International Journal of
Molecular Sciences

MDPI

Review

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

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

[1] College of Pharmacy, University of Illinois, Chicago, IL 61820, USA
[2] Zamara Mariam, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST), Islamabad 24090, Pakistan; zmariam.msbi21rcms@student.nust.edu.pk
* Correspondence: niazi@niazi.com; Tel.: +1-312-297-0000

Abstract: In modern drug discovery, the combination 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 structure–activity relationships (SARs), are pivotal 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. Due to 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; chemoinformatics; small molecules; AI/ML; molecular descriptors; biological activity; SAR; predictive modeling; computational validation

Citation: Niazi, S.K.; Mariam, Z. Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. *Int. J. Mol. Sci.* **2023**, *24*, 11488. https://doi.org/10.3390/ijms241411488

Academic Editor: Hanoch Senderowitz

Received: 9 June 2023
Revised: 30 June 2023
Accepted: 12 July 2023
Published: 15 July 2023

1. Introduction

In 1998, the term "chemoinformatics", coined by Frank K. Brown, was intended to hasten drug discovery and development; however, now, chemoinformatics is crucial in biology, chemistry, and biochemistry. The general process of drug discovery took 12 to 15 years and involved investments of around $500 million in 1998. New developments in machine learning (ML) and artificial intelligence (AI) have revolutionized chemoinformatics and drug discovery to a great degree. Market revenue for small-molecule drug discovery was $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, allowing computers to analyze data and draw conclusions and 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 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. This review discusses the current state of research, the potential

integration of ML-driven chemoinformatics tools, techniques in drug discovery, and the challenges and limitations of using 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.

2. 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 increasing accessibility 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. Compared to a few decades ago, it has expedited the process by many folds. It focuses on molecular engineering, molecular manipulation, library design, compound database searching, chemical space exploration, molecular graph mining, pharmacophore, and scaffold analysis [5–9].

3. 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 searching, 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.

3.1. 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 enable searching for specific molecules or analyze large chemical datasets. 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 techniques based on structural similarities can be leveraged to convey knowledge regarding the mechanism. Various similarity scores, as mentioned before, can be computed, considering the similarity of 1D structures (e.g., SMILES- or SELFIES-based similarity [17]), 2D structures (e.g., 2D fingerprints or topological similarity), and even 3D structures (e.g., 3D geometric shape-based similarity). Previous studies have identified several metrics suitable for molecular similarity calculations, including the Tanimoto index, Manhattan distance, Dice index, overlap coefficient, cosine coefficient, and Soergel distance [18–20]. Furthermore, chemical bioactivity and structural data can be acquired from drug databases like ChEMBL [21], BindingDB [22], DrugBank [23], Inxight [24], and Protein Data Bank [25]. Despite the availability of extensive databases, utilizing machine learning and deep learning techniques offers significant potential to enhance the creation of molecules and focused libraries, enabling the discovery of potent bioactive compounds through targeted design and generation strategies in QSAR studies.

Generative models like recurrent neural networks (RNN) have been employed to generate novel chemical structures predicted to have desirable properties, such as high potency or low toxicity. RNN models have been previously used to generate focused

molecule libraries and have implicitly learned chemical knowledge to create molecules with combined characteristics of both bioactive natural products and synthetic compounds, such as DeepMGM. Besides this, generative models have been used for inverse QSAR/QSPR, which involves generating molecules that meet specific target properties.

The DeepMGM model was trained using drug-like molecules and produced a general model (g-DeepMGM) capable of generating scaffold-focused libraries. A target-specific model (t-DeepMGM) for the cannabinoid receptor 2 (CB2) using transfer learning was also developed. A discriminator was incorporated into DeepMGM for in silico molecular design and testing. The generated molecule XIE9137 was identified as a potential CB2 allosteric modulator, highlighting the effectiveness of deep learning in de novo molecular design and chemical library generation [26,27].

3.2. Chemical Data Representation

Advancements in ML modeling and the availability of a vast pool of chemical and biological data have led to a dire need for data to be translated into computer-understandable form before models are trained on them. Chemical data representation can be empirical, molecular, and structural data represented in molecular graphs, fingerprints, descriptors, etc. [28,29]. A multivariate random forest model generated for genomic characterization was trained on genomic sequencing data given in numerical representation in one study [30]. In another, a Naïve Bayesian (NB) model was developed on numeric-based activity data, representing antagonists' binding on estrogen receptors [31]. An ML-based model was trained on 31 chemical numerical datasets obtained from Merck to predict the properties of small compounds based on ADMET (absorption, distribution, metabolism, excretion, and toxicity) [32]. Similarly, molecular fingerprint data have 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 [33]. Furthermore, the graph neural networks (GNNs) function with the graph structure data of 3D molecules and have been used to identify potential drug molecules [34]. 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 addressed 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. The principal components analysis (PCA) and the partial least squares (PLS) methods generally transform massive datasets with correlated variables into smaller 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. The original data matrix from a different investigation was examined using PCA, in which molecules are represented by several predictor variables (molecular descriptors). PCA has also been used to design features for estrogen receptor binding prediction. Furthermore, observations revealed enhanced performance in therapeutic activity predictions against a diverse range of pharmacological protein targets identified by the kernel–principal components (kernel-

PCA) analysis and a nonlinear PCA variation, surpassing the predictive capabilities of LASSO regression.

Similarly, the partial least squares (PLS) method has 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 [6,35,36].

3.3. Molecular Descriptors

Molecular descriptors are quantifiable representations that capture chemical compounds' structural, physicochemical, and biological properties. These descriptors are quantitative measures used for similarity analysis, virtual screening, and predictive modeling. Chemical molecular descriptors are categorized as 0D, 1D, 2D, 3D, and 4D (Table 1) [37–40].

