Review

Recent Advances in Machine Learning-Based Chemoinformatics: A Comprehensive Review

Sarfaraz K. Niazi,[1*] Zamara Mariam[2]

1 College of Pharmacy, University of Illinois, Chicago, Illinois USA. niazi@niazi.com

2 Zamara Mariam, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST); Zmariam.msbi21rcms@student.nust.edu.pk

* Correspondence: niazi@niazi.com; +1-312-297-0000

Abstract: In the realm of modern drug discovery, the integration of chemoinformatics and quantitative structure-activity relationship (QSAR) modeling has emerged as a formidable alliance, enabling researchers to harness the vast potential of machine learning (ML) techniques for predictive molecular design and analysis. This review delves into the fundamental aspects of chemoinformatics, elucidating the intricate nature of chemical data and the crucial role of molecular descriptors in unveiling the underlying molecular properties. Molecular descriptors, including 2D fingerprints and topological indices, in conjunction with the structural-activity relationships (SARs), play a pivotal role in unlocking the pathway to small-molecule drug discovery. Technical intricacies of developing robust ML-QSAR models, including feature selection, model validation, and performance evaluation, are discussed herewith. Various ML algorithms, such as regression analysis and support vector machines, are showcased in the text for their ability to predict and comprehend the relationships between molecular structures and biological activities. This review serves as a comprehensive guide for researchers, providing an understanding of the synergy between chemoinformatics, QSAR, and ML. By embracing these cutting-edge technologies, predictive molecular analysis holds promise for expediting the discovery of novel therapeutic agents in the pharmaceutical sciences.

Keywords: QSAR/QSPR; cheminformatics; small-molecules; AI/ML; molecular descriptors; biological activity; SAR; predictive modeling; computational validation

Introduction

In 1998, the term "chemoinformatics," coined by Frank K. Brown, was aimed to expedite the drug discovery and development process; however, now, cheminformatics has a pivotal role in many areas of biology, chemistry, and biochemistry. The general process of drug discovery took 12 to 15 years and involved investments of around $500 million at that time. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized chemoinformatics and drug discovery by many folds. The global small molecule drug discovery market amounted to $75.96 billion in 2022 and is projected to hit around $163.76 billion by 2032.[1] [2]

In contrast to previously well-established statistics, mathematics, and physics-based stand-alone models, ML has introduced a paradigm shift enabling computers to learn from data and make predictions without relying solely on explicit rules or predefined mathematical equations. These algorithms can discover complex patterns and relations in 3D chemical structures and biological activity data, adaptively adjust their models based on feedback, and generalize from training examples to make accurate predictions on new, unseen data. This data-driven approach has opened new avenues for optimizing drug-target interactions, empowering target-based drug discovery, chemical library screening, molecular modeling, mechanics, and dynamics, prioritizing potential drug candidates, and predicting possible toxicological responses of biologics with improved accuracy and efficiency. In this review, we will explore the current state of research, the potential integration of ML-driven chemoinformatics tools and techniques in drug discovery, and the challenges and limitations of using

 doi:10.20944/preprints202306.0803.v1

these methods. Through a comprehensive analysis of recent studies and developments, we aim to provide insights into the exciting possibilities this integration holds for the future of small molecule drug discovery and design.

Exploration of chemoinformatics

At the intersection of chemistry and informatics, chemoinformatics has emerged as a potent field in drug discovery employing inductive learning to predict chemical phenomena.[3] [4] With the exponentially growing availability of chemical data, the application of ML in chemoinformatics has revolutionized the way researchers now explore, analyze, and predict the properties and activities of molecules, and then a few decades ago, expediting the process by many folds. It focuses on molecular engineering, molecular manipulation, library design, compound database searching and chemical space exploration, molecular graph mining, pharmacophore, and scaffold analysis. [5] [6] [7] [8] [9]

Fundamentals of Chemoinformatics

ML models perform prediction tasks based on chemical training data provided in the form of mathematical equations or a numerical representation. This transformation of compound structures into machine-learning-ready chemical data involves a complex, multilayer computational process. The process encompasses descriptor generation, molecular graphs, fingerprint construction, similarity analysis, chemical-space search, molecular dynamic simulations, etc. Each layer is interwoven with the preceding layers, significantly influencing the interpretation of the chemical data by the machine learning models and enhancing their predictive capabilities.

Data Mining and Chemical Databases

Training ML models requires chemical data, and chemoinformatics involves using chemical databases to store and retrieve chemical information. These databases can search for specific molecules or analyze large chemical data sets. The training of models relies heavily on managing and utilizing chemical databases that store vast amounts of chemical information, including compound structures, biological activities, and other relevant physiochemical properties. These databases facilitate data mining, knowledge discovery, and information retrieval for target prediction. Specialized databases of naturally existing compounds, including LOTUS,[10] COCONUT,[11] SuperNatural-II,[12] NPASS,[13] SymMap,[14] TCMSP, [15] and TCMID,[16] provide valuable resources. These databases contain comprehensive information on compound structures, molecular physicochemical properties, and molecular descriptors.

Utilizing the known structures of these compounds, abductive methods to transfer knowledge regarding mechanisms based on structural similarities can be leveraged. Various similarity scores, as mentioned before, can be computed, considering the similarity of 1D structures (e.g., SMILES-based similarity), 2D structures (e.g., 2D fingerprints-based similarity), and even 3D structures (e.g., 3D shape similarity). Previous studies have identified several metrics suitable for molecular similarity calculations, including the Tanimoto index, Dice index, Cosine coefficient, and Soergel distance. [17] [18] Furthermore, chemical bioactivity and structural data can be acquired from drug databases like ChEMBL,[19] BindingDB, [20] DrugBank, [21] Inxight, [22] Protein Data Bank[23] , etc.

Chemical Data Representation

Chemical data representation is subdivided into empirical, molecular, and structural data and can be represented in molecular graphs, fingerprints, descriptors, etc. [24] [25] Chemical and biological data is translated into computer-understandable form before models are trained on them. Genomic characterization through a multivariate Random Forest model was trained on genomic sequencing data given as numbers in one study.[26] In another, Naive Bayesian (NB) model was developed on numeric-based activity data, representing antagonists' binding on estrogen receptors.[27] For the prediction of the properties of small molecules based on ADMET (absorption, distribution, metabolism, excretion, and toxicity), an ML-based model was trained on 31 chemical numerical data sets acquired

 doi:10.20944/preprints202306.0803.v1

from Merck.[28] Similarly, molecular fingerprint data has also been used to train such models for ADMET properties prediction. NB and QSAR integrated models have been used to predict active compounds against Human Immunodeficiency virus type-1 trained on descriptors including extended-connectivity fingerprint data.[29] Furthermore, Graph Neural Networks (GNNs) operate on the graph-structure-data of 3D molecules and have been used to identify potential drug molecules.[30] Besides the choice of representation, data augmentation, and pre-processing, the twin curse of dimensionality and collinearity must be tackled.

When encountered in these data representations and modeling approaches, the twin curse of dimensionality and collinearity is catered through Principal Components Analysis (PCA), Partial Least Squares (PLS), and other available techniques. The data often involve many genomic or chemical descriptors in genomic characterization and small molecule property prediction. This high-dimensional feature space can lead to overfitting, decreased model interpretability, and increased computational complexity. In studies involving activity data, binding assays, or molecular fingerprints, collinearity can arise from strong correlations or dependencies among these input variables. Highly correlated variables can introduce redundancy and multicollinearity issues, leading to unstable model estimates and difficulties in interpreting the contributions of individual variables.

To address these challenges, dimensionality reduction techniques such as feature selection, feature extraction, data regularization, penalization, and genetic algorithms can help mitigate these issues by imposing constraints and encouraging sparsity. Principal Components Analysis (PCA) and Partial Least Squares (PLS) are generally used to transform large sets of correlated variables into smaller sets of uncorrelated ones. PCA has been used to explore complex datasets in QSAR and dimensionality reduction. A study investigating PCA's different applications in QSAR uses a dataset including CCR5 inhibitors. PCA has been used to detect outliers in the datasets as well. Another study used PCA to analyze the original data matrix in which molecules are represented by several predictor variables (molecular descriptors). PCA has also been used to decorate features for estrogen receptor binding prediction. Furthermore, observations revealed enhanced performance in predicting activity against a diverse range of pharmacological protein targets using kernel-principal components analysis (kernel-PCA) and a nonlinear variant of PCA, surpassing the predictive capabilities of LASSO regression.

Similarly, Partial Least Squares (PLS) have been employed to discern significant structural patterns that contribute to the biological activity of a molecule. The efficiency and accuracy of PLS in combination with unsupervised dimensionality reduction techniques surpass the approach of explicitly combining unsupervised dimensionality with multivariate regression. PLS is also widely utilized in the field of 3D-QSAR modeling. [31] [32] [33]

Molecular Descriptors

Molecular descriptors are numerical representations that capture chemical compounds' various structural, physicochemical, and biological properties. These descriptors are quantitative measures used for similarity analysis, virtual screening, and predictive modeling. Chemical molecular descriptors can be 0D, 1D, 2D, 3D or 4D: [34] [35] [36] [37]

- 0D Descriptors: These are constitutional or count descriptors, scalar values that describe several atoms, bonds, or functional groups in the molecule. i.e., molecular weight.
- 1D Descriptors: These descriptors capture molecular properties in one dimension along a linear sequence or chain of atoms. i.e., Structural fragments or fingerprints
- 2D Descriptors: These descriptors provide information about the molecular structure and properties within a 2D plane. i.e., Topological polar surface area (TPSA) and graph invariants.
- 3D Descriptors: These descriptors describe the molecular properties in 3D space, considering the spatial arrangement of atoms. i.e., autocorrelation descriptors, substituent constants, surface: volume descriptors, quantum, chemical descriptors, 3D-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 June 2023 doi:10.20944/preprints202306.0803.v1

MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, size, steric, surface, and volume descriptors, etc.

- 4D Descriptors: These descriptors encompass properties that change over time or involve spatiotemporal aspects. i.e., Drug dissolution rate, Volsurf, and GRID or CoMFA methods.

Table 1. Common chemical 0D to 4D descriptors for QSAR/QSPR analysis.

Descriptor Dimension	Descriptor Type	Example
0D	Number of atoms, bonds, and functional groups in the molecule	Molecular weight, LogP (partition coefficient)
1D	Molecular properties in a linear manner	Molecular Formula, SMILES (Simplified Molecular Input Line Entry System) Strings
2D	Topological polar surface area (TPSA)	Molecular fingerprint (e.g., Morgan fingerprint), Constitutional descriptors (e.g., number of atoms, bonds, and rings)
3D	Special properties of a molecule	Molecular shape descriptors (e.g., volume, surface area), Pharmacophore features
4D	Electrostatic potential descriptors with spatiotemporal aspects	Molecular dynamics descriptors, radius of gyration (Rg), solvent accessible surface area (SASA), Time-dependent properties (e.g., dynamic polar surface area (dPSA), time-dependent dipole moment.

These molecular descriptors have been used to select the most relevant properties. MoDeSus is an ML-based tool to determine the most informative molecular descriptors for QSAR studies. Molecular descriptors have been used to compare QSAR and QSPR models and can hasten early drug discovery by enabling ligand-based scaffold hopping for hit and lead optimization. Although each type of descriptor plays a vital role, 3D and 4D descriptors have shown the most significant contribution to identifying active molecules and potential drug targets. 4D descriptors like CoMFA and GRID have been used to identify active sites of receptors and characterize interactions providing insight into the functional properties of small molecules. [38] [39] [40]

Structure-Activity Relationship (SAR) Analysis

SAR analysis investigates the relationship between the chemical structure of a compound and its biological activity or property. It plays a key role in investigating potential bioactivity on changes in the chemical structure of drugs. Similarity analysis is interested in quantifying the degree of the structural or chemical similarity between compounds and extrapolating chemical properties from molecular similarity. Structure-activity relationship (SAR) models assume a correlation between calculated similarity and specific properties of small molecules like biological activities. This similarity analysis is used to transform and analyze chemical data, determine the relationship between the chemical structure and the function of molecules, and design new compounds or optimize the compounds of interest.[41]

Similarity search mainly aims to find compounds with similar bioactivity to a reference molecule but with different chemotypes. This results in scaffold hopping - derivatives acquired from a reference compound with a novel core structure. Computational exploration for scaffold hops includes 3D shape-based similarity search, fingerprint-based similarity search, pharmacophore matching, and fragment replacement techniques. Multimodal Deep transformer neural approach for scaffold hopping aids in designing molecules of novel scaffolds with improved pharmaceutical activity and similar 3D structure but dissimilar 2D structure.[42] [43] [44] [45]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 June 2023 doi:10.20944/preprints202306.0803.v1

SARs is also used in diversity quantification, outlier and novelty analysis, clustering, and intermolecular comparisons. AIMSim, a unified platform, performs similarity-based tasks using binary similarity metrics and molecular fingerprints on molecular datasets.[46] Another tool, Similarity Ensemble Approach (SEA), was employed in a study to predict accuracies of k-Nearest Neighbors (kNN) QSAR models built for known ligands of each GPCR target independently to identify active and inactive compounds. [47] Similarly, ChemSAR is an online pipelining platform for molecular SAR modeling that provides an integrated web-based platform for generating SAR classification models. The SAR Matrix (SARM) concept is a method used in multiple studies for the identification and structural organization of analog series, SAR analysis, and compound design. [48] [49] [50] The SARM methodology was expanded with the introduction of DeepSARM, which incorporated deep learning and generative modeling. This enhancement allowed for target-based analog design, considering compound information from related targets to enhance structural novelty and diversity.[51] By examining SAR, chemoinformatics helps understand the structure-activity patterns, identify critical structural features, and predict the activity of new compounds through QSAR models.

QSAR

QSAR analysis enables the prediction of the biological response or activity of a ligand from its physicochemical properties using QSAR models.[52] QSAR modeling tools have been utilized to identify potential drug candidates and have evolved into AI-based QSAR approaches.[53] Modern machine learning techniques can be used to model QSAR or quantitative structure–property relationships (QSPR) and develop artificial intelligence-based predictive models. [54] [55] Cheminformatics, QSAR, and machine learning applications have been used to showcase different structure-based, ligand-based, and machine learning-based approaches for drug development. QSAR/QSPR models employ information on multiple levels, i.e., chemical data, descriptors, molecular graphs, fingerprints, similarity analyses, molecular dynamic simulations, etc., to predict the most optimal properties of a potential drug.

The current approach in constructing QSAR models typically involves generating descriptors for the compounds in the training set, applying descriptor selection algorithms, and employing statistical fitting methods to build the model. Nevertheless, there have been investigations into the potential of developing high-quality, interpretable QSAR models for large and diverse datasets without relying on pre-calculated descriptors. To achieve this objective, these studies explore using deep learning techniques, precisely long short-term memory neural networks.[56]

QSAR Modeling

The standardized procedure for building QSAR models in drug discovery encompasses a series of modular steps that incorporate the afore-discussed chemoinformatics and machine learning techniques. By following the protocol, QSAR modeling aided by ML and DL (deep learning) can predict the properties or activities of chemical compounds, toxicity, and other related physiochemical properties.

Molecular Encoding

Molecular encoding is like chemical data representation transformation, as discussed before. The chemical features and properties of compounds are derived either directly from their chemical structures or through a lookup of experimental results. This process involves extracting relevant information from the molecular structure, such as atom types, bond types, functional groups, and physicochemical properties.

Feature Selection

Feature selection in QSAR aims to identify the most informative and relevant features from a larger set. It involves techniques such as univariate analysis, filter methods, wrapper methods, and embedded methods. Hybridizing feature selection, feature learning approaches, and unsupervised

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 June 2023 doi:10.20944/preprints202306.0803.v1

learning techniques are used to identify the most relevant properties and reduce the dimensionality and collinearity of the feature vector. These methods have proved effective in maintaining a reasonable computational effort without losing the accuracy of the final QSAR models.[57]

Model training

During the model training and learning phase of QSAR modeling, a supervised machine learning model is generally employed to uncover an empirical function that effectively maps input feature vectors to biological responses. This function is optimized to achieve the best possible mapping. It is crucial to carefully select and consider the SAR datasets and descriptors used for training and model validation to ensure the development of accurate QSAR models. [58]

Machine Learning-based QSAR Modeling

Machine learning models can be categorized into supervised and unsupervised learning. Supervised learning involves training a model with labeled data to make predictions based on known input-output relationships (e.g., linear regression and support vector machines). Unsupervised learning analyzes unlabeled data to discover underlying patterns and relationships without explicit guidance (e.g., clustering and dimensionality reduction).

QSAR involves training supervised learning models using labeled datasets, where the input features represent the chemical structures and the output labels represent the corresponding biological activities, toxicity, or other properties. Furthermore, unsupervised learning techniques can be applied in QSAR to uncover hidden patterns or relationships within the chemical data, such as clustering similar compounds based on their structural similarities or reducing the dimensionality of the dataset. QSAR models can be built using traditional methods like Random Forest, Multiple Linear Regression, Naïve Bayes, K-nearest Neighbours, Support Vector Machine, or Deep Neural Network (DNN). A comparative study found that the DNN approach outperformed traditional QSAR methods in Triple-negative breast cancer (TNBC) inhibitors and G-Protein Coupled Receptors (GPCR) agonist discovery.[59]

Regression Analysis

Regression analysis is a statistical method to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit line that minimizes the sum of squared residuals. The relationship between variables can be inferred by estimating the regression equation coefficients. Early QSAR techniques like Hansch and Free Wilson analysis extensively use multivariate linear regression. Since QSAR deals with multidimensional data, the twin curses must be tackled before further processing chemical data. Many variations and ensembles of regression analysis are now employed for predictive modeling in QSAR.

Network-based linear regression has been used to build interpretable QSAR models by combining elements of network analysis and piecewise linear regression. In a study on Polo-Like Kinase-1 inhibitors, linear regression QSAR models were developed to search for predictive models on a large and structurally diverse dataset of 530 compounds. Another variation of the regression model, a Discriminant-Regression Model (DIREM), an ordinary discrete-continuous QSAR approach, combines discriminant and regression analyses to explore structure-activity relationships for compounds. A comparative analysis study conducted on 5-nitrofuran-2-yl derivatives as inhibitors of Mycobacterium tuberculosis H37Rv, evaluated the performance of PLS-based QSAR models and compared them with results obtained from Multi Linear Regression (MLR) and Principal Component Regression (PCR). The results of PLS and MLR analysis showed significant predictive power and reliability compared to PCR, vouching for the reliability of these methods. [60] [61] [62] [63] [64] Although linear regression analysis and its derivates have been successfully employed in many drug optimization studies, underlying linearity, overfitting, limited interpretability, the need for high-quality data, and invalid vector space assumptions are significant limitations.

 doi:10.20944/preprints202306.0803.v1

K-nearest Neighbor

The k-nearest neighbors (kNN) algorithm represents labeled and unlabeled data nodes in a multiple-dimensional feature space. The kNN methodology relies on a simple distance learning approach whereby an unknown member is classified according to most of its k-nearest neighbors. Using a majority-voting rule, it assigns labels to query points by transferring them from the nearest neighbors. This approach leverages the proximity of data points in the feature space to make predictions. [65]

Determining the optimal number of nearest neighbors to use in the kNN algorithm is challenging, as selecting values that are too high or too low can result in undesirable false-positive or false-negative rates. To address this, the similarity ensemble approach (SEA) was introduced that compares chemical similarity values to a randomized background score, like the approach used in a BLAST sequence similarity search, proving to be a more principled approach for choosing the appropriate number of neighbors in kNN analysis. [66]

A study used a kNN model to develop a 3D QSAR model for 30 compounds with anti-HIV activity. This kNN model classified the compounds according to the majority of their nearest neighbors in the training set. The method identified the key structural features that contribute to the anti-HIV activity of the compounds. [67] Another study used the kNN-Molecular Field Analysis method to develop QSAR models for a set of 50 compounds with anti-HIV activity, unveiling the significant role played by electrostatic and steric interactions in determining the anti-HIV activity of the compounds. [68] In a study, consensus kNN QSAR, a versatile method for predicting the estrogenic activity of organic compounds in silico, was employed using a diverse set of compounds and was a feasible method for rapid screening of the estrogenic activity of organic compounds. [69] A deep neural network in conjunction with the kNN method to develop QSAR models for a set of 1,000 compounds with anti-cancer activity was also reported to be useful. The study found that the kNN method could identify the key structural features that contribute to the anti-cancer activity of the compounds. [70]

Naïve Bayes

Naive Bayes is a probabilistic classifier commonly assuming that features are independent, simplifying the modeling process. It is the probability of correct label assignment by considering the prior probability distribution of labels in the training set. It assumes conditional independence between multiple labels and calculates probabilities for each label individually. The PASS program, a notable example, utilizes this approach for predicting drug activities.[71]

A study compared the ability of Pipeline Pilot Naïve Bayes (PLPNB) and random forest to make accurate predictions on 18 large, diverse in-house QSAR datasets. The study found that PLPNB was computationally efficient and able to make accurate predictions on binary and multicategory activities. They have shown efficacy in large-scale virtual screening for significant pharmacological properties, including cytochrome P450 inhibition, human plasma protein binding, and animal model bioavailability. [72] [73] Another study demonstrated that the Naïve Bayes model gives minimal mean error over uniform dispersion of the data points in QSAR modeling. This study used QSAR as an illustration to show the optimality of the Naïve Bayes model. [74] In a comparative study to choose the best learning algorithm and optimal feature selection, Naïve Bayes was shown to be one of the best-performing algorithms for small datasets.[75]

Support Vector Machine

Support Vector Machines (SVM) are widely used in QSAR due to their ability to handle high-dimensional data and nonlinear relationships. They construct a hyperplane that maximally separates different classes in the feature space. SVMs have demonstrated excellent performance in various QSAR applications, such as predicting compound activities, toxicity, and bioavailability. Their versatility and robustness make them valuable tools in QSAR modeling. A study analyzed the application of machine learning algorithms in QSAR modeling and introduced a framework called 'ML-QSAR.' The framework was designed to facilitate the selection of proper strategies among existing algorithms according to the application area requirements and to help develop and improve current

 doi:10.20944/preprints202306.0803.v1

methods and found that SVM was one of the most commonly used machine learning algorithms in QSAR modeling. [76]

A study developed multiple QSAR methods using several ML algorithms, including SVM, to predict the activity of active substances against Pseudomonas aeruginosa. The study found that SVM could better predict the compounds' activity among other models accurately. [77] Another study investigated SVM's performance and predictive capability in QSAR modeling of HEPT derivatives. This study compared SVM with other methods, such as artificial neural networks, and found that SVM achieved good predictive performance. [78] SVM was also used to model phenethylamines' structure-activity relationships (SAR). The study aimed to classify antagonists and agonists and predict their activities and found that SVM was a robust tool in the SAR/QSAR field. [79]

Another study assessed the performance of 16 machine learning algorithms, including SVM, on 14 QSAR datasets and concluded that different ML algorithms provided different QSAR modeling methods to reveal the relationships between structures and properties of compounds. [80] When used for large-scale ligand-based predictive modeling, SVM predicts the properties of new, unknown compounds and can achieve good predictive performance for large-scale QSAR modeling. [81] SVMs have also been applied in a QSAR investigation involving ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolidinyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate derivatives, targeting the transcription factors activator protein (AP)-1 and nuclear factor (NF)-kB. [82] To identify the structural features responsible for a strong vascular endothelial growth factor (VEGF)-2 inhibition activity in aminopyrimidine-5-carbaldehyde oxime derivatives, a genetic variable selection approach was combined with SVMs. This integrated approach successfully identified several critical structural features associated with the desired biological activity, proving SVM helpful in QSAR modeling. [83]

Convolutional Neural Networks, Recurrent Neural Networks, Deep Neural Networks, and Ensemble methods

By leveraging the power of neural networks with multiple hidden layers, deep learning models can effectively learn complex relationships between molecular structures and their related biological activities. In QSAR, deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Neural Networks (DNNs), have been utilized to analyze and predict various properties of molecules, including binding affinity, activity, toxicity, and bioavailability. These models and their Ensemble methods have been applied in QSAR studies to enhance models' accuracy and predictive power.

CNNs have successfully captured molecular features and patterns from 2D chemical structures and search spaces. RNNs have been utilized to model sequential data, such as molecular fingerprints and SMILES strings. DNNs have effectively learned complex relationships between 3D and 4D molecular descriptors and their respective bioactivity data. Ensemble methods combining CNN, RNN, and DNN, have been employed to improve prediction performance. These advanced neural network architectures and ensemble techniques have been widely applied to all branches of chemoinformatics, including modeling QSAR/QSPR properties of small molecules and performing pharmacokinetic and pharmacodynamic analysis. Specifically, the unique ability of CNN to analyze images enabled the viewing of protein structures as '3D images' with four different atom-type channels. These 3D-CNNs were used to analyze amino acid microenvironment similarities and to predict the effects of mutations on the structure of proteins. [84] A study proposed a Transformer-CNN architecture for QSAR modeling and interpretation. The architecture replaced all recurrent units with convolutional and element-wise feed-forward layers and found that the Transformer-CNN architecture provided good results for small datasets and required less than a hundred iterations to converge for QSAR tasks.[85]

Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), are designed to capture both short-term and long-term dependencies in sequential data. In the context of QSAR, LSTM networks have been employed for tasks such as de novo drug design, where they learn the structural patterns and rules from SMILES strings to generate novel molecules. Other advanced techniques such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 June 2023 doi:10.20944/preprints202306.0803.v1

deep reinforcement learning have been utilized to learn latent representations of molecules and facilitate the generation of compounds with specific molecular properties. These approaches contribute to the exploration of new chemical space and the discovery of potential drug candidates. [86] [87] [88] [89] [90] [91] A study that proposed an ensembled RNN-CNN architecture, DeepCpG, for DNA methylation analysis concluded that combining RNN and CNN improved the performance of the QSAR model. [92] A novel DL-based technique called DeepSnap was developed to conduct QSAR analysis using three-dimensional images of chemical structures. This technique can also predict the potential toxicity of many chemicals to various receptors without extracting descriptors. [93] CNN, RNN, and Deep learning-based methods have also shown promising results in QSAR modeling.

Validation of ML-QSAR Models

ML-QSAR models are typically assessed using established metrics like sensitivity, specificity, precision, and recall. In cases where the dataset is unbalanced, the area under the curve (AUC) obtained from receiver-operating-characteristic (ROC) curves can be employed. QSAR models can also be evaluated by various methods, such as external validation, conformal prediction methods, and evaluation of QSAR equations for virtual screening. External validation is the main point to check the reliability of developed models for the prediction activity of not yet synthesized compounds. QSAR models can inform on the correlation between activities and structure-based molecular descriptors, essential for understanding the factors that govern molecular properties and designing new compounds with favorable properties.[94] [95] [96]

While 3D-QSAR methods, such as CoMFA, incorporate structural conformation, they require significant computational resources and can introduce uncertainties stemming from conformation prediction, ligand orientation, and structural alignment. As a result, 2D-QSAR models can offer a competitive alternative and, in certain instances, even outperform 3D-QSAR approaches.[97] The goal of QSAR analysis is the development of validated models for accurate and precise prediction of the biological activities of compounds. Metrics such as R^2 and QCV2 are typically optimized in deriving QSAR models. Similar metrics, calculated on an external data set, are used to evaluate the performances of the final models.[98]

A comparative study on 5-nitrofuran-2-yl derivatives as inhibitors of Mycobacterium tuberculosis H37Rv, used statistical parameters, including squared correlation coefficient, cross-validated correlation coefficient, and Fischer's value for statistical importance, to assess the quality of the generated QSAR models. Another study compared various statistical parameters of external validation of 44 reported QSAR models for biologically active compounds reported in scientific papers. They concluded that employing the coefficient of determination (R^2) alone could not indicate the validity of a QSAR model. These established criteria for external validation have advantages and disadvantages that should be considered in QSAR studies. [99]

Interpretability and Explainability of ML-QSAR Models

The ultimate goal of QSAR analysis is the development of validated models for accurate and precise prediction of the biological activities of compounds. The interpretability and explainability of ML-QSAR models promote transparency, reproducibility, and trust in the model's predictions, allowing researchers and stakeholders to make informed decisions regarding drug discovery and development. A study created six synthetic datasets of three complexity levels for benchmarking QSAR model interpretation methods. Using these datasets, the study investigated an extensive set of the descriptor and algorithm combinations and the universal interpretation approach, Structure-Property Correlation Index (SPCI). The study established that interpretation performance might decrease faster than predictivity, and in some cases, models with acceptable predictivity may have poor interpretation performance. [100]

Various techniques can enhance the explainability and interpretability of ML-QSAR models. Feature importance analysis can identify the most influential molecular descriptors or features contributing to the model's predictions. Visualization methods, such as heat maps or feature importance plots, can aid in understanding the relationships between features and the predicted outcomes.

 doi:10.20944/preprints202306.0803.v1

Additionally, model-agnostic techniques like LIME (Local Interpretable Model-Agnostic Explanations)[101] or SHAP (Shapley Additive Explanations)[102] can provide insights into individual predictions by highlighting the contributions of each feature. A paper describes a new QSAR model visualization approach that simplifies the analysis by introducing a new similarity measure between two models. The approach is based on mapping models onto a two-dimensional plane, where the distance between two models is proportional to the difference in their predicted activities.[103] Another study combines direct kernel-based PLS with Canvas 2D fingerprints to arrive at predictive QSAR models that can be projected onto the atoms of a molecule. The study provides a visualization of the model that can be used to identify the most critical atoms for predicting activity. [104]

Conclusions

Applying machine learning techniques in chemoinformatics has contributed significantly to discovering and designing highly effective drugs. This paper highlights the significant role of chemoinformatics and ML-based QSAR in drug discovery and development. Integrating computational approaches with large-scale data analysis has revolutionized the field, enabled the efficient exploration of chemical space and predicting biological activities. Multiple algorithms built for QSAR modeling play a significant role in highlighting features necessary for further designing small molecules. They have demonstrated their effectiveness in predicting molecular properties and activities, aiding in compound prioritization and optimization.

The future of chemoinformatics and QSAR modeling holds promising opportunities for further advancements. Integrating QSAR models with molecular docking techniques can enhance the accuracy of binding affinity predictions and provide valuable insights into the interaction between ligands and target proteins. Fragment-based design approaches can benefit from QSAR models by guiding the selection and optimization of fragments to develop novel drug candidates. Additionally, integrating QSAR models with de novo drug generation methods, such as deep learning and generative modeling, opens up possibilities for computer-assisted design and discovering new molecules with desired properties.

This convergence of QSAR models with molecular docking, fragment-based design, and de novo drug generation methods holds great potential in accelerating the drug discovery process, reducing costs, and increasing the success rates of identifying novel therapeutic agents. Continued research and development in this area will undoubtedly pave the way for more efficient and precise drug design strategies, ultimately benefiting patients and advancing the field of pharmaceutical sciences.

Conflict of Interest: Authors declare no conflict of interest.
Funding: No funding

---

[1] Small Molecule Drug Discovery Market Size, Report By 2032. (n.d.). https://www.precedenceresearch.com/small-molecule-drug-discovery-market (Accessed on: 24th May 2023)

[2] Brown, F. K. (1998). Chapter 35 – Chemoinformatics: What is it and How does it Impact Drug Discovery. In J. A. Bristol (Ed.), Annual Reports in Medicinal Chemistry (Vol. 33, pp. 375–384). Academic Press. https://doi.org/10.1016/S0065-7743(08)61100-8

[3] Polanski, J. (2013). Chemoinformatics. In Elsevier eBooks (pp. 635–676). https://doi.org/10.1016/b978-0-12-409547-2.14327-6

[4] Gasteiger, J. (2016). Chemoinformatics: Achievements and Challenges, a Personal View. Molecules, 21(2), 151. https://doi.org/10.3390/molecules21020151

[5] Polanski, J. (2009). Chemoinformatics. In Elsevier eBooks (pp. 459–506). https://doi.org/10.1016/b978-044452701-1.00006-5

[6] Gasteiger, J. (2003). Handbook of Chemoinformatics. In Wiley eBooks. https://doi.org/10.1002/9783527618279

[7] Varnek, A., & Baskin, I. I. (2011). Chemoinformatics as a Theoretical Chemistry Discipline. Molecular Informatics, 30(1), 20–32. https://doi.org/10.1002/minf.201000100

[8] Chemoinformatics and Computational Chemical Biology. (2011). In Methods in molecular biology. Springer Science+Business Media. https://doi.org/10.1007/978-1-60761-839-3

[9] Kapetanovic, I. M. (2008). Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. Chemico-Biological Interactions, 171(2), 165–176. https://doi.org/10.1016/j.cbi.2006.12.006

[10] Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., ... & Allard, P. M. (2021). The LOTUS initiative for open natural products research: knowledge management through Wikidata. BioRxiv, 2021-02.

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

[12] Banerjee, P., Erehman, J., Gohlke, B. O., Wilhelm, T., Preissner, R., & Dunkel, M. (2015). Super Natural II—a database of natural products. Nucleic acids research, 43(D1), D935-D939.

[13] Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., ... & Chen, Y. Z. (2018). NPASS: natural product activity and species source database for natural product research, discovery and tool development. Nucleic acids research, 46(D1), D1217-D1222.

[14] Wu, Y., Zhang, F., Yang, K., Fang, S., Bu, D., Li, H., ... & Chen, J. (2019). SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic acids research, 47(D1), D1110-D1117.

[15] Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., ... & Yang, L. (2014). TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. Journal of cheminformatics, 6, 1-6.

[16] Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., & Shi, T. (2012). TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic acids research, 41(D1), D1089-D1095.

[17] Chemoinformatics: Basic Concepts and Methods. (2018, August 1). Wiley.com. https://www.wiley.com/en-dk/Chemoinformatics:+Basic+Concepts+and+Methods-p-9783527331093

[18] Xue, H., Stanley-Baker, M., Kong, A. W. K., Li, H., & Goh, W. W. B. (2022). Data considerations for predictive modeling applied to the discovery of bioactive natural products. Drug Discovery Today, 27(8), 2235–2243. https://doi.org/10.1016/j.drudis.2022.05.009

[19] Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic acids research, 47(D1), D930-D940.

[20] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1), D1045-D1053.

[21] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 46(D1), D1074-D1082.

[22] Siramshetty, V. B., Grishagin, I., Nguyễn, Đ. T., Peryea, T., Skovpen, Y., Stroganov, O., ... & Southall, N. T. (2022). NCATS Inxight Drugs: a comprehensive and curated portal for translational research. Nucleic Acids Research, 50(D1), D1307-D1316.

[23] Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallographica Section D: Biological Crystallography, 54(6), 1078-1084.

[24] Haghighatlari, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X., & Head-Gordon, T. (2020). Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. Chem, 6(7), 1527–1542. https://doi.org/10.1016/j.chempr.2020.05.014

[25] David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. Journal of Cheminformatics, 12(1). https://doi.org/10.1186/s13321-020-00460-5

[26] Rahman, R.; Dhruba, S.R.; Ghosh, S.; Pal, R. Functional random forest with applications in dose-response predictions. Sci. Rep. 2019, 9, 1628.

[27] Pang, X.; Fu, W.; Wang, J.; Kang, D.; Xu, L.; Zhao, Y.; Liu, A.L.; Du, G.H. Identification of Estrogen Receptor α Antagonists from Natural Products via In Vitro and In Silico Approaches. Oxid. Med. Cell. Longev. 2018, 2018, 6040149.

[28] Feinberg, E. N., Joshi, E., Pande, V. S., & Cheng, A. (2020). Improvement in ADMET Prediction with Multitask Deep Featurization. Journal of Medicinal Chemistry, 63(16), 8835–8848. https://doi.org/10.1021/acs.jmedchem.9b02187

[29] Wei, Y.; Li, W.; Du, T.; Hong, Z.; Lin, J. Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method. Int. J. Mol. Sci. 2019, 20, 3572.

[30] Xiong, J., Xiong, Z., Chen, K., Jiang, H., & Zheng, M. (2021). Graph neural networks for automated de novo drug design. Drug Discovery Today, 26(6), 1382–1393. https://doi.org/10.1016/j.drudis.2021.02.011

[31] Kubinyi, H. (1996) Evolutionary variable selection in regression and PLS analyses. J. Chemom. 10, 119–133
[32] Eriksson, L., Jaworska, J., Worth, A., Cronin, M. T. D., McDowell, R., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environmental Health Perspectives, 111(10), 1361–1375. https://doi.org/10.1289/ehp.5758
[33] Gasteiger, J. (2003b). Handbook of Chemoinformatics. In Wiley eBooks. https://doi.org/10.1002/9783527618279
[34] Dehmer, M., Varmuza, K., & Bonchev, D. (2012). Statistical Modelling of Molecular Descriptors in QSAR/QSPR. In Wiley-VCH Verlag GmbH & Co. KGaA eBooks. https://doi.org/10.1002/9783527645121
[35] Lo, Y., Rensi, S. E., Torng, W., & Altman, R. B. (2018b). Machine learning in chemoinformatics and drug discovery. Drug Discovery Today, 23(8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010
[36] Chandrasekaran, B., Abed, S. N., Al-Attraqchi, O., Kuche, K., & Tekade, R. K. (2018). Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In Elsevier eBooks (pp. 731–755). https://doi.org/10.1016/b978-0-12-814421-3.00021-x
[37] Basic overview of chemoinformatics. (2006). PubMed. https://doi.org/10.1021/ci600234z
[38] Ash, J., & 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. https://doi.org/10.1021/acs.jcim.7b00048
[39] Concepts and Experimental Protocols of Modelling and Informatics in Drug Design. (n.d.). ScienceDirect. https://www.sciencedirect.com/book/9780128205464/concepts-and-experimental-protocols-of-modelling-and-informatics-in-drug-design
[40] Machine learning descriptors for molecules. (2021, January 5). ChemIntelligence. https://chemintelligence.com/blog/machine-learning-descriptors-molecules (Accessed on: 14th May 2023)
[41] Bajorath J. (2017). Molecular Similarity Concepts for Informatics Applications. Methods in molecular biology (Clifton, N.J.), 1526, 231–245. https://doi.org/10.1007/978-1-4939-6613-4_13
[42] Sun, H., Tawa, G. J., & Wallqvist, A. (2012). Classification of scaffold-hopping approaches. Drug Discovery Today, 17(7–8), 310–324. https://doi.org/10.1016/j.drudis.2011.10.024
[43] Zheng, S., Lei, Z., Haitao, A., Chen, H., Deng, D., & Yang, Y. (2021). Deep scaffold hopping with multimodal transformer neural networks. Journal of Cheminformatics, 13(1). https://doi.org/10.1186/s13321-021-00565-5
[44] Jenkins, J. L., Glick, M., & Davies, J. (2004). A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. Journal of Medicinal Chemistry, 47(25), 6144–6159. https://doi.org/10.1021/jm049654z
[45] Grisoni, F., Merk, D., Consonni, V., Hiss, J. A., Tagliabue, S. G., Todeschini, R., & Schneider, G. (2018). Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. Communications Chemistry, 1(1). https://doi.org/10.1038/s42004-018-0043-x
[46] Bhattacharjee, H., Burns, J., & Vlachos, D. G. (2023). AIMSim: An accessible cheminformatics platform for similarity operations on chemicals datasets. Computer Physics Communications, 283, 108579. https://doi.org/10.1016/j.cpc.2022.108579
[47] Luo, M., Wang, X. S., & Tropsha, A. (2016). Comparative Analysis of QSAR-based vs. Chemical Similarity Based Predictors of GPCRs Binding Affinity. Molecular informatics, 35(1), 36–41. https://doi.org/10.1002/minf.201500038
[48] Dong, J., Yao, Z., Zhu, M., Wang, N., Lu, B., Chen, A. F., Lu, A., Miao, H., Zeng, W., & Cao, D. (2017). ChemSAR: an online pipelining platform for molecular SAR modeling. Journal of Cheminformatics, 9(1). https://doi.org/10.1186/s13321-017-0215-1
[49] Yoshimori, A., & Bajorath, J. (2020). The SAR Matrix Method and an Artificially Intelligent Variant for the Identification and Structural Organization of Analog Series, SAR Analysis, and Compound Design. Molecular Informatics, 39(12), 2000045. https://doi.org/10.1002/minf.202000045
[50] Hu, H., & Bajorath, J. (2021). Systematic assessment of structure-promiscuity relationships between different types of kinase inhibitors. Bioorganic & Medicinal Chemistry, 41, https://doi.org/10.1016/j.bmc.2021.116226.
[51] Yoshimori, A., Hu, H., & Bajorath, J. (2021). Adapting the DeepSARM approach for dual-target ligand design. Journal of computer-aided molecular design, 35(5), 587–600. https://doi.org/10.1007/s10822-021-00379-5
[52] Lo, Y., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discovery Today, 23(8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010
[53] Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opinion on Drug Discovery, 16(9), 949–959. https://doi.org/10.1080/17460441.2021.1909567

[54] Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. Drug Discovery Today, 26(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010
[55] Priya, S., Kumar, A., Singh, D. B., Jain, P., & Tripathi, G. (2022). Machine learning approaches and their applications in drug discovery and design. Chemical Biology & Drug Design, 100(1), 136–153. https://doi.org/10.1111/cbdd.14057
[56] 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. https://doi.org/10.3389/frai.2019.00017
[57] Ponzoni, I., Sebastián-Pérez, V., Requena-Triguero, C., Roca, C. P., Martínez, M. J., Cravero, F., Díaz, M. P. M., Páez, J. A., Arrayás, R. G., Adrio, J., & Campillo, N. E. (2017). Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-02114-3
[58] Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics, 29(6–7), 476–488. https://doi.org/10.1002/minf.201000061
[59] Tsou, L. K., Yeh, S. H., Ueng, S., Chang, C., Song, J., Wu, M., Chang, H. T., Chen, S., Shih, C., Chen, C., & Ke, Y. (2020). Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-73681-1
[60] Duchowicz, P. R. (2018). Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors. Cells, 7(2), 13. https://doi.org/10.3390/cells7020013
[61] Cardoso-Silva, J., Papageorgiou, L. G., & Tsoka, S. (2019). Network-based piecewise linear regression for QSAR modelling. Journal of Computer-Aided Molecular Design Volume, 33(9), 831–844. https://doi.org/10.1007/s10822-019-00228-6
[62] Dudek, A. Z., Arodz, T., & Galvez, J. (2006). Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. Combinatorial Chemistry & High Throughput Screening, 9(3), 213–228. https://doi.org/10.2174/138620706776055539
[63] Raevsky, O. A., Sapegin, A., & Zefirov, N. S. (1994). The QSAR Discriminant-Regression Model. Quantitative Structure-activity Relationships, 13(4), 412–418. https://doi.org/10.1002/qsar.19940130406
[64] Doreswamy, & Vastrad, B. (2013). Predictive Comparative Qsar Analysis of as 5-Nitrofuran-2-Yl Derivatives Myco Bacterium Tuberculosis H37RV Inhibitors. Healthcare Informatics: An International Journal, 2(4), 47–62. https://doi.org/10.5121/hiij.2013.2404
[65] Ajmani, S., Jadhav, K. M., & Kulkarni, S. A. (2006). Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. Journal of Chemical Information and Modeling, 46(1), 24–31. https://doi.org/10.1021/ci0501286
[66] Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. Nature Biotechnology, 25(2), 197–206. https://doi.org/10.1038/nbt1284
[67] Ajmani, S., Jadhav, K. M., & Kulkarni, S. A. (2006b). Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. Journal of Chemical Information and Modeling, 46(1), 24–31. https://doi.org/10.1021/ci0501286
[68] Raj, N., & Jain, S. (2011). 3d QSAR studies in conjunction with k-nearest neighbor molecular field analysis (k-NN-MFA) on a series of. . . ResearchGate. https://www.researchgate.net/publication/294708142_3d_QSAR_studies_in_conjunction_with_k-nearest_neighbor_molecular_field_analysis_k-NN-MFA_on_a_series_of_substituted_2-phenyl-benzimidazole_derivatives_as_an_anti_allergic_agents
[69] Asikainen, A. H., Ruuskanen, J., & Tuppurainen, K. A. (2004). Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. Environmental science & technology, 38(24), 6724–6729. https://doi.org/10.1021/es049665h
[70] Nigsch, F., Bender, A., Van Buuren, B. N., Tissen, J., Nigsch, E. A., & Mitchell, J. C. (2006). Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. Journal of Chemical Information and Modeling, 46(6), 2412–2422. https://doi.org/10.1021/ci060149f
[71] Poroikov, V. V., Filimonov, D. A., Borodina, Y. V., Lagunin, A. A., & Kos, A. (2000). Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. Journal of chemical information and computer sciences, 40(6), 1349–1355. https://doi.org/10.1021/ci000383k
[72] Chen, B., Sheridan, R. P., Hornak, V., & Voigt, J. H. (2012). Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions. Journal of chemical information and modeling, 52(3), 792–803. https://doi.org/10.1021/ci200615h
[73] Kupervasser, O. (2019). Quantitative Structure-Activity Relationship Modeling and Bayesian Networks: Optimality of Naive Bayes Model. In IntechOpen eBooks. https://doi.org/10.5772/intechopen.85976

[74] Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). Choosing Feature Selection and Learning Algorithms in QSAR. Journal of Chemical Information and Modeling, 54(3), 837–843. https://doi.org/10.1021/ci400573c
[75] Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., & Davies, J. W. (2006). "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept?. Journal of chemical information and modeling, 46(6), 2445–2456. https://doi.org/10.1021/ci600197y
[76] Keyvanpour, M. R., & Shirzad, M. B. (2021). An Analysis of QSAR Research Based on Machine Learning Concepts. Current drug discovery technologies, 18(1), 17–30. https://doi.org/10.2174/1570163817666200316104404
[77] Bugeac, C. A., Ancuceanu, R., & Dinu, M. (2021). QSAR Models for Active Substances against Pseudomonas aeruginosa Using Disk-Diffusion Test Data. Molecules, 26(6), 1734. https://doi.org/10.3390/molecules26061734
[78] Darnag, R., Schmitzer, A. R., Belmiloud, Y., Villemin, D., Jarid, A., Chait, A., Seyagh, M., & Cherqaoui, D. (2009). QSAR Studies of HEPT Derivatives Using Support Vector Machines. Qsar & Combinatorial Science, 28(6–7), 709–718. https://doi.org/10.1002/qsar.200810166
[79] Niu, B., Lu, W., Yang, S., Cai, Y., & Li, G. (2007). Support vector machine for SAR/QSAR of phenethylamines. Acta Pharmacologica Sinica, 28(7), 1075–1086. https://doi.org/10.1111/j.1745-7254.2007.00573.x
[80] Wu, Z., Zhu, M., Kang, Y., Leung, E. L., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D., & Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Briefings in Bioinformatics, 22(4). https://doi.org/10.1093/bib/bbaa321
[81] Alvarsson, J., Lampa, S., Schaal, W., Andersson, C., Wikberg, J. E. S., & Spjuth, O. (2016). Large-scale ligand-based predictive modelling using support vector machines. Journal of Cheminformatics, 8(1). https://doi.org/10.1186/s13321-016-0151-5
[82] Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., & Fan, B. T. (2003). QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF-kappa B mediated gene expression based on support vector machines. Journal of chemical information and computer sciences, 43(4), 1288–1296. https://doi.org/10.1021/ci0340355
[83] Nekoei, M., Mohammadhosseini, M., & Pourbasheer, E. (2015). QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach. Medicinal Chemistry Research, 24(7), 3037–3046. https://doi.org/10.1007/s00044-015-1354-4
[84] Torng, W., & Altman, R. B. (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics, 18(1). https://doi.org/10.1186/s12859-017-1702-0
[85] Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics, 9(1). https://doi.org/10.1186/s13321-017-0235-x
[86] Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech Recognition with Deep Recurrent Neural Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1303.5778
[87] Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Publications, 4(1), 120–131. https://doi.org/10.1021/acscentsci.7b00512
[88] Kingma, D. P. (2013, December 20). Auto-Encoding Variational Bayes. arXiv.org. https://arxiv.org/abs/1312.6114
[89] Goodfellow, I. J. (2014, June 10). Generative Adversarial Networks. arXiv.org. https://arxiv.org/abs/1406.2661
[90] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M. F., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236
[91] Kusner, M. J. (2017, March 6). Grammar Variational Autoencoder. arXiv.org. https://arxiv.org/abs/1703.01925
[92] Matsuzaka, Y., & Uesawa, Y. (2019). Optimization of a Deep-Learning Method Based on the Classification of Images Generated by Parameterized Deep Snap a Novel Molecular-Image-Input Technique for Quantitative Structure–Activity Relationship (QSAR) Analysis. Frontiers in Bioengineering and Biotechnology, 7. https://doi.org/10.3389/fbioe.2019.00065
[93] Karpov, P., Godin, G., & Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. Journal of Cheminformatics, 12(1). https://doi.org/10.1186/s13321-020-00423-w

 doi:10.20944/preprints202306.0803.v1

[94] Xu, Y. (2023). Development and Evaluation of Conformal Prediction Methods for QSAR. arXiv.org. https://arxiv.org/abs/2304.00970

[95] Shayanfar, S., & Shayanfar, A. (2022). Comparison of various methods for validity evaluation of QSAR models. BMC chemistry, 16(1), 63. https://doi.org/10.1186/s13065-022-00856-4

[96] Shayanfar, S., Shayanfar, A. Comparison of various methods for validity evaluation of QSAR models. BMC Chemistry 16, 63 (2022). https://doi.org/10.1186/s13065-022-00856-4

[97] Lo, Y., Rensi, S. E., Torng, W., & Altman, R. B. (2018c). Machine learning in chemoinformatics and drug discovery. Drug Discovery Today, 23(8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010

[98] Golbraikh, A., Wang, X., Zhu, H., & Tropsha, A. (2017). Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment. In Springer eBooks (pp. 2303–2340). https://doi.org/10.1007/978-3-319-27282-5_37

[99] Spiegel, J., & Senderowitz, H. (2020). Evaluation of QSAR Equations for Virtual Screening. International journal of molecular sciences, 21(21), 7828. https://doi.org/10.3390/ijms21217828

[100] Matveieva, M., & Polishchuk, P. G. (2021). Benchmarks for interpretation of QSAR models. Journal of Cheminformatics, 13(1). https://doi.org/10.1186/s13321-021-00519-x

[101] C3.ai. (2022). LIME: Local Interpretable Model-Agnostic Explanations. C3 AI. https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/#:~:text=What%20is%20Local%20Interpretable%20Model,to%20explain%20each%20individual%20prediction.

[102] Molnar, C. (2023, March 2). 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/shap.html

[103] Izrailev, S., & Agrafiotis, D. (2004). A method for quantifying and visualizing the diversity of QSAR models. Journal of Molecular Graphics & Modelling, 22(4), 275–284. https://doi.org/10.1016/j.jmgm.2003.10.001

[104] An, Y., Sherman, W., & Dixon, S. L. (2013). Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. Journal of Chemical Information and Modeling, 53(9), 2312–2321. https://doi.org/10.1021/ci400250c