IRJSE International Research Journal of Science & Engineering
Int. Res. J. of Science & Engineering, Special Issue A14, June, 2024
ISSN: 2322-0015
SJIF Impact Factor 6.70
https://doi.org/10.5281/zenodo.12699237
RESEARCH ARTICLE OPEN ACCESS

Machine / Deep Learning Technique for Faster Analyzing Molecular Data and Predicting Bioactivity Profiles for Drug Discovery Process

Sahebrao B. Pawar[1], Nilesh Kailasrao Deshmukh[2], Shivling Shankarrao Patil[3], Sunil Madhavrao Lomte[4], and Nayak Sunil Kashibarao[5]

[1,2,3,4,5]School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded,
Email: [1]Shpawar04@gamil.com | [2]nileshkd.srt@gmail.com

Manuscript Details

Available online on https://www.irjse.in
ISSN: 2322-0015

Editor: Dr. Arvind Chavhan

Cite this article as:
Sahebrao B. Pawar, Nilesh Kailasrao Deshmukh, Shivling Shankarrao Patil, Sunil Madhavrao Lomte and Nayak Sunil Kashibarao. Machine / Deep Learning Technique for Faster Analyzing Molecular Data and Predicting Bioactivity Profiles for Drug Discovery Process, *Int. Res. Journal of Science & Engineering*, 2024, Special Issue A14: 39-48.
https://doi.org/10.5281/zenodo.12699237

Article published in Special issue of National Conference on Machine Learning and Data Science (NCMLDS-2024)" organized by College of Computer Science and Information Technology (COCSIT) Ambajogai Road, Latur, Maharashtra, India on date April 16th to 17th 2024

Abstract

The field of drug discovery stands at the cusp of a revolutionary transformation motivated by the combination of deep learning (DL) and machine learning (ML) methodologies. Conventional drug discovery procedures are frequently costly, laborious, risky, and have a high failure rate. However, the advent of ML and DL methodologies offers unprecedented opportunities to expedite the discovery of novel therapeutic compounds, enhance target identification, optimize lead compounds, and streamline preclinical and clinical trials. This paper gives a comprehensive overview of the ML and DL techniques in various stages of the drug discovery pipeline, highlighting their potential to revolutionize the pharmaceutical industry and improve patient outcomes.

Keywords: Machine Learning, Deep Learning method, Drug Discovery Process, Computational biology

1. Introduction

Drug discovery is a long and labor-intensive complex process that involves finding Active Chemical Molecule with therapeutic potential, testing their efficacy and safety, and optimizing their properties for clinical use. Traditional methods rely heavily on experimental approaches, which are resource intensive and time-consuming. ML and DL techniques offer alternative strategies to accelerate drug discovery by leveraging computational methods for analyzing big datasets and predicting molecular properties with unprecedented accuracy. Several phases of the drug development process have seen the

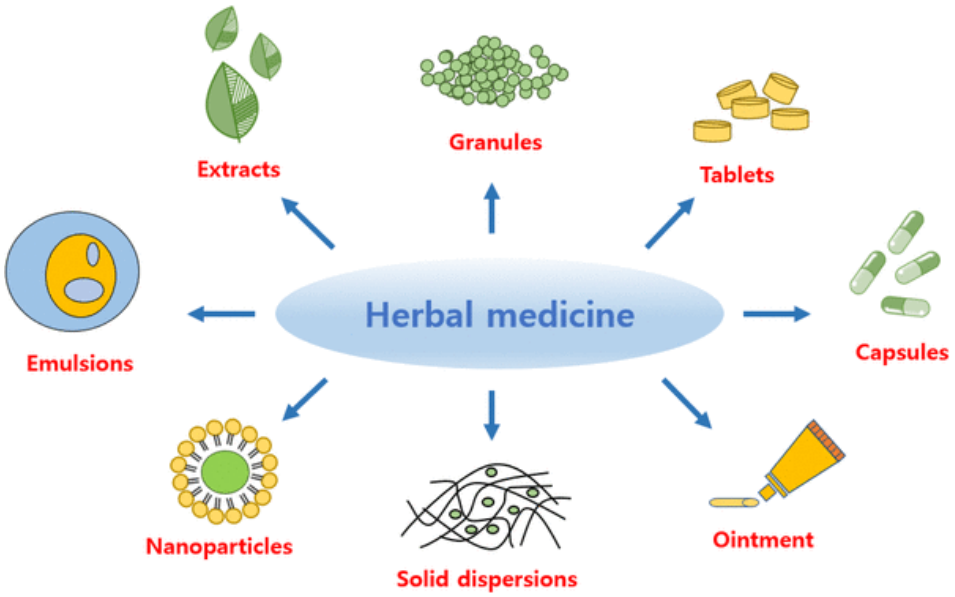

Figure 1. Traditional Drug Finding Method

role of ML/DL techniques, encompassing lead optimization, compound assessment, and target selection. These algorithms can analyze biological data, including protein structures and gene expression levels, to determine possible therapeutic targets and estimate the effectiveness of candidate compounds. Moreover, ML models can prioritize compounds for experimental validation based on their predicted properties, thereby reducing the quantity of substances that must be synthesized and tested in the laboratory[1]. Deep learning, a subset of ML, has become an effective instrument for deciphering intricate biological data and producing precise forecasts. Deep neural networks (DNNs) can learn intricate patterns and representations from high-dimensional data, such as genomic sequences and chemical structures, enabling them to generate novel insights into drug-target interactions and identify new drug candidates. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have demonstrated remarkable efficacy in bioactivity profile prediction and molecular data analysis. For some molecular targets and patient subgroups, the process of finding new drugs can be extremely complex due to the stringent requirements set forth by regulatory bodies and the procedure itself. The process of finding and developing new medications is still exceedingly laborious and costly in today's world. Conventional Drug discovery need typically takes 10 to 15 years for finding new drug with testing and different trial as well.

However, innovative in silico screening approaches for large drug libraries have emerged in the last 10 years due to the development of Information and Communication Technologies and the rise in computer capacity available. This stage before preclinical research lowers the financial burden and expands the search area. Phases of novel drug discovery within the framework of precision medicine are depicted in Figure 2 Given this, machine learning (ML) techniques have become increasingly popular in the pharmaceutical sector, where they can be used to speed up and calculate automatically the processing of the vast amounts of data that are now available. A subfield of artificial intelligence (AI) called (ML/DL) Learning seeks to create and implement computer system that can learn from unprocessed, raw data in order to carry out particular tasks in the future. Within a big data set, the primary functions of the AI algorithms are pattern recognition, grouping, regression, and classification. The pharmaceutical sector has employed a wide range of ML/DL techniques to anticipate novel chemical features, biological activity, interactions, and side effects

of pharmaceuticals. Naive Bayes, SVM, Random Forests, and, more recently, DNL technique (Deep Neural Learning Method) [2]. This work has been conceived and produced to examine the state of the art in this sector. It compiles the most pertinent papers from the previous five years about the role of machine learning methods to early drug development. The works that were found for this study are then divided into various areas, with a focus on examining the ML/DL method, the biological issue that needs to be resolved, and the descriptors that were employed[3].

2. Typical Machine Learning Drug Discovery Process

In the subject of computational intelligence, and particularly in When it comes to machine learning, the experimental phase's design is crucial. It is crucial to initially specify the methodology that will be used for this. Figure 2. Even though several steps in the experimental design are shared by many research domains, the application of ML approach needs to be transversal [4]. More specifically, we may distinguish between the subsequent processes in the ML/DL Technique used in drug discovery: 1) gathering data; 2) creating fingerprint descriptors; 3) identifying the optimal selection of variables; 4) training the model; and 5) validating the model. Figure 2 shows a schematic of the machine learning technology that is frequently applied to drug development (Figure.3). Getting the data set is the initial stage, and it needs to have a few requirements. It must possess physical-chemical properties that aid in absorption, selectivity, and low toxicity in addition to qualities that make it easy to develop and configured in the laboratory. This is since big proteins or exceptionally complicated compounds are not used in the pharmaceutical sector. It typically interacts with peptides and small molecules as its primary targets. The SMILES formats reflect the molecular composition and organizing of peptides and small molecules, as well as facilitate the handling and analysis of these substances. A lot of valuable queried the dataset required for drug discovery process is currently stored in several public repositories 1, including DrugBank[5], PubChem[6], ChEMBL [7], and ZINC[7]. Figure 2). Another crucial aspect is the labeling of the various chemicals (see target in Fig. 2). While some machine learning models can be used without labeling, Drug development is a field that regularly uses supervised learning algorithms. In this case, the success of the experiment will depend on the categorization definitions established by the researchers. The process of creating mathematical descriptors results in a set of data that ML/DL model can process. This dataset is divided into Test and Training dataset ML/DL Process (shown in Fig. 2) is used for training the model, and the smaller subset (illustrated in Fig. 2) is used for testing the model.

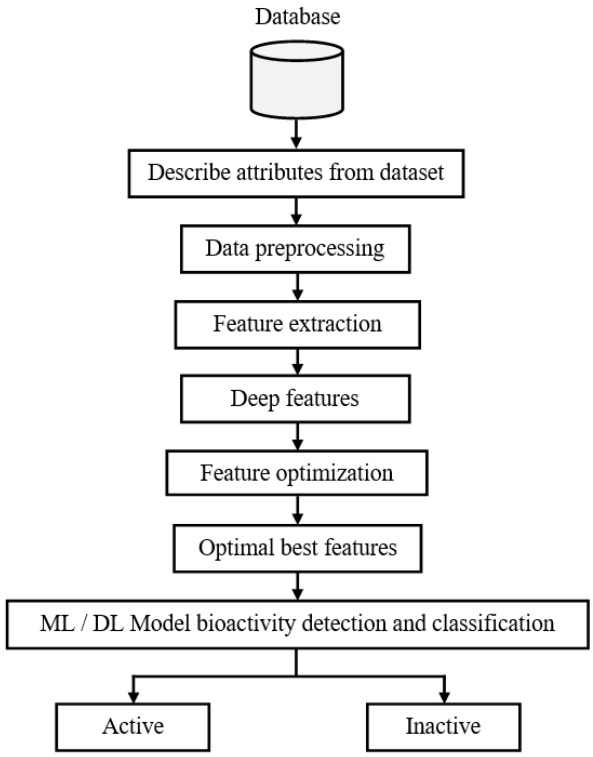

Figure 2. ML / DL Architecture for Drug Discovery Process

The ideal subset of the training set's variables is identified using the pertinent and necessary data. Creating mathematical descriptors usually requires a large number of numerical variables. Reducing the amount of unnecessary or redundant variables is the primary goal of this technique. There are various methods for achieving this, including PCA, t-SNE, FS,

Table 1. Common online database used in ML / DL Learning model training.

Database	Molecule	Uses Type	URL	Ref.
ChEMBL	2.4M	Drug Discovery	https://www.ebi.ac.uk/chembl/	[5]
DrugBank	14k	Drug Discovery	https://go.drugbank.com/	[6]
PubChem	118M	Drug Discovery	https://pubchem.ncbi.nlm.nih.gov/	[7]
ZINC	750M	High Virtual Screening	https://zinc.docking.org/	[8]

Autoencoder, etc. The variable content is preserved by FS techniques by obtaining a subset of features from the original set. Most scientists adopt these approaches in their experimental designs because they provide an explanation that is understandable from a biological perspective [8].

The model gets trained after the best set of parameters has been identified. The algorithms and their parameters need to be chosen first. To make sure they are suitable for the issue at hand as well as the quantity and kind of data accessible, these must be carefully selected. In these situations, the application of strategies like cross-validation (CV) is typical. The CV enables assessing the model's performance, assessing the model's performance using unidentified data and tracking the level of generalization as the model is being trained. For every execution of the experiment, the original data set is divided into two categories: the training set and the validation set. Fig. 2 illustrates the progression of the CV technique over ten runs. The goal of the CV process is to identify the best possible set of parameters for each method. These parameters define the performance of any model. The optimal model is the one that achieves the highest performance value at the lowest total cost.

Ultimately, a final validation procedure is conducted on the Top Algorithm that emerged from the CV process, and the A test set is retrieved that was extracted from the original set. A novel predictive drug model might be declared created if the validation results show statistical significance[9]. Numerous fields have employed machine learning techniques, and especially in The past several years have seen an increase in the quantity of papers published in this area. On open access websites, Still, not a lot of articles on machine learning pertaining to medication development. Works like [9][10] facilitate this information by providing a summary of machine learning techniques and the current status of drug discovery applications in academia and industry settings.

3. ML/DL Predication

The description of the molecules by descriptors that can capture their properties and structural traits is a crucial stage in the model's training process. Numerous molecular descriptors, ranging from basic molecule attributes to intricate, three-dimensional molecular fingerprint formulations, have been documented in the literature. Descriptor having 1000-bit sting for ML/DL as input for predication.

3.1. QSAR Model

According to the tenets that "A molecules structure determines its biological activity" and "Molecularly similar molecules have similar biological activities," quantitative structure activity relationship (QSAR) models, which quantitatively link molecules chemical structures with their biological activities, enable the prediction of a new compounds physicochemical and biological fate properties using mathematical systems based on the compounds known chemical structure and prior experimental research. Significance Valid Data set for Machine learning Predication QSAR models combine computer and statistical techniques to produce a hypothetical forecast of biological activity that permits the theoretical development of possible new medications in the future without having to go through the organic synthesis method of error-by-trials. It permits the elimination of some resources, including tools, materials, personnel, and equipment because it is

a science that only exists virtually. By concentrating on the connections between molecular structure and biological action, novel drug candidates can be designed considerably more quickly and affordably. When sufficient experimental data and facilities are unavailable, one of the best ways to accomplish compound prediction is through modeling studies like QSAR [11]. Three different kinds of information are required in order to conduct a QSAR study [12]

1. The molecular makeup of several substances sharing a same mode of action.
2. Information about each ligands biological action that is part of the study.
3. Physicochemical qualities, derived from the molecular structure that is computationally created and characterized by a collection of numerical variables Significance.

Valid Data set for Machine learning Predication Predicting the biological activity of synthetic substances that are virtually generated in a short amount of time is made possible by the results of the QSAR model or equation in the prospective type; however, these compounds must share structural characteristics with the ligands included in the study in order to stay within the parameters, biological structure, or desired values of the descriptors. The other kind, known as a retrospective, examines molecules that have already been created (such as synthesis and bioassays). To comprehend the subtle relationships between biological processes and structures. Getting the input data ready is the most important stage because the outcome is automated and solely dependent on the input Because of its interdisciplinary nature, the QSAR methodology draws information from the fields of pharmacology and organic chemistry. This scenario, which is the goal in order to, is rewarded by The produced calculations have shown a high likelihood of pharmacological efficacy since, as previously noted, they provide a forecast of the biological activity. QSAR through the directed creation of compounds that do not yet exist. A statistical method for analyzing data obtained from lab or published sources is multiple linear regression. It takes the estimated descriptors as an independent variable and the biological activity levels of ligands as a dependent variable. The duration of a chemical simulation performed using computational tools is far shorter than what would be required to synthesize and test novel chemicals in bioassays, which may be weeks, months, or even years. This benefit makes it possible to take a number of molecules and, because of the speed at which the findings are obtained, immediately feed the synthesis lab in the project's ongoing process. As a result, QSAR predicts previously undiscovered structures and suggests that organic chemists take them to bioassays, the outcomes of which either support or refute the values indicated by the QSAR mode. If everything goes according to plan, this operational cycle will produce better prospects than purely trial and error. This helps people who create novel treatments succeed by saving time, money, and resources. The benefits of QSAR include its low cost due to its lack of need for chemical reagents or laboratory equipment, as well as the availability of free software for model creation that offers user-friendly interfaces for handling and designing. Furthermore, the descriptor calculations and molecule synthesis can be completed quite quickly[12].

3.2. Molecular Descriptors

Descriptors of molecules in numerous scientific fields, or MDs, are essential input. They are characterized as numerical depictions of the molecule whose physicochemical information is quantitatively described. However, only a portion of a molecule's information may be obtained through experimental measures. In order to establish QSAR and properties, biological-activities, and other experimental data, there has been an increasing focus in recent decades on how to theoretically take hold of and translate the data contained in the chemical structure into one or more values. Because they may locate molecules with comparable physical qualities based on how close they are to the values of the computed descriptors, MDs have thus emerged as a highly helpful tool for doing similarity searches in molecular repositories. Molecular descriptors, which have been defined since the beginning of their application, have encoded [13] molecules in various ways. One type of description that they can offer is a one-dimensional (1D) descriptor,

which is simpler to calculate than 2-D and 3-D, dimensional descriptors, define more and detailed characteristics but are more difficult to calculate. There are two primary categories into which the molecular descriptors fall. The following experimental measurements can be divide into the different kinds of molecular format: log P, molecular refractivity, a dipole moment, polarization, and, in general, additive physical-chemical properties and theoretical molecular descriptors. All of these measurements are obtained from a visual presentation of the chemical compound. Theoretical ones are further divided into:

1. **Constitutional**: represent the general characteristics of molecules
2. **Topological**: graph theory is used to calculate it.
3. **Geometric**: These are based on empirical schemes and represent a molecule's capacity to engage in various kinds of interactions.
4. **Electronics**: See the characteristics of the electronic
5. **Physicochemical**: describe how a molecule behaves when exposed to outside reactions

3.2.1. 0D Molecular-Descriptors

This Molecular Descriptors type includes all molecular descriptors that can be compute without the necessity for molecular structure optimization or the knowledge of the molecule's connection between atoms. As a result, these descriptors are not affected by optimization issues or constraint problems. Typically, they exhibit an extremely high degeneration, meaning that several compounds, like isomers, have similar values. Despite having a minimal information level, they can nevertheless be quite useful in modeling a variety of physicochemical features or taking part in more intricate models. These descriptors include things like the Molecule all atom the quantity of a particular type of bond, the molecular weight, the average atomic weight, and the sum of atomic properties like Van der Waals volumes[14].

3.2.2. 1D-Molecular Descriptor

This category can contain all biomarkers for molecules that allow information to be calculated from molecule fractions. Typically, they take the form of fingerprints, which are essentially just binary vectors with 1 denoting the presence of a substructure and 0 denoting its absence. This kind of representation has several advantages, chief among them being the speedy computation of molecular similarities. Similar to 0D, these descriptors are simple to compute, readily comprehended, independent of conformational issues, and do not require molecular structure optimization. They often exhibit a medium-to-high degeneration and are frequently highly helpful in simulating biological and physicochemical characteristics. We also examine known as atom-centered fragments, which are based on the number of distinct molecule fragments, in addition to the 1D descriptors based on the number of chemical functional groups, e.g., total number of main carbon molecules, number of cyanates, number of nitriles, etc. The final three instances include hydrogen bonding to a main carbon, an alpha carbon, and a heteroatom.[14].

3.2.3 2D-Molecular Descriptor

They explain characteristics that can be computed using two-dimensional molecular models. They are derived using graph theory, regardless of the molecule's confor-mation. They describe theoretical structural features that are retained by isomorphism, i.e., properties with same values for isomorphic graphs, based on a visual representation of the molecule. Applying algebraic operators to molecules that reflect molecular structures yields an invariant component that can be represented as a feature polynomial, a sequence of integers, or a single numerical index. The values of these components are independent of how the vertices are labeled or numbered. Usually, they come from a molecular structure that has been broken down in hydrogen. They might be responsive to one or more of the characteristics that make up the molecule, including its size, shape, symmetry, branching, and cyclical nature. Additionally, they might be able to encode chemical data regarding the many types of bonds and atoms. [14]. Actually, they fall into one of two categories:

1. Structural-Topology index: only stores data pertaining to the proximity and separation of atoms inside a molecular structure.
2. Topochemical index: measures data on both topology and certain atom characteristics, such as identity or hybridization state.

3.2.4 3D Molecular Descriptors

The conformation of the molecular structure, which takes into account bond distances, dihedral angles, and other factors, is one of the three-dimensional descriptors that relate to the three-dimensional representation of the molecule and can be used to characterize the stereochemical properties of the molecules. Compared to the previous ones, its calculation is more intricate and could call for the examination of several conformations for molecular. The pharmacophore type atom, It is characterized as a combination of steric and electronic properties required to guarantee ideal supramolecular binding to a particular biological receptor and either cause or prevent its biological action, is represented by the most commonly used 3D molecular descriptors. Hydrophobic centers and hydrogen bond donors, for example, are features that are mapped into positions in molecules and are thought to be accountable for biological compound activity. Next, we calculate and note the distances between these locations that vary on conformation. With the development of more potent four-point pharmacophores, analyzing millions of possible pharmacophores for a test molecule can become necessary. Three-point pharmacophores are still often utilized. For instance, complex 3D descriptors are computed to determine a compound's active conformations or to pinpoint crucial features that account for variations in activity over a range of analogs. In addition, this kind of computation is required to create a query molecule's "pharmacophore shape" so that databases can be searched for compounds with comparable three-dimensional properties. Additionally, the generation of 3D-QSAR or 4D-QSAR models requires the application of pharmacophore type descriptors [14].

3.3. Types of fingerprints for Molecule Representation

A specific type of molecular descriptor known as fingerprints (FP) makes it possible to represent a molecule structure using a chain or vector of bits effectively and quickly that have a set length and show whether internal sub-structures or functional groups are present. The dataset has in the SMILE String that includes the molecular information that can be processed, stored, and compared with great efficiency using this type of molecular coding. The biological context is ignored by fingerprints formed from chemical structures, which creates a discrepancy between biological activity and molecular structure. As a result, little variations in the formant can result in significant variations in bioactivity. FP comes in several forms, ranging from the most basic that simply provides a list of 2D substructures (MACCS, for example) to more sophisticated ones that incorporate 3D data on molecule conformation. The most popular ones are enumerated in the list below. An overview of the descriptors from the consulted papers is displayed in Figure 3.

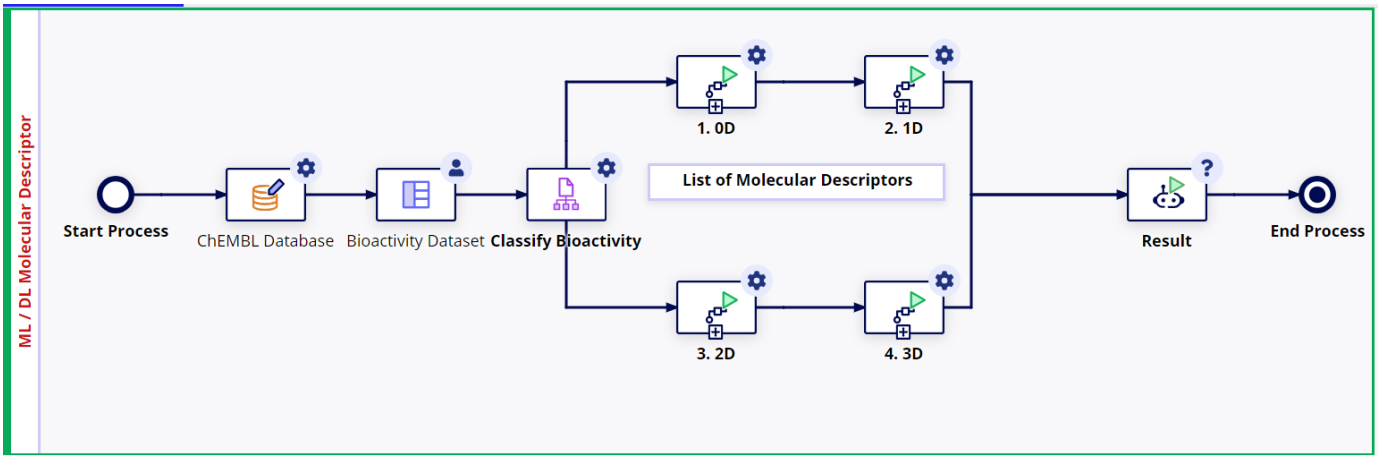

Figure 3. Typical Molecular Descriptor Using in ML/DL Drug Discovery Process

It shows how frequently a description occurs on its own in a publication and how much research has utilized multiple instances of it. Comparing multiple of them is typical in studies of this kind. In terms of absolute usage, ExtFP fingerprint is mostly used. Second on the list are the MACCS. Their widespread use is mostly due to their simplicity in calculation and the favorable outcomes they consistently produce when applied to various challenges. We can observe how these descriptors are still utilized in research today, given that the example of publications consulted is typical of the past five years of study, and that the emergence of other, more advanced analysis tools has not lessened their usefulness [15].

3.3.1. Extended Connectivity Fingerprint

A group of topo-logical fingerprints for molecular features is called Extended Connectivity Fingerprints (ECFP) [16] [17]. Traditionally, topological fingerprints were created to look for commonalities and substructures; nevertheless, they were created especially for structure–activity modeling [18]. Circular fingerprints known as ECFPs have several advantageous characteristics.

1. They are very fast to calculate.
2. They can show a wide range of distinct chemical properties, consisting of stereochemical data, and are not recommended.
3. Its features indicate the existence of specific substructures, making it simpler to evaluate the analysis's findings.
4. Because both are essential for assessing molecular activity, they are made to depict both the presence and lack of functioning.
5. The ECFP technique can be modified to produce various circular fingerprints that are tailored for applications.

3.3.2. Fingerprints-MACCS

The MACCS (Molecular ACCess System) key is another widely used kind of structural key [36,37,39,38]. Because of the company that made them, they are occasionally called MDL keys. Of the two sets of MACCS keys [19], one including 960 keys and the other a subset of 166 keys, the public can only see the shortest fragment definitions. These 166 public keys are used by open-source cheminformatics software packages such as CDK, RDKit, and others.

3.3.3. Fingerprints-PubChem

A vast amount of molecular data is available for free consultation and download from the PubChem library. Substructures are segment of a composition that a list of bits known as a fingerprint is produced by PubChem [19]. PubChem employs structural keys with a length of 881 bits, known as PubChem Fingerprints, to carry out similarity searches [20]. Additionally, it is used to nearby structures, which for every molecule compute a list of chemical structures that are similar in advance. The Compound Summary page provides access to this pre-calculated list.

3.3.4. AtomPaires Fingerprints

These are topological route-based fingerprints, which depict every connection path that may be defined by a certain fingerprint via an input compound [22]. Their primary focus is on the chemical connectivity data of artificial substances can we Classification differentiate type:

1. AtomPairs2DFingerprint (APFP) is defined by the shortest path separations between each pair of atoms within a composite structure's topological representation and the atomic environment. It stores 780 atom pairs at different topological separations.
2. The Chemical Development Kit's (CDDK) Graph Only Fingerprint (GraphFP) is a customized molecular fingerprint that records a fragment's 1024 path inside the composite structure without accounting for the binding order.

3.3.5. Fingerprints-CDK

The Chemical Development Kit (CDK) is a collection of popular freely available chemoinformatic tools (drug discovery, toxicology, etc.) that include methods for modifying and execute computations on data structures representing a chemical terms, such as 2D and 3D representations of chemical structures. The library applies a broad range of methods, from molecular descriptor computations and pharmacophore detection to the canonicalization of the structure of molecules [21].

The CDK offers techniques for standard molecular activities, such as the creation of structure diagrams, SMILES, ring searches, isomorphism verification, and 2D and 3D representations of Molecular atom structures [22].

4. Algorithms for graph-based machine learning

As mentioned in the preceding section, the majority of cheminformatics prediction models use molecular descriptors that are computed and coded in numerical vectors as their basis for input data. When these descriptors are used, high dimensionality matrices are produced, which may then be used with traditional machine learning algorithms like Random Forest, SVM, ANN, NB, etc. These methods consist of not being able to use the entire details about molecules depicted as a mathematical matrix network; instead, they are only made to handle data that is structured in matrices or vectors. In terms of graph theory, the graphical representation of a molecular network is a chemical compound's structural formula. Every compound is shown as a graph (G) in terms of representation. A node in the network represents each atom within it and shown all Supervised and unsupervised algorithm below figure.

5. Conclusion

In conclusion, the integration ML/DL Methods has the innovative capability the field of drug discovery by enabling faster, more efficient, and cost-effective methods for identifying novel therapeutics. While challenges remain, ongoing research efforts and technological advancements are expected to further enhance the capabilities of ML and DL in drug discovery, finally resulting in the creation of more secure and effective medical therapies for a variety of illnesses. Despite the challenges, With ML and DL approaches, the process of discovering drugs appears to have a bright future. Advances in data generation technologies, such as high-throughput screening and single-cell sequencing, will continue to expand the scope and quality of available data for training ML models. Moreover, the development of interpretable ML algorithms and hybrid approaches that combine computational predictions with experimental validation will enhance the reliability and utility of ML/DL-driven drug discovery platforms.

6. References

1. P. Carracedo-Reboredo *et al.*, "A review on machine learning approaches and trends in drug discovery," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021, doi: 10.1016/j.csbj.2021.08.011.
2. J. L. Blanco, A. B. Porto-Pazos, A. Pazos, and C. Fernandez-Lozano, "Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection," *Sci. Rep.*, vol. 8, no. 1, p. 15688, Oct. 2018, doi: 10.1038/s41598-018-33911-z.
3. C. R. Munteanu *et al.*, "Drug Discovery and Design for Complex Diseases through QSAR Computational Methods," *Curr. Pharm. Des.*, vol. 16, no. 24, pp. 2640–2655, Aug. 2010, doi: 10.2174/138161210792389252.
4. C. Fernandez-Lozano, M. Gestal, C. R. Munteanu, J. Dorado, and A. Pazos, "A methodology for the design of experiments in computational intelligence with multiple regression models," *PeerJ*, vol. 4, p. e2721, Dec. 2016, doi: 10.7717/peerj.2721.\
5. "DrugBank Online | Database for Drug and Drug Target Info." Accessed: Mar. 24, 2024. [Online]. Available: https://go.drugbank.com/
6. "PubChem." Accessed: Mar. 24, 2024. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/
7. "ChEMBL Database." Accessed: Mar. 24, 2024. [Online]. Available: https://www.ebi.ac.uk/chembl/
8. Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: 10.1093/bioinformatics/btm344.
9. P. Gramatica and A. Sangion, "A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology," *J. Chem. Inf. Model.*, vol. 56, no. 6, pp. 1127–1131, Jun. 2016, doi: 10.1021/acs.jcim.6b00088.
10. J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nat. Rev. Drug Discov.*, vol. 18, no. 6, pp. 463–477, Jun. 2019, doi: 10.1038/s41573-019-0024-5.
11. P. Gramatica, "Principles of QSAR Modeling: Comments and Suggestions From Personal Experience," *Int. J. Quant.*

Struct.-Prop. Relatsh., vol. 5, no. 3, pp. 61–97, Jul. 2020, doi: 10.4018/IJQSPR.20200701.oa1.

12. M. T. D. Cronin and T. W. Schultz, "Pitfalls in QSAR," *J. Mol. Struct. THEOCHEM*, vol. 622, no. 1–2, pp. 39–51, Mar. 2003, doi: 10.1016/S0166-1280(02)00616-4.
13. G. Piir, S. Sild, and U. Maran, "Interpretable machine learning for the identification of estrogen receptor agonists, antagonists, and binders," *Chemosphere*, vol. 347, p. 140671, Jan. 2024, doi: 10.1016/j.chemosphere.2023.140671.
14. N. Schaduangrat, A. A. Malik, and C. Nantasenamat, "ERpred: a web server for the prediction of subtype-specific estrogen receptor antagonists," *PeerJ*, vol. 9, p. e11716, Jul. 2021, doi: 10.7717/peerj.11716.
15. T. Yu, C. Nantasenamat, N. Anuwongcharoen, and T. Piacham, "Machine Learning Approaches to Investigate the Structure–Activity Relationship of Angiotensin-Converting Enzyme Inhibitors," *ACS Omega*, vol. 8, no. 46, pp. 43500–43510, Nov. 2023, doi: 10.1021/acsomega.3c03225.
16. M. Lee, H. Kim, H. Joe, and H.-G. Kim, "Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery," *J. Cheminformatics*, vol. 11, no. 1, p. 46, Dec. 2019, doi: 10.1186/s13321-019-0368-1.
17. H.-M. Lee *et al.*, "Computational determination of hERG-related cardiotoxicity of drug candidates," *BMC Bioinformatics*, vol. 20, no. S10, p. 250, May 2019, doi: 10.1186/s12859-019-2814-5.
18. A. L. Perryman *et al.*, "Naïve Bayesian Models for Vero Cell Cytotoxicity," *Pharm. Res.*, vol. 35, no. 9, p. 170, Sep. 2018, doi: 10.1007/s11095-018-2439-9.
19. P. Di *et al.*, "Prediction of the skin sensitising potential and potency of compounds via mechanism-based binary and ternary classification models," *Toxicol. In Vitro*, vol. 59, pp. 204–214, Sep. 2019, doi: 10.1016/j.tiv.2019.01.004.
20. J. Mendenhall and J. Meiler, "Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout," *J. Comput. Aided Mol. Des.*, vol. 30, no. 2, pp. 177–189, Feb. 2016, doi: 10.1007/s10822-016-9895-2.
21. S. Lim *et al.*, "A review on compound-protein interaction prediction methods: Data, format, representation and model," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1541–1556, 2021, doi: 10.1016/j.csbj.2021.03.004.
22. J. Dong *et al.*, "ChemSAR: an online pipelining platform for molecular SAR modeling," *J. Cheminformatics*, vol. 9, no. 1, p. 27, Dec. 2017, doi: 10.1186/s13321-017-0215-1.

Conflicts of interest: The authors stated that no conflicts of interest.

Publisher's Note
IJLSCI remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Correspondence and requests for materials should be addressed to Sahebrao B. Pawar.

Peer review information
IRJSE thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.