- 0D Descriptors: These are constitutional or count descriptors, scalar values that describe several atoms, bonds, or functional groups in the molecule, e.g., molecular weight.
- 1D Descriptors: These descriptors capture molecular properties in one dimension along a linear sequence or chain of atoms, e.g., structural fragments or fingerprints.
- 2D Descriptors: These descriptors provide information about the structure on a molecular level and its properties within a 2D plane, e.g., 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, e.g., autocorrelation descriptors, substituent constants, surface:volume descriptors, quantum, chemical descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, size, steric, surface, and volume descriptors.
- 4D Descriptors: These descriptors encompass properties that change over time or involve spatiotemporal aspects, e.g., drug dissolution rate, Volsurf, and GRID or CoMFA methods.

Table 1. The most common 0D to 4D chemical descriptors for QSAR/QSPR analysis.

Descriptor Dimension	Descriptor Type	Example
0D	The molecule's atoms, bonds, and functional groups count	Molecular weight, LogP (partition coefficient)
1D	Molecular properties in a linear manner	Molecular Formula, SMILES & SELFIES
2D	Topological polar surface area (TPSA)	Molecular fingerprint (e.g., Morgan fingerprint), Constitutional descriptors (e.g., atoms, bonds, and rings count)
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, solvent accessible surface area (SASA), radius of gyration (Rg), 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 used to determine the most informative molecular descriptors for QSAR studies. Molecular descriptors allow for ligand-based scaffold hopping for hit and lead optimization, which speeds up the early stages of drug development and has been used to compare QSAR and QSPR models. 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. Furthermore, 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 [41–43].

4. QSAR

Based on its physicochemical characteristics, a ligand's biological response or activity can be predicted using QSAR analysis [38]. QSAR modeling techniques have been used to find prospective drug candidates, and these have developed into AI-based QSAR methods [44]. Modern machine learning approaches can be applied to model QSAR or quantitative structure–property relationships (QSPR) and create predicative models based on artificial intelligence [45,46]. Chemoinformatics, 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, e.g., chemical data, descriptors, molecular graphs, fingerprints, similarity analyses, and molecular dynamic simulations, to predict the most optimal properties of a potential drug.

Structure–activity relationship (SAR) analysis investigates how the chemical structure of a compound relates to its biological activity or properties and plays a crucial role in exploring potential effects of bioactivity on changes in the chemical structure of drugs. Quantifying the degree of the structural or chemical similarity between molecules and extrapolating chemical attributes from molecular similarity are the goals of similarity analysis [47]. Similarity search mainly aims to identify 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. Fragment replacement approaches, fingerprint-based similarity search, pharmacophore matching, and 3D shape-based similarity search are all examples of computational research for scaffold hops. Designing molecules of novel scaffolds with increased pharmacological activity and identical 3D structure but a multimodal deep transformer neural technique for scaffold hopping aids the distinct 2D structure [48–51].

SARs are also employed in clustering, inter-molecular comparisons, outlier and novelty analysis, diversity quantification, and outlier analysis. Molecular datasets are used by AIMSim, a unified platform, to carry out similarity-based tasks using binary similarity metrics and molecular fingerprints [52]. In a study, a tool called the similarity ensemble approach (SEA) was used to estimate the accuracy of k-nearest neighbors (kNN) QSAR models constructed for known ligands of each GPCR target individually to discover active and inactive molecules [53]. ChemSAR, another tool, offers an integrated web-based platform for creating SAR classification models, and it is also an online pipelining platform for molecular SAR modeling. For the identification and structural organization of analog series, SAR analysis, and compound design, various researchers have employed the SAR Matrix (SARM) concept [54–56]. DeepSARM, which combined deep learning and generative modeling, was introduced, expanding the scope of the SARM technique. This improvement made it possible to create target-based analogs by considering the chemical information from related targets to increase structural uniqueness and variety [57].

The current approach to 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 for 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, specifically long short-term memory neural networks [58].

4.1. 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.

4.2. Molecular Encoding

Molecular encoding is like chemical data representation transformation, as discussed before. Compounds' chemical characteristics and attributes are directly deduced from their chemical structures or by looking up experimental findings. This process involves extracting relevant information from the molecular structure, such as atom types, bond types, functional groups, and physicochemical properties.

4.3. 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. To determine the most pertinent attributes and lessen the dimensionality and collinearity of the feature vector, hybrid feature selection, feature learning methodologies, and unsupervised learning techniques are applied. These techniques have successfully preserved a fair computing effort without reducing the precision of the final QSAR models [59].

4.4. 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 [60].

5. Machine-Learning-Based QSAR Modeling

Unsupervised learning and supervised learning are two categories of machine learning models. In supervised learning, a model is trained with labeled data to produce predictions based on known input–output correlations (for example, support vector machines and linear regression). 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 neighbors, support vector machine, or deep neural network (DNN).

The hit-to-lead optimization and hit-to-hit identification can be performed using ML-based QSAR models. The automated hit identification and optimization tool (A-HIOT), among other sophisticated virtual screening frameworks, can be employed to find and improve hits for fixed protein receptors. A-HIOT uses numerous open-source methods to combine chemical and protein space and produce high-quality predictions [61]. To show that deep neural networks (DNN) and random forests (RF) were superior in hit prediction efficiency, comparison studies between DNN and other ligand-based virtual screening (LBVS) approaches were conducted. A scan of an in-house library of 165,000 chemicals revealed numerous triple-negative breast cancer (TNBC) inhibitors as powerful hits using DNN. