e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

Predicting Bioactivity of Molecules by using Machine Learning Models

V L Naga Sreya Chitiprolu, Kanaka Durga Devi Nelluri*
KVSR Siddhartha College of Pharmaceutical Sciences, Vijayawada, Andhra Pradesh, India
Corresponding author: nelluriss@rediffmail.com

Received: 07 July 2025 | **Revised & Accepted:** 26 October 2025

Keywords: Machine learning (ML), Bioactive compounds, Drug discovery, Virtual screening, Cheminformatics

Abstract: Machine Learning has transformed the process of drug discovery and development by predicting bioactive compounds. ML models have been driven by advanced algorithms, that analyze vast chemical datasets to identify potential bioactive molecules with precision and efficiency. These models leverage molecular structures, physicochemical properties, and biological activity to predict a compound´s pharmacological potential. Techniques like regression, classification and deep learning enhance the predictive power by identifying patterns that traditional methods often overlook. Furthermore, ML accelerates the virtual screening of compound libraries, reducing time and cost in experimental validation. Integrating cheminformatics and ML ensures reliable predictions, aiding researchers in prioritizing lead compounds for further testing. As this field evolves, combining ML with tools like molecular docking and quantum computing holds promising for even more accurate predictions. This synergy fosters innovation, making ML a transformative force in identifying bioactive compounds critical to therapeutic advancements.

1. Introduction:

Machine learning (ML) has become a transformative approach in drug discovery and development, especially in the prediction of bioactive compounds. Medicinal plants, known for their rich diversity of natural compounds, have long served as a valuable source of novel therapeutic agents. These plant-derived bioactives offer significant potential in treating a wide range of diseases. By combining ML techniques with phytochemical information, researchers can efficiently identify and prioritize promising compounds, accelerating the development of plant-based drugs and enhancing the overall drug discovery process[1]. Bioactive compounds are naturally occurring chemical substances present in minute quantities in products derived from plants and animals. These compounds are integrated into dietary supplements and hold significant potential as starting points for pharmaceutical development[2,3]. They serve as key molecular frameworks that interact with various proteins in biological systems, making them valuable as foundational elements for numerous therapeutic agents [4,5]. Many existing drugs share structurally similarities with these natural bioactive compounds [6]. These compounds play a critical role in regulating cellular signaling and maintaining homeostasis, underscoring the importance of exploring their medicinal applications [7]. The concept of chemical space, a virtual domain encompassing over 10^{60} theoretically possible small organic

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

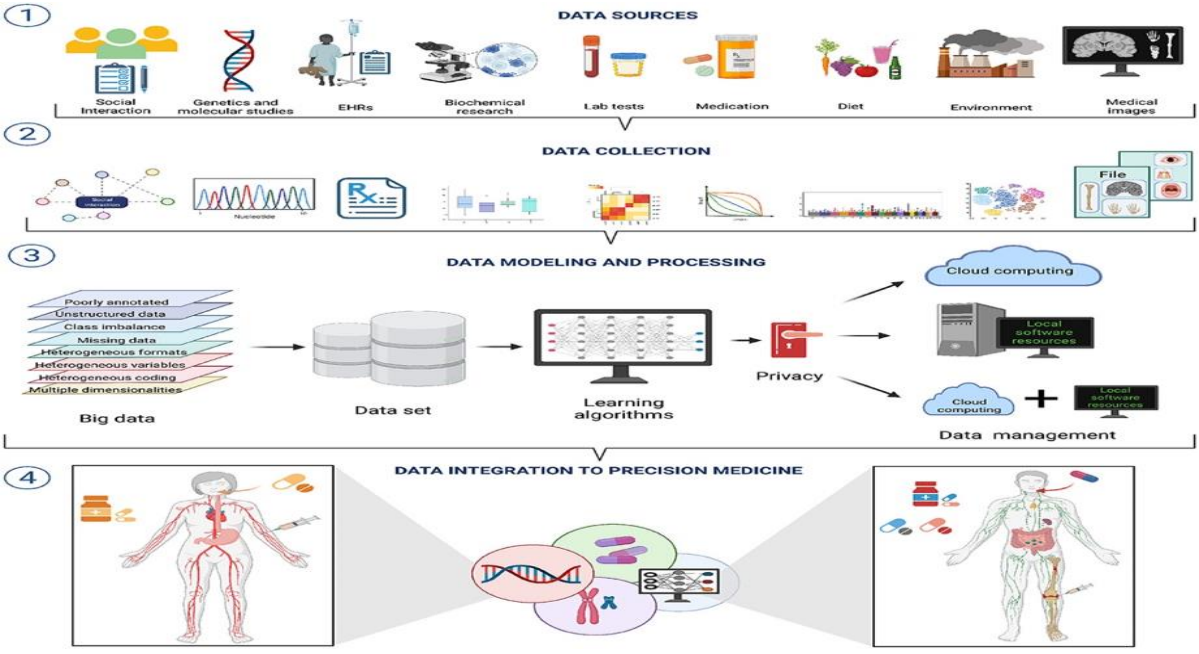
Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

