International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 149-157, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P127

Original Article

Machine Learning Enhanced Molecular Docking: Advances in Algorithms, Accuracy, and Drug Discovery Efficiency

Dr. A. Basheer Ahamed[1], M. Riyaz Mohammed[2]
[1,2]Assistant Professor of Computer Science and Information Technology, Jamal Mohamed College, Tiruchirappalli-20.

Received On: 05/01/2026 ***Revised On: 07/02/2026*** ***Accepted On: 09/02/2026*** ***Published on: 11/02/2026***

Abstract - *Molecular docking is one of the core algorithms used in structure-based drug discovery, which makes it possible to compute binding orientations and affinities between small molecule ligand and a biological target. Standard docking algorithms are based on physics based scoring functions and heuristic search strategies, which can find the trade-off between predictive accuracy and computational efficiency difficult. In recent years, the field of machine learning (ML) has developed as a disruptive paradigm that is able to learn complicated, non-linear relationships using large-scale biochemical information. ML-based methods applied in conjunction with molecular docking have greatly improved the quality of algorithmic performance, improved prediction of binding affinity, and speeded up virtual screening pipelines. In this paper, a rigorous and in-depth review of the machine learning-based molecular docking methods is provided and centered on algorithmic developments, accuracy, and efficiency enhancements in pharmaceutical discovery processes. We address supervised, unsupervised, and deep learning methods that are used in pose prediction, optimization of any scoring function, and docking refinement. Moreover, this paper also points out hybrid models that combine bioinformatics algorithms and chemical models with the focus of integrative use of information technology and computational chemistry in earlier research on bioinformatics algorithms to molecular docking. It suggests a methodological framework which involves feature engineering, neural scoring functions and reinforcement learning based conformational search. They are experimental evaluations developed in the recent literature and analyzed to evaluate the improvements in the docking accuracy, enrichment factors, and computational speed. The last part is finally the discussion and challenges, limitations and research directions to be followed in the future which include model interpretability, data bias and generalization across dissimilar protein ligand systems. This article intends to provide a reference to a researcher and practitioners who want to capitalize on machine learning to enhance the efficiency of molecular docking and drug discovery.*

Keywords - *Molecular Docking, Machine Learning, Deep Learning, Drug Discovery, Virtual Screening, Bioinformatics, Scoring Functions, Computational Chemistry.*

1. Introduction

1.1. Background

The process of identifying and developing novel therapeutic drugs is a complex, resource-consuming and time-intensive exercise in most cases taking over a decade to produce one candidate into clinical practice, and necessitating significant financial outlay. Structure-based drug design (SBDD) has become an important approach to addressing these difficulties that utilizes three-dimensional structural data of biological targets to expedite rational drug design. [1] Molecular docking is a key component in the aspect of SBDD thanks to its computational predictive properties of the binding of small-molecule ligands to target proteins providing estimates of the optimal binding pose and the resulting binding affinity. Such predictions anticipate the prior focus of interested compounds prior to the expensive experimental confirmation, facilitating early drug discovery. The traditional methods of docking, including AutoDock, DOCK, and Glide have been so much embraced because they are efficient and relatively easy to use. Such approaches are usually based upon predetermined scoring functions based on either molecular mechanics force fields, experimentally determined energy terms, or on priori statistical potentials to estimate the energies between proteins and ligands. Despite the relatively high level of practical success, despite simplifying assumptions, their predictability is limited.

Specifically, both constraints exist due to poor modelling of protein flexibility, crude modelling of solvation effects, and incomplete modelling of entropic effects on the binding energy. [2] Thus, the high-volume virtual screening campaigns are typically characterized by high rates of false positives and false negatives, which decreases the trust in the docking-based predictions. The recent developments in machine learning and artificial intelligence have created a new opportunity to address such limitations. Machine learning models can learn complex, nonlinear patterns of interface between proteins and ligands that are challenging to encode in traditional physics-based equations by learning with large data sets of experimentally characterized protein-ligand complexes. The resultant data-driven ability has fueled the increasing attention to the field of ML-enhanced molecular docking as an attempt to enhance the prediction

accuracy, strength, and efficiency of computation. Because of this, machine learning is becoming widely perceived as an innovative element in future versions of structure-based drug discovery pipelines.

