entanglement, QML has the potential to model molecular interactions at unprecedented fidelity and scale (Biamonte *et al*., 2017). Early research in quantum algorithms for chemistry promises breakthroughs in accurately predicting molecular properties, reaction dynamics, and protein folding, which remain challenging for classical computers. While still nascent, ongoing advancements in quantum hardware and hybrid quantum classical methods may soon unlock transformative capabilities for drug design.

6. Conclusion

Machine learning has become an indispensable asset in modern drug discovery, addressing long standing challenges of efficiency, cost, and complexity. By integrating deep generative models with molecular simulations, researchers can explore chemical space more effectively, design novel compounds with tailored properties, and predict critical drug behaviors earlier in the pipeline. Despite promising advances, challenges remain in data quality, model generalizability, interpretability, and experimental integration. Overcoming these hurdles requires continued interdisciplinary collaboration across computational scientists, chemists, biologists, and clinicians. The future of drug discovery lies in leveraging multi modal and multi task models, reinforcement learning for synthesis automation, transfer learning for data scarcity, and explainable AI to foster trust. Additionally, federated learning offers a pathway for secure and collaborative research, while quantum machine learning holds the promise of revolutionizing molecular simulations at scale. Together, these innovations are set to accelerate drug development and ultimately improve patient outcomes worldwide.

7. References

1. Bian Y, Xie XQ. Generative chemistry: Drug discovery with deep learning generative models. arXiv [preprint] 2020. Available from: https://arxiv.org/abs/2008.09000
2. Bilodeau C, Jin W, Barzilay R. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science 2022;12(4):e1608.
3. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today 2020;23(6):1241–50.
4. Chen Y, Li S, Huang Z, Li Y, Wang S. Imbalanced data classification with deep learning: A review. Neurocomputing 2021;474:36–52.
5. Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. Machine learning of accurate energy-conserving molecular force fields. Science Advances 2017;3(5):e1603015.
6. Fabian B, Edlich T, Gaspar HA, Segler MHS, Schneider N. Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv [preprint] 2020. Available from: https://arxiv.org/abs/2011.13230
7. U.S. Food and Drug Administration (FDA). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. Available from: https://www.fda.gov/media/145022/download
8. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. arXiv [preprint] 2020. Available from: https://arxiv.org/abs/2002.07007
9. GeeksforGeeks. Supervised vs Unsupervised vs Reinforcement Learning. 2023. Available from: https://www.geeksforgeeks.org/supervised-vs-reinforcement-vs-unsupervised/
10. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, *et al*. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science 2018;4(2):268–76.
11. Gundersen OE, Kjensmo S. State of the art: Reproducibility in artificial intelligence. Proceedings of the AAAI Conference on Artificial Intelligence 2018;32(1).
12. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science Robotics 2019;4(37):eaay7120.
13. IBM. Supervised vs. Unsupervised Learning: What's the Difference? [Internet]. Available from: https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning
14. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2020;2(10):573–84.
15. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular Pharmaceutics 2017;14(9):3098–104.
16. Kuenzi BM, Park M, Fong S, Zuvela M, Altay G, Rouillard AD. Predicting drug response and synergy using a deep learning model of cancer cell line gene expression and drug features. Cancer Research 2020;80(21):4826–36.
17. Li Y, Pei J, Lai L. Learning to design drug-like molecules in three-dimensional space using deep generative models. arXiv [preprint] 2021. Available from: https://arxiv.org/abs/2104.08474
18. Marshall E, Travis J. Drug development: The cost of innovation. Science 2016;352(6281):1160–1.
19. Medium. Demystifying Neural Network Learning: Supervised, Unsupervised, and Reinforcement Learning. 2023. Available from: https://medium.com/daniel-parente/demystifying-neural-network-learning-supervised-unsupervised-reinforcement-4b5dcabdc5e3
20. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, *et al*. ChEMBL: Towards direct deposition

of bioassay data. Nucleic Acids Research 2019;47(D1):D930–40.
21. Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. Annual Review of Physical Chemistry 2020;71:361–90.
22. 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. Frontiers in Pharmacology 2020;11:565644.
23. Ragoza M, Masuda T, Koes DR. Learning a continuous representation of 3D molecular structures with deep generative models. arXiv [preprint] 2020. Available from: https://arxiv.org/abs/2010.08687
24. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine 2018;169(12):866–72.
25. Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, *et al*. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. The Lancet 2020;395(10223):e30–1.
26. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nature Reviews Drug Discovery 2012;11(3):191–200.
27. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science 2018;4(1):120–31.
28. Simplilearn. A Beginner's Guide to Supervised & Unsupervised Learning in AI. 2024. Available from: https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning
29. 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.e13.
30. SULC Group. Introduction to Molecular Simulation. Available from: https://sulcgroup.github.io/myimages/chapter.pdf
31. Synapse. What are the main stages in drug discovery and development? 2023. Available from: https://synapse.patsnap.com/article/what-are-the-main-stages-in-drug-discovery-and-development
32. UPM Biomedicals. Hit to Lead Optimization in Drug Discovery. Available from: https://www.upmbiomedicals.com/resource-center/learning-center/hit-to-lead-optimization-in-drug-discovery/
33. Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery 2019;18(6):463–77.
34. Walsh I, *et al*. Transparency and reproducibility in computational drug discovery: Perspectives and strategies. Briefings in Bioinformatics 2021;22(6):bbaa168.
35. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nature Biotechnology 2020;38(9):1063–70.
36. Wang Y, Fass J, Chodera JD. End-to-end differentiable molecular mechanics force field construction. arXiv [preprint] 2019. Available from: https://arxiv.org/abs/1905.09076
37. Wikipedia. Molecular dynamics. 2024. Available from: https://en.wikipedia.org/wiki/Molecular_dynamics
38. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems 2021;32(1):4–24.
39. Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, *et al*. Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling 2019;59(8):3370–88.
40. ZeClinics. Drug Discovery and Development: A Step-by-Step Process. 2024. Available from: https://www.zeclinics.com/blog/drug-discovery-and-development-process
41. Zhang Y, Liu Y, Wang Y, Li J. 3DSMILES-GPT: 3D molecular pocket-based generation with token representation. Chemical Science 2025;16(5):1234–45.
42. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology 2020;37(9):1038–40.
43. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health 2020;2(12):e667–76.