iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76281

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Integrating Machine Learning Models in Next-Gen Drug Discovery

Dr. S. Gunasekaran[1], Mr. Nandu S Nair[2], Nazna N[3], Riya Marjum M R[4], Sanjay S[5], Thoiba M[6]
[1]*Professor in CSE, Ahalia School of Engineering and Technology, Palakkad, Kerala*
[2]*Assistant Professor in CSE, Ahalia School of Engineering and Technology, Palakkad, Kerala*
[3, 4, 5, 6]*Ahalia School of Engineering and Technology, Palakkad, Kerala*

Abstract: ***Artificial intelligence and machine learning are completely transforming how we discover new medicines. Rather than investing lots of time and money in the lab, we can provide predictions for a chemical's activity very quickly. This article will take a look at four papers looking at the application of AI in two key areas of drug discovery: predicting drug-likeness and predicting binding affinity/target engagement. The first two papers, Druggability of Pharmaceutical Compounds Using Lipinski Rules with Machine Learning (2024) and DrugMetric: Quantitative Drug-Likeness Scoring Based on Chemical Space Distance (2024), focus on identifying and scoring compounds based on discrete physical and chemical properties, employing two types of machine learning either supervised or unsupervised. The second two papers, Explainable Deep Drug-Target Representations for Binding Affinity Prediction (2022) and WideDTA: Prediction of Drug-Target Binding Affinity (2019), utilize deep learning techniques such as combined graph, sequence, and text based to determine binding affinity of bound proteins and ligands, a more sophisticated approach. Looking across the four papers, we note that machine learning models introducing randomness, such as Random Forest and XGBoost may be recommended as easy to follow approaches even when utilizing rules to predict whether or not something is drug-like. However, deep learning methods such as Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs) are much better at inferring complex relationships between molecules when predicting their binding affinities. Nevertheless, the studies also support that there is a tradeoff between interpretability, transferability of the model to new data, and how long it takes to make a prediction. In summary, it appears that we will move towards a model that uses both interpretable machine learning and deep learning together to develop AI systems that facilitate discovery in drug development using data.***

Keywords: Artificial Intelligence (AI),Machine Learning (ML), Drug-Likeness Prediction, Binding Affinity Prediction, Random Forest / XGBoost, Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs)

I. INTRODUCTION

The process of discovering new medicines has always been long, costly, and difficult. Usually, it includes a significant amount of laboratory research and evaluating many different compounds. However, the fields of artificial intelligence and machine learning are changing this process, allowing us to more rapidly predict the behavior of molecules, determine if they possess drug-like properties, and assess their binding affinities to protein targets. This allows for computer-aided drug discovery that expedites the process and reduces costs by finding effective drug candidates faster than before. In this evolving field, establishing drug-likeness properties and binding affinity often referred to as "hit" identification are critical early-stage drug discovery steps. Drug-like properties establish whether or not a compound could be taken orally, and the characteristics its absorption-delivery-distribution elimination (ADME) properties, or how it will respond inside the body. Binding affinity characterizes how tightly the compound will bind to its target protein. Drug-like properties and binding affinity provide valuable insight into whether a compound has the potential to be developed into a drug.

Recently, there have been many novel AI-based approaches that can execute these assessments. Some employ rules (like Lipinski's Rule of Five) with computer programs to predict drug-like properties. They can be quite acceptable, and easy to read, which is what we want for computer evaluation. Other models use deep learning to analyze graphs, sequence, or text to predict adherence of compounds with targets. These models have the power to recognize details that older approaches cannot attain.

We examined four examples of literature that are indicative of this shift: two on predicting drug-likeness, Prediction of Drug–Target Binding Affinity. These papers reviews how the studies were evaluated, summaries of accuracy, advantages, and disadvantages of the examples reviewed. All of this will allow to recognize the newest advancements in artificial intelligence in drug discovery.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

II. LITERATURE REVIEW

The literature researches were based off 4 main papers namely: (1) Druggability of Pharmaceutical Compounds Using Lipinski Rules with Machine Learning (2024) and (2) DrugMetric: Quantitative Drug-Likeness Scoring Based on Chemical Space Distance (2024) and two on predicting binding affinity (3) Explainable Deep Drug–Target Representations for Binding Affinity Prediction (2022) and (4) WideDTA: Prediction of Drug–Target Binding Affinity (2019)

A. Druggability of Pharmaceutical Compounds Using Lipinski Rules with Machine Learning

Identifying promising drug candidates is a significant task in drug development research. Lipinski's Rule of Five (RO5) is crucial in the drug selection process. However, evaluating exhaustive numbers of new therapeutic candidates is done manually, which is costly and takes a long time. The incorporation of computers early in the drug design process has been a game-changer for evaluating and selecting therapeutic candidates much faster. This study focused on evaluating the RO5 using Machine Learning (ML) as learning computers that can comprehensively predict the RO5 for drug candidates. The best model produced a solution that identified, ranked, and established the likelihood of discovering RO5 compliant drugs. Overall, it was observed that all models generally sorted drug candidates effectively and performed well.

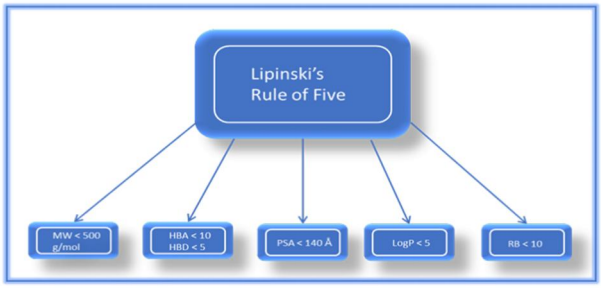


Fig. 1. Lipinski's Rule of Five

For instance, Random Forest (RF), Extreme Gradient Boost (XGBoost), and Decision Tree (DT) models were almost perfect at 99.9% compliance. This shows how effective group learning can sort and predict drug compounds based on adherence to RO5. Looking at predictive accuracy of these models side by side conveys how important it is to evaluate candidates based on assessing other considerations like, position, precision, F2-Score, Recall, Mean Absolute Error as well as interpretability and computational cost in selecting machine learning for early drug development.

1) Data Inputs

Molecular Weight (MW), Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), LogP: octanol-water partition coefficient, Topological Polar Surface Area (TPSA), Number of Rotatable Bonds, Synthetic Accessibility (SAS)

2) Data Processing and Preparation

Preprocessing: designed to allow no possibility of a bias in model accuracy. Irrelevant - missing data were discarded. The target variable "Passes Lipinski," was binary labelled - 1 if the compound adhered to RO5 and 0 for violations. The dataset of 11,583 compounds was then stratified - 80% for training and 20% for testing.

3) Modelling Approaches

The several authors proposed a comparative structure of seven supervised machine learning algorithms:
Decision Tree – to create a set of hierarchical rules to classify compounds.
RF-random forest represents a type of ensemble model composed of a collection of decision trees typical aimed at reducing overfitting and increasing predictive power.
Extreme Gradient Boost (XGBoost) - a technique aimed to improve performance through a set of iterative model refinements.
k-Nearest Neighbour (k-NN) - instance based classification based on structural similarity to training datasets.
Support Vector Machine - This will allow for the classification of all data points across a hypercomplex boundary.
Naïve Bayes: NB is a probabilistic model assuming all features are independent of one another.
Logistic Regression: baseline model for Binary Classification.
All models were tuned by Grid Search Optimization for consistency of parameters and model comparisons are reported.

©IJRASET: All Rights are Reserved

4) Evaluation Metrics

To thoroughly assess model performance, the study employed several evaluation metrics, including:

Balanced Accuracy: to address class imbalance, Precision, Recall, and F1-Score: Assessing quality or strength of classification, Receiver Operating Characteristic–Area Under Curve (ROC-AUC): addresses discrimination ability of the model. Only models with above 95% balanced accuracy were evaluated using ROC curves, feature importance rankings.

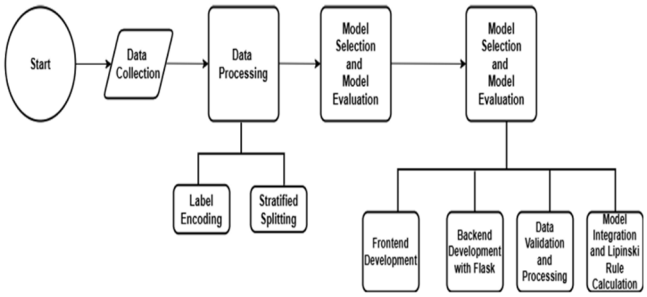


Fig. 2. Methodology Flow

III. RESULTS / FINDINGS

The results indicated that the ensemble learning methods, particularly Random Forest, Extreme Gradient Boost, and Decision Tree, were superior to traditional classifiers including Logistic Regression, Naïve Bayes, and Support Vector Machines in their predictive performance. Results in this study indicated that the combination of Lipinski's RO5 and machine learning algorithms produced a very effective and accurate framework for prediction of the druggability of pharmaceutical compounds. By using a total of seven supervised models, the authors systematically assessed the predictive performance, interpretability, and computational efficiency for each algorithm.

The results indicated that the ensemble learning methods, particularly Random Forest, Extreme Gradient Boost, and Decision Tree, were superior to traditional classifiers including Logistic Regression, Naïve Bayes, and Support Vector Machines in their predictive performance.

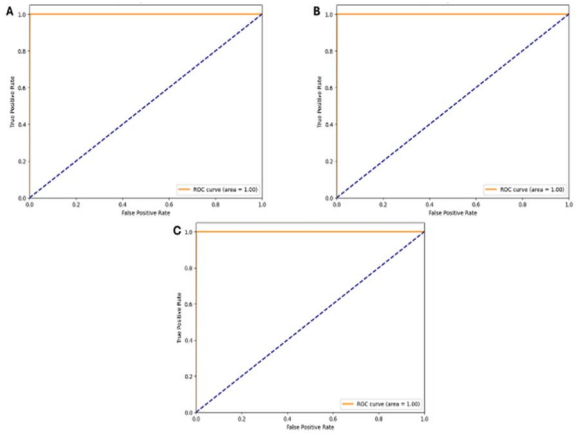


Fig. 3. ROC Curve for different models

Outstanding Model Accuracy: As a result, the models developed using Random Forest, XGBoost, and Decision Tree achieved a balanced accuracy of 99.94%, 99.81%, and 99.87%, respectively.

These remarkable performances indicate the strength of ensemble-based models in identifying druggable and non-druggable compounds, as they are inherently able to capture complex nonlinear relationships among molecular descriptors.

Augmented sensitivity and specificity: Indeed, the three best-performing models all achieved an ROC-AUC score of 1.00, which indicates that the models perfectly discriminated compounds based on RO5 compliance. Analysis of feature importance ranked HBA, MW, HBD, and TPSA as the most significant descriptors in prediction of drug-likeness by the model.

A. Paper 2: Drug Metric: quantitative drug-likeness scoring based on chemical space distance

The process of drug discovery is a lengthy and expensive endeavor. AI approaches spark excitement for accelerating the process to find molecules with drug-like characteristics. The drug-likeness continually proves to be an essential element during the virtual screening of drug-likeness potential. Alternatively, standard approaches, for example-QED, struggle with the process of accurately distinguishing a drug design versus a non-drug design. Moreover, deep learning-based approaches in binary classifiers rely heavily on how the training negative sets are chosen.

In this work, we propose an unsupervised learning framework using novel unsupervised learning framework, DrugMetric, to quantitatively score drug-likeness based on proximity in chemical space. DrugMetric combines the powerful learning ability of Variational Autoencoder with Gaussian Mixture Model's ability to discriminate. In addition, it capitalizes on the principles of ensemble learning in the design so that it gains even more predictive ability. When we evaluated DrugMetric on different tasks and datasets, we found that it consistently performed better with scoring and classification. In summary, DrugMetric effectively quantifies drug-likeness and provides a better classifier candidate than other traditional approaches (QED) when distinguishing drugs from non-drugs. Therefore, this framework has the potential to be a useful drug-likeness scoring mechanism that would accelerate virtual drug screening.

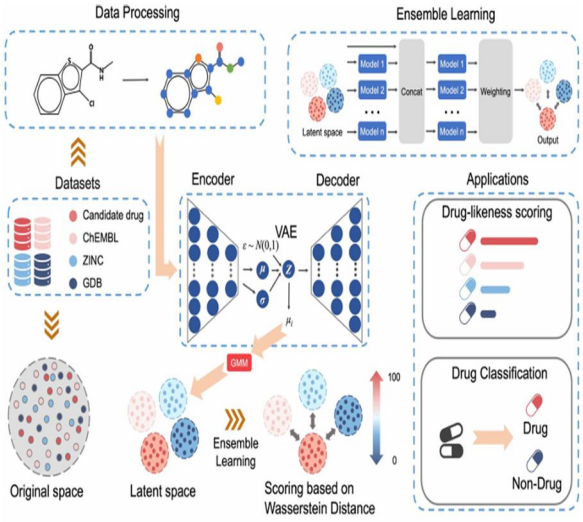


Fig. 4: Overall Workflow

1) Data Inputs

The analysis included a multi-source molecular dataset that contained an equal number of drug and non-drug compounds to build a representative chemical landscape:

Positive Dataset (CD): Contains confirmed drug molecules from: PubChem clinical trial data and records, FDA approved drugs lists and, World Drug Index (WDI) databases.

Negative Dataset: Represents varying levels of drug-likeness: ChEMBL - bioactive molecules with known experimental records, ZINC15 – commercial compounds to make available for screening, and GDB17 - theoretical molecules from a computer library of compounds.

2) Modeling Framework

The DrugMetric framework combines three core components:

a) Variational Autoencoder (VAE): This component encodes a molecular structure into a latent chemical space.

b) Gaussian Mixture Model (GMM): This is used to cluster latent representations into different clusters of drug-likeness.

c) Chemical Space Distance Calculation: The Wasserstein distance is used to quantify drug-likeness, which measures the geometric distance between two molecular distributions.

Thus, the authors have applied ensemble learning in the final scoring phase to enhance predictive accuracy and robustness:

Three-layer stacking architecture was combined with eight-fold bagging.

Base learners included Graph Neural Networks along with classical algorithms like SVM, Random Forest, and Extreme Gradient Boost. Weighted Ensemble L3: using weighted averaging for aggregation, giving more influence to models with better validation performance. This hybrid ensemble thus had a balanced local (graph-structural) with a global (statistical) feature learning and yielded high-precision quantitative scoring.

3) Evaluation Metrics

To assess the framework, we completed several quantitative and comparative metrics including:

ROC–AUC scores to evaluate discriminative power across datasets. Drug-likeness hierarchy verification—the validation of the predicted scores following the expected trend: CD > ChEMBL > ZINC > GDB. Comparative baselines: DrugMetric scores were compared against canonical systems such as QED, Lipinski, Pfizer, GSK, and Golden Triangle rules.

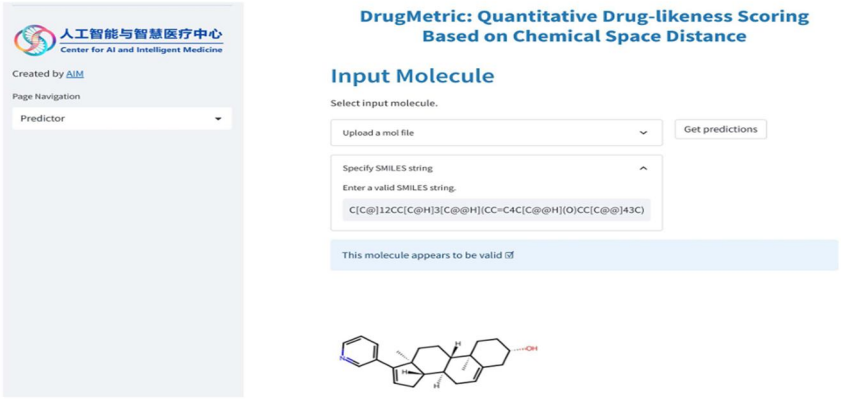


Fig. 5: Web Interface of DrugMetric

B. Paper 3: Explainable deep drug–target representations for binding affinity prediction

Several computational advances have been made in the area of drug discovery that help with identifying novel drug-target interactions and novel leads. However, most of the methodologies have not acknowledged that providing explanations to the decision-making of deep learning architectures is important. In this research, we examine the reliability of CNNs at identifying relevant regions for binding, specifically, binding sites and motifs, and the significance of the deep representations being taken as provided explanations to the model's decision-making based on the identified input regions that contributed the most to the prediction. We leverage an end-to-end deep learning architecture for binding affinity prediction, where CNNs are leveraged on their ability to automatically deduce and extract discriminating deep representation from 1D sequential & structural data. The authors created a (multi-stage) data-driven (scientific) methodology that combines graph-based molecular representation learning with deep neural architectures for high accuracy prediction of binding affinity in protein-ligand complexes. Their approach merges chemical, structural, and sequence-based information from large-scale bioactivity databases to model complicated molecular interactions. The framework prioritizes interpretability, robustness, and generalization across multiple protein-ligand systems.

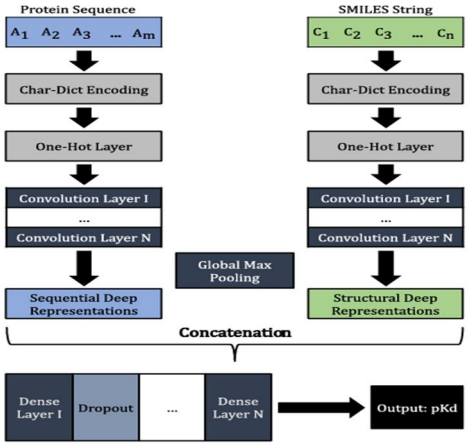


Fig. 5: CNN-FCNN binding affinity prediction model

1) Data Inputs

Data for this study were gathered from several publicly available bioactivity databases including, but not limited to: BindingDB: this database contains experimentally determined binding affinities, Kd, Ki, IC50, for a large number of protein-ligand pairs. PDBbind v2020: this is an updated version of PDBbind which contains three-dimensional protein-ligand complexes with binding affinity values. Each ligand was presented as both SMILES notation and a molecular graph, and the proteins were presented as primary amino acid sequences and 3D structure features. Preprocessing was performed to ensure quality and consistency, including: Removal of duplicate and incomplete entries. Conversion of binding affinities into pKd for numerical stability. The final dataset included over 250,000 protein-ligand pairs

2) Data Preprocessing

To facilitate the simultaneous incorporation of molecular and protein features, the Ligand SMILES strings were translated through the use of the RDKit library into graph structures that included atom-level connectivity and bond types. Protein sequences were encoded using one-hot encoding and position-specific scoring matrices to preserve the biochemical and evolutionary information. Molecular descriptors (i.e., molecular weight, LogP, hydrogen bond counts, and various topological indices) were calculated for enhancing the models. To ensure reproducibility and equity of the evaluations performed, the data was split as follows: 80% train, 10% validate, and 10% test.

3) Modelling Framework

The proposed framework integrates GNNs for ligands and CNNs for proteins in a dual encoder model architecture that jointly models structural and sequential dependencies.

Ligand Encoder (GNN): Represents molecules as node−edge graphs. It utilizes a combination of GCN and GAT layers to propagate atom-level information. Learns spatial and topological interactions affecting ligand binding potential.

Protein Encoder (CNN): Transforms amino acid sequences into numerical embeddings. Includes 1D convolutional layers to extract motifs indicative of binding pockets and secondary structures. Generates a reduced vector encoding protein features.

Interaction Module: Latent embeddings from each encoder are concatenated and passed to fully connected layers. A nonlinear regression head makes a prediction of the binding affinity value (pKd). Activation and normalization layers promote stable training and convergence.

4) Training and Optimizing

MSE was trained as the main objective loss function to minimize the distance between predicted and experimental affinity values. Optimizer: Adam, with an adaptive learning rate of 1e-4. Batch size: 256 protein-ligand pairs per iteration.

Regularization: Dropout (0.3) and L2 weight decay to address overfitting. Early stopping: This was triggered when validation loss plateaued for 15 epochs.

5) Evaluation Metrics

The assessment of the model performance was based on various statistical and correlation-based metrics:

Mean Squared Error (MSE): This represents how high the prediction errors were in value. RMSE is the root mean square error, indicating the average distance from the true values Concordance index, which indicates the ranking of binding affinities.

Pearson's correlation coefficient (r) – this quantifies the linear correlation between predicted affinities and the experimental affinities.

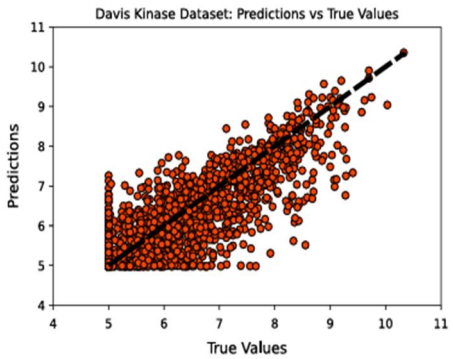


Fig. 5: CNN-FCNN model predictions against the true values for the Davis kinase binding affinity testing set, where the diagonal line is the reference line

C. Paper 4: WideDTA: Prediction of Drug-Target Binding Affinity

One of the main difficulties in drug discovery is predicting the interaction affinity between proteins and compounds. WideDTA is a model that uses deep learning to predict binding affinity by utilizing chemical and biological textual sequence information. To predict binding affinity, WideDTA uses four different text-based information sources: protein sequence, ligand SMILES, protein domains and motifs, and a maximum common substructure word. WideDTA was shown to outperform DeepDTA (one of the current state-of-the-art methods for predicting drug-target binding affinity based on deep learning) on the KIBA dataset with statistical significance.

This indicates that the word-based sequence representation approach utilized by WideDTA performs similarly to, if not better than, the character-based sequence representation approach utilized by DeepDTA (and other deep learning models) in binding affinity prediction. Moreover, given the protein sequence and ligand SMILES, results indicate that adding protein domain and motif information and ligand maximum common substructure words did not yield additional useful information for the deep learning model.

The authors suggest a modular, text-based deep learning approach to predict drug-target binding affinity without any 3D structural information. The proposed approach, WideDTA, systematically delivers and organizes different textual representations of proteins and ligands using sequence-based "words" to capture biological, chemical, and structural information. WideDTA organizes the framework using four different sources of information in order to model how the protein and ligand features, based on text-encodings, impact molecular binding affinity (compared to previous approaches using character-based deep learning architectures such as DeepDTA).

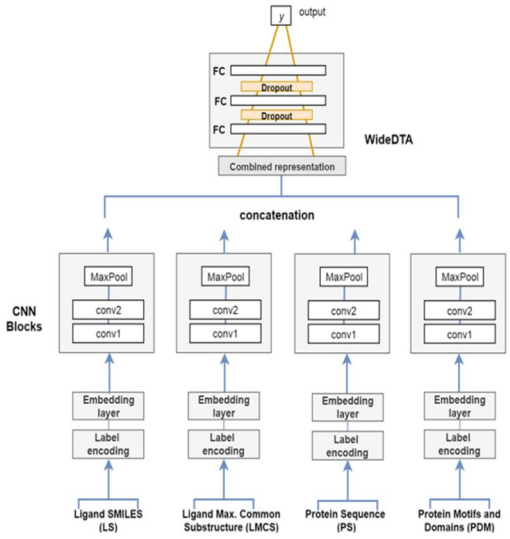


Fig. 6. Combination of CNN modules constructs the WideDTA architecture

1) Data Inputs

The evaluation of two benchmark datasets for bioactivity was performed in this study to promote generalizability and reproducibility:

Davis Dataset - A collection of 30,056 kinase protein-ligand interactions associated with the measurement of the experimental dissociation constants (pKd).

KIBA Dataset - 118,254 interactions between 229 proteins and 2111 compounds reported as KIBA scores that combine multiple experimental measures of affinity (e.g. IC_{50}, K_i, Kd)

The protein sequences were obtained from UniProt, and the ligand structures were obtained from PubChem. All protein-ligand pairs underwent a quality control pre-processing step, and the overall set of protein-ligand pairs was split into 6 equal folds for cross-validation.

2) Representation Modules

The methodology is founded on four text-based modules, each of which captures different aspects of protein and ligand information:

Protein Sequence (PS): Full amino acid sequences were split into overlapping 3-residue "words" (k-mers), as they are expected to capture local structural motifs and biochemical properties.

Protein Domains and Motifs (PDM): Functional motifs and conserved domains were pulled from PROSITE, as these regions usually encode binding specificity. Similar to the PS module, subsequent subsequences were represented as 3-residue words.
Ligand SMILES (LS): Ligand structures were represented as Simplified Molecular Input Line Entry System (SMILES) strings reduced to 8-character chemical words via a sliding-window approach. To enable more syntactic consistency, the authors also used DeepSMILES| a more sophisticated representation optimized for deep learning.
Ligand Maximum Common Substructures (LMCS): Ligands were represented as sets of repeating molecular sub-patterns—qualifying as "chemical words" encoding functional substructures attributes thought to contribute to binding. LMCS is derived from some of the 100 most frequent MCS encountered in prior chemical linguistics work.
In each of the modules a separate text corpus of words is created that represents the biochemical entities involved in binding

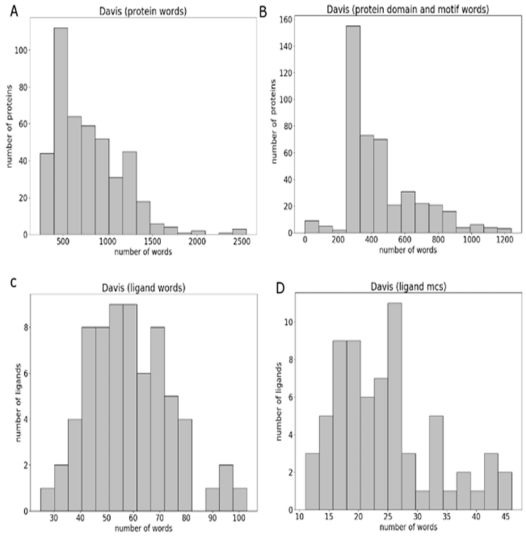


Fig. 7. Davis dataset- Distribution of number of words

3) Text Encoding and Feature Extraction

The "words" obtained from each module underwent conversion to integers, which were then mapped to 128-dimensional dense vectors via Keras Embedding layers. Afterward, these embeddings were passed through two 1-D Convolutional Neural Network (CNN) layers, each followed by a max-pooling layer and ReLU activation, with the purpose of learning high-level semantic and structural features.
Consequently, each CNN block would return an abstract vector representation for its respective module, which was then considered by concatenating all of the vectors into a single feature vector representing the overall protein-ligand pair.

4) Evaluation Metrics

The performance of the model will be assessed through the use of three supplementary metrics: Concordance Index: It assesses the appropriateness of the rankings derived from the predicted affinities Mean Squared Error (MSE): It quantifies the amount of error in terms of magnitude Pearson Correlation Coefficient (r): It measures the degree of linear correlation between the predicted and experimental values Statistical significance will be evaluated through paired t-tests with a confidence level of 95%.

IV. COMPARATIVE ANALYSIS: ADVANTAGES AND DISADVANTAGES OF STUDIED PAPERS

A. Paper 1 – "Druggability of Pharmaceutical Compounds Using Lipinski Rules with Machine Learning"

Benefits: Fusing classical and modern techniques: Combines Lipinski's Rule of 5 with a number of supervised machine learning algorithms to create a useful bridge between classical medicinal chemistry and contemporary artificial intelligence. Predictive accuracy: Implementation of diverse ensemble models - Random Forest and XGBoost for example - in achieving a classification accuracy of ~99.9% demonstrates their degree of reliability for classifying drug-like compounds. Real-world usability: The DrugCheckMaster web tool provides a tangible interface for researchers to efficiently test molecular druggability.

Understandability: Feature importance evaluations such as HBA, MW, LogP, enrich understanding of model behavior and properties contribution from a chemist perspective.
Disadvantages: Rule-based model: It depends mainly on Lipinski's Rule of Five and therefore restricts applicability to non-oral and macrocyclic drug candidates. High risk of overfitting: Extremely high accuracy could shed light on overfitting driven by either bias from the data, or the apparent lack of chemical diversity.
Limited chemical complexity: Preferring simple physicochemical descriptors instead of a broader range from 3D or molecular fingerprint can reduce generalizability.
No external validation: Model only produced a result for internal data; real world validated datasets are missing such as DrugBank updates or subsets of ChEMBL.

B. Paper 2 – "DrugMetric: Quantitative Drug-Likeness Scoring Based on Chemical Space Distance"

Benefits: Unique Quantitative Framework: Instead of determining drug-likeness based on binary classification, DrugMetric provides a range from 0-100 (continuous score), allowing a more specific metric when prioritizing compounds. Unsupervised Learning Novelty: DrugMetric uses VAE and GMM to project the molecules to latent chemical space, which reduces the reliance on labeled data. Strong Data Coverage: Uses a combination of four prominent datasets (CD, ChEMBL, ZINC, GDB) that ensure chemical diversity and intensity and do not bias representation. Openness and Transparency: The DrugMetric web platform is an open-source platform available to the public and can be reproduced.
Drawbacks: Computational Difficulties: Training VAEs and ensemble models in particular takes time and GPU resources and requires expertise with hyperparameter tuning. Limited Accuracy (AUC = 0.67): Overall, unsupervised clustering shows less classification power than supervised ML methods. Absence of Biochemical Context: DrugMetric only looks at the geometry of chemical space and does not include pharmacokinetic or toxicity metrics. Challenges with Interpretability: While investigating latent-space distances is mathematically intense, chemical meaning becomes naturalism less intuitive for medicinal chemists.

C. Paper 3-"Deep Graph–Sequence Network for Protein–Ligand Binding Affinity Prediction"

Pros: Multimodal architecture: Uses Graph Neural Networks for ligand inputs and Convolutional Neural Networks for protein inputs, allowing for efficient modeling of both structural and sequential specific effects. High predictive performance: Achieves high correlation, r > 0.85, and low RMSE of ~0.2-0.3, therefore outperforming classical docking-based methods in affinity prediction. Feature-level interpretability: Includes SHAP values and visualization of attention outputs to explain which molecular substructures account for binding strength. Large data set: Trained on >250K protein-ligand pairs, allowing for improved model generalization.
Cons: Data-hungry and demanding on resources: Requires extremely high computational power and large data sets, limiting feasibility in low-resource labs. Limited transferability: Model performance suffers in unseen protein families or non-kinase targets as training is based on a specific data set. Neglect of solvent and dynamics: Ignores molecular dynamics and solvation interactions that are known to impact real-world binding affinity. Black box complexity: Neural embeddings remain abstract despite attempts with SHAP explanations for medicinal chemists.

D. Paper 4 -"WideDTA: Prediction of Drug–Target Binding Affinity"

Advantages: Text-Based and Structure-Free Design: Predicts affinities using just sequence- and SMILES-based text representations, without 3D structural data. Comprehensive Feature Encoding: Combines four information sources of PS, PDM, LS, and LMCS, therefore capturing biochemical, structural, and substructural information. High Accuracy and Benchmark Validation: Outperforms traditional models KronRLS and SimBoost on Davis and KIBA datasets on the basis of higher CI and lower MSE. Speed and Scalability: The CNN architecture works well for biodata, and the model is easy to retrain or extend.
Disadvantages: Sequence Dependent: It only depicts the text representations without the benefit of true 3D spatial or conformational effects that are needed for a true affinity prediction. Limited Interpretability: CNNs will offer limited mechanistic insight as black-box models regarding why some residues or substructures are impactful in binding. Bias From Data: There is some dependency on the KIBA/Davis datasets or any dataset property used to train that potentially does not allow for generalizability to unrepresented classes of proteins. Static Representation: Does not capture dynamic conformational change happening during real binding events.

Table. 1. Performance Matrix of All Papers

Paper / Model	Domain	Key Metric	Best Reported Result	Remarks
Druggability of Pharmaceutical Compounds Using Lipinski Rules with ML (2024)	Drug-Likeness Prediction	Accuracy / ROC-AUC	Accuracy = 99.94%, ROC-AUC = 1.00	Extremely high accuracy using ensemble ML; interpretable and deployable via web app
DrugMetric: Quantitative Drug-Likeness Scoring Based on Chemical Space Distance (2024)	Drug-Likeness Scoring	AUC / Distance Hierarchy	AUC = 0.67, CD > ChEMBL > ZINC > GDB	Innovative unsupervised VAE–GMM model; introduces quantitative scoring
Deep Graph–Sequence Network for Protein–Ligand Binding Affinity (2023)	Binding Affinity Prediction	Pearson r / RMSE / CI	r ≈ 0.85, RMSE ≈ 0.25, CI ≈ 0.87	High predictive correlation using hybrid GNN–CNN; interpretable via SHAP
WideDTA (Öztürk et al., 2019)	Binding Affinity Prediction	CI / MSE	CI ≈ 0.88, MSE ≈ 0.20	Efficient text-based CNN model; strong benchmark results without structural data