molecules, has become a focal point in drug discovery. Advances in machine learning (ML), particularly deep learning (DL), supported by powerful hardware like graphics processing units (GPUs) and tensor processing units (TPUs), have revolutionized the exploration of chemical supspaces with desirable properties. ML algorithms are now widely utilized in various stages of drug development, including HIT identification, lead optimization and clinical development, with the aim of streamlining the discovery process and accelerating the creation of new therapeutics [8,9,10].

1.1 Applying Cheminformatics and Bioinformatics Insights to optimize Machine Learning in Drug Discovery:

Cheminformatics is a multidisciplinary field that leverages computational tools to analyze chemical information, including the 2D and 3D structures of molecules. Proteochemometric (PMC) modelling is a key approach in this domain, used to predict the bioactivity of compounds and their target proteins by Simplified Molecular-Input Line- Entry System (SMILES) and amino acid sequence data, without relying on 3D structures or dynamic properties [11]. SMILES strings can be stored in digital databases and can be utilized through a variety of computational algorithms to identify, assess and predict chemical properties, facilitating their integration into drug discovery pipelines to accelerate development. In the drug discovery process, initial hits- molecules exhibiting desired activity in experimental screening are identified and optimized into lead compounds through computer-aided drug design (CADD). A prominent example of CADD in the development of commercial drugs like Zanamivir (Relenza), it is a neuraminidase inhibitor which is approved by United States Food and Drug Administration (FDA) in 1999 for the prophylactic treatment of influenza symptoms [12]. For the identification of new bioactive compounds for specific target proteins, it requires chemical and molecular information about the target protein, such as amino acid sequence and 3D structural information is essential.



© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

FIGURE 1. Visual representation of workflow of machine learning modelling in precision medicine.

The data process involves:

1. **Data Sources:** Utilizing diverse datasets like DNA sequences, protein sequences, gene expression profiles, molecular descriptors, SMILE strings or fingerprints.
2. **Data Pre-processing:** Pre- processing of data is required to scale the data for machine learning algorithms.
3. **Data Integration:** Integration of heterogenous data is done by combining chemical, biological and clinical data by using methods to combine data.
4. **Precision Medicine:** Leveraging integrated data to improve targeted therapeutic outcomes.

While cheminformatics and bioinformatics are interchangeable as these two fields deal with computer- readable data from small chemical compounds, metabolites, RNAs, DNAs and proteins. Cheminformatics depends on chemical data, while bioinformatics deals with DNA, RNA and protein information. Several publicly accessible databases provide valuable resources, including the Nucleotide Database maintained by the National Center for Biotechnology Information (NCBI)[13] (https://www.ncbi.nlm.nih.gov/nucleotide/; accessed on 10 December 2021), the UniProt database (**https://www.uniprot.org/** (accessed on 10 December 2021)) [14], and the Protein Data Bank (PDB) database (**https://www.rcsb.org/** (accessed on 10 December 2021)) [15]. Additionally, numerous web-based databases for bioactive compounds have been described in a detailed literature reviews [16], provide readily available molecular sequence information to support research and development.

1.2 Mapping Chemical Space to Discover Unidentified Bioactive Compounds Using Machine learning Models

Chemical space encompasses a vast, multidimensional framework that includes all possible small carbon-based molecules, each defined by a unique set of characteristics such as molecular structure, physiochemical properties and biological activities.[17]. This concept is fundamental to drug discovery as it provides a structured approach for exploring the connections between molecular structures and their potential biological effects. Traditional compounds screening methods often depend on databases of known bioactive molecules, which restricts exploration to familiar compounds and limits the discovery of novel entities. Machine learning is a transformative tool for mapping chemical space, enabling the discovery of unrecognized bioactive compounds that are not previously identified through conventional screening methods. Various algorithms are present in computational and medicinal chemistry which allows the numerical calculation of chemical properties. The key metrics involved are Quantitative Estimate of Drug-likeness (QED) [18], assess a compound´s potential as a drug, while absorption, distribution, metabolism, elimination and toxicity (ADMET) properties which provide critical insights into a molecule´s pharmacokinetic and safety profile [19,20,21,22,23,24,25,26]. Proper utilization of these quantitative metrics, researchers can navigate targeted regions of chemical space, particularly those resembling natural products, to uncover novel hits for specific protein targets.

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

2. Material and methods:

2.1. Data Set Collection–

To predict the bioactivity of chemical compounds, a dataset of bioactive molecules should be selected from publicly available chemical databases such as CHEMBL, PUBCHEM, BindingDB. The dataset should consist of experimental bioactivity data, with molecules for specific biological targets. This bioactivity can be represented as continuous values (Like IC50, EC50) or categorical labels (Like active or inactive), depending on the task like regression or classification.

2.1.1Data Collection: Molecules with reliable and experimental bioactivity values are selected. The data should be curated to remove duplicates, ensure uniformity in bioactivity measurements.

2.1.2. Data Preprocessing: The dataset should be preprocessed by standard bioactivity measurements and normalizing molecular features. If there is any missing value, any extreme outliners should be handled via removal or capping method

.

2.2. Molecular Representation-

In machine learning to process chemical information, compounds are represented as numerical vectors, graphs. They are some methods used in machine learning for molecular representation like:

2.2.1. Simplified Molecular Input Line Entry System (SMILES): Compounds are represented by their SMILES strings, which concise textual representation of molecular structures.

2.2.2. Molecular descriptors: There are several molecular descriptors like LogP, TPSA, Polar surface area which can be computed using cheminformatics tools like RDkit. They are widely used descriptors to capture structural and chemical properties which influence bioactivity [27].

2.2.3. Fingerprints: A part from descriptors fingerprints like MACCS keys and Extended Connectivity Finger prints (ECFP) can be considered as potential feature vectors for representation of chemical structure of the molecules [28].

2.2.4. Graph-based Representation: It offers a distinct method for depicting molecules, where atoms are represented as nodes and chemical bonds as edges. Models such as Graph Neural Networks (GNNs) have demonstrated considerable promise in capturing complex structural patterns within molecular frameworks, enabling effective modelling of molecular structures.

2.3 Machine Learning Models-

There are several machine learning algorithms which can be applied to predict bioactivity of the compounds based on molecular features. They include both traditional machine learning models and deep learning tactics:

2.3.1. Traditional Machine Learning Models:

- Random Forest (RF): This method builds multiple decision trees and combines their predictions. It is often due to its robustness and ability to handle large number of features [29].
- Support Vector Machines (SVM): A labelled learning algorithm that can be used for both active or inactive compounds as well as for regression models for prediction of bioactivity value. They are effective in high-dimensional feature spaces [30].
- Gradient Boosting Machines (GBM): In this model like XGBoost has been applied to improve predictive accuracy by combining weak learners (trees) into a stronger learner [31].