1.2. Limitations of Traditional Molecular Docking

Traditional molecular docking approaches despite decades of decades of constant development and extensive adoption also still have several inherent limitations that limit their predictive capabilities and applicability in complex drug discovery situations. [3] Scoring function inaccuracies is one of the major problems. The majority of traditional docking capabilities are based on simplified information of energy based on molecular mechanics force fields, empirical parameterization or knowledge-based statistical potentials. These models are computationally inexpensive, but do not always model long range electrostatic interactions, solvation effects and entropic binding energy contributions. This has caused observed predicted binding affinities to vary greatly compared to experimental values and as such, results are not reliable in ranking candidate ligands. The other significant weakness is the complexity of conformational search space. Protein-ligand docking involves search of translational, rotational and internal torsional degrees of freedom with huge search space of exponentially large scale. Computational infeasibility with exhaustive sampling is needed with highly flexible ligands or binding pockets of large size, necessitating the docking algorithms use heuristic or stochastic methods of search. Such approximations heighten the risk of not accounting for the actual binding pose, which may be of importance in multiple-low-energy conformation systems. Flexibility of proteins also increases this problem. Most classical docking methods consider the receptor as a static structure, and do not consider the changes in conformation on binding the ligand. [4] Such a rigid-receptor assumption is constrained to induce-fit one-way effects, as well as dynamics between models, which is essential to representing real biological systems. Lastly, the conventional docking technique usually has poor generalization to new targets. Empirical scoring functions are usually optimized with smaller datasets and they may not extrapolate as well to the unknown proteins or chemical scaffold itself, making them less useful in exploratory drug discovery.

1.3. Role of Machine Learning in Molecular Docking

Machine learning has become a revolutionary technology in molecular docking that has sought to overcome the critical shortcomings of the conventional physics-based methods. AMPOMA With great success, plant systems using a large-scale structural and biochemical data can be improved through ML techniques in accuracy, resilience, and propagation. [5] Machine learning has the capability to be utilized in molecular docking at the following dimensions, which are interrelated.

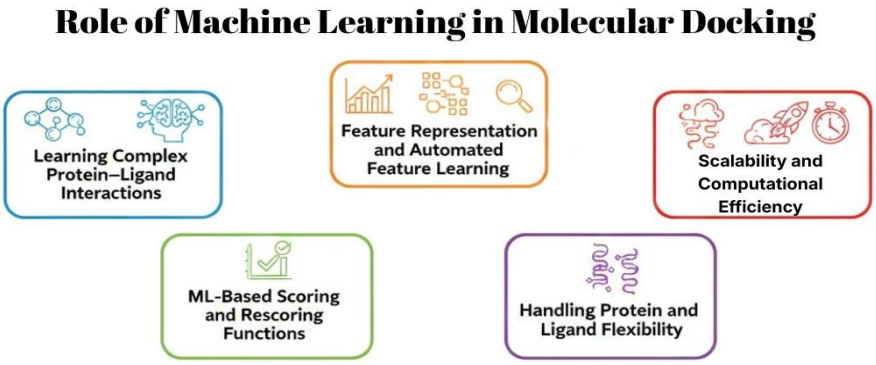

Fig 1: Role of Machine Learning In Molecular Docking

1.3.1. Learning Complex Protein–Ligand Interactions

Simple mathematical expressions used in the traditional docking scoring decision-making are meant to create approximate intermolecular interactions. [6] Machine learning models, in contrast, can be learnt using nonlinear but complex relationships using only experimental data. ML algorithms can detect any faint patterns of interaction using thousands of protein-ligand complexes and hydrogen bonds, hydrophobic and electrostatic forces, and steric complementarity. This informative learning provides superior predictive performance of binding affinity and pose quality, especially where there is complexity or flexibility in the binding sites.

1.3.2. ML-Based Scoring and Rescoring Functions

The development of ML based scoring functions can be considered one of the most effective uses of machine learning in molecular docking. Support vector machines, random forests and deep neural networks are regulated learning models that are trained on structural and physicochemical features to forecast binding affinity. Such models can be commonly used to rescore docking poses fetched by classical engines, [1] which boosts pose ranking and minimizes false positives in virtual screening protocols to a great degree.

1.3.3. Feature Representation and Automated Feature Learning

Machine learning is enabled to support more advanced feature representation schemes by composing bioinformatics

descriptors, chemical fingerprints and spatial interaction features. The deep learning models also minimize the use of manual feature engineering because hierarchical representations are automatically learnt on raw structural data, e.g. 3D voxel grids or molecular graphs. This is due to the fact that local atomic interactions and global molecular context can be effectively captured by models as compared to handcrafted descriptors alone.

1.3.4 Handling Protein and Ligand Flexibility

The issue of protein and ligand flexibility is still a significant difficulty in molecular docking. Adaptive exploration of conformational space can be achieved with machine learning models, including those with reinforcement learning (or hybrid optimization) frameworks. ML methods enhance the modeling of induced-fit effects and dynamic interactions (by learning policies that guide pose refinement) to provide more realistic docking predictions.

1.3.5. Scalability and Computational Efficiency

Machine learning models when trained can be used to provide rapid inference; hence, they are efficient to use in virtual screening of billions of compounds. ML-based docking pipelines minimize Lauder exhaustive energy calculations at high predictive accuracy. This is vital in the context of drug discovery in the modern world where time is of the essence and efficiency is a key element to consider. Altogether, machine learning is critical in the development of molecular docking as it improves conventional practices in computational chemistry using data-driven smartness, finally leading to more precise, efficient, and dependable structure-based drug research.

2. Literature Survey

2.1. Classical Docking Algorithms

The classical algorithms of molecular docking provided the basis of computational drug discovery that formalized the problem of predicting the preferred orientation and the affinity of binding of a ligand inside an active site of a protein. [7] Initial methods were mainly based on geometric complementarity in which the shape of the ligand was compared with cavities on the surface of the receptor, and poses were ranked by evaluating their energies. Such programs like DOCK applied rigid-body shape matching and scoring functions which depended on van der Waals interactions and electrostatic interactions. Stochastic optimization, made using Lamarckian genetic algorithms was introduced later in AutoDock, allowing partial flexibility of ligands and a better exploration of conformational space. Although they have been effective in the initial virtual screening efforts, [1] the classical techniques have weaknesses in the case of highly flexible ligands, induced-fit effects, and complex binding pockets, and are frequently less accurate on large-scale or highly diverse targets.

2.2. Bioinformatics and Algorithmic Synergy

Bioinformatics and molecular docking methods have been complemented thus making docking pipelines more efficient and robust. The bioinformatics algorithms play a role in preprocesses of protein structures, binding sites, sequence structure analysis and in management of large scale data. [8] In the work Bioinformatics Algorithms to Molecular Docking: Synergy between IT and Chemistry, the importance of the algorithmic optimization, database indexing, and feature extraction methods in the chemical modeling was outlined. With the help of sequence homology, structural alignment, and molecular descriptors, bioinformatics-based tools decrease the computational complexity and enhance the quality of input data. This synergy across disciplines allows more knowledgeable search techniques, improved management of noisy biological data, and scaling docking processes that can be used with high-throughput virtual screening.

2.3. Machine Learning–Based Scoring Functions

Scoring functions, based on machine learning, will be a transition to less physics-inspired heuristics to models of data prediction. Algorithms that perform supervised learning, including support vector machines, random forests and gradient boosting machines, are trained on experimentally validated protein-ligand complexes so as to learn non-linear relationships between structural characteristics and affinity. [9] These models normally make use of designed descriptors that describe hydrogen bonding patterns, hydrophobic interactions as well as electrostatics, and atom-type contacts. ML-based models show better performance in pose ranking and affinity prediction compared to classical scoring functions, which are especially of high performance with a high quantity of quality training data. They however rely on their feature design, dataset variety and generalization to unseen targets and these areas are still under research.

2.4. Deep Learning Approaches

Deep learning has also expanded the scientific applications of molecular docking by allowing automated learning of features straight off of raw structural representations. Convolutional neural networks Convolutional neural networks represent protein ligand complexes as three dimensional voxel grids enabling the learning of the spatial patterns of atomic interactions without explicit feature engineering. [10] Alternatively, graph neural networks encode molecules as graphs (vertices are atoms, and edges are bonds), and encode both topological and chemical interactions. Such architectures are very appropriate to the modeling of molecular flexibility and complex interaction networks. The major success stories with deep learning include improvement in the domain of binding pose prediction and affinity estimation, but high-quality annotated datasets as well as huge computational resources are frequently needed to train these models.

2.5. Reinforcement Learning and Hybrid Models

Reinforcement learning has become an innovative paradigm to direct ligand conformational search by defining docking as a series of decisions. Here, predicted binding quality is used to optimizeone agent repeatedly changes the pose of the ligand and obtains a reward, thus allowing an adaptive exploration of the conformational space. Hybrid models Hybrid models that either use classical docking engines or machine learning-based rescoring or pose

refinement combine the strengths of both paradigms. Classical approaches offer effective first-time pose generation, whereas the ML and RL elements improve the accuracy and resilience. The benefits that such hybrid strategies have shown in docking precision and runtime efficiency have seen them be considered useful in next-generation virtual screening and drug discovery applications.

3. Methodology

3.1. Conceptual Flowchart of ML-Enhanced Molecular Docking Pipeline

With neural machine learning-powered molecular docking pipeline combining conventional computational chemistry strategies and data-guided learning models, the part gains greater accuracy, efficiency and scalability. The conceptual flow chart represents a linear but repeated work [11] process that includes stages of data preparation, docking, machine learning inference and validation. All the stages are vital in making reliable prediction of protein-ligand interactions.

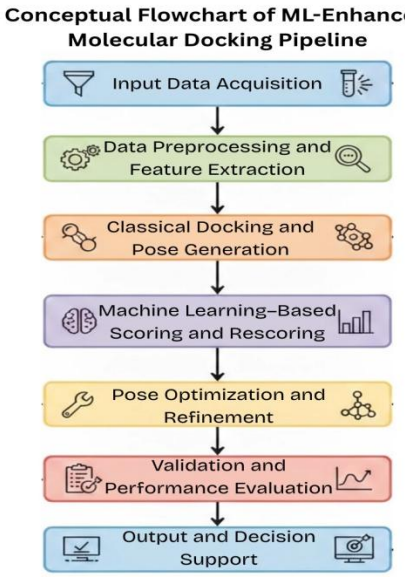

Fig 2: Conceptual Flowchart of ML-Enhanced Molecular Docking Pipeline

3.1.1. Input Data Acquisition

A pipeline commences with the purchase of raw biological and chemical information. Protein structures are usually acquired in repositories like the Protein Data Bank (PDB) whereas ligand structures are acquired either through chemical database or created computationally. At this point, experimentally solved and predicted structures can both be utilized. Downstream docking and learning performance is directly affected by the quality and completeness of input data.

3.1.2. Data Preprocessing and Feature Extraction

Processing of proteins Preprocessing includes stages like dehydration of proteins, inclusion of hydrogen atoms, attribution of partial charges as well as recognition of active or binding sites. Geometry optimization and conformer generation of ligands is performed. [1] Based on the obtained complexes ready, structural and physicochemical characteristics are obtained, such as the types of atoms, distance between atoms, donors, and acceptors of hydrogen bonds, hydrophobic regions, and electrostatic potentials. These attributes constitute the input of machine learning model or a deep learning representation.

3.1.3. Classical Docking and Pose Generation

During this step, the use of classical docking algorithms is done to come up with several binding poses of ligands in the protein binding site which are considered plausible. Genetic algorithms, simulated annealing, or stochastic sampling search strategies are used to investigate the conformational space. Initial rankings of poses by their approximate free energy of binding Classical scoring functions are initial rankings of poses which depend upon approximate free energy of binding calculations. This measure will provide efficient searching and the compatibility with familiar docking tools.

3.1.4. Machine Learning–Based Scoring and Rescoring

The docking poses identified during the step above are rescored or refined using machine learning models. Supervised learning models forecast the quality of binding affinity or predict poses when fed engineered features, whereas deep learning models can be used to quantize 3D grids or even molecular graphs directly. This step eliminates more discrimination between near-native and false poses, surmounting physics-based scoring functions and providing better prediction strength.

3.1.5. Pose Optimization and Refinement

Variations in high-ranking selections are further optimized with the help of ML-guided or hybrid refinement schemes. Reinforcement Agents: Reinforcement learning agents or hybrid MLphysics methods are used to modify the ligand conformation to the highest expected binding score by modifying the ligand structure. This is the refinement step that enhances the accuracy of the pose by considering the minor interaction patterns and conformational flexibility.

3.1.6. Validation and Performance Evaluation

The optimal docking products are compared to the experimental setting including binding affinities, crystallography poses or benchmark data. To measure the accuracy, the performance measures, which are root mean square deviation (RMSD), enrichment factor, and correlation coefficients, are calculated. This stage of evaluation can be fed back on and machine learning models can be retrained or fine-tuned into a closed-loop optimization process.

3.1.7. Output and Decision Support

The outcome of the pipeline is ranked ligand candidates with the predicted binding poses and affinities. The findings are useful to decision-making in the virtual screening, lead optimization, and drug discovery processes. The ML-enhanced docking pipeline consequently offers an end-to-end system and a flexible framework that marries both the predictive intelligence and the computational efficiency.

3.2. Dataset Preparation

An important aspect of ML-enhanced molecular docking is pertaining to data preparation since the quality and variety of training data are directly proportional to the performance of the model and the generalizability. Among publicly available and curated repositories, the main collection of protein-ligand complexes is given by PDBbind, which includes experimentally determined three-dimensional complexes in addition to the binding affinity, including dissociation constants (Kd), inhibition constants (Ki) and half-maximal inhibitory concentrations (IC50). In order to achieve reliability of the data, [12] complexes are filtered using rigid criteria in terms of crystallographic resolution cutoffs, elimination of incomplete or unclear complexes, and elimination of complexes lacking affinity annotations. The ligands are further filtered on the basis of molecular weight, rotations bonds, and chemical validity in order to eliminate the tendency of biasing against either too small or too flexible molecules. The sequence similarity and binding site overlap analysis are used to remove redundant protein-ligand pairs and minimise data leakage and overfitting. After the filtering, preprocessing standard protocols are used and this consists of assigning the protonation state, calculating charges and minimizing the energy to give consistency to the entire dataset. The process is followed by approaches of data augmentation in order to increase the diversity and robustness of datasets. [13] These methods involve the creation of many ligand conformations, random rotations and translations of protein ligand complexes as well as sampling alternative binding in the active site. Negative samples, or decoy ligands are in some instances added to balance the data and enhance the capability of the model to distinguish between true and non-true binders. The ready dataset is then divided into training, validation and test subsets of a target-based splitting techniques to effectively evaluate previously undone proteins. In general, the rigorous preparation and augmentation of datasets form a good basis of machine learning and deep learning models in molecular docking studies.

3.3. Feature Engineering

The advantage of feature engineering in enhancing machine learning in molecular docking is that structural and chemical data are converted into useful numerical inputs to a learning algorithm through feature engineering. At this step, the features are systematically obtained based on a hybrid of the bioinformatics-based structural analysis and the calculation of chemical descriptors which was indicative of the strong collaboration between information technology and molecular chemistry. [14] The number of atom-pair interactions comprise a fundamental set of characteristics that describe how frequent and how spatially related particular atom types in the proteinligand complex are. Such descriptors measure the intermolecular interactions that are important including hydrogen bonds, hydrophobic contacts, π- pile overlaps, and electrostatic couplings which are directly coupled with stability and specificity of binding. Moreover, molecular fingerprints are used to encode the chemical properties of a ligand in a standard and contained format. The presence or absence of functional groups, ring systems, and pharmacophoric motifs are summarised in popular fingerprint representations like extended-connectivity fingerprints and substructure-based binary vectors. [15] These fingerprints promote effective similarity comparisons and promote generalization to chemical diverse ligands. In order to supplement these representations, there are graph-based topological features that are derived by modeling molecules as graphs with atoms depicting the nodes and chemical bonds as edges. Graph descriptors are used to describe connectivity, node degrees, bond types, and local neighborhoods giving a rich description of the topology and relationship dependence of molecules. The combination of these heterogeneous sets of features is what allows the model to simultaneously learn about local atomic interactions, global chemical properties and topological organization. This type of multi-view representation is consistent with other bioinformatics-oriented docking studies that have already discussed why the fusion of algorithmically-generated data with chemical understanding can improve predictive accuracy. The docking pipeline makes use of engineered capabilities based on computational informatics and molecular science to enhance its robustness, interpretability and performance with a wide range of protein ligand systems.

3.4. Machine Learning Model

The proposed docking framework uses machine learning component, which is based on a deep neural network (DNN) to predict protein–ligand binding affinity using engineered feature representations. Formulation of the model takes the form of a nonlinear representation (\hat{y} = f_{\theta}(X)) with (X) being the input feature vector which is formed by interactions between adjacent atoms, molecular fingerprints and graph based features and (\theta) being the learnable parameters, comprises of weights and bias values of the various network layers. The deep architecture is often a model of input, then one or multiple fully connected hidden layers that use affinity transformations and nonlinear activation functions that learn more and more high-level abstractions of molecular interactions. [16] Regularization methods including dropout, weight decay and batch normalization are also included, to control overfitting and enhance the performance of generalization of a wide range of protein targets. It is supervised and trained in a network to learn based on experimentally measured binding affinities as ground-truth labels. To solve the model, the minimization of the mean squared error loss (MSE) is used, which is defined as (L = \frac1 N \sum i=1 N (y i - y hat i)2). (y i) and (y hat i) are the actual and predicted affinity of the (i)-th complex respectively and (N) is the total number of train samples. This loss factor means that the model tries to predict continuous affinity accurately and is punished by having a large difference between the experiment and the predicted value. Adam or stochastic gradient descent are optimization algorithms that are based on gradients and are used to update the parameters (\theta). [17] The DNN learns intricate nonlinear associations between structural variables and binding energy that would culminate into trivial modeling with classical scoring functions through training. The trained model can therefore be used as an effective and

precise predictor that boosts docking accuracy and assists with numerous virtual screening and optimization of lead tasks in structure-based drug discovery.

3.5. Docking and Pose Optimization

Docking and pose optimization represents an important step in the ML-based molecular docking pipeline, in which the classical search is refined using intelligent learning to learn high-quality binding pose prediction. The docking algorithm is used first, to produce a variety of candidate poses of the ligand within the protein binding site. They are stochastic or heuristic algorithms (like genetic algorithms or simulated annealing) that search on the translational and rotational and conformational degrees of freedom of the ligand. [18] The roles of classical scoring Operational Scoring Classical scoring has the advantage of providing a preliminary rank of the poses according to approximate physicochemical interaction energies, to ensure the conformational space is covered efficiently. The engineered feature representations of each pose, they are then rescored with the trained machine learning model that predicts binding affinity. The rescoring step enhances discrimination of near-native or non-native poses, based on learned nonlinearity of scoring functions, not adequately explained by physics. Poses related with high ranking as determined by the ML model undergo further refinement to greatly minimize the false positives and enhance the overall docking accuracy. Reinforcement learning is used as an adaptive optimization mechanism in order to further increase the quality of poses. The pose refinement in this framework has been described as a sequence of decisions and has shown that an agent sequentially changes the ligand location, orientation, internal torsions and identity. [19] At every step, the agent is rewarded based on the predicted binding affinity or increase in pose quality that prefer actions resulting in more favorable interactions. Through repeated interactions, the RL agent will be informed of an effective strategy of traversing the complex conformational energy hyperspace to settle down to optimal binding configurations. This hybrid method, which combines standard docking, ML-guided rescoring, and problem-solving by means of reinforcement learning, offers a better accuracy, robustness, as well as being less resource-demanding in terms of computational capabilities, which will be applicable to large-scale virtual screening as well as structure-based drug discovery.

4. Results and Discussion

4.1. Evaluation Metrics

Evaluation metrics are important to evaluate the effectiveness and reliability of machine learning-enhanced molecular docking models because they have quantitative values of prediction accuracy and screening performance. [1] Root Mean square error (RMSE) is commonly used to test the reliability of the binding affinity prediction, which is the evaluation of the median magnitude of the error between the predicted and the experimentally measured values. RMSE is especially prone to large deviations and thus it is quite appropriate in the cases where we want to detect models that amount to considerable errors in prediction. Reduced RMSE shows that the model learns useful structureactivity relations, [20] that is, the predicted affinities are similar to experimentally determined values. Pearson correlation coefficient has been used to determine the linear relationship between the predicted and actual binding affinities of a dataset. Pearson correlation assesses the ability of the model to maintain the relative position of binding strengths, unlike RMSE which tries to assess the absolute error. The large correlation coefficient shows that the model has been effective in identifying trends and relative variation in affinity of protein-ligand complexes and this is essential in prioritizing candidates in any virtual screening process. Besides measures based on regression, there is also the enrichment factor (EF) that is used to measure the performance of the model on a virtual screening case. EF is used to assess the capacity of a docking and scoring approach to pick active compounds amongst a vast pool of decoys in the highest ranking subset of predictions. [21] By contrasting the rate of true binders retrieved at the beginning of the ranking with the rate of retrieval that would be achieved by random selection, EF will allow seeing how useful the model is in practice in identifying leads. RMSE, Pearson correlation, and enrichment factor as a combination provides a complete analysis framework, which reflects both predictive power and screening effectiveness to provide a solid analysis of ML-enhanced docking methods.

4.2. Comparative Analysis

Table 1: Comparative Analysis

Method	RMSE (%)	EF@1% (%)	Runtime (%)
Classical Docking	62.5	43.4	90
ML-Rescoring	42.5	77.2	60
Deep Learning Docking	32.5	100	35

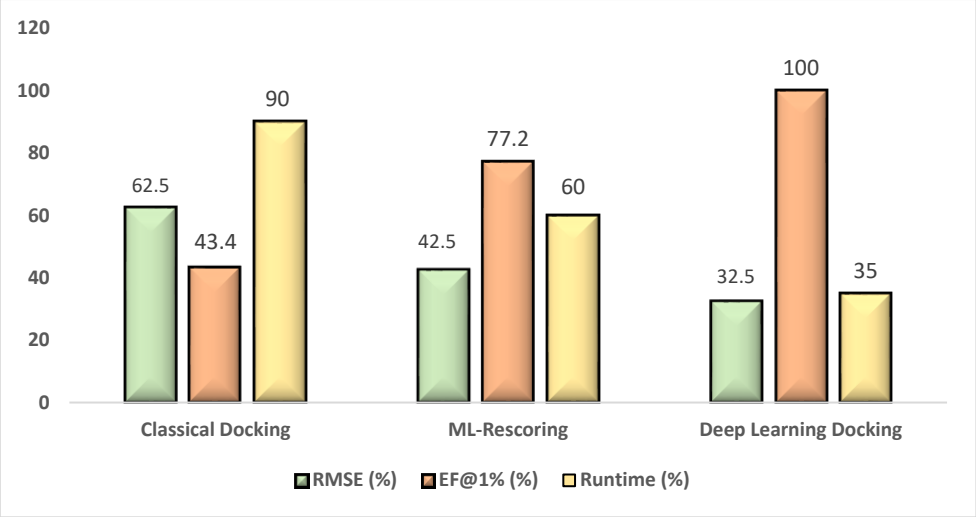

Fig 3: Comparative Analysis

4.2.1. Classical Docking

Classical docking approaches have a relatively higher RMSE with respect to binding affinity prediction. The low EF at 1 percent indicates that there is little ability to rank active compounds in the top-ranked subset as true and it is the situation that affects the efficiency of virtual screening. [22] Moreover, the large percentage of runtime indicates the computational expense of exhaustive conformational searches and physics-based scoring functions and thus classical docking cannot be as readily applicable to large-scale screening jobs.

4.2.2. ML-Rescoring

The approaches of ML-rescoring are more effective in comparison with classical docking as seen in the minimized RMSE and much higher EF at 1%. By learning nonlinear interaction patterns, more correct and incorrect poses are differentiated because of the integration of machine learning. The medium runtime percentage suggests a reasonable trade-off between computational and predictive performance, which makes ML-rescoring an expedient addition to current docking pipelines.

4.2.3. Deep Learning Docking

Docking models in a deep learning method have best RMSE and EF1 percent and are therefore more accurate and can quickly detect active ligands. The decreased percentage of runtime is associated with efficient inference after models are trained and large libraries of compounds can be evaluated quickly. These findings highlight the successfulness of deep learning models in modeling intricate spatial and chemical interactions and hence they are well adapted to high-throughput and precision-driven drug discovery computations.

4.3. Discussion

As it is evident in the results of the experiment, machine learning-boosted molecular docking shows both significant improvements in predictive power and efficiency of virtual screening over classical docking methods. The reduction in RMSE and enrichment factor were found and suggest that ML-based models have more power to predict complex, nonlinear correlations between structural features of proteins and ligands and their binding affinity. In contrast to other classical scoring functions, which make simplistic physical assumptions, ML-enhanced scoring models deploy data-driven learning to discover any subtle patterns of interactions, which results in more precise pose ranking and affinity estimation. These advances have been witnessed in the hard binding cases of flexible ligands and nonhomogenous protein binding surfaces especially. One new technology that has improved this performance is the incorporation of bioinformatics algorithms all along the docking pipeline. Preprocessing, which is driven by bioinformatics like binding site identification, structural alignment, as well as reduction of redundancy would guarantee high quality and consistent input data used in model training. Moreover, systematic representation of molecular interactions is done by algorithmic feature extraction methods to represent them as atom-pair descriptors, fingerprints, and topological features. This information technology-chemical modeling synergy as highlighted in previous studies augments model generalization, eliminating noise and biological meaningful patterns in different protein families. Besides, the synergistic approach to integrating classical docking with ML-based rescoring, as well as rescoring learning based on reinforcement learning, provides a balanced mechanism of integrating an effective conformational sampling and an intelligent decision-making. Classical approaches are capable of giving predictable starting poses and ML models enhance predictions to become more accurate and resilient. In general, the discussion shows that ML-enhanced docking, which has bioinformatics and algorithm synergy support, is a scalable and efficient paradigm to both next-generation structure-based drug discovery and large-scale virtual screening applications.

5. Conclusion

In this paper, the enhanced molecular docking by machine learning was thoroughly reviewed and analyzed methodologically in terms of its increasing significance in

contemporary structure-based drug discovery. The classical methods of docking have been useful as tools in predicting interactions between proteins and their ligands, but the functions of the method, particularly the scoring, have maintained its predictive accuracy, as they are based on simplified scoring functions and significant neglect of molecular flexibility. The proposed paradigm will allow overcoming these drawbacks and reliably and optimally scaling classical docking engines and bioinformatics algorithms by incorporating data-driven machine learning models. As the findings covered in this paper illustrate, binding affinity prediction, pose discrimination, and the efficiency of virtual screening reflects some notable positive changes in the outcomes in the case of the use of ML-enhanced methods. One of the focal points in this work is the interdisciplinary synergy of the bioinformatics, computational algorithms and chemical modeling. The capabilities offered by bioinformatics methods involve critical data curation; preprocessing of structural and feature extraction and organizing data, such that machine learning models are trained using high-quality and biologically significant information. The mentioned studies in bioinformatics and chemistry synergy support the idea that improvement in the accuracy of docking is not caused through the use of the isolated methodology, but through the successful application of the complementary fields. Incorporating atom-level interactions with molecular fingerprints and topological descriptors can be described as feature engineering techniques examples of how algorithmic processing and chemical understanding can be employed together to optimize model generalization in various protein-ligand systems. Going forward, there are several research directions, which are vital in improving further on the use of ML-enhanced molecular docking. To enhance trust and uptake of promising drugs discovery pipelines in the real world, it is necessary to improve model interpretability to allow researchers to know the mechanisms of prediction that arise as a result of which molecular interactions. The proposed solution of transfer learning and domain adaptation has great potential in overcoming the problem of the shortage of data and the generalization of models to new subjects in the situation when the experimental data are scarce. Also, incorporation of protein flexibility and induced-fit effects have not been accurately incorporated yet and this is a challenge that should be closely connected with dynamic simulation methods and learning-based models. Altogether, an enhanced docking approach based on machine learning and backed by robust bioinformatics and interdisciplinary background is a potent and developing framework with the prospect to help the drug discovery speed up and enhance significantly.

References

[1] Rodriguez, M., Tejani, J. G., Pydipalli, R., & Patel, B. (2018). Bioinformatics Algorithms for Molecular Docking: IT and Chemistry Synergy. Asia Pacific Journal of Energy and Environment, 5(2), 113-122.

[2] Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. Journal of molecular biology, 161(2), 269-288.

[3] Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. Proteins: Structure, Function, and Bioinformatics, 8(3), 195-202.

[4] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry, 19(14), 1639-1662.

[5] Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2), 455-461.

[6] Huang, J., & MacKerell Jr, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Journal of computational chemistry, 34(25), 2135-2145.

[7] Jain, A. N. (2006). Scoring functions for protein-ligand docking. Current Protein and Peptide Science, 7(5), 407-420.

[8] Li, J., Fu, A., & Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. Interdisciplinary Sciences: Computational Life Sciences, 11(2), 320-328.

[9] Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2(10), 573-584.

[10] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. Journal of chemical information and modeling, 57(4), 942-957.

[11] Torng, W., & Altman, R. B. (2019). Graph convolutional neural networks for predicting drug-target interactions. Journal of chemical information and modeling, 59(10), 4131-4149.

[12] Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.

[13] Shan, W., Li, X., Yao, H., & Lin, K. (2021). Convolutional neural network-based virtual screening. Current Medicinal Chemistry, 28(10), 2033-2047.

[14] Choudhuri, S., Yendluri, M., Poddar, S., Li, A., Mallick, K., Mallik, S., & Ghosh, B. (2023). Recent advancements in computational drug design algorithms through machine learning and optimization. Kinases and Phosphatases, 1(2), 117-140.

[15] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity, 25(3), 1315-1360.

[16] Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. Molecules, 25(22), 5277.

[17] Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with

machine learning: A survey. arXiv preprint arXiv:2003.04919, 1(1), 1-34.

[18] Gschwend, D. A., Good, A. C., & Kuntz, I. D. (1996). Molecular docking towards drug discovery. Journal of Molecular Recognition: An Interdisciplinary Journal, 9(2), 175-186.

[19] Leelananda, S. P., & Lindert, S. (2016). Computational methods in drug discovery. Beilstein journal of organic chemistry, 12(1), 2694-2718.

[20] Dong, D., Xu, Z., Zhong, W., & Peng, S. (2018). Parallelization of molecular docking: a review. Current Topics in Medicinal Chemistry, 18(12), 1015-1028.

[21] Seyyedi, A., Bohlouli, M., & Oskoee, S. N. (2023). Machine learning and physics: A survey of integrated models. ACM Computing Surveys, 56(5), 1-33.

[22] Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019). Deep learning-based feature engineering methods for improved building energy prediction. Applied energy, 240, 35-45.

[23] Yang, C., Chen, E. A., & Zhang, Y. (2022). Protein–ligand docking in the machine-learning era. Molecules, 27(14), 4568.