Revolutionizing Natural Product Discovery: The Role Of Generative AI

Sanyogita Shahi[1*], Shirish Kumar Singh[2]

[1*]Department of Chemistry, Kalinga University, Raipur, Chhattisgarh, 492101, India.
[2]Regional Science Centre, Saddu, Raipur, Chhattisgarh, 492014, India.

***Corresponding author**- Sanyogita Shahi
Email: drsanyogitashahi@gmail.com

ABSTRACT
The traditional, resource-intensive process of de novo natural product design, a cornerstone of drug discovery requiring extensive interdisciplinary knowledge, has been revolutionized by the advent of generative artificial intelligence (AI). This review comprehensively explores the integration of generative AI methodologies, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Recurrent Neural Networks (RNNs), in designing novel natural products with desired biological activities. These AI models facilitate efficient molecule generation and optimization, marking a significant shift from conventional approaches like combinatorial chemistry and high-throughput screening. We examine key methodologies encompassing data preparation, model training, and rigorous molecule validation, highlighting successful case studies in the discovery of antibiotics, anticancer agents, and antiviral compounds where AI-designed molecules demonstrate notable advantages. The review critically analyzes existing challenges such as data quality, model interpretability, the crucial integration with experimental validation, and ethical and regulatory considerations. Looking ahead, we discuss potential advancements in AI algorithms, the enhanced incorporation of biological data, collaborative research paradigms, and the potential impact of quantum computing on this field. In conclusion, generative AI offers transformative capabilities for de novo natural product design, significantly accelerating the discovery and optimization of new therapeutic compounds. Despite ongoing challenges, continuous progress in AI and computational chemistry positions generative AI as a pivotal tool in the future development of innovative therapeutics.

Keywords: Virtual screening, De novo design, Chemoinformatic, Drug discovery, AI in chemistry, Bioactive molecules.

1. INTRODUCTION: THE DAWN OF AI IN NATURAL PRODUCT DESIGN

1.1 The Enduring Legacy and Limitations of Traditional Natural Product Discovery

Natural products have long been an invaluable reservoir of therapeutic agents, forming the bedrock of numerous pharmaceuticals across diverse medical fields. Their structural complexity and biological activities have provided inspiration and direct leads for countless drugs. However, the conventional journey of discovering and designing these intricate molecules is fraught with challenges. It demands substantial interdisciplinary expertise spanning chemistry, biology, and pharmacology, coupled with significant investments in time and resources. Traditional methods, such as isolating compounds from natural sources, combinatorial chemistry, and high-throughput screening, while successful, often suffer from limitations in efficiency, scope of chemical space exploration, and the ability to precisely design molecules with desired properties.

1.2 The Generative AI Revolution in Natural Product Science

The recent integration of generative artificial intelligence (AI) has ushered in a transformative era for natural product design. This paradigm shift offers innovative and powerful avenues for the rapid and efficient creation of novel bioactive molecules. Generative AI encompasses a diverse suite of sophisticated computational techniques, prominently including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and recurrent neural networks (RNNs). These advanced models possess the remarkable capability to generate and optimize molecular structures based on predefined or learned property landscapes.

1.3 Overcoming Traditional Bottlenecks: AI-Driven Efficiency and Expanded Chemical Space

This transition from traditional, often laborious methods to AI-driven approaches has yielded significant advantages. AI-powered tools have not only accelerated the pace of molecule discovery but have also dramatically expanded the horizons of chemical space exploration. By learning from vast datasets of existing molecules and their properties, generative AI models can propose and design novel structures that might be difficult or impossible to access through conventional synthetic or screening methodologies.

1.4 Scope and Objectives of this Review

This comprehensive review aims to thoroughly explore the application of generative AI in the realm of de novo natural product design. We will delve into the fundamental principles of key AI methodologies, examine the diverse tools and applications currently employed, and showcase compelling case studies where AI has demonstrably facilitated the discovery of bioactive compounds with significant therapeutic potential. Furthermore, we will critically discuss the inherent challenges and limitations associated with AI-driven approaches, including crucial aspects such as data quality, the interpretability of AI models, and the essential integration with experimental validation techniques. Our ultimate goal is to provide a clear understanding of the current landscape and future trajectory of this rapidly evolving field.

2. LITERATURE REVIEW: THE EVOLUTION OF NATURAL PRODUCT DESIGN

2.1 Traditional Paradigms in Natural Product Discovery and Optimization

Historically, the quest for therapeutic molecules from natural sources has relied on established yet demanding methodologies.

2.1.1. Combinatorial Chemistry: Generating Molecular Diversity

Combinatorial chemistry emerged as a powerful strategy for generating vast libraries of structurally diverse chemical compounds. By systematically varying molecular building blocks, researchers aimed to explore a broad spectrum of potential drug candidates. This approach has been crucial in identifying novel chemical entities. However, its effectiveness is often tempered by the sheer scale of synthesis and the subsequent exhaustive screening required to pinpoint bioactive molecules within these large libraries. The accessible chemical space is also inherently limited by the chosen building blocks and reaction chemistries.

2.1.2. High-Throughput Screening (HTS): Empirical Identification of Bioactivity

High-throughput screening revolutionized the pace of drug discovery by automating the testing of large compound collections against biological targets or assays. This technology enables the rapid assessment of thousands to millions of compounds for their potential pharmacological activity. While HTS significantly accelerates the identification of initial "hit" compounds exhibiting desired biological effects, it often operates on an empirical basis. The process typically lacks a strong rational design component in the initial screening phase, potentially leading to hits with suboptimal properties requiring extensive downstream optimization.

2.1.3. Rational Drug Design: Structure-Based and Ligand-Based Approaches

Rational drug design represents a more targeted approach, integrating computational modelling, insights from structural biology, and the principles of medicinal chemistry. This strategy aims to design molecules with a higher probability of interacting with specific biological targets or pathways implicated in disease. By leveraging knowledge of target structures (structure-based design) or existing ligands (ligand-based design), researchers can optimize drug candidates for improved efficacy, safety, and pharmacokinetic properties. Although this approach has yielded numerous successful therapeutics, it can be constrained by the complexity of biological systems, the availability of detailed structural information, and the inherent limitations in designing entirely novel scaffolds.

2.1.4 The Bottlenecks of Traditional Methods

Despite their individual contributions, these traditional approaches are inherently time-consuming, labour-intensive, and face limitations in comprehensively exploring the vastness of chemical space. The synthesis and screening of large compound libraries can be resource-prohibitive, and rational design often requires significant prior knowledge and may struggle with novel target mechanisms or the design of first-in-class molecules.

2.2 The Emergence of AI-Driven Approaches: A Paradigm Shift in Molecular Innovation

AI-driven approaches in chemistry signify a fundamental change in how we approach molecular design, discovery, and optimization. By harnessing the power of artificial intelligence and advanced computational techniques, these methods offer the potential to overcome many of the limitations inherent in traditional strategies.

2.2.1 Data-Driven Design: Learning from Chemical and Biological Information

A cornerstone of AI-driven approaches is data-driven design. This leverages the wealth of existing chemical structures and their associated biological activities to train sophisticated predictive models. Machine learning algorithms, including deep neural networks and support vector machines, learn intricate patterns and relationships within these large datasets. This enables the prediction of crucial molecular properties such as bioactivity, solubility, and toxicity for novel, unseen compounds. By integrating these predictions early in the design process, researchers can prioritize compounds with a higher likelihood of success, significantly streamlining the drug discovery pipeline and reducing the need for extensive experimental trial-and-error.

2.2.2 AI-Powered Virtual Screening: Accelerating Lead Identification

AI-powered virtual screening represents a significant advancement in the efficiency of lead identification. Algorithms can rapidly evaluate vast libraries of virtual compounds – molecules that exist only in computer models – against specific biological targets or disease models. This computational pre-screening dramatically reduces the number of compounds that need to be synthesized and experimentally tested. Techniques such as molecular docking (predicting binding interactions), quantitative structure-activity relationship (QSAR) modelling (correlating structure with activity), and pharmacophore-based screening (identifying key structural features responsible for activity) are enhanced by AI to provide more accurate and efficient filtering of virtual libraries, guiding researchers towards the most promising lead candidates for further development.

2.2.3 Generative Models: Unleashing "Chemical Creativity"

Furthermore, AI-driven approaches have introduced the exciting capability of generating and optimizing entirely novel molecular structures through generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These models can be trained to understand the underlying rules and patterns of chemical structures and their properties, allowing them to "imagine" and create new compounds with specified characteristics. This opens up the possibility of exploring regions of chemical space that might be inaccessible or overlooked by traditional methods, potentially leading to the discovery of first-in-class drugs or molecules with unprecedented properties. The iterative refinement of these generated structures, guided by feedback from predictive models and experimental data, allows for a continuous optimization of compound potency, selectivity, and safety profiles.

2.2.4 The Promise and Challenges of AI Integration

While the transformative potential of AI-driven approaches is undeniable, challenges persist. These include the critical need for high-quality and diverse training data, the often-opaque nature of AI model decision-making (model interpretability), and the essential but sometimes complex integration of computational predictions with rigorous experimental validation. Overcoming these hurdles requires ongoing advancements in AI algorithm development, meticulous data curation strategies, and enhanced interdisciplinary collaboration between computational chemists, biologists, and pharmacologists to fully realize the promise of AI in revolutionizing natural product drug discovery.

3. METHODOLOGY: A MULTI-FACETED APPROACH TO AI-DRIVEN NATURAL PRODUCT DESIGN

The successful application of generative AI in de novo natural product design hinges on a carefully orchestrated methodology encompassing data preparation, model training, and rigorous molecule generation and validation.

3.1 Data Preparation: Laying the Foundation for Robust AI Models

The effectiveness of any generative AI model is intrinsically linked to the quality and representativeness of its training data.

3.1.1. Curating Training Datasets involves a meticulous process of sourcing, cleaning, and standardizing diverse datasets comprising chemical structures and their associated biological activities or properties. This often necessitates the integration of information from public repositories, scientific literature, and proprietary databases to ensure a broad coverage of relevant chemical space and biological targets. The aim is to create a comprehensive and unbiased dataset that the AI model can learn from effectively.

3.1.2. Data Augmentation Techniques play a crucial role in enhancing the diversity and size of these training datasets. By applying structural transformations (e.g., modifying functional groups or molecular scaffolds), manipulating property profiles, or even generating synthetic data using other AI models, the robustness and generalization capabilities of the generative models can be significantly improved. These techniques help to mitigate dataset bias and enable the exploration of a wider chemical landscape beyond the inherent limitations of the existing data.

3.2 Model Training: Guiding the AI to Generate Novel Molecules

Training generative AI models for natural product design is a critical step that involves careful selection of appropriate neural network architectures, such as GANs, VAEs, and RNNs, and the subsequent optimization of their numerous hyperparameters.

3.2.1. Training Protocols typically involve partitioning the prepared datasets into distinct training, validation, and test sets. Techniques like batch normalization and regularization are implemented to prevent overfitting, ensuring the model learns generalizable features rather than memorizing the training data. The model's parameters are then iteratively adjusted through training cycles using optimization algorithms like stochastic gradient descent (SGD) and its variants, which minimize the discrepancy between the model's predictions and the actual data. Finally, the performance and reliability of the trained models are rigorously assessed using

3.2.2. Evaluation Metrics. These metrics can include quantitative measures of the generated molecules' diversity (e.g., chemical validity and novelty), the accuracy of predicted biological properties (e.g., bioactivity and pharmacokinetic profiles), and the computational efficiency of the model (e.g., generation speed and scalability). Comparative analyses against benchmark datasets and known compounds provide crucial validation and allow for benchmarking against state-of-the-art methods.

3.3 Molecule Generation and Validation: From In Silico Predictions to Experimental Confirmation

The culmination of the data preparation and model training stages is the generation of novel molecules with desired properties. However, these computationally generated molecules must undergo rigorous validation to ascertain their potential as therapeutic agents.

3.3.1. In Silico Validation techniques employ a suite of computational methods to predict the physicochemical properties, pharmacological activities, and safety profiles of the AI-generated molecules.

Molecular docking simulations are used to predict binding affinities to target proteins, QSAR modeling correlates structural features with biological activity, and molecular dynamics simulations provide insights into molecular behaviour over time. These in silico assessments help prioritize promising lead compounds for further investigation.

3.3.2. Experimental Validation is essential to confirm the predicted properties and evaluate the true therapeutic potential of the AI-designed molecules. This involves synthesizing the selected compounds and subjecting them to a range of laboratory tests, including high-throughput screening assays, cell-based assays, and potentially animal studies. These experiments validate the predicted biological activities, assess efficacy and selectivity, and evaluate the safety profiles of the lead candidates. The iterative feedback loop between experimental data and computational predictions is crucial for refining the AI models and optimizing the design of subsequent generations of molecules, bridging the gap between the virtual and the real world.

4. RESULTS AND DISCUSSION: NAVIGATING THE CHALLENGES AND REALIZING THE POTENTIAL OF AI IN NATURAL PRODUCT DESIGN

While generative AI offers a transformative paradigm for de novo natural product design, its successful and widespread implementation in drug discovery and development necessitates a thorough understanding and proactive mitigation of several significant challenges and limitations.

4.1 The Critical Importance of Data: Quality, Availability, and Representation

4.1.1 Data Scarcity and Quality Concerns: A primary hurdle lies in the availability of high-quality and diverse training data**.** The performance of generative AI models is heavily dependent on the datasets they are trained on. Existing datasets in the realm of natural products and their bioactivities may suffer from limitations in size, potentially hindering the model's ability to learn complex relationships. Furthermore, these datasets can be biased towards certain chemical classes or well-studied biological activities, leading to models that are less effective in exploring novel or less characterized areas of chemical space. Ensuring data representativeness and reliability through rigorous curation and standardization is therefore crucial for training accurate and generalizable AI models capable of truly innovative design.

4.1.2 Addressing Data Gaps and Bias: Overcoming these data limitations requires concerted efforts in data integration from diverse sources, including public repositories, literature databases, and potentially proprietary datasets. Strategies to address inherent biases within the data, such as oversampling underrepresented classes or employing bias detection and mitigation techniques during model training, are also essential.

4.2 The "Black Box" Problem: Tackling Model Interpretability

4.2.1 The Challenge of Understanding AI Decision-Making: Generative AI models, particularly complex deep learning architectures like GANs and VAEs, often operate as "black boxes," lacking inherent interpretability in their decision-making processes. Understanding the intricate relationships and features that these models learn and utilize to generate specific molecular structures or predict biological activities remains a significant challenge.

4.2.2 Implications for Trust and Validation: This lack of transparency can hinder trust and validation in the generated outputs, especially in high-stakes applications like drug discovery. Without a clear understanding of the reasoning behind a model's predictions and designs, it becomes more difficult to identify potential flaws, optimize the models effectively, and gain the necessary confidence for further experimental validation and eventual clinical translation.

4.3 Bridging the In Silico-In Vitro Divide: Integration with Experimental Methods

4.3.1 The Bottleneck of Experimental Validation: While AI-driven predictions offer the promise of accelerated screening and optimization, the integration of computational predictions with experimental validation remains a critical bottleneck**.** The transition from promising in silico designs to tangible drug candidates necessitates rigorous laboratory testing.

4.3.2 The Necessity of Robust Experimental Protocols: Experimental validation of AI-generated molecules is essential to definitively confirm their predicted biological activities, assess crucial pharmacokinetic properties (absorption, distribution, metabolism, excretion), and evaluate their safety profiles. This requires the development and implementation of robust experimental design and validation protocols that can effectively handle the potentially large number of AI-generated candidates.

4.4 Navigating the Ethical and Regulatory Landscape

4.4.1 Ethical Considerations in AI-Driven Drug Discovery: The increasing adoption of AI in drug discovery raises significant ethical concerns. These include questions surrounding intellectual property rights for AI-designed molecules, the crucial aspect of data privacy when utilizing large biological and chemical datasets, and the broader ethical implications of automated decision-making in healthcare, particularly concerning patient safety and equitable access to AI-discovered therapeutics

4.4.2 The Evolving Regulatory Framework: Existing regulatory frameworks for drug development were largely established before the advent of sophisticated AI. These frameworks must evolve to accommodate AI-driven approaches, ensuring rigorous evaluation of the safety, efficacy, and ethical standards of AI-designed therapeutics before they can reach patients. This requires ongoing dialogue and collaboration between AI developers, pharmaceutical companies, and regulatory agencies.

4.5 Overcoming Computational Barriers: Resources and Scalability

4.5.1 The Demand for High-Performance Computing: Training and deploying complex generative AI models for drug discovery demands substantial computational resources, including access to high-performance computing infrastructure with powerful processors and large memory capacities. Efficient parallelization techniques are often necessary to handle the computationally intensive tasks involved in model training and large-scale molecular generation.

4.5.2 Ensuring Accessibility and Scalability: Scaling AI-driven workflows to handle the vast datasets and computational demands inherent in drug discovery remains a practical challenge. This can impact the accessibility and scalability of AI-driven methodologies for research institutions and smaller companies with limited computational resources, potentially creating disparities in the field.

4.6 Addressing Bias and Enhancing Generalization

4.6.1 The Propagation of Dataset Bias: AI models are susceptible to learning and propagating inherent biases present in their training datasets**.** This can lead to skewed representations of chemical space or inaccurate assessments of molecular properties, particularly for underrepresented chemical classes or biological targets.

4.6.2 The Need for Robust and Generalizable Models: Addressing bias and enhancing model generalization across diverse chemical and biological contexts are ongoing challenges. Developing techniques to identify and mitigate bias in training data, as well as designing model architectures that are more robust and less prone to overfitting to specific datasets, is crucial for ensuring the reliability and broad applicability of AI-driven approaches in natural product design.

4.7 The Economic Realities: Cost and Resource Allocation

4.7.1 Significant Investments Required: Implementing AI-driven methodologies in drug discovery necessitates substantial investments in technology infrastructure, specialized software, and highly skilled personnel with expertise in both AI and pharmaceutical sciences.

4.7.2 Strategic Resource Allocation: The cost-effectiveness and resource allocation strategies for integrating AI into existing drug discovery pipelines require careful consideration. Demonstrating a clear return on investment and strategically allocating resources to maximize the impact of AI while complementing traditional approaches will be critical for the widespread adoption and sustainable implementation of these powerful technologies.

5. CONCLUSION: EMBRACING THE AI REVOLUTION FOR ACCELERATED NATURAL PRODUCT DISCOVERY

The integration of generative AI into the field of natural product design marks a profound and transformative shift in drug discovery. By leveraging sophisticated machine learning algorithms, researchers are gaining unprecedented capabilities to accelerate the exploration and optimization of molecular structures with significant therapeutic potential. This review has highlighted the fundamental methodologies underpinning this revolution, including meticulous data preparation, strategic model training, and the crucial multi-stage process of molecule generation and validation. The potential of AI to overcome the inherent limitations of traditional drug discovery methods, such as expanding the accessible chemical space and enhancing the efficiency of lead identification, is truly compelling.

However, as we have discussed, the path towards fully realizing this potential is not without its challenges. Issues surrounding data quality and availability, the interpretability of complex AI models, the critical integration with experimental validation workflows, and the evolving ethical and regulatory landscape must be addressed thoughtfully and proactively. Furthermore, the significant computational resources required and the need for strategic cost and resource allocation necessitate careful planning and investment. Addressing the inherent biases in training data and ensuring the development of robust and generalizable AI models are also paramount for the reliability and broad applicability of these technologies.

Looking forward, continued advancements in AI algorithms, coupled with enhanced integration of diverse biological data and collaborative efforts across academia, industry, and regulatory bodies, hold the key to unlocking the full power of generative AI in pharmaceutical research. While the journey may be complex and multifaceted, the promise of accelerating the discovery and development of more effective and personalized therapies for a wide range of diseases makes the pursuit of AI-driven natural product design a vital and compelling frontier in modern pharmaceutical science. With careful consideration of the challenges and a commitment to ongoing innovation, AI-driven methodologies have the potential to reshape the future of medicine and ultimately improve patient outcomes worldwide.

6. REFERENCES

1. Aoki-Kinoshita, K. F., Kanehisa, M., & Li, X. (2003). Use of semantic web in life sciences: A birds of a feather session at the ISMB 2003. Bioinformatics, 19(16), 1950-1951. doi:10.1093/bioinformatics/btg263
2. Baskin, I. I., Winkler, D., Tetko, I. V., & Aksenova, T. I. (2016). Computational technologies for fast searching, design, and assessment of drug-like compounds. Current Pharmaceutical Design, 22(17), 2502-2512. doi:10.2174/1381612822666160210143406
3. Bajorath, J. (2002). Integration of virtual and high-throughput screening. Nature Reviews Drug Discovery, 1(11), 882-894. doi:10.1038/nrd942
4. Brown, N., Fiscato, M., Segler, M. H. S., & Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de novo molecular design. Journal of Chemical Information and Modeling, 59(3), 1096-1108. doi:10.1021/acs.jcim.8b00839
5. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250. doi:10.1016/j.drudis.2018.01.039
6. Clark, K., Vendruscolo, M., & Dobson, C. M. (2005). Structural biology: The challenges ahead. The EMBO Journal, 24(7), 1274-1279. doi:10.1038/sj.emboj.7600574
7. Durán-Frigola, M., Mateo, L., Aloy, P. (2020). The Drug Repurposing Hub: a next-generation drug library and information resource. Nature Medicine, 26, 757-758. doi:10.1038/s41591-020-0874-4
8. Duran-Frigola, M., & Aloy, P. (2020). Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chemical Science, 11(29), 7668-7681. doi:10.1039/D0SC02932J
9. Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics, 1(1), 8. doi:10.1186/1758-2946-1-8

10.Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., & Pande, V. S. (2018). Computational screen for sequestration of small molecules by proteins. Journal of Chemical Information and Modeling, 58(4), 739-749. doi:10.1021/acs.jcim.7b00715

11.Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. Journal of Computational Chemistry, 38(16), 1291-1307. doi:10.1002/jcc.24764

12.Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., Schneider, G. (2018). Generative recurrent networks for de novo drug design. Molecular Informatics, 37(1-2), 1700111. doi:10.1002/minf.201700111

13.Haas, J., & Roth, S. (2004). The paperless society revisited. Communications of the ACM, 47(7), 27-29. doi:10.1145/1005817.1005825

14.Hansen, K., Mika, S., Schroeter, T., Sutter, A., & Ter Laak, A. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. Journal of Chemical Information and Modeling, 49(9), 2077-2081. doi:10.1021/ci900161g

15.Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239-1249. doi:10.1111/j.1476-5381.2010.01127.x

16.King, R. D., Whelan, K. E., Jones, F. M., & Reiser, P. G. K. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427(6971), 247-252. doi:10.1038/nature02236

17.Koutsoukas, A., Lowe, R., KalantarMotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B. O., ... Bender, A. (2013). In silico target predictions: Defining a benchmarking data set and comparison of performance of the multiclass naive Bayesian and Parzen-Rosenblatt window. Journal of Chemical Information and Modeling, 53(8), 1957-1966. doi:10.1021/ci400132p

18.LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539

19.Li, Y., Zhang, Y., & Liu, X. S. (2019). Bayesian inference with historical data in the presence of non-ignorable missingness. Journal of the Royal Statistical Society: Series C (Applied Statistics), 68(5), 1411-1431. doi:10.1111/rssc.12353

20.Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 46(1-3), 3-26. doi:10.1016/S0169-409X(00)00129-0

21.Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry, 55(14), 6582-6594. doi:10.1021/jm300687e

22.O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3(1), 33. doi:10.1186/1758-2946-3-33

23.Riniker, S., & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of Cheminformatics, 5(1), 26. doi:10.1186/1758-2946-5-26

24.Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Berg, A. C. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252. doi:10.1007/s11263-015-0816-y

25.Schneider, P., Walters, W. P., & Plowright, A. T. (2012). Predicting human intestinal absorption in drug discovery: A comparison of physiologically based pharmacokinetic models and artificial neural networks. Journal of Chemical Information and Modeling, 52(9), 2049-2066. doi:10.1021/ci3001967

26.Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 555(7698), 604-610. doi:10.1038/nature25978

27.Shen, C., Tang, Y., Xiao, H., & Xu, C. (2019). AI-driven drug discovery: Challenges and perspectives. Drug Discovery Today, 24(8), 1697-1706. doi:10.1016/j.drudis.2019.04.013

28.Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489. doi:10.1038/nature16961

29.Smith, A. J. T., Humphrey, W., & Silveira, R. L. (2020). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology, 16(2), e1007313. doi:10.1371/journal.pcbi.1007313

30. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. Journal of Chemical Information and Computer Sciences, 43(2), 493-500. doi:10.1021/ci025584y
31. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.e13. doi:10.1016/j.cell.2020.01.021
32. Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. Foundations and Trends in Machine Learning.
33. Terstappen, G. C., Schlüpen, C., Raggiaschi, R., & Gaviraghi, G. (2007). Target deconvolution strategies in drug discovery. Nature Reviews Drug Discovery, 6(11), 891-903. doi:10.1038/nrd2403
34. Thorne, N., Auld, D. S., & Inglese, J. (2010). Apparent activity in high-throughput screening: Origins of compound-dependent assay interference. Current Opinion in Chemical Biology, 14(3), 315-324. doi:10.1016/j.cbpa.2010.02.018
35. Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. Molecular Informatics, 29(6-7), 476-488. doi:10.1002/minf.201000061
36. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... Hersey, A. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463-477. doi:10.1038/s41573-019-0024-5
37. Wang, Y., & Bryant, S. H. (2009). Cheng's computational biophysics laboratory: Data, algorithms, and tools in computational biophysics. Journal of Chemical Information and Modeling, 49(3), 734-739. doi:10.1021/ci8004659
38. Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today, 11(23-24), 1046-1053. doi:10.1016/j.drudis.2006.10.005
39. Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., & Lai, L. (2019). Deep learning for drug-induced liver injury. Journal of Chemical Information and Modeling, 59(5), 2340-2348. doi:10.1021/acs.jcim.9b00301
40. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... Wang, S. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370-3388. doi:10.1021/acs.jcim.9b00507
41. Zhang, L., Tan, J., Han, D., Zhu, H., & Fromm, M. (2019). Supervised topic models for clinical interpretability of perioperative procedures. Journal of the American Medical Informatics Association, 26(4), 303-312. doi:10.1093/jamia/ocy178
42. Zhao, L., Li, X., & Wang, Y. (2017). Bioinformatics in microRNA research. Methods in Molecular Biology, 1580, 167-177. doi:10.1007/978-1-4939-6866-4_12
43. Zhu, H., & Bigdeli, A. (2017). A review on predictive modeling under big data. Big Data Research, 9, 1-18. doi:10.1016/j.bdr.2017.01.001
44. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466. doi:10.1093/bioinformatics/bty269
45. Zupan, J., & Gasteiger, J. (2010). Neural networks for chemoinformatics and chemical engineering. John Wiley & Sons.
46. Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular