Chemoinformatics- Revolutionizing Drug Discovery

Amara Arshad[1], Muhammad Hussnain Tariq[2], Rohina Azam[1], Iqra Fatima[3], Hafsa Naeem[1] and Javeria Arshad[1,*]

[1]Department of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Government College University, Faisalabad-38000, Punjab, Pakistan
[2]Faculty of Pharmaceutical Sciences, Government College University, Faisalabad-38000, Punjab, Pakistan
[3]Department of Chemistry, University of Agriculture Faisalabad-38000, Punjab, Pakistan
*Corresponding author: Pharmacistj.arshad@yahoo.com

Abstract

Chemoinformatics involves employing informatics to tackle challenges within the field of chemistry. By leveraging computational tools and techniques such as virtual screening of chemical libraries, predictive modeling, and lead optimization approaches, it significantly reduces the need for extensive experimental work in wet labs. Predicting pharmacokinetic parameters helps eliminate unsuitable compounds, reducing synthesis-evaluation cycles and minimizing costly failures later. The integration of machine learning (ML) and artificial intelligence (AI) has significantly advanced the field, particularly in QSAR modeling. Various algorithms have been effective in predicting molecular properties, aiding compound refinement. Additionally, ML techniques tackle major challenges in disease research, facilitating personalized treatments. AI based models have significantly enhanced the accuracy of early predictions related to drug safety and efficacy. Nonetheless, there remains a continuous need to create explainable models that offer a high degree of interpretability. A substantial amount of high-quality data, uncertainty estimation, and justification of predictions are vital for enhancing the efficacy of these models in future drug design. This chapter examines the important contributions of chemoinformatics to drug discovery and development, recent advancements, and current challenges that must be addressed to enhance the reliability of these methodologies.

Keywords: Chemoinformatics; Molecular docking; Predictive modeling; Virtual screening; Machine learning; Artificial intelligence

Cite this Article as: Arshad A, Tariq MH, Azam R, Fatima I, Naeem H and Arshad J, 2025. Chemo-informatics- revolutionizing drug discovery. In: Farooqi SH, Kholik K and Zaman MA (eds), One Health Horizons: Integrating Biodiversity, Biosecurity, and Sustainable Practices. Unique Scientific Publishers, Faisalabad, Pakistan, pp: 233-242. https://doi.org/10.47278/book.HH/2025.144

A Publication of
Unique Scientific
Publishers

Chapter No:
25-032

Received: 18-Jan-2025
Revised: 25-Apr-2025
Accepted: 12-May-2025

Introduction

"Chemoinformatics" is a term first introduced by Frank Brown in 1998 who defined it as combining various information sources to convert data into usable information, then ultimately into knowledge, to accelerate decision-making in the field of identification and optimization of lead compound (Brown, 1998). Gasteiger and Funatsu offered a wider interpretation, describing it as 'the use of informatics to address challenges or problems in chemistry (Gasteiger & Funatsu, 2006). Cheminformatics has been practiced for over 40 years. While it is mainly used in drug discovery, it has many other applications across different fields, highlighting its broad and interdisciplinary use (Bajorath, 2004).

Chemoinformatics involves the use of computational methods that can analyze data and make predictions without relying solely on fixed rules. This helps improve the drug discovery, molecular modeling, drug–target interactions, and chemical screening. ML also aids in prioritizing drug candidates and predicting potential harmful effects of biologics accurately and efficiently (Niazi & Mariam, 2023).

1. The Traditional Approach to Drug Discovery

Although there are significant advancements in biotechnology and fundamental life sciences, the processes involved in drug discovery and development (DDD) continue to be lengthy and costly, averaging about 15 years and around US$2 billion for a drug molecule to develop (Sadybekov & Katritch, 2023). Recent advancements in artificial intelligence (AI) and machine learning (ML) have greatly transformed cheminformatics and thereby, drug discovery. Xu and Hagler in 2002, described how chemoinformatics has revolutionized the drug discovery process (Xu & Hagler, 2002).

Figure1 shows a schematic view of the evolution of the drug discovery process by incorporation of chemoinformatics in comparison to the classical approach. Chemo-informatics has sped up the process of drug development, from finding leads to optimizing them.

3. Role of Chemoinformatics in Drug Design

Chemoinformatics has revolutionized the drug discovery process in recent years. By utilizing computational tools and data analysis, it streamlines research efficiency. It has fundamentally changed the drug design process in several important ways

3.1. Data Integration and Management

Chemo-informatics utilizes chemical databases to manage and access chemical information.

Databases support information retrieval to predict targets, knowledge discovery, and data mining. Specialized databases that focus on naturally occurring compounds, such as COCONUT, SuperNatural-II, LOTUS, SymMap, and NPASS, serve as valuable assets. Additionally, data regarding chemical structures and bioactivity can be obtained from drug databases like BindingDB, the Protein Data Bank, Inxight, DrugBank and ChEMBL. Although there are numerous extensive databases, the application of ML and deep learning methodologies presents considerable opportunities to improve molecule creation and the development of focused libraries (Niazi & Mariam, 2023).

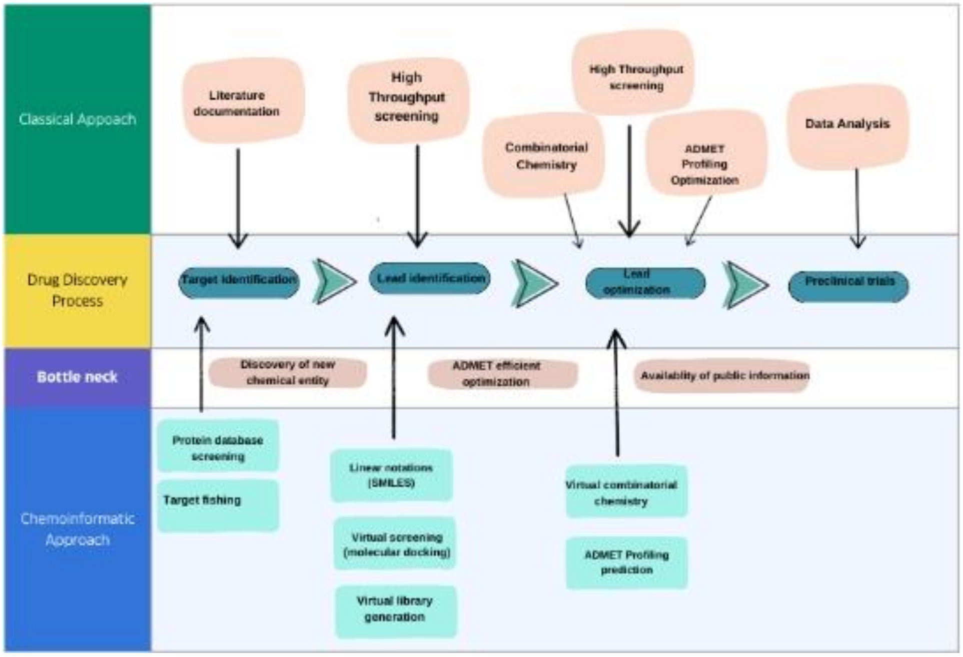

Classical Appoach
Literature documentation
High Throughput screening
High Throughput screening
Combinatorial Chemistry
ADMET Profiling Optimization
Data Analysis
Drug Discovery Process
Target identification
Lead identification
Lead optimization
Preclinical trials
Bottle neck
Discovery of new chemical entity
ADMET efficient optimization
Availablity of public information
Chemoinformatic Approach
Protein database screening
Target fishing
Linear notations (SMILES)
Virtual screening (molecular docking)
Virtual library generation
Virtual combinatorial chemistry
ADMET Profiling prediction

Fig. 1: Schematic view of evolution in the drug discovery process

3.2. Molecular Modelling and Simulations

The primary aim of molecular modeling simulations is to assist in identifying the most suitable compound to synthesize, ultimately conserving time, effort, and resources. It can also illustrate properties such as electrophilicity, nucleophilicity, and electrostatic potentials, while predicting molecular and biological attributes to comprehend the structure-activity relationships, thereby providing insights for drug design (Cohen, 1996). Depending on the arrangement of atoms and molecules within a specific system, these models allow for energy calculations, indicating how the system's energy fluctuates with changes in atomic and molecular positions. The subsequent phase in molecular modeling computations involves selecting the type of calculation, which could involve energy minimization, Monte Carlo simulation, or conformational searching and molecular dynamics (MD) (Saleh et al., 2017).

3.2.1. Structure Representation

The structural representations are generally classified as one-dimensional, two-dimensional, or three-dimensional, known as 1-D, 2-D, and 3-D, respectively.

1-D representations represent the physicochemical characteristics of molecules, with their values forming the vector components. Consequently, they use various scaled descriptors that encompass macroscopic properties like logP, heat of vaporization, solubilities, as well as individual molecular characteristics like the count of aromatic rings, molecular weight, the number of hydrogen bond acceptors and donors, and numerous graph-theoretical indices.

In contrast, 2-D representations predominantly focus on sub-structural fragments obtained from the two-dimensional structures of molecules. Typically, the collection of sub-structural features is treated as elements of classical sets, commonly referred to as molecular fingerprints.

3-D representations are the most intricate. They represent the approximate electron density usually in a “ball and stick model” or “stick model” (Mestres et al., 1997).

4D descriptors represent spatiotemporal elements, such as Volsurf, drug dissolution rates, or the properties that change with respect to time or the methods like CoMFA and GRID (Engel & Thomas, 2006; Matthias Dehmer, 2012; Chandrasekaran et al., 2018; Lo et al., 2018).

Although each descriptor has its own importance, 3D and 4D descriptors have significantly contributed to identifying potential drug targets and active molecules. Additionally, 4D descriptors like GRID and CoMFA are used to identify binding sites of receptors and describe interactions (Jeremy Ash & Fourches, 2017; Silakari & Singh, 2020).

3.2.2. Molecular Docking

Molecular docking is a computational technique used to identify the appropriate binding orientation of a protein-ligand complex. Different scoring functions are used to assess the binding strength of each pose and rank the most favorable poses produced by each molecule (Irwin D. Kuntz et al., 1982). The docking methods involve fitting a ligand into the target protein binding site as well as optimizing factors such as steric effects, and the electrostatic and hydrophobic compatibility, thus calculating their binding free energy (Diller & Merz Jr, 2001). The docking process is shown in Figure 2.

Tools such as AFT(Arakaki et al., 2004) , CATALYTIC SITE ATLAS (Porter et al., 2004), PATCH-SURFER (Sael & Kihara, 2012), POCKET SURFER (Chikhi et al., 2010), and SURFACE (Fabrizio et al., 2004), are used to identify active sites (Adelusi et al., 2022).

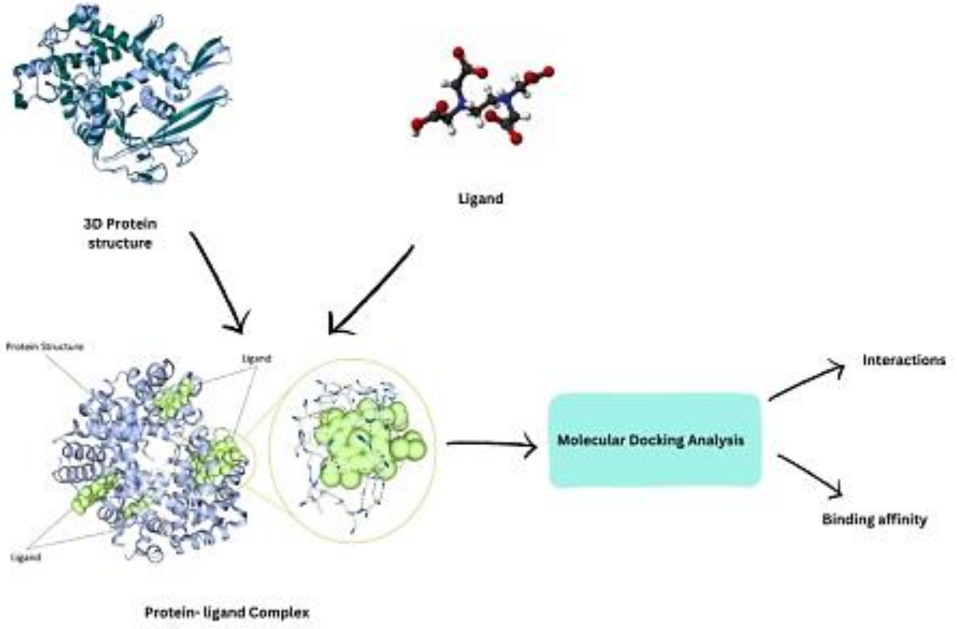


Fig. 2: Molecular docking.

3.2.2.1. Search Functions

A search algorithm is involved in creating an optimal number of geometries for ligand- protein interactions, including the determination of binding conformation experimentally (Taylor et al., 2002). Docking methods utilize three primary techniques based on the flexibility of the protein and the ligand

1) Treating the protein as partially or entirely flexible while examining all torsional degrees of freedom (TORSDOF) of the ligand,
2) Protein is considered as a rigid structure while assessing the ligand's TORSDOF, which include conformational, translational, and rotational aspects,
3) Viewing the protein as a rigid structure without any torsional degrees of freedom for the ligands (Guedes et al., 2014).

FRED, Surflex, Ligand fit, AUTODOCK 4.0, and Schrodinger's Glide are examples of docking algorithms (Adelusi et al., 2022)

3.2.2.2. Scoring Function

Scoring functions are the algorithms created to assess the binding affinity of a protein-ligand complex (Taylor et al., 2002). When a ligand attaches to its corresponding receptor, the scoring function predicts the binding pose that most closely reflects the actual protein-ligand complex structure. Currently, there are three main types of scoring functions: Force-Field, Empirical, and Knowledge-Based Scoring Functions.

Force Field scoring functions rely on physical interactions, as well as electrostatic interactions, van der Waals forces, as well as bond lengths, torsions, and angles (Huang et al., 2006). The DOCK program serves as a well-known example.

In Empirical scoring functions, the score of binding energy of a ligand- protein complex is determined by summing a series of weighted empirical energy terms that include hydrogen bonding energy, van der Waals energy, hydrophobic interactions, de-solvation energy, entropy, and electrostatic energy.

$$\Delta G = \Sigma iWi\cdot\Delta Gi$$

In this equation, **"ΔGi"** represents individual empirical energy terms, while the associated coefficients **"Wi"** are calculated by compiling the binding affinity data from a training set of ligand- protein complexes that have known 3D structures through least squares fitting (Head et al., 1996; Jain, 1996). Numerous empirical scoring functions have been created, including Rank Score, Chem Score, Glide Score and SCORE2.

Knowledge-Based scoring functions utilize the potential mean force (PMF) principle, where the energy of the complex is calculated as the total of all interaction terms between the atoms of protein-ligand complex. Knowledge-based functions consist of PMF, DrugScore, SMoG, MScore, ITScore/SE, and BLEEP, among others (Adelusi et al., 2022).

3.3. Quantitative Structure-Activity Relationship (QSAR) Models:

QSAR serves as a computational tool that quantifies the link between the physicochemical characteristics of a drug and its biological activity, thereby producing a mathematical model that informs how the structural or physicochemical attributes of molecules should be altered to enhance the activity of these compounds. The physicochemical properties of molecules are characterized by their steric constants (Es), hydrophobicity (log P), molar refractivity (MR), and various electronic properties that can be theoretically assessed using quantum mechanics (Saleh et al., 2017).

The current methodology for developing QSAR models generally includes generating compound descriptors in the training set, implementing descriptor selection algorithms, and utilizing statistical fitting techniques to create the model. Deep learning techniques are aimed at creating high-quality, interpretable QSAR models for extensive datasets without depending on precalculated descriptors (Chakravarti & Alla, 2019).

3.4. Pharmacokinetics and Pharmacodynamics Modelling

Pharmacokinetics (PK) and Pharmacodynamics (PD) modeling have evolved over the last few years from dose-response relationship to the discovery of drugs. The term "pharmaco" is derived from the Greek word "pharmackon," which means drug, while "dynamics" pertains to the variations in intensity of a phenomenon. PD focuses on the extent of drug responses.

On the other hand, "Kinetics" comes from the Greek word "kinetikos," which relates to movement. PK investigates how drugs move into, though, and out of the body. This area of study encompasses the processes of drug absorption, distribution, metabolism, excretion, and possible toxicity (Rosenbaum, 2016).

Comprehensive ADMET prediction platforms can effectively eliminate unsuitable compounds by concentrating on various pharmacokinetic parameters, thereby reducing the quantity of synthesis evaluation iterations and decreasing the likelihood of costly late-stage failures (Ferreira & Andricopulo, 2019).

4. High-Throughput Screening (HTS) and Virtual Screening (VS)

4.1. Modern Screening Methods in HTS

High-throughput screening (HTS) emerged in the mid-1980s to serve the pharmaceutical industry. It involves testing various substances in a shared assay, qualifying as high-throughput when it exceeds 10,000 wells daily. Ultra HTS refers to processes handling over 100,000 wells daily. This method relies on automation, liquid handling, and detection. Improvements in automation and miniaturization have enabled in vitro assays to be performed in 384-well and 1536-well microtiter formats, leading to efficient HTS processes capable of assessing hundreds of thousands of compounds each day. Different dispensing mechanisms such as air displacement, positive displacement, direct transfer acoustic transducer, a peristaltic pump, and solenoid syringe are used for fluid transfer. For detection, absorbance, fluorescence, luminescence, and radiometric methods are used (Wildey et al., 2017)

4.2. Virtual Screening (VS)

The VS uses computational techniques to filter chemical databases, optimize combinatorial libraries, and evaluate large chemical structures to select potential drug candidates. Yielding a unique pharmacological profile is the major objective of this technique. Captopril (antihypertensive drug), Saquinavir, Ritonavir, and Indinavir (fight against human immunodeficiency virus (HIV)) are some of the medications that have successfully reached the market with the assistance of VS Figure 3 shows steps involved in virtual screening**.** Receptor-based, or structure-based VS methods focus on ligand-receptor interactions and require a 3D structure of the target, which can be obtained through crystallography, X-ray imaging, or homology modeling (F Sousa et al., 2010).

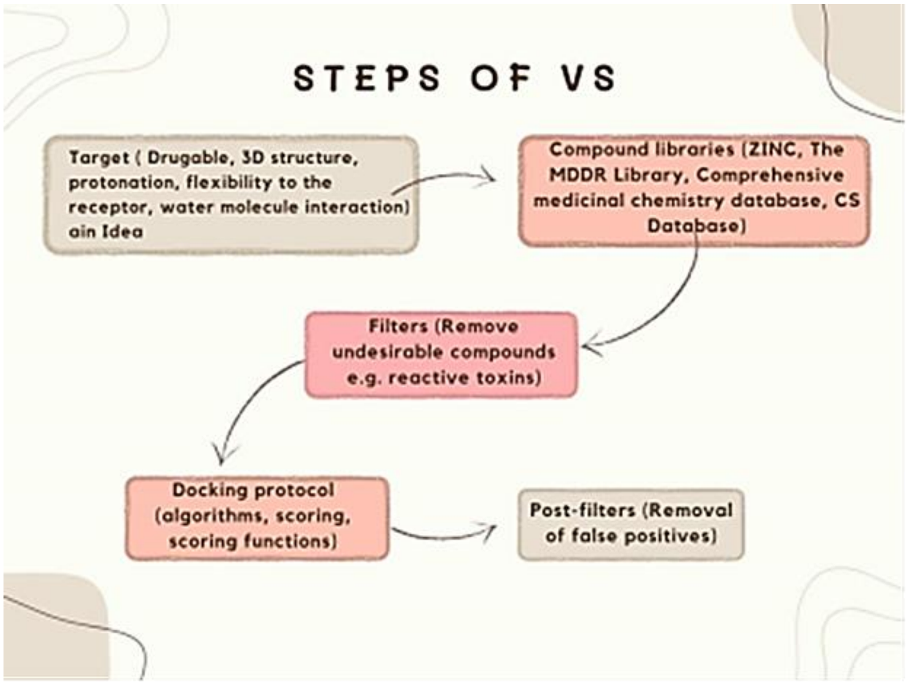
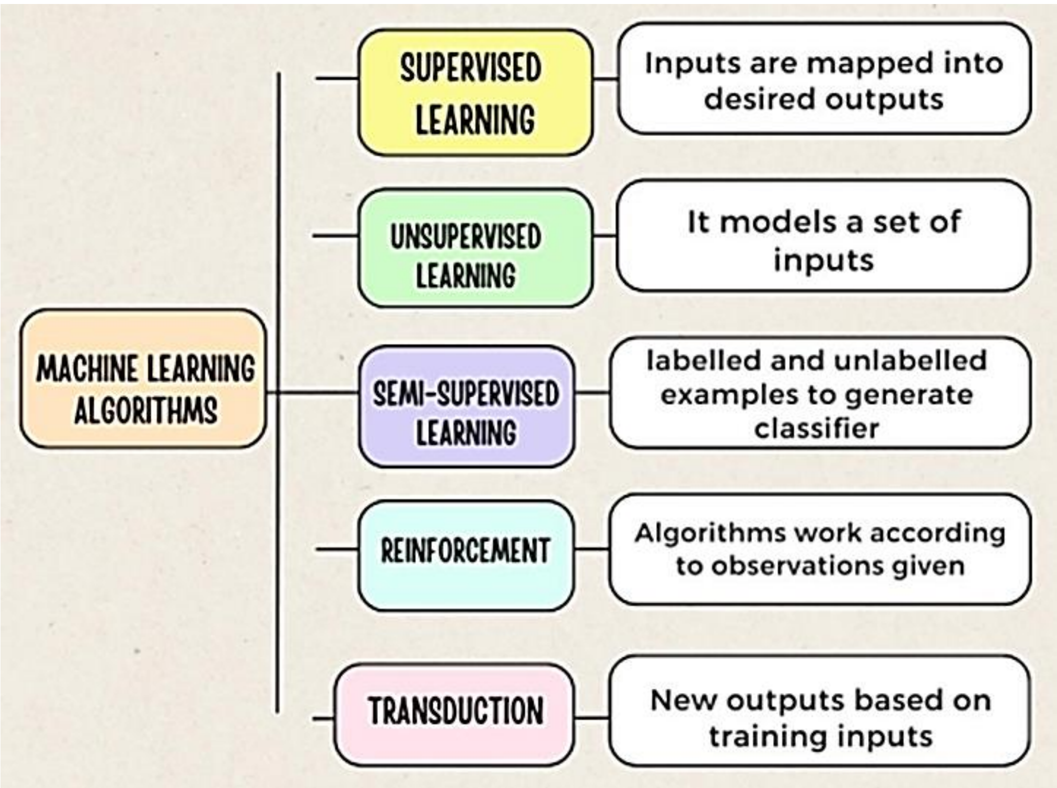

STEPS OF VS
Target (Drugable, 3D structure, protonation, flexibility to the receptor, water molecule interaction) ain Idea
Compound libraries (ZINC, The MDDR Library, Comprehensive medicinal chemistry database, CS Database)
Filters (Remove undesirable compounds e.g. reactive toxins)
Docking protocol (algorithms, scoring, scoring functions)
Post-filters (Removal of false positives)

Fig. 3: Steps of Virtual Screening

It is a resource-saving technique and is used to identify biologically active compounds (Oprea & Matter, 2004). The TOSS-MODE Approach can uncover anti-cancer leads, while feature-based pharmacophores can identify various compounds, including aldose reductase and retinoic-acid ligands (Langer & Hoffmann, 2001).

5. Integration of Machine Learning (ML) and Data Mining

Machine learning (ML) is one of the top emerging sciences and has a broad range of applications as mentioned in figure 4 (Mohammed et al., 2016) and it is an evolved branch of computational algorithms (El Naqa & Murphy, 2015) which is designed to emulate human intelligence. These techniques are used for pattern recognition, finance, entertainment, drug development, and other fields.


MACHINE LEARNING ALGORITHMS
SUPERVISED LEARNING
Inputs are mapped into desired outputs
UNSUPERVISED LEARNING
It models a set of inputs
SEMI-SUPERVISED LEARNING
labelled and unlabelled examples to generate classifier
REINFORCEMENT
Algorithms work according to observations given
TRANSDUCTION
New outputs based on training inputs

(**Fig. 4:** Machine learning algorithms)

Data Mining Techniques has been a part of AI since the 1960s, facilitating the searching for valuable information and handling vast amounts of data (Liao et al., 2012). Figure 5 shows the process involved in data mining technique.

DM has been classified based on verification and description of data as associations (to find all associations in a database), classifications (Developing the profile of different groups), sequential patterns (User-set baseline requirement), and clustering (dividing a database into clusters) (Pujari, 2001). Other DM Techniques are based on verification-driven or discovery-driven rules. Verification-driven data mining approaches involve the user formulating and testing hypotheses, while discovery-driven data mining techniques focus on automatically uncovering insights from the data.

6. Chemoinformatics Tools and Software

Computer-aided drug design (CADD) refers to computational techniques that are used for discovering, developing, and analyzing drugs and active molecules that contain comparable biochemical properties (Surabhi & Singh, 2018; Sabe et al., 2021). Some of the main software and tools are as follows:

ð ChEMBL (https://www.ebi.ac.uk/chembl/) is a high-quality, open database of bioactive molecules with drug-like properties, updated in the 2012, 2014, 2017, and 2019 Nucleic Acids Research Database Issues (Zdrazil et al., 2024).

ð AutoDock-GPU is a software of state-of-the-art docking, that by reducing scoring function, measures the geometrical conformation of a docked ligand-protein complex (Schieffer & Peng, 2023).

ð KNIME (Konstanz Information Miner) is a popular public data analytics platform offering a wide range of tools for ligand and structure-based drug design, supported by a large community of contributors (Mazanetz et al., 2020).

ð MolAICal software generates 3D drugs in protein target pockets by combining deep learning models and classical algorithms for improved results (Bai et al., 2021).

ð Quantitative structure-activity relationship (QSAR), molecular docking, homology modeling, virtual screening (Jiříčková et al.), virtual high-throughput screening (vHTS), and 3D pharmacophore mapping are key techniques in drug discovery. Among these, virtual screening is the most significant and widely approved (Hassan Baig et al., 2016; Sabe et al., 2021)

7. Collaborative Approaches in Drug Discovery

Various strategic methods have been proposed and applied to enhance efficacy in drug discovery and development across pharmaceutical R&D projects (Kiriiri et al., 2020).

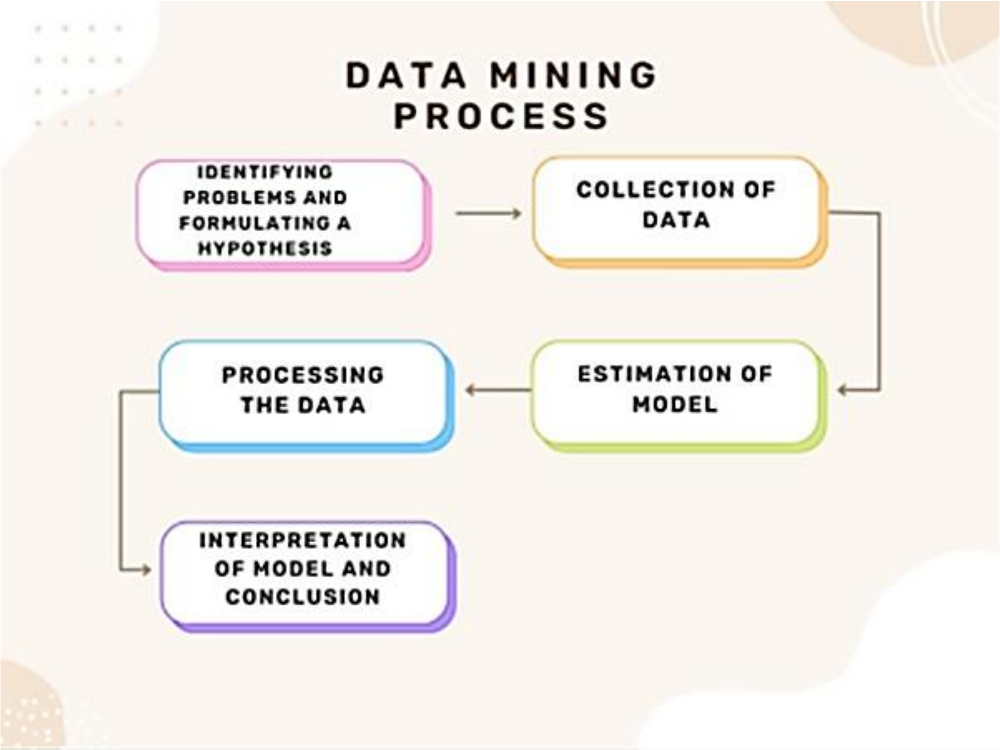

DATA MINING PROCESS
IDENTIFYING PROBLEMS AND FORMULATING A HYPOTHESIS
COLLECTION OF DATA
ESTIMATION OF MODEL
PROCESSING THE DATA
INTERPRETATION OF MODEL AND CONCLUSION

Fig. 5: Data Mining Process

7.1. Role of Public Databases and Resources

Over the past decade, public databases have been progressively reinforced by funding agencies, research networks, and governments (Wood, 2015). Chemo-informatics uses these databases to stock, recover, and examine massive chemical information, including structures, physicochemical properties, and biological activities, helping dataset analysis and molecular searches for effective results. These sources play a critical role in tasks such as data mining and target prediction. Specialized databases like TCMID, NPASS, COCONUT, LOTUS, SymMap, Supernatural-II, and TCMSP provide inclusive data on naturally occurring compounds, their molecular descriptors, and structural properties (Sorokina, 2020).

Abductive techniques use structural similarities of known compounds to deduce mechanisms. Similarity scores can be calculated for 1D (e.g., SMILES/SELFIES), 2D (e.g., fingerprints/topology), and 3D structures (e.g., geometric shapes). Common metrics include Dice index, Manhattan distance, Tanimoto index, and cosine coefficient. The model, created for producing scaffold-focused libraries is DeepMGM, a general model that is developed with drug-like molecules (Aathmanathan et al., 2022). The increasing trend of public databases (Open Data sharing) and code in field of computational drug discovery is supposed to enhance the progress in this field, as by taking previous work new information is built on that creates a snowball effect (Kanwal et al., 2017).

7.2. Collaborative Platforms and Their Impact

Open Science is being supported by government and funding agencies since the last decade. It has impacted drug discovery by promoting open scientific data sharing platforms in a significant manner.

7.2.1. Data Sharing in Open Science

Data sharing is necessary to improve or advance scientific research. This is inferred as allowing data reuse across studies and disciplines. Open access to resources supports reproducibility and also fosters innovation. There are certain databases like PubChem and ChEMBL, with platforms such as GitHub and Figshare that facilitate data exchange. Data papers with detailed explanations of data collection and methods also enhance transparency (Randles et al., 2017).

7.2.2. Addressing Reproducibility Challenges

In fields like Bioinformatics and computational research reproducibility of data is an important aspect. If there is no reproducibility it can lead to complex workflows inconsistencies. Some challenges include software versioning, hardware differences, and pipeline complexity (Kanwal et al., 2017). Sharing of data and code can improve reproducibility as suggested by Sandve et al. (2013). Some tools like BaseSpace, Galaxy, and Bioconductor support these principles, but some issues like platform customization and ethical concerns in cloud-based environments persist (Kim, Poline, & Dumas, 2017).

7.2.3. Container-Based Platforms for Computational Research

The use of container-based platforms like Docker provides isolated environments for application installation. BioContainers, Bio-Docklets, and Dugong are community initiatives that update the use of bioinformatics tools. Jupyter Notebooks facilitate reproducible workflows.

7.2.4. Promoting Open Science and Collaborative Research

Sharing of research findings, tools, and data promote global collaboration thus making scientific discoveries transparent and accessible. This approach is key in drug discovery and relevant fields for promoting reproducibility thus advancing scientific knowledge (Borgman, 2017).

8. Future Trends in Chemoinformatics

8.1. Machine Learning (ML)-based QSAR Modeling

The use of ML methods in cheminformatics has played a crucial role in discovering and designing potent pharmaceuticals. Furthermore, ML enhances biomarker analysis and genomics by identifying mutations and genes linked to diseases, through which the disease progression can be predicted, leading to precision medicine.

QSAR models are being developed by using classical ML techniques, such as support vector machine, linear regression, Naïve Bayes, and k-nearest neighbor.

8.1.1. Support Vector Machine

Support vector machines (SVM) are frequently utilized in QSAR studies because they are proficient in managing nonlinear interactions and multidimensional data. They establish a hyperplane that best separates various classes within the feature space. SVMs have shown outstanding effectiveness in predicting the biological activities of compounds, their bioavailability, and toxicity (Keyvanpour & Shirzad, 2020).

8.1.2. Regression Analysis

Regression analysis, a statistical technique which is used to show a relationship between two variables, a dependent variable and one or more independent variables. The objective is to find the best-fitting line that minimizes squared residuals, to allow for an understanding of variable relationships through regression coefficients. Early QSAR methods, like Free Wilson and Hansch analysis, were based on multivariate linear regression. Predictive modeling in QSAR uses various forms and combinations of regression analysis.

However, significant limitations including overfitting, limited interpretability, assumptions of linearity, and the requirement for high-quality data, still exist (Cardoso-Silva et al., 2019).

8.1.3. Naïve Bayes

Naïve Bayes is a probability classifier that typically assumes the independence of features, by which modeling process is carried out. The model predicts that the labels are conditionally independent and computes the probabilities for each label separately. A prominent application of this method is the PASS program, which uses it to forecast drug activities (Poroikov et al., 2000).

8.1.4. K-Nearest Neighbor

The k-nearest neighbors (kNN) algorithm illustrates both unlabeled and labeled data points within a multi-dimensional feature space. The kNN approach is a simple distance-based learning technique where an unknown instance is classified according to the majority of its k-nearest neighbors. It applies a majority-voting mechanism in which query points are transferred from the closest neighbors and labeled (Ajmani et al., 2005)

8.2. Artificial Intelligence (AI) Applications in Drug Designing

AI models have significantly enhanced early predictions of drug safety and efficacy by leveraging extensive ADME-Tox data. some of the key examples are as follows;

Chemoinformatics strategy was employed to develop a tetracycline analogue known as iodocycline, which has shown greater activity as a bacteriostatic agent compared to tetracycline, thereby exhibiting less potential for bacterial resistance (Kassab, 2022). Also, a research initiative focused on the synthesis and evaluation of antibacterial properties of ten compounds derived from benzimidazole and pyrazole against two Staphylococcus aureus strains, MRSA USA300 and MSSA ATTC6538 was carried out. The results indicated that three of the compounds demonstrated moderate bactericidal activity against MSSA, VRSA and MRSA (Shalaby et al., 2019).

9. Limitations and Areas of Improvement

Although, there has been a noticeable increase in the use of artificial intelligence in drug discovery, and this trend is continuing to grow. However, several challenges remain to be addressed. Even with the success of deep learning models, it's noteworthy that the quality of data remains fundamental to both the development and assessment of these models (Walters & Murcko, 2020). To enhance the utility of these models, whether they are predictive or generative, it is crucial to have a high-quality data in substantial amount.

When assessing these models, it's necessary to acquire suitable datasets and also to implement data balancing techniques alongside appropriate evaluation metrics (Walters & Murcko, 2020). Another issue is that, despite the advantages of deep learning, these models often lack transparency, making them difficult to interpret. Therefore, there is a continual need to create explainable models that offer high levels of interpretability. Specifically, this entails addressing four key aspects.

(i) Transparency, which involves understanding how the system arrives at a specific conclusion
(ii) Informativeness, which is offering fresh insights to human decision-makers;
(iii) Uncertainty estimation, which involves measuring the reliability of a prediction.
(iv) Justification, it justifies why the answer given is right

Conclusion

Chemoinformatics has simplified the drug design process, decreasing the time and expenses involved. The dependence on experimental wet lab procedures has diminished due to virtual screening, predictive modeling, and lead optimization techniques. Models based on machine learning and AI have significantly enhanced the precision of early predictions concerning drug safety and effectiveness by leveraging extensive data from various ADME-Tox datasets. Nonetheless, these models typically lack transparency, which makes them challenging to interpret. As a result, there is an ongoing need to develop explainable models that provide high levels of interpretability. A considerable amount of high-quality data, uncertainty estimation, and justification of predictions are crucial for improving the usefulness of these models for future drug design.

References

Aathmanathan, V. S., Arumugam, V., & Krishnan, M. (2022). Computational approach to explore the inhibitory potential of biologically derived compounds against Spodoptera litura vitellogenin receptor (VgR) using structure based virtual screening and molecular dynamics. *Journal of Biomolecular Structure and Dynamics*, 40(11), 4954-4960.

Adelusi, T. I., Oyedele, A.-Q. K., Boyenle, I. D., Ogunlana, A. T., Adeyemi, R. O., Ukachi, C. D., Idris, M. O., Olaoba, O. T., Adedotun, I. O., & Kolawole, O. E. (2022). Molecular modeling in drug discovery. *Informatics in Medicine Unlocked*, 29, 100880.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, 12(1), 307-328.

Arakaki, A. K., Zhang, Y., & Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, 20(7), 1087-1096.

Bai, Y., Santos, D. A., Rezaei, S., Stein, P., Banerjee, S., & Xu, B.-X. (2021). A chemomechanical damage model at large deformation: numerical and experimental studies on polycrystalline energy materials. *International Journal of Solids and Structures*, 228, 111099.

Bajorath, J. (2004). Understanding chemoinformatics: a unifying approach. *Drug Discovery Today,* 9(1), 13-14.

Brooijmans, N., & Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32(1), 335-373.

Brown, F. K. (1998). Chemoinformatics: what is it and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry*, 33, 375-384.

Chakravarti, S. K., & Alla, S. R. M. (2019). Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. *Frontiers in Artificial Intelligence, 2,* 17.

Chandrasekaran, Balakumar, Abed, Nidal, S., Al-Attraqchi, O., Kuche, Kaushik, Tekade, & K, R. (2018). Computer-aided prediction of pharmacokinetic (ADMET) properties. In Dosage form design parameters (pp. 731-755). Elsevier.

Chauhan, J., Luthra, T., Gundla, R., Ferraro, A., Holzgrabe, U., & Sen, S. (2017). A diversity oriented synthesis of natural product inspired molecular libraries. *Organic & Biomolecular Chemistry*, 15(43), 9108-9120.

Chemoinformatics: A Comprehensive Review. *International Journal of Molecular Sciences*, 24(14).

Chikhi, R., Sael, L., & Kihara, D. (2010). Real-time ligand binding pocket database search using local surface descriptors. *Proteins: Structure, Function, and Bioinformatics*, 78(9), 2007-2028.

Cohen, N. C. (1996). Guidebook on molecular modeling in drug design. Gulf Professional Publishing.

Csajka, C., & Verotta, D. (2006). Pharmacokinetic–pharmacodynamic modelling: history and perspectives. *Journal of Pharmacokinetics and Pharmacodynamics*, 33, 227-279.

Davis, A. M., & Riley, R. J. (2004). Predictive ADMET studies, the challenges and the opportunities. *Current Opinion in Chemical Biology*, 8(4), 378-386.

Derendorf, H., Lesko, L. J., Chaikin, P., Colburn, W. A., Lee, P., Miller, R., Powell, R., Rhodes, G., Stanski, D., & Venitz, J. (2000). Pharmacokinetic/pharmacodynamic modeling in drug research and development. The Journal of Clinical Pharmacology, 40(12), 13991418.

Diller, D. J., & Merz Jr, K. M. (2001). High throughput docking for library design and library prioritization. *Proteins: Structure, Function, and Bioinformatics*, 43(2), 113-124.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? Springer.

Engel, & Thomas. (2006). Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling*, 46(6), 2267-2277.

F Sousa, S., MFSA Cerqueira, N., A Fernandes, P., & Joao Ramos, M. (2010). Virtual screening in drug design and development. *Combinatorial Chemistry & High Throughput Screening, 13*(5), 442-453.

Fabrizio, F., Gabriele, A., Andreas, Z., & Manuela, H. C. (2004). SURFACE: a database of protein surface regions for functional annotation. Nucleic acids research, 32(suppl_1), D240-D244.

Fauman, E. B., Rai, B. K., & Huang, E. S. (2011). Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Current Opinion in Chemical Biology*, 15(4), 463-468.

Ferreira, L. L., & Andricopulo, A. D. (2019). ADMET modeling approaches in drug discovery. *Drug Discovery Today*, 24(5), 1157-1165.

Gasteiger, J., & Funatsu, K. (2006). Chemoinformatics–an important scientific discipline. *Journal of Computer Chemistry, Japan*, 5(2), 53-58.

Guedes, I. A., de Magalhães, C. S., & Dardenne, L. E. (2014). Receptor–ligand molecular docking. *Biophysical Reviews*, 6, 75-87.

Hajduk, P. J., Huth, J. R., & Tse, C. (2005). Predicting protein druggability. *Drug Discovery Today*, 10(23-24), 1675-1682.

Hassan Baig, M., Ahmad, K., Roy, S., Mohammad Ashraf, J., Adil, M., Haris Siddiqui, M., Khan, S., Amjad Kamal, M., Provazník, I., & Choi, I. (2016). Computer aided drug design: success and limitations. *Current Pharmaceutical Design*, 22(5), 572-581.

Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M., & Marshall, G. R. (1996). VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *Journal of the American Chemical Society*, 118(16), 3959-3969.

Hillisch, A., & Hilgenfeld, R. (Eds.). (2002). *Modern methods of drug discovery* (Vol. 93). Springer Science & Business Media.Pp 292. ISBN 376436081X, 9783764360818.

Huang, N., Kalyanaraman, C., Irwin, J. J., & Jacobson, M. P. (2006). Physics-based scoring of protein– ligand complexes: Enrichment of known inhibitors in large-scale virtual screening. *Journal of Chemical Information and Modeling*, 46(1), 243-253.

Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2), 269-288.

Jain, A. N. (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *Journal of Computer-aided Molecular Design*, 10, 427-440.

Jeremy Ash, & Fourches, D. (2017). Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *Journal of Chemical Information and Modeling*, 57(6), 1286-1299.

Jiříčková, A., Jankovský, O., Sofer, Z., & Sedmidubský, D. (2022). Synthesis and Applications of Graphene Oxide. *Materials*, *15*(3), 1-21. https://www.mdpi.com/1996-1944/15/3/920

Jun Lee, S., & Siau, K. (2001). A review of data mining techniques. *Industrial Management & Data Systems*, 101(1), 41-46.

Kanwal, S., Khan, F. Z., Lonie, A., & Sinnott, R. O. (2017). Investigating reproducibility and tracking provenance–a genomic workflow case study. *BMC Bioinformatics*, 18, 1-14.

Kiriiri, G. K., Njogu, P. M., & Mwangi, A. N. (2020). Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences*, 6, 1-12.

Lage, O. M., Ramos, M. C., Calisto, R., Almeida, E., Vasconcelos, V., & Vicente, F. (2018). Current screening methodologies in drug discovery for selected human diseases. *Marine Drugs*, 16(8), 279.

Langer, T., & Hoffmann, R. (2001). Virtual screening an effective tool for lead structure discovery. *Current Pharmaceutical Design*, 7(7), 509-527.

Leach, A. R. (2001). Molecular modelling: principles and applications. Pearson education.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12), 11303-11311.

Lo, Yu-Chen, Rensi, E, S., Torng, Wen Altman, & B, R. (2018). Machine learning in chemoinformatics and drug discovery. Drug discovery today, 23(8), 1538-1546.

Martinez-Mayorga, Karinam Madariaga-Mazon, Abraham, Medina-Franco, L, J., Maggiora, & Gerald. (2020). The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opinion on Drug Discovery*, 15(3), 293-306.

Matthias Dehmer, K. V., DanailBonchev. (2012). Statistical modelling of molecular descriptors in QSAR/QSPR. Wiley Online Library.

Mazanetz, M. P., Goode, C. H., & Chudyk, E. I. (2020). Ligand-and structure-based drug design and optimization using KNIME. *Current Medicinal Chemistry*, 27(38), 6458-6479.

Mestres, J., Rohrer, D. C., & Maggiora, G. M. (1997). MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *Journal of Computational Chemistry*, 18(7), 934-954.

Mike Hann, R. G. (August 1999). Chemoinformatics — a new name for an old problem? *Current Opinion in Chemical Biology*, 3(4), 379-383.

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.

Saleh, N. A., Elhaes, H., & Ibrahim, M. (2017). Design and development of some viral protease inhibitors by QSAR and molecular modeling studies. In *Viral proteases and their inhibitors* (pp. 25-58). Academic Press.

Oprea, T. I., & Matter, H. (2004). Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology*, 8(4), 349-358.

Porter, C. T., Bartlett, G. J., & Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic acids research, 32(suppl_1), D129-D133.

Pujari, A. K. (2001). Data mining techniques. Universities press.

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). Using the Jupyter notebook as a tool for open science: An empirical study. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL),

Rosenbaum, S. E. (2016). Basic pharmacokinetics and pharmacodynamics: An integrated textbook and computer simulations. John Wiley & Sons.

Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E., Govender, T., Naicker, T., & Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, 224, 113705.

Sadybekov, A. V., & Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature*, 616(7958), 673-685.

Sael, L., & Kihara, D. (2012). Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins: Structure, Function, and Bioinformatics*, 80(4), 1177-1195.

Niazi, S. K., & Mariam, Z. (2023). Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. *International Journal of Molecular Sciences*, *24*(14), 11488. https://doi.org/10.3390/ijms241411488

Schieffer, G., & Peng, I. (2023). Accelerating drug discovery in AutoDock-GPU with tensor cores. European Conference on Parallel Processing,

Silakari, O., & Singh, P. K. (2020). Concepts and experimental protocols of modelling and informatics in drug design. Academic Press.

Singh, R., Rathore, S. S., Khan, H., Karale, S., Chawla, Y., Iqbal, K., Bhurwal, A., Tekin, A., Jain, N., & Mehra, I. (2022). Association of obesity with COVID-19 severity and mortality: an updated systemic review, meta-analysis, and meta-regression. *Frontiers in Endocrinology*, 13, 780872.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213(4), 859-883.

Sorokina, M., & Steinbeck, C. (2020). Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics*, *12*(1), 20.

Surabhi, S., & Singh, B. (2018). Computer aided drug design: an overview. *Journal of Drug delivery and Therapeutics*, 8(5), 504-509.

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., & Mosquera, J. F. (2024). The

ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1), D1180-D1192.

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207-235.

Talele, T. T., Khedkar, S. A., & Rigby, A. C. (2010). Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current Topics in Medicinal Chemistry*, 10(1), 127-141.

Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2002). A review of protein-small molecule docking methods. Journal of computer-aided molecular design, 16, 151-166.

Verkhivker, G., Appelt, K., Freer, S., & Villafranca, J. (1995). Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligandprotein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering, Design and Selection*, 8(7), 677-691.

Wildey, M. J., Haunso, A., Tudor, M., Webb, M., & Connick, J. H. (2017). High-throughput screening. *Annual Reports in Medicinal Chemistry*, 50, 149-195.

Xu, J., & Hagler, A. (2002). Chemoinformatics and drug discovery. *Molecules*, 7(8), 566600.

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., & Mosquera, J. F. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1), D1180-D119.