Application And Limitations of Artificial Intelligence in Natural Product Discovery

[1]Rutuja M. Bobade, [2]Revansiddha S. Sindagi, [3]Dr.Md. Javeed Y. Manure, [4]Bahubali N. Patil

[13]Department of Pharmaceutical Quality Assurance, [2]Department of Human health care and management, [4]Department of Pharmaceutics

[124]Appasaheb Birnale College of Pharmacy Sangli, Maharashtra India-416416, [2]Dr. D Y Patil Vidyapeeth (Deemed to be University), Pune, Maharashtra, India-411018

***Abstract*—In pharmaceutical sector, AI is significantly influential as the younger generation shifts towards natural remedies, appreciating their safety and effectiveness with minimal side effects. However, isolation of pure and specific phytoconstituents from a wide array of options is challenging. Artificial intelligence is utilized in several capacities, including the structural elucidation of chemical compounds, predicting the bioactivity of targeted substances, designing novel drugs with unique structures, and assessing the absorption, distribution, metabolism, efficacy, and toxicity of key compounds. Although it reduces human errors, it encounters difficulties such as managing the complexities of data, variability, quality, and biases. Issues related to the model, including overfitting and underfitting of datasets, disrupt the training of models on these datasets. Even the biological and experimental research, challenges arise from incomplete gene annotations and the failure to accurately replicate in-vitro and in-vivo models. This review article discusses the challenges faced by artificial intelligence during natural product drug discovery.**

***Index Terms*—Dereplication, Bottlenecks, Drug discovery, Data integration, Cheminformatics, Natural products**

I. Introduction

The emergence of natural remedies can be attributed to the extensive discovery of chemical compounds, leading to the isolation of new components for enhanced results. The natural products offer possibilities for uncovering the new compounds that can be developed into pharmaceuticals[(1)].

Although the intricate structures of natural products provide benefits, developing structural analogs to explore structure-activity relationships and enhance natural product candidates can be complex, particularly when the synthesis pathways are complicated. Moreover, phenotypic screening is often utilized to discover therapeutic leads from natural products, and ascertaining chemical structure is a lengthy process. The main obstacles for upcoming discoveries associated with natural products include dereplication and structural analysis, particularly in determining the relative and absolute configurations of secondary metabolites with stereogenic centers [(2)].

Typically, multiple compounds interact synergistically and concurrently, contributing effectiveness to treat disease with minimizing side effects plant extracts. Consequently, the process for developing plant-derived medications should commence with a combinatorial approach when assessing potential molecules and technological advancements[3].

II. SCOPE

To improve the systematic approach in the search for natural product drug discovery, it is crucial to utilize various emerging artificial intelligence technologies to create innovative and pure compounds, ensuring that new treatments along with its safety, efficacy, and quality. During these procedures, a variety of challenges can surface when integrating and training AI with large datasets. To address these issues, new AI tools can effectively identify precise compounds along with their structural screening. All new drugs should be designed using robust methods that prioritize safety; AI can assist in overcoming many difficulties faced when sourcing pure compounds from extensive collections and in developing reliable biological models, thus decreasing the necessity for experimental animals and expediting the identification of new compounds[4].

III. LITERATURE SEARCH STRATEGY

Different scientific databases like **Scopus, PubMed, Google scholar** are utilized for information gathering regarding the challenges faced by AI during natural products drug discovery, employing keywords such as "natural product drug discovery", "the use of AI in natural product drug discovery", "issues concerning artificial intelligence", "data-related challenges", "model-related challenges", and "biological and experimental challenges. Research articles concerning the applications and drawbacks of artificial intelligence in the discovery of natural products drug discovery were examined, specifically those published from 2010 to 2025, regardless of their publication date. In cases where two papers presented identical information, the more recent one was chosen.

IV. CASE STUDIES

4.1. Case study based on deep learning model for antibiotic [5].

4.1.1. Background and rationale

The quest for natural products is currently impeded by the challenge of dereplication, where identical molecules are often rediscovered. Furthermore, the rapid expansion of available chemical spaces through the modification of complex scaffolds results in a greater number of failures than successes when developing next-generation antibiotics based on existing ones. Consequently, Stokes et al. (2020) employed deep learning models across different chemical libraries to identify promising lead compounds.

4.1.2. Key identification

The model observed demonstrated efficacy against bacteria due to its structurally diverse compounds compared to current antibacterial agents. The study highlights the success of diverse approaches in expanding the array of antibiotics through the identification of structurally distinct antibacterial compounds.

4.1.3. Issues

In this instance, the deep learning model created for antibiotic discovery performed well; however, it had some limitations due to the restricted datasets provided, which mainly concentrated on the phenotypic data of bacterial strains that were biologically inadequate. This issue led to problems with the model's ability to demonstrate consistent and reliable results, it potentially produced biased reports, and it struggled to comprehend the biosynthetic pathways of natural products, as the models were provided synthetic libraries datasets. For further enhancement, deep learning models should undergo more thorough training on particular task to improve detection accuracy.

4.1.4. Improvement

To ensure effectiveness in future applications, deep learning models should be trained using chemical libraries that encompass of molecules which have molecular properties. similar to those of antibacterial drugs but also be diverse enough to enable tools to identify the molecules. Additionally, by conducting multiple

training cycles across a variety of Phytoconstitute, we can analyze particular chemical constitutes which exhibits activity against bacteria.

4.2. Case study based on mass spectroscopy based molecular networking [6].

4.2.1. Background and rationale

After millions of years of evolutionary fine-tuning, natural products possess remarkable pharmacological characteristics and have traditionally been the primary resource for drug development. A considerable portion of small-molecule drugs and their derivatives have shown effectiveness in the treatment of diverse ailments. Nevertheless, it is clear that many pharmaceutical companies have discontinued their initiatives to screen new chemical entities from natural products since the year 2000. The reasons given point to the swift progress in biopharmaceuticals.

4.2.2. Key identification

The initial report on classical MS-based techniques was published by Dorrestein's team in 2012, aimed at examining the metabolic profiles of active microbial communities. In 2013, worldwide products originated from natural sources, serves as collaborative web platform for the storage, analysis, sharing, and comparison of MS data, along with the generation of molecular networks, further advanced the application of molecular networking across various research teams.

4.2.3. Issues

It's important to emphasize that MS-based analysis is a detection method that can be biased, depending on the effectiveness of compound fragmentation, as certain natural products do not ionize as effectively as others. Despite the use of various innovative techniques for screening and analyzing the molecular data of compounds, errors still occur, highlighting the necessity for manual skills and expertise.

4.2.4. Improvement

While there are conceptual approaches for molecular networking that rely on extensive datasets, the processes for data handling, algorithm development, and annotation compilation are essential. In recent years,

there has been a growing demand for innovations in the field of molecular analysis, detection, and the dereplication of datasets. This case study discusses various tools utilized for molecular networking, including CLMN, MS2LDA, FBMN, and IIMN, highlighting their wide-ranging applications in practical work and showcasing their advantages.

Table no. 1 provides a brief overview of case studies to help readers understand how frequently used approaches in NP drug development operate and the challenges they encounter.

Table 1. Observations of applications and limitations identified in the previously discussed case studies.

Study	AI approach	Objective	Key Findings	Major Limitations	Ref.
Case study 1	Deep Learning Model for Halicin Antibiotic Discovery.	To identify structurally novel antibacterial compound	Discovered eight antibacterial compounds including halicin with unique scaffolds.	Dataset biased towards *E. coli* Phenotypic data; Limited natural product representation.	Stokes et al.,2020 (5).
Case study 2	MS based Molecular Networking.	To improve dereplication and visualization of natural product chemical space	Enabled rapid comparison of MS spectra; improved annotation and collaborative data sharing.	Ionization bias, fragmentation variability, manual intervention required.	Qiang et al.,2021 (6).

V. CURRENT APPLICATIONS OF AI IN NATURAL PRODUCTS DRUG DISCOVERY

5.1. Structure elucidation

Artificial intelligence, particularly through machine learning and deep learning, has transformed drug development by improving data analysis and predictive modeling. AI algorithm toolkits exhibit strong performance, such as multitarget drug target interaction, which assesses the binding affinity between specific proteins and ligands. **Table no. 2** lists a few of the toolkits along with their intended usage. The NMR based model categorizes specific compounds using NMR data, whereas multitarget profile prediction identifies small molecules with the desired multitarget profiles for disease-related targets by employing the SMILES notation of a chemical compound[7].

AI Technique	Application Area	AI Tools used	Purpose	Ref.
Visualization models	Dereplication	Convolutional Neural Networks (CNN). Deep Belief Network (DBN).	For rapid identification of known compounds from complex extracts.	(10)
Deep learning	Structure Elucidation	Neural Networks, MS2DeepScore, Micro ED	Prediction of molecular structure from MS OR NMR data.	(7)
QSAR models	Activity Prediction	k-NN, Decision trees, Gradient boosting	Correlation of molecular descriptors with biological activity.	(8)
virtual screening	Target based screening	Molecular Docking + ML scoring	Screening large NP libraries against biological targets.	(11)
generative models	Drug Deigning by De novo.	Variational Autoencoder	Designing novel NP inspired scaffolds.	(12)

Table 2. Utilizing AI toolkits in the process of drug discovery.

5.2. Bioactivity prediction

In order to detect the biological activity of specific compounds, it is necessary to comprehend how these chemical compounds interact with targeted biological molecules. Several machine learning, deep learning, and natural language processing techniques have been advanced for the prediction of biological activity. Computational tools are essential, as demonstrated by SEA (similarity ensemble approach), which categorizes proteins by the chemical similarity of their ligands. TiGER (target inference generator) qualitatively identifies up to 331 targets. DEcRyPT (drug target relationship predictor) reveals the targets of phenotypic hits and accurately predicts their affinities. STarFish (stacked ensemble target fishing) assesses how small molecules interact with 1907 targets, with a particular emphasis on predicting the targets of natural products[8].

5.3. Drug designing

The integration of AI into computational chemistry has resulted in the creation of de novo design approaches that focus on developing new molecules. These methods aim to investigate and navigate vast regions of chemical space, extending beyond the limits of current databases of drug-like compounds. De novo molecular design primarily utilizes two approaches. The first, termed "Holistic Generation", involves constructing an entire molecule from the ground up, which is especially effective during the early discovery stages and for broadly exploring chemical spaces. The second method, "Iterative Generation," involves assembling molecules in a sequential manner, making it better suited for refining existing molecules or customizing them for specific applications[9].

VI. MAJOR ISSUES AND CHALLENGES

Artificial intelligence has made significant progress in discovering new natural compounds and identifying optimal methods for their extraction. Meanwhile, AI is addressing various challenges to enhance the efficiency of natural product discovery. The range and diversity of natural products are immense. The difficulties that develop during the various phases of drug discovery are many and stem from a number of causes, which are listed in **figure no.1** below.

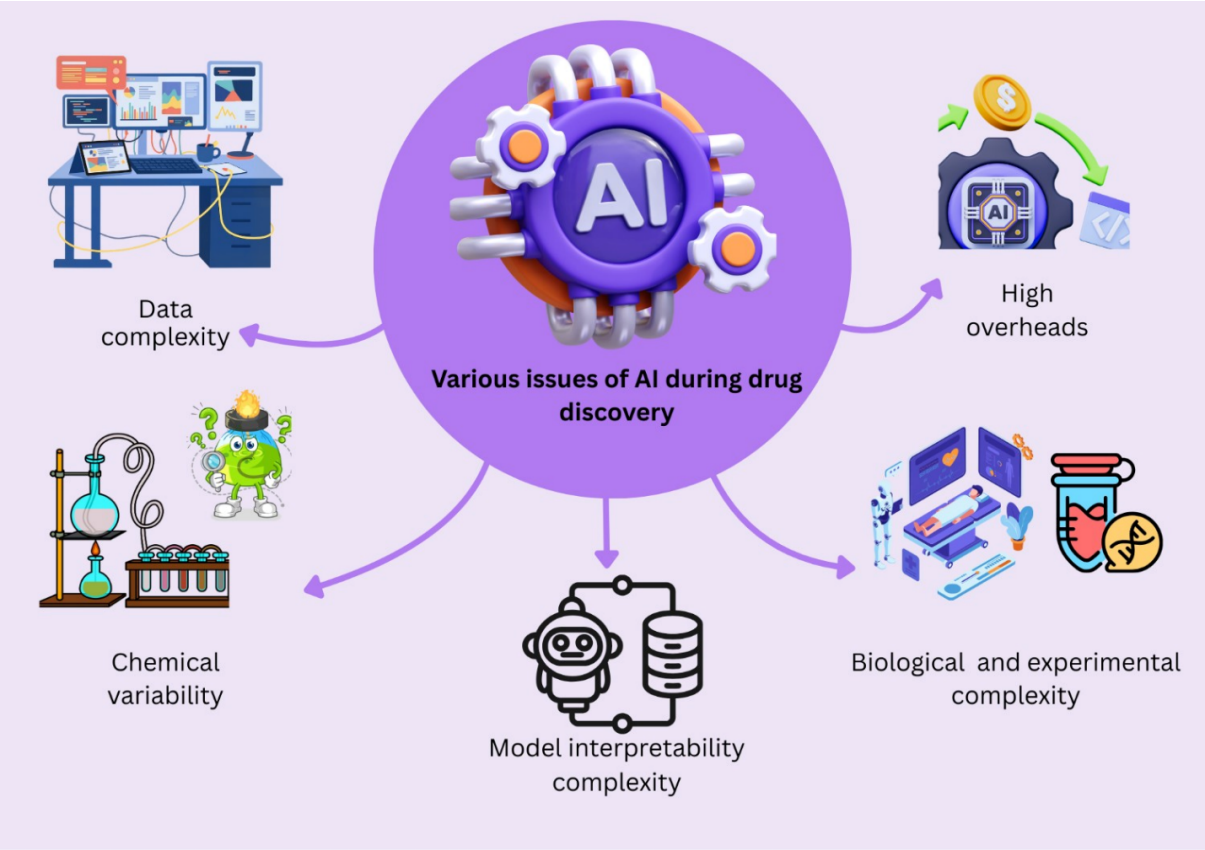

Figure 1. Issues faced by AI during natural product drug discovery

6.1. Data related issues

6.1.1. Scarcity and biasness of data

The main obstacle faced by AI in NP drug discovery is the challenge of incorporating a large variety of chemical moieties. However, early limitations in methods for determining structure and inaccuracies in data collection frequently result in incomplete datasets, which may lack stereochemical specifics, leading to the unnecessary storage of duplicate substances due to differences in the level of structural detail. Consequently, it is wise for users to confirm the accuracy of the information obtained[10].

Developing a community-oriented, openly available database for natural product structures would enhance advancements in technologies aimed at discovering natural products and strengthen the connections among current natural product data resources. To foster collaboration and integration, it is crucial to compile data from a variety of sectors, such as academic studies, industry research, and regulatory guidelines, to achieve validated results[11].

As scientists receive extensive responses from gene banks to evaluate the activity of chemical moieties for predicting their effects. Blood samples are gathered from various blood banks for analysis of m-RNA strains, which are then utilized to develop and train AI models aimed at predicting drug efficacy on these blood samples. However, the collected blood samples do not provide insights for organ-specific studies, as treating particular diseases like Alzheimer's, diabetes, and cancer necessitates m-RNA strains from the corresponding organs, and such data is often limited and of questionable quality. This poses a significant challenge for AI models in the drug discovery process for organ-specific diseases[12].

6.1.2. CHEMICAL COMPLEXITY

Natural products display an impressive variety in structural composition compared to synthetic compound libraries; they generally possess greater molar masses, a higher number of carbon and oxygen atoms, an increased quantity of hydrogen bond donors and acceptors, and enhanced rigidity in their molecular frameworks. This diversity however, diminishes their target affinity and specificity, presenting a challenge in the field of drug discovery[13].

The generalizability of machine learning models used for predicting toxicity is hindered by the variety of toxicity endpoints and the intricate nature of biological systems in vivo. These endpoints assess the toxicity profiles of compounds and highlight the diverse interactions that occur within biological systems. Generally, models are created for individual endpoints, each of which necessitates distinct data parameters. As a result, a model that is successful for one specific type of toxicity may not consistently yield reliable results for another[14].

Natural products frequently exhibit intricate structures that include multiple stereocenters, a variety of functional groups, and different isomers, which makes it challenging for AI algorithms to accurately forecast their bioactivity, toxicity, and other traits. To tackle this complexity, advanced computational tools such as automated statistical analysis, in silico screening, and multivariate data analysis are essential[15].

6.1.3. VARIABILITY IN HERBAL RAW MATERIALS

The composition, yield, and bioactivity of natural products can be significantly affected by variations related to age, the season of harvest, and the methods of extraction. Such inconsistencies can introduce bias in the data collected. Numerous challenges can impact the effectiveness, safety, and feasibility of developing new medications, despite the advantages of using phytochemicals in pharmacology. The first challenge arises from the complex and varied phytochemical compositions of plants, which can lead to inconsistent characterization and isolation of these compounds[16].

Environmental factors such as climate, geographic location, and soil conditions have a significant effect on a plant's phytochemical profile, complicating efforts for standardization. The second challenge is that extraction and purification processes for pure compounds are often time-consuming and sometimes yield extremely low amounts, which can hinder the ability to conduct thorough pharmacological studies and further development[17].

6.1.4. BIASED DATA

The difficulty is exacerbated by the lack of high quality, annotated datasets related to NPs, which complicates effective model training. This shortfall leads to issues such as overfitting, where models may perform exceptionally well on the training dataset but struggle to generalize to new or unseen data. The resultant fragmentation of data impedes AI systems from identifying cross-modal patterns that link structural, functional, and pharmacological aspects, thus limiting their capability to predict new chemotypes or reveal emerging bioactivities[18].

A common challenge in large data modeling is the issue of missing information. In earlier studies, a typical strategy was to create QSAR models for specific tests and then apply these models to predict target drugs that had not been tested against those assays. This approach was effective only when the data used in model-building demonstrated clear biological processes (such as log values or rigid target bindings, for instance). Nevertheless, due to prediction errors from QSAR models, this method introduced additional uncertainty into the modeling process. When dealing with diverse and intricate datasets (like clinical data), advanced statistical

methods such as multiple imputations become necessary. During the modeling phases, it is essential to prioritize active findings over inactive ones to account for the biased nature of high throughput screening data[19].

The reliability of AI models in natural product discovery is notably constrained by the "garbage in, garbage out" concept, where the quality of the input data poses a significant challenge. Systematic biases within publicly accessible databases (like ChEMBL), which predominantly focus on synthetic compounds and singular target assays, cause a significant underrepresentation of NP chemical variety and poly pharmacology. Furthermore, the diverse nature of data from various sources including raw extracts, different bioassays, and traditional records lacks uniform annotation, generating noise and complicating the integration of machine learning[20].

6.2. Model related issues

6.2.1. Overfitting due to small databases

Overfitting in machine learning refers to a situation where a model is trained to large number of datasets but it fails to identify unknown data. Data scientists usually develop a machine learning model using a particular dataset to make predictions. Once the model is trained, it applies this knowledge to predict outcomes for new data. A model that is overfit can yield inaccurate predictions and may have difficulty performing well across different types of new data. This issue can arise when the training dataset is minor and not enough to represent the required output. Overfitting can emerge for several reasons, such as having an insufficiently sized training dataset with a lack of adequate samples to represent all potential input data values. Furthermore, an overly complex model can lead to the capture of noise in the training data[21].

The initial problem in QSAR modeling is that imprecise training data can lead a QSAR model to capture the patterns present in the noise instead of the actual trends in the data, a commonly recognized issue known as overfitting[22].

6.2.2. Insufficient representation of stereochemistry and 3d geometry

Computers have allowed for the swift digital storage and retrieval of chemical substances and their representations, along with rapid updates to digital data and improved efficiency in physical storage. Algorithms emerged to visualize compounds in two dimensions, and the computational representation of these compounds in a three-dimensional format gained popularity with the launch of specialized software. Throughout that time, the efficiency of memory was vital in shaping the development of chemical notations. The representations that are prevalent today were primarily formulated in the 1970s to depict small molecules, macromolecules, and chemical processes. These representations are frequently employed in the fields of artificial intelligence, particularly within cheminformatics and drug development. The emphasis is confined to domains actively being explored in the fusion of AI and ML [(23)].

Figure No. 2 illustrates how recently developed techniques for performing QSPR and QSAR analysis use a simplifying molecular input-line entry system (SMILES) to describe chemical structures for two-dimensional modelling.

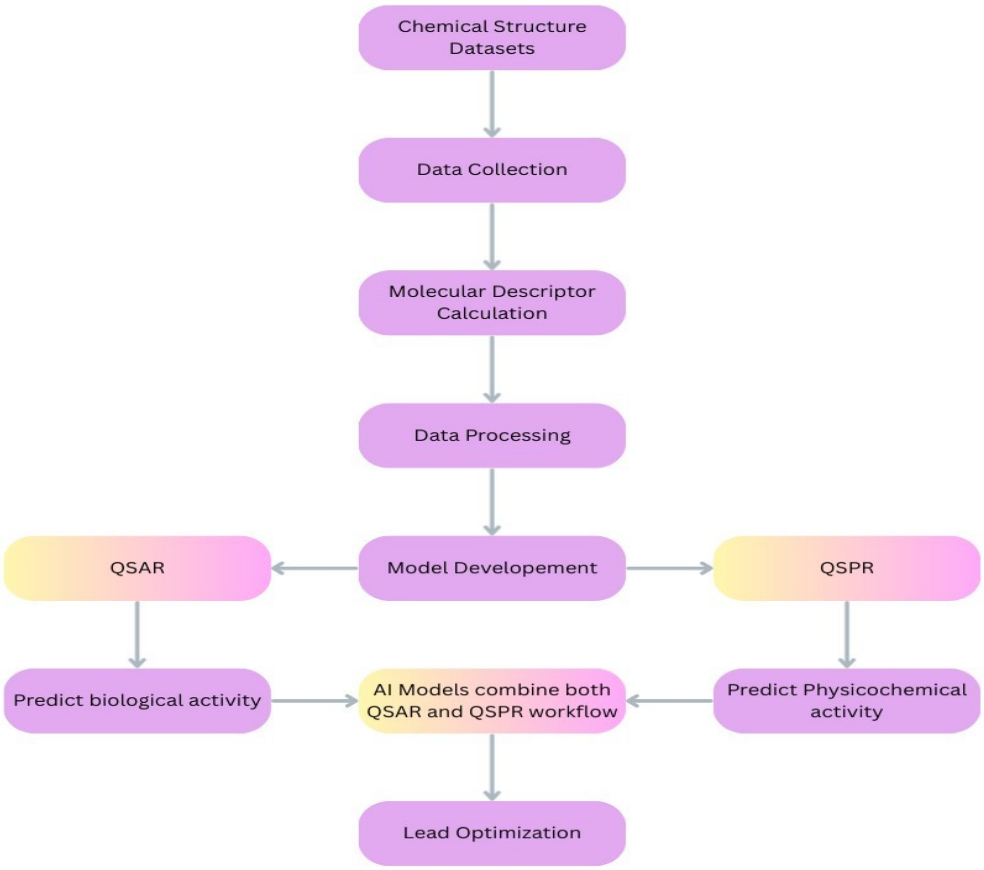

Figure 2. QSAR and QSPR models working processes

Unlike standard SMILES, quasi-SMILES consists of a series of special symbols and codes that represent molecular characteristics and specific experimental conditions. Both standards are the foundation methods to analyze the structures[24].

CoMFA and CoMSIA are two of the most commonly employed 3D QSAR techniques. CoMFA evaluates the forces related to steric and electrostatic interactions between the compound and the probe atom at various locations within a structured 3D grid, utilizing Lennard-Jones and Coulomb potentials. CoMSIA investigates the influence of steric fields, electrostatic fields, hydrogen bond donors and hydrogen bond acceptors on the effectiveness of inhibitors[25].

6.2.3. Difficulty in integrating multiomics datasets

Multi-omics research represents a revolutionary approach in life sciences, integrating information from genomic studies and various other omics technologies to provide a comprehensive perspective on biological systems[26]. The presence of unmeasured data in multi-omics datasets increases data sparsity, making the identification of patterns or predictions more challenging. Combining different types of data, such as binary genomic information and continuous gene expression levels, is complicated due to their differing scales. Variations in data from various laboratories, batches, or platforms can obscure true biological signals. Problems such as missing data, outliers, and noisy information in actual datasets can skew results if not properly addressed. Different omics technologies carry inherent biases, including variations in sensitivity or dynamic range, which must be corrected before integrating data. This highlights the heightened issues stemming from the lack of standardization, resulting in challenges in the integration of multi-omics data[27].

The number of gene expression datasets that are publicly accessible and stored in the Gene Expression Omnibus is increasing rapidly. These datasets are highly valuable for discovering new knowledge, especially when they are integrated. These datasets are inherently multimodal, including genomic sequences, metabolomic signatures, spectral readouts, and unstructured text records; however, they do not have standardized formats or consistent ontologies. The lack of standardization and interoperability greatly limits their compatibility with modern deep learning frameworks, which are designed to process clean, structured, and uniform data[28].

6.2.4. Blackbox

The challenges hindering the use of AI in natural product research go beyond simply data-related issues and include significant algorithmic and methodological hurdles. A notably important concern is the “black box” characteristic of sophisticated deep learning models, in which the processes behind their decision-making are unclear and lack transparency. This lack of clarity poses a considerable obstacle in a field that requires robust, transparent, and verifiable evidence to satisfy the strict criteria for regulatory approval and

clinical application. Without a clear understanding of the underlying mechanisms, researchers cannot fully trust the predictions made by the models[29].

A major limitation of AI technologies, such as neural networks, is their perception as black boxes. These methods are dependent on the training data, creating a constant risk of overgeneralization to scenarios not represented in that dataset. A drawback of genetic algorithm techniques is the absence of a guarantee for achieving optimal outcomes. The performance of deep learning models declines when the dataset is not adequately large. Reliable and sufficient data is fundamental for the success of AIDD models. Deep Neural Networks provide highly accurate predictions for AI models through extremely complex non-linear statistical models with numerous parameters, which obscures the transparency of these techniques. As a result, AI algorithms face issues of transparency, commonly referred to as opacity, which is a condition where a system does not provide justifications or logical explanations for its actions, often termed "the black box problem." Grasping black box systems presents a considerable difficulty for humans. Depending on black-box models for important decisions creates an urgent requirement for AI algorithms to be clear in their decision-making processes[30].

6.3. Biological and experimental issues

6.3.1. Complexity in np biosynthetic pathway

Although there have been considerable improvements in target identification, virtual screening, and compound optimization through AI, the synthesis of biosynthetic compounds continues to pose a challenge. While additional advancements are needed to completely leverage AI’s capabilities in natural product synthesis, innovations like Chemotic employ expert-like thinking to develop viable synthetic pathways for intricate compounds. A proficient and well-trained workforce presents another challenge. There is a demand for additional software engineers who understand AI technology, as well as for more skilled data scientists. With in depth knowledge in these areas combined with expertise in pharmaceuticals, the workforce can effectively engage in AI driven drug discovery. Nevertheless, there is still a need for a more skilled and trained workforce[31].

6.3.2. Incomplete annotations of genes

This presents a challenging issue, as a significant portion of genetic data in microbial and plant systems is classified as "hypothetical proteins" with unclear functions. AI models that are based on incomplete or insufficient annotations may generate predictions that lack innovation or reinforce existing flaws. Structural annotations are essential for understanding a wide range of biological data, which makes the annotation of small molecules a particularly challenging and critical aspect of untargeted mass spectrometry research. Nevertheless, the dissemination of these annotations requires manual scrutiny of mass spectrometry spectra within spectral networks, as the structural and natural diversity of natural products makes it difficult for AI models to accurately identify compounds. This type of analysis can only be conducted when a spectrum from a reference library is at hand. A major challenge in in silico annotation stems from the ambiguity regarding the accurate structure within the proposed candidate list. One key problem with annotation techniques based on machine learning is that they primarily concentrate on annotating molecules one by one. Nevertheless, structural hypotheses generated by interpretation of compounds discovered through MS can be strengthened by molecular connections based on spectrum similarity[23].

Now, there are two primary challenges in creating annotated datasets. It is impractical to integrate annotations from different sources since most datasets may be labeled in various ways, necessitating research into a unified training framework. The accuracy and structure of models are influenced by annotation techniques that often introduce biases and misleading categorization. While it is crucial for subject matter experts to annotate datasets for AI developers, the field should prioritize the development and implementation of annotation standards for fundamental data categories[32].

The diversity among these annotation databases is quite remarkable, and enrichment testing may not have been a primary aim in their original creation. Instead, they focus on specific contexts, including detailed metabolic pathway illustrations from KEGG, curated human pathways from Reactome, comprehensive protein function annotations from UniProt, identification of protein families and domains through SMART and InterPro, and structured ontologies related to gene functions or cellular locations across different levels

of detail in the Gene Ontology. These variations result in substantial disparities in gene representation, with a small number of genes predominantly affecting pathway annotations[33].

6.3.3. In vivo and in vitro

Modern methods that primarily rely on HTS datasets encounter practical difficulties because they are fundamentally based on less biosystem complexity or excessively easy experimental situation, especially when there are insufficient controls to accurately replicate in vivo features. They also have shortcomings in their ability to elucidate protein interactions. The effectiveness of implementations focused on structural-based drug design is limited by the existing knowledge of protein topology, which is often biased towards known protein conformations, leading to difficulties in selecting diverse or innovative chemical structures. This reliance on previously published information, while neglecting areas beyond the explored or reported experimental data ranges, significantly affects the creation of new medications or scaffolds an issue known as implementation bias. Furthermore, the success of utilizing data influences drug repurposing efforts and the capacity to achieve results for specific targets, while limitations in the reliability of data sets hinder both existing and prospective applications, failing to consider factors that govern clinical responses or patient demographics due to insufficient methodologies for assessing new medication applications in repurposing initiatives and flawed data evaluations. Consequently, to ensure that transforms and application in NP innovation are reliable, the existing emphasis on "structural analysis and virtual assays" ought to transition towards models that integrate various validation parameters, which should include real in vivo models and clinical endpoint information.

Neurodegenerative disorders are a category of long-term, advancing illnesses marked by the deterioration of neuronal structure and functionality. Traditional drug development approaches that focus on single molecular targets have often fallen short in preventing or halting the progression of these diseases, highlighting the need for more holistic strategies. In this context, Multi-Target Directed Ligands have emerged as a promising threating option[34].

6.4. Practical and commercial issues

6.4.1. Workflow and dereplication problem

Assessing the biological activity of extracts or extract fractions is the first step in creating a medication from natural matrices. A phytochemical profile of the samples under examination should be supplied to ensure study repeatability, as the phytochemical profile of natural samples differs by each batch. As an alternative, the specific substances that cause biological effects have to be recognized, separated, and their activity verified. The absence of effective methods for locating and separating physiologically active substances is one of the main obstacles to screening complicated mixtures for pharmacological leads[35].

Conventional screening techniques frequently call for the chemical separation of specific chemicals from mixtures, integrated and evaluated. These procedures usually expensive, labor intensive, and time-consuming. Synthetic libraries of individual chemicals are the primary focus of high throughput screening methods. To find the pharmacologically active compounds in mixtures, time-consuming isolation and characterization are required. The most generally performed strategy relies by separating of individual chemicals followed by evaluating them in their pure forms. Thus often results in the isolation of the most prevalent new compounds or molecules that have previously been thoroughly studied[36].

6.4.2. High computational cost for deep learning models

Deep learning models requires a large financial investment to meet the computational demands of both hardware and software assets. In addition to GPUs and TPUs, which are typically unavailable to many startups and academic institutions, training complicated models on big datasets requires an effective computational infrastructure. Moreover, the energy consumption associated with drug development makes these deep learning models environmentally unsustainable[37].Additionally, the process of model creation is ongoing, which adds to the computational load as models are continuously refined. Keeping costs under control necessitates well-organized workflows and structured algorithms. Nevertheless, developing such solutions requires knowledge in both pharmaceutical development and computational sciences[38].

6.4.3. Limited collaboration between phytochemistry and computational scientist

The combination of ethnopharmacological insights with computational drug discovery methods marks a significant change in the initial phases of drug development, particularly in the quest for bioactive substances derived from natural sources. Ethnopharmacology, which includes the traditional understanding of plant-based treatments in various cultures, offers a valuable repository of structurally varied bioactive substances. However, the process of converting these traditional leads into clinically applicable drug candidates has often been hindered by lengthy and resource-demanding experimental processes.(39) Although computers have made it easier to choose and refine lead compounds, fully automating drug development remains unfeasible because there are no advanced scientific robots powered by artificial intelligence or their virtual equivalents. The goal of fully automated AI drug development continues to be a hopeful ambition(40).

6.5. Regulation and ethical aspect

The main regulatory challenge is the absence of unified international standards for AI products. While the FDA has released its AI/ML Action Plan and white papers, and the EMA has published reflection papers regarding AI applications during the product life cycle, regulatory frameworks for adaptive learning algorithms and ongoing maintenance after market entry are still being developed. Submission requirements also differ greatly among agencies, leading to uncertainty for industry participants.

Another challenge is the question of accountability. When an AI system produces an incorrect result, it's often ambiguous who is responsible whether it's the developer, the pharmaceutical company utilizing it, or the regulatory agency relying on it. This ambiguity in accountability leads to legal and ethical uncertainties within the regulatory landscape(41).

VII. FUTURE PERSPECTIVE

7.1. Data standardization

The integration of AI with the Internet of Things could signify a significant advancement in pharmaceutical manufacturing, effectively bridging the gap between the digital and physical worlds. By incorporating AI algorithms into IoT sensors, both the production process and quality assurance are optimized,

resulting in greater overall efficiency. This synergy facilitates machine learning and deep learning, which allows for real-time analysis, predictive maintenance, and automation, as it continuously monitors vital manufacturing parameters[42].

7.2. Ethical concerns [43].

To fully harness the potential of AI while ensuring regulatory integrity is preserved, enhanced collaboration among global organizations, industry stakeholders, and technology developers is essential. Recommended best practices should involve:

- **Implementing ALCOA+ principles:** Attributable, Legible, Contemporaneous, Original, Accurate, Complete, Consistent, Enduring, and Available in AI-driven systems and Adhering to data protection laws such as GDPR and HIPAA[44].
- **Collaborative governance:** Developing worldwide regulatory structures that cater to the requirements of researchers, developers, and policymakers will aid in the incorporation of AI in natural product research[45].

VIII. Advanced Docking Methodologies

Future studies should concentrate on expanding these generative docking frameworks and giving priority to improvements in precision, computer effectiveness, and smooth compilation in NP discovery[46].

- **Four-dimensional (4D) docking** - technique reduces sampling time while preserving the accuracy of traditional assemblage dockage and allows for the seamless integration of receptor conformational ensembles into a single docking simulation[47].
- **Molecular dynamic -** The efficiency of evaluating metric depend on machine learning for interaction strength prediction may be improved by refining. The assessment was carried out by a number of angles, including feature selection, conformation sampling and collecting, descriptor development and combination, machine learning techniques, and the targets' structural flexibility. The most accurate prediction results were obtained for feature selection when Smina and NNscore were integrated, and each target's final gain features included a certain amount of strength[48].

- **Molecular dynamics simulations -** provide crucial information on the workings of macromolecules, enabling more adaptable sampling than conventional docking studies[49].
- **Convolutional neural network** - A CNN scoring function autonomously identifies essential characteristics of conjugate interactions that are associated by interacting. The development of CNN ranking to distinguish among accurate and inaccurate docking mode, as well as between known binders and non-binders[50].

IX. CONCLUSION

The incorporation of Artificial Intelligence into natural product research represents a major shift in drug discovery and development. By leveraging AI technology, researchers can address numerous challenges, with advancements in structure elucidation, De novo design, and more. However, despite these advancements, AI cannot yet be regarded as a completely reliable and robust approach for drug discovery, as several issues related to data reliability, AI models, clinical validity, computational complexity, as well as regulatory and ethical concerns must be acknowledged in this discussion. For further advancements in AI tools and techniques, it is essential to explore the specific issues that arise and to adopt new integrations in this field through ethical practices and collaboration with innovative methods to enhance the reliability of natural product drug discovery.

X. Acknowledgements

The authors would like to thank the Appasaheb Birnale College of Pharmacy, Sangli, for providing the resources to carry out a literature review and for editing figures.

Authors Contribution

Rutuja M Bobade: Conceptualization, literature review and writing. Revansiddha S Sindagi: Review and editing. Nishant Patil: Writing and visualization. Bahubali N Patil: Supervision, review, and editing. Javeed Y Manure: Conceptualization, Supervision, review and editing.

Funding

No funding was received for this study.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics Approval

None to declare

References

[1] Dzobo K. The Role of Natural Products as Sources of Therapeutic Agents for Innovative Drug Discovery. Compr Pharmacol. 2022;408–22. doi:10.1016/B978-0-12-820472-6.00041-4 PubMed PMID: null; PubMed Central PMCID: PMC8016209.

[2] Moseley HNB. Error Analysis and Propagation in Metabolomics Data Analysis. Comput Struct Biotechnol J. 2013 Feb 19;4:e201301006. doi:10.5936/csbj.201301006 PubMed PMID: 23667718; PubMed Central PMCID: PMC3647477.

[3] Li FS, Weng JK. Demystifying traditional herbal medicine with modern approach. Nat Plants. 2017 Jul 31;3(8):17109. doi:10.1038/nplants.2017.109

[4] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021 Jan;26(1):80–93. doi:10.1016/j.drudis.2020.10.010

[5] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020 Feb 20;180(4):688-702.e13. doi:10.1016/j.cell.2020.01.021 PubMed PMID: 32084340.

[6] Qiang B, Lai J, Jin H, Zhang L, Liu Z. Target Prediction Model for Natural Products Using Transfer Learning. Int J Mol Sci. 2021 Apr 28;22(9):4632. doi:10.3390/ijms22094632 PubMed PMID: 33924898; PubMed Central PMCID: PMC8124298.

[7] Structural characterisation of natural products by means of quantum chemical calculations of NMR parameters: new insights. Org Chem Front. 2021 Jan 1;8(9):2019–58. doi:10.1039/d1qo00034a

[8] Wright AD, de Nys R, Angerhofer CK, Pezzuto JM, Gurrath M. Biological activities and 3D QSAR studies of a series of Delisea pulchra (cf. fimbriata) derived natural products. J Nat Prod. 2006 Aug;69(8):1180–7. doi:10.1021/np050510c PubMed PMID: 16933872.

[9] Crucitti D, Pérez Míguez C, Díaz Arias JÁ, Fernandez Prada DB, Mosquera Orgueira A. De novo drug design through artificial intelligence: an introduction. Front Hematol. 2024 Jan 25;3:1305741. doi:10.3389/frhem.2024.1305741

[10] Guo K, Shen Y, Gonzalez-Montiel GA, Huang Y, Zhou Y, Surve M, et al. Artificial Intelligence in Spectroscopy: Advancing Chemistry from Prediction to Generation and Beyond [Internet]. arXiv; 2025 [cited 2025 Dec 21]. Available from: http://arxiv.org/abs/2502.09897 doi:10.48550/arXiv.2502.09897

[11] Gangwal A, Lavecchia A. Artificial Intelligence in Natural Product Drug Discovery: Current Applications and Future Perspectives. J Med Chem. 2025 Feb 7;68(4):3948–69. doi:10.1021/acs.jmedchem.4c01257 PubMed PMID: 39916476; PubMed Central PMCID: PMC11874025.

[12] GEO Accession viewer [Internet]. [cited 2026 Jan 29]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94873

[13] V. Simoben C, B. Babiaka S, A. Moumbock AF, T. Namba-Nzanguim C, Bekindaka Eni D, L. Medina-Franco J, et al. Challenges in natural product-based drug discovery assisted with in silico -based methods. RSC Adv. 2023;13(45):31578–94. doi:10.1039/D3RA06831E

[14] Abou Hajal A, Al Meslamani AZ. Overcoming barriers to machine learning applications in toxicity prediction. Expert Opin Drug Metab Toxicol. 2024 Jul 2;20(7):549–53. doi:10.1080/17425255.2023.2294939 PubMed PMID: 38088128.

[15] Gangwal A, Lavecchia A. Artificial Intelligence in Natural Product Drug Discovery: Current Applications and Future Perspectives. J Med Chem. 2025 Feb 7;68(4):3948–69. doi:10.1021/acs.jmedchem.4c01257 PubMed PMID: 39916476; PubMed Central PMCID: PMC11874025.

[16] Kumar S, Yadav A, Yadav M, Yadav JP. Effect of climate change on phytochemical diversity, total phenolic content and in vitro antioxidant activity of Aloe vera (L.) Burm.f. BMC Res Notes. 2017 Jan 25;10(1):60. doi:10.1186/s13104-017-2385-3

[17] Martinez-Mayorga K, G. Rosas-Jiménez J, Gonzalez-Ponce K, López-López E, Neme A, L. Medina-Franco J. The pursuit of accurate predictive models of the bioactivity of small molecules. Chem Sci. 2024;15(6):1938–52. doi:10.1039/D3SC05534E

[18] Paerhati Y, Aikebaier A, Dilimulati D, Baishan A, Yusufujiang N, Qiu X, et al. Rethinking Nature's Pharmacy: AI Era and Natural Product Drug Discovery. Pharmaceuticals. 2026 Feb 11;19(2). doi:10.3390/ph19020301

[19] Zhu H. Big Data and Artificial Intelligence Modeling for Drug Discovery. Annu Rev Pharmacol Toxicol. 2020 Jan 6;60(Volume 60, 2020):573–89. doi:10.1146/annurev-pharmtox-010919-023324

[20] Paerhati Y, Aikebaier A, Dilimulati D, Baishan A, Yusufujiang N, Qiu X, et al. Rethinking Nature's Pharmacy: AI Era and Natural Product Drug Discovery. Pharmaceuticals. 2026 Feb 11;19(2). doi:10.3390/ph19020301

[21] Zantvoort K, Nacke B, Görlich D, Hornstein S, Jacobi C, Funk B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. Npj Digit Med. 2024 Dec 18;7(1):361. doi:10.1038/s41746-024-01360-w

[22] Kolmar SS, Grulke CM. The effect of noise on the predictive limit of QSAR models. J Cheminformatics. 2021 Dec 1;13(1):92. doi:10.1186/s13321-021-00571-7

[23] Da Silva RR, Wang M, Nothias LF, Van Der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, et al. Propagating annotations of molecular networks using in silico fragmentation. Schlessinger A, editor. PLOS Comput Biol. 2018 Apr 18;14(4):e1006089. doi:10.1371/journal.pcbi.1006089

[24] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminformatics. 2020 Mar 18;12(1):17. doi:10.1186/s13321-020-00423-w

[25] Xie H, Qiu K, Xie X. 3D QSAR Studies, Pharmacophore Modeling and Virtual Screening on a Series of Steroidal Aromatase Inhibitors. Int J Mol Sci. 2014 Nov 14;15(11):20927–47. doi:10.3390/ijms151120927

[26] Luo Y, Zhao C, Chen F. Multiomics Research: Principles and Challenges in Integrated Analysis. BioDesign Res. 2024;6:0059. doi:10.34133/bdr.0059

[27] Ali N, Hanif N, Khan HA, Waseem MA, Saeed A, Zakir S, et al. Deep learning and artificial intelligence for drug discovery, application, challenge, and future perspectives. Discov Appl Sci. 2025 May 23;7(6):533. doi:10.1007/s42452-025-06991-6

[28] Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. Biophys Rev. 2018 Dec 29;11(1):103–10. doi:10.1007/s12551-018-0490-8 PubMed PMID: 30594974; PubMed Central PMCID: PMC6381352.

[29] Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery | Chemical Reviews [Internet]. [cited 2026 Feb 20]. Available from: https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00728

[30] Alizadehsani R, Oyelere SS, Hussain S, Calixto RR, Albuquerque VHC de, Roshanzamir M, et al. Explainable Artificial Intelligence for Drug Discovery and Development -- A Comprehensive Survey [Internet]. arXiv; 2023 [cited 2025 Nov 28]. Available from: http://arxiv.org/abs/2309.12177 doi:10.48550/arXiv.2309.12177

[31] Meijer D, Beniddir MA, Coley CW, Mejri YM, Öztürk M, Van Der Hooft JJJ, et al. Empowering natural product science with AI: leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025;42(4):654–62. doi:10.1039/D4NP00008K

[32] Mullowney MW, Duncan KR, Elsayed SS, Garg N, Van Der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023 Nov;22(11):895–916. doi:10.1038/s41573-023-00774-7

[33] Gable AL, Szklarczyk D, Lyon D, Matias Rodrigues JF, von Mering C. Systematic assessment of pathway databases, based on a diverse collection of user-submitted experiments. Brief Bioinform. 2022 Sep 1;23(5):bbac355. doi:10.1093/bib/bbac355

[34] Katsoulaki EE, Dimopoulos D, Hadjipavlou-Litina D. Multitarget Compounds Designed for Alzheimer, Parkinson, and Huntington Neurodegeneration Diseases. Pharmaceuticals. 2025 Jun 1;18(6). doi:10.3390/ph18060831

[35] Corson TW, Crews CM. Molecular understanding and modern application of traditional medicines: triumphs and trials. Cell. 2007 Sep 7;130(5):769–74. doi:10.1016/j.cell.2007.08.021 PubMed PMID: 17803898; PubMed Central PMCID: PMC2507744.

[36] Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015 Feb;14(2):111–29. doi:10.1038/nrd4510

[37] Jeon W, Ko G, Lee J, Lee H, Ha D, Ro WW. Deep learning with GPUs. In: Advances in Computers. 2021. doi:10.1016/bs.adcom.2020.11.003

[38] Dhudum R, Ganeshpurkar A, Pawar A. Revolutionizing Drug Discovery: A Comprehensive Review of AI Applications. Drugs Drug Candidates. 2024 Mar;3(1):148–71. doi:10.3390/ddc3010009

[39] Babalola OO, Bridget K, Oyubu G, Waheed SA, Ajiboye SA, Fakayode AE, et al. Integrating phytochemicals and in silico methods for modern drug discovery: a comprehensive review. Discov Chem. 2025 Dec 1;2(1):297. doi:10.1007/s44371-025-00373-y

[40] Wei J, Zhuo L, Fu X, Zeng X, Wang L, Zou Q, et al. DrugReAlign: a multisource prompt framework for drug repurposing based on large language models. BMC Biol. 2024 Oct 8;22(1):226. doi:10.1186/s12915-024-02028-3

[41] Pawar HB, Neelanshu N, Bhat GM, Srikanth A, Muragundi PM. Challenges of using artificial intelligence (Natural Language Processing) tools for regulatory documentation. J Appl Pharm Sci. 2025. doi:10.7324/JAPS.2026.267918

[42] Kodumuru R, Sarkar S, Parepally V, Chandarana J. Artificial Intelligence and Internet of Things Integration in Pharmaceutical Manufacturing: A Smart Synergy. Pharmaceutics. 2025 Feb 22;17(3):290. doi:10.3390/pharmaceutics17030290 PubMed PMID: 40142954; PubMed Central PMCID: PMC11944607.

[43] Craddock M. Data-Inclusive Open Source AI: Building a Fair and Collaborative Future. AI Created Strategy Reports [Internet]. 2024 Oct 20 [cited 2026 Jan 4]. Available from: https://medium.com/ai-created-strategy-reports/data-driven-open-source-ai-the-imperative-for-inclusive-definitions-and-practices-29bd30a76758

[44] Pawar HB, Neelanshu N, Bhat GM, Srikanth A, Muragundi PM. Challenges of using artificial intelligence (Natural Language Processing) tools for regulatory documentation. J Appl Pharm Sci. 2025. doi:10.7324/JAPS.2026.267918

[45] Trump BD, Cummings CL, Ellinport B, Galaitsi S, Janisko T, Pinigina E, et al. Governing the AI–biotech convergence. EMBO Rep. 2026 Jan 3;27(2):259–64. doi:10.1038/s44319-025-00628-w PubMed PMID: 41484382; PubMed Central PMCID: PMC12852783.

[46] Lam JH, Katritch V. Navigating structure-based drug discovery with emerging innovations in physics- and knowledge-based approaches. Npj Drug Discov. 2025;2(1):29. doi:10.1038/s44386-025-00031-4 PubMed PMID: 41341471; PubMed Central PMCID: PMC12669043.

[47] Bottegoni G, Kufareva I, Totrov M, Abagyan R. Four-dimensional Docking: a Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. J Med Chem. 2009 Jan 22;52(2):397–406. doi:10.1021/jm8009958 PubMed PMID: 19090659; PubMed Central PMCID: PMC2662720.

[48] Gu S, Shen C, Yu J, Zhao H, Liu H, Liu L, et al. Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning? [Internet]. [cited 2026 Jan 30]. Available from: https://dx.doi.org/10.1093/bib/bbad008

[49] Vidal-Limon A, Aguilar-Toalá JE, Liceaga AM. Integration of Molecular Docking Analysis and Molecular Dynamics Simulations for Studying Food Proteins and Bioactive Peptides. J Agric Food Chem. 2022 Feb 2;70(4):934–43. doi:10.1021/acs.jafc.1c06110

[50] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model. 2017 Apr 24;57(4):942–57. doi:10.1021/acs.jcim.6b00740 PubMed PMID: 28368587; PubMed Central PMCID: PMC5479431.