2.3.2. Deep Learning Models:

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

- Feedforward Neural Networks (DNN): Deep neural networks, with fully connected layers, can be used for predicting continuous bioactivity measurements or classifications on chemical descriptors or fingerprints [32].
- Graph Neural Networks (GNN): In case of molecular graph representations, GNNs is applied directly to predict bioactivity from graph-based inputs, capturing more complex relationships between molecular structure and activity [33].

2.4 Feature Selection and Dimensionality Reduction-

In machine learning there is high dimensionality of the data especially with the molecular descriptors and fingerprints, feature selection and dimensionality reduction techniques should be used to enhance model interpretability:

- Principal Component Analysis (PCA): It is a method to reduce dimensionality of descriptor data while retaining the variance, which is useful in models predicting continuous bioactivity values.
- Correlation- based Feature Selection: This was proposed to identify the most relevant molecular features for predicting bioactivity, reducing redundancy and focusing on the most informative descriptors.

2.5 Model Training and Optimization:

In a significant approach, machine learning models are trained on training set and evaluated on a separate test to assess their predictive performance. Cross- validation techniques such as K- fold cross- validation to ensure the model generalizes well and does not overfit to the training data.

- Hyperparameter Tuning: There are optimization methods such as grid search or random search would be used to identify the combination of hyperparameters such as number of trees in random forest, kernel type in support vector machines or learning rate in neural networks to maximize model performance [34].

2.6 Model Evaluation:

To analyse the predictive power of the models, various performance metrics are utilized, tailored to the specific type of p rediction task, such as regression or classification:

- Classification Tasks:

It depends on –

a) Accuracy: It is nothing but the overall percentage of correct predictions.
b) Precision, Recall, F1-Score: They are used for evaluating performance in binary classification tasks like predicting active vs inactive.
c) AUC-ROC: This area under the receiver operating characteristic curve is to assess the ability of the model to distinguish between positive and negative classes [35].

- Regression Tasks:

It depends on –

a) Mean Absolute Error (MAE): It measures the average magnitude of errors between predicted and actual bioactivity values.
b) R-Squared (R^2): In this the proportion of variance in the target variable that is predictable from input features [36].

2.7 Software and Tools:

For the Machine Learning models to be implemented they would have involved the use of several software packages and libraries:

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

- Python: It is the primary programming language used for data processing, feature extraction and model training.
- Cheminformatics Tools: RDKit or Open Babel is used for feature extraction that is descriptors, fingerprints and SMILES parsing, molecular property calculation.
- Machine learning Libraries:

a) Scikit- Learn – It is used for traditional machine learning models like Random Forest, Support Vector Machines, XGBoost.
b) TensorFlow or PyTorch- These are used for deep learning models like Neural Networks, Graph Neural Networks.

2.8 Validation and Testing:

Predictive models are validated using independent test set, with performance metrics calculated to evaluate the ability of the models. When external datasets are available then validation is performed to determine the model´s robustness across diverse chemical spaces.

3. Conclusion:

Machine Learning models are a transformative tool in modern pharmaceutical research, insisting new avenues for the discovery of bioactive compounds and accelerating the drug development towards safer, effective ones. This field provides integration of ML models with experimental biology, cheminformatics, and systems pharmacology to enhance precision of drug discovery for better therapeutic advancements.

ACKNOWLEDGEMENT:

The authors sincerely express their gratitude to the Management and Principal of KVSR Siddhartha College of Pharmaceutical Sciences, Vijayawada, for their unwavering support and encouragement.

References:

1. Hanson, J. R. (2003). *Natural products: the secondary metabolites*. London: Royal Society of Chemistry.
2. Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. 2016;79(4):629-661.
3. Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, Dzobo K. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. Int J Mol Sci. 2018;19(5):1578.
4. Atanasov AG, Waltenberger B, Pferschy-Wenzig E-M, Linder T, Wawrosch C, Uhrin P, Temml V, Wang L, Schwaiger S, Heiss EH, et al. Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv. 2015;33(8):1582-1614.
5. Waltenberger B, Mocan A, Šmejkal K, Heiss EH, Atanasov AG. Natural products to counteract the epidemic of cardiovascular and metabolic disorders. Molecules. 2016;21(7):807. MacDougall C, Chambers HF. Aminoglycosides. In: Brunton LL, Chabner BA, Knollman BJ, eds. Goodman & Gilman's the pharmacological basis of therapeutics. 12th ed. New York: McGraw-Hill; 2011. p. 1505-20.
6. Hu Y, Lounkine E, Bajorath J. Many approved drugs have bioactive analogs with different target annotations. AAPS J. 2014;16(5):847-859.
7. Pellavio G, Rui M, Caliogna L, Martino E, Gastaldi G, Collina S, et al.

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

Regulation of aquaporin functional properties mediated by the antioxidant effects of natural compounds. IJMS. 2017;18(12):2665. doi:10.3390/ijms18122665.

8. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93.
9. Mak KK, Balijepalli MK, Pichika MR. Success stories of AI in drug discovery—Where do things stand? Expert Opin Drug Discov. 2022;17(1):79-92.
10. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: Recent advances and future perspectives. Expert Opin Drug Discov. 2021;16(9):949-959.
11. Cortés-Ciriano I, Ain QU, Subramanian V, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. Medchemcomm. 2015;6(1):24-50.
12. Woods JM, Bethell RC, Coates JA, et al. 4-Guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid is a highly effective inhibitor both of the sialidase (neuraminidase) and of growth of a wide range of influenza A and B viruses in vitro. Antimicrob Agents Chemother. 1993;37:1473-1479.
13. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;46(D8-D13).
14. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al. UniProt: The Universal Protein Knowledgebase. Nucleic Acids Res. 2018;46:2699.
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.
16. Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12:20.
17. Dobson CM. Chemical space and biology. Nature. 2004;432:824–828.p 23].
18. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012;4:90–8.
19. Durán-Iturbide NA, Díaz-Eufracio BI, Medina-Franco JL. In silico ADME/tox profiling of natural products: A focus on BIOFACQUIM. ACS Omega. 2020;5:16076–84.
20. Dong J, Wang NN, Yao ZJ, Zhang L, Cheng Y, Ouyang D, Lu AP, Cao DS. ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. J Cheminform. 2018;10:29.
21. Bocci G, Carosati E, Vayer P, Arrault A, Lozano S, Cruciani G. ADME-space: A new tool for medicinal chemists to explore ADME properties. Sci Rep. 2017;7:6359.
22. Banerjee P, Eckert AO, Schrey AK, Preissner R. ProTox-II: A webserver for the prediction of toxicity of chemicals. Nucleic Acids Res. 2018;46:W257–63.
23. Van de Waterbeemd H, Gifford E. ADMET in silico modelling: Towards prediction paradise? Nat Rev Drug Discov. 2003;2:192–204.
24. Feinberg EN, Joshi E, Pande VS, Cheng AC. Improvement in ADMET

© BTS Pharma Empire All Rights Reserved

e-ISSN: 3108-3218

BTS International Advanced Pharmaceutical Sciences Journal

Vol 01 | Issue 05 | May 2026 Journal homepage: www.btsempire.com

ARTICLE ID: BTS0526001

prediction with multitask deep featurization. J Med Chem. 2020;63:8835–48.
25. Daina A, Michielin O, Zoete V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7:42717.
26. Cáceres EL, Tudor M, Cheng AC. Deep learning approaches in predicting ADMET properties. Future Med Chem. 2020;12:1995–9.
27. O'Boyle NM, Banck M, James CA, et al. Open Babel: An open, collaborative approach to chemical informatics. J Cheminform. 2011;3(1):33.
28. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. J Chem Doc. 1965;5(2):107-113.
29. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
30. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
31. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-94.
32. Gilmer J, Schoenholz SS, Riley PF, et al. Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6-11; Sydney, Australia. JMLR Workshop and Conference Proceedings; 2017. p. 1263-72.
33. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations (ICLR); 2017 Apr 24-26; Toulon, France. 2017.
34. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427-437.
35. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
36. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281-305.

© BTS Pharma Empire All Rights Reserved