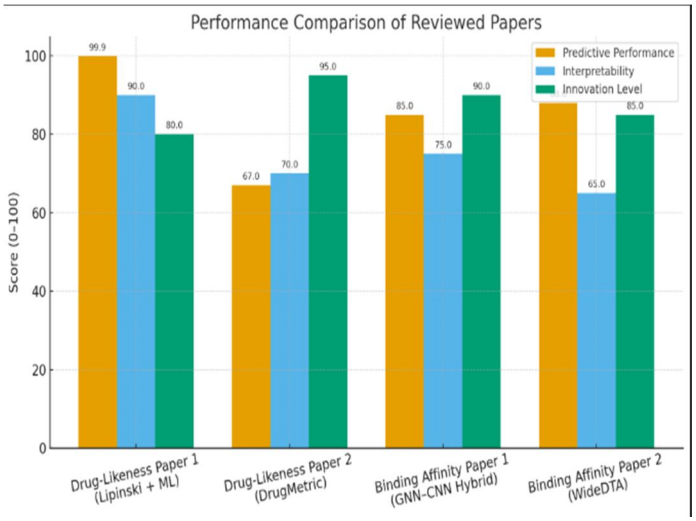


Fig. 8. Performance Comparison of Reviewed Papers

V. CONCLUSION

A collective review of the four studies illustrates the transformative role of AI in modern drug discovery, ranging from drug-likeness prediction to binding affinity estimation. Machine learning models, notably the ensemble framework based on Lipinski, achieved very high predictive accuracy and interpretability, allowing for early-stage compound screening. In contrast, deep learning models, like DrugMetric's latent-space model, the GNN–CNN hybrid model, and WideDTA, provide evidence of data-driven learning of representation exposing complex protein–ligand interaction. These methods will be used together to showcase a substantial shift from traditional rule-based screening to automated data-centric prediction systems capable of modeling the chemical and biological relationship with unmatched precision.

Additionally, the findings included a balanced compromise between performance and interpretability: classical ML models are interpretable, while deep architectures tend to better generalization and scalability. Inclusion of graph and sequence-based encoders further demonstrates rich molecular and biochemical representations as an important route to improving predictive depth by capturing true interaction dynamics.

Taken together, these four articles communicate a common idea: if we are hoping to balance both scientific insight and computational efficiency for drug discovery in the future, drug discovery will eventually require synthesizing interpretable, rule-based principles with high-capacity deep learning paradigms.layer for the subsequent trajectory models. The Grad- CAM++ framework (2021) provides the missing piece of the puzzle in interpretability, as it presents visual justifications that clarify the reasoning behind the projections of neural networks and simultaneously increase the trust of the user.

When these studies are compared, it becomes apparent that they have mutualistic capabilities: the object detectors are very accurate and provide a very good visual comprehension quickly but they cannot forecast very well; on the other hand, the trajectory models are very accurate in their movement predictions but they often lack transparency which is a disadvantage; and the explainability techniques such as Grad- CAM++ provide the understanding of the networks that's the reason why they expose the area of the networks' focus for the forecasts. Overall, the literature indicates that the future of self-driving vehicles will be based on the merging of highly accurate perception, contextually aware prediction, and transparent reasoning that is all done in one deep-learning pipeline. To conclude, the papers reviewed provide a clear way for the research intended to be carried out—the creation of an explainable trajectory-prediction framework that will not only be able to estimate the motion accurately but also effectively communicate the reasoning behind every prediction. This kind of combination is needed for making autonomous systems that technically dependable and valuable of human trust.

VI. FUTURE WORK

Looking forward, research can strive for a synthesis of drug-likeness and binding affinity prediction as an end-to-end AI pipeline to automate evaluation, prioritization, and optimization of compounds. Developing explainable and transparent deep learning frameworks will likely be key in bridging high predictive power and interpretability for domain experts to offer science validation for model outputs. Research also needs to motivate greater diversity in chemical representations through the contributions of physicochemical, graph, 3D conformational, and quantum descriptors in multi-modal learning frameworks.

Future AI models will also need to cross-dataset benchmark and validate against a standard annotation data system, as standard repositories (i.e., ChEMBL, DrugBank, etc.) to reinforce generalization. The goal will be the eventual advancement of scalable, interpretable, fully automated AI platforms that internally integrate drug-likeness scoring, binding affinity prediction, toxicity filtering, and simulated validation, which starts to characterize the era of intelligent data-driven drug discovery.to handle uncertainty in a better way, along with the use of transformer or attention-based networks coupled with visual- explanation tools like Grad-CAM++ for real-time generation of interpretable heatmaps.

REFERENCES

[1] Macal, C. M., & North, M. J. (2009). Agent-based modeling and simulation. Proceedings of the 2009 Winter Simulation Conference, 86–98. IEEE. https://doi.org/10.1109/WSC.2009.5429318

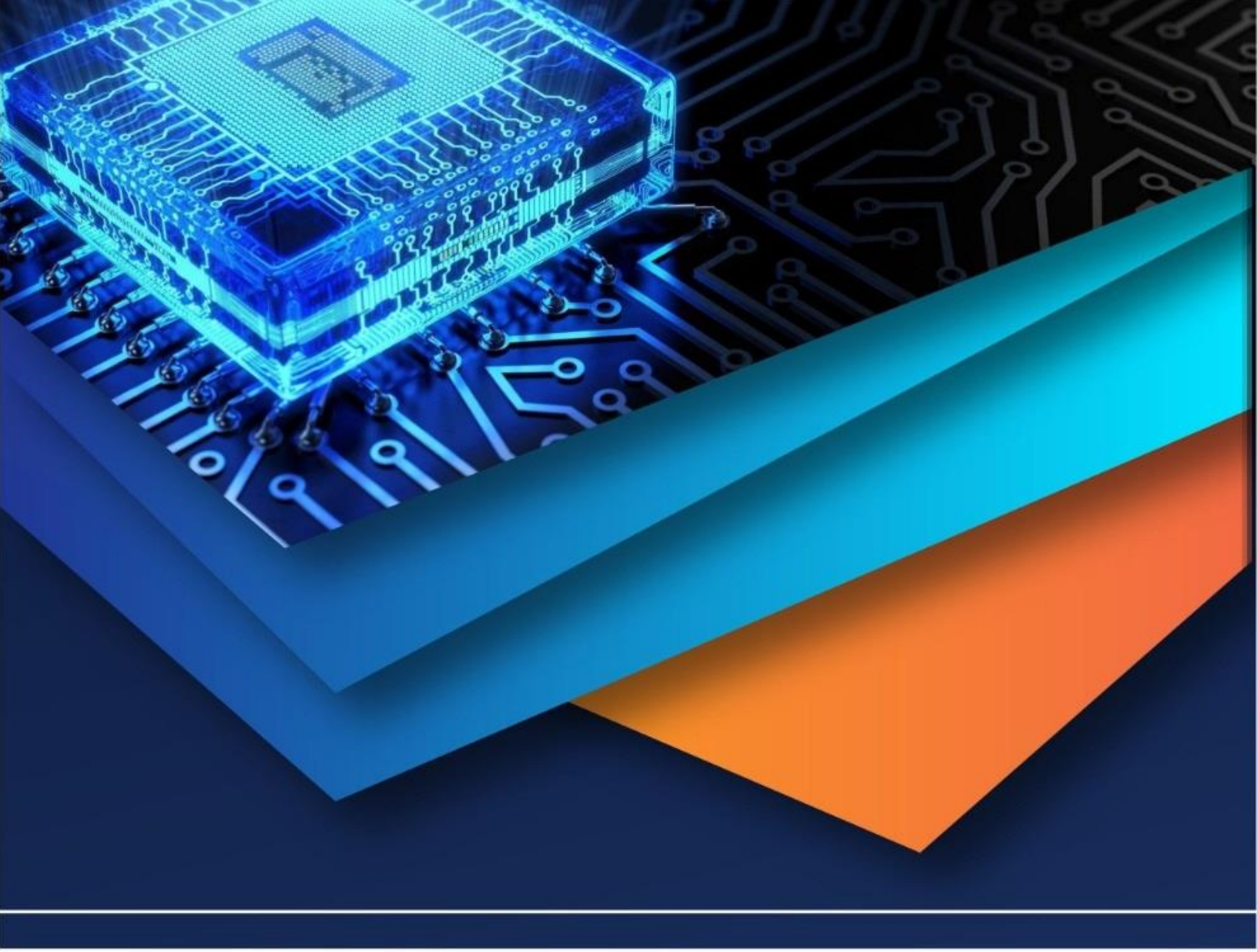
[2] An, G., Mi, Q., Dutta-Moscato, J., & Vodovotz, Y. (2009). Agent-based models in translational systems biology. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 1(2), 159–171. https://doi.org/10.1002/wsbm.45

[3] Vodovotz, Y., & An, G. (2019). Agent-based models of inflammation: A decade later. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 11(3), e1460. https://doi.org/10.1002/wsbm.1460

[4] Tong, X., Chen, J., Miao, H., Li, T., & Zhang, L. (2015). Development of an agent-based model (ABM) to simulate the immune system and integration of a regression method to estimate key ABM parameters by fitting experimental data. PLOS ONE, 10(11), e0141295. https://doi.org/10.1371/journal.pone.0141295

[5] Sivakumar, S., Mura, C., & Peirce, S. M. (2022). Integrating machine learning and agentbased modeling to study complex biological systems. Frontiers in Systems Biology, 2, 959665. https://doi.org/10.3389/fsysb.2022.959665

[6] Öztürk, H., Özgür, A., & Ozkirimli, E. (2019). DeepDTA: Deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829. tps://doi.org/10.1093/bioinformatics/bty593
[7] Öztürk, H., Özgür, A., & Ozkirimli, E. (2020). WideDTA: Prediction of drug–target binding affinity using word-based deep learning. Bioinformatics, 36(17), 4627–4635. https://doi.org/10.1093/bioinformatics/btaa626
[8] Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., & Zhang, P. (2018). Interpretable deep neural networks for predicting drug–target interactions. Frontiers in Pharmacology, 9, 1426. https://doi.org/10.3389/fphar.2018.01426
[9] Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., & De Fabritiis, G. (2018). DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics, 33(19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350
[10] Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: Interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics, 35(18), 3329–3338. https://doi.org/10.1093/bioinformatics/btz111
[11] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A webaccessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Research, 35(Database issue), D198–D201. https://doi.org/10.1093/nar/gkl999 74
[12] Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. Nature Reviews Drug Discovery, 1(9), 727–730. https://doi.org/10.1038/nrd892
[13] Lipinski, C. A. (2004). Lead- and drug-like compounds: The rule-of-five revolution. Drug Discovery Today: Technologies, 1(4), 337–341.
[14] Walters, W. P., & Namchuk, M. (2003). Designing screens: How to make your hits a hit. Nature Reviews Drug Discovery, 2(4), 259–266. https://doi.org/10.1038/nrd1063
[15] Vamathevan, J., et al. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5
[16] Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., & Luo, X. (2020). TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics, 36(16), 4406–4414. https://doi.org/10.1093/bioinformatics/btaa524
[17] Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Dogan, T. (2020). Deep learning and machine intelligence in drug discovery: A comprehensive survey. Artificial Intelligence Review, 53, 295–347. https://doi.org/10.1007/s10462-018-09605-9
[18] Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301
[19] Chou, C. H., & Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Mathematical Biosciences, 219(2), 57–83. https://doi.org/10.1016/j.mbs.2009.03.004
[20] 20. Xiang, Y., & Gong, Y. (2021). Hybrid agent-based and machine learning approaches for biological system modeling: Challenges and opportunities. Computational and Structural Biotechnolog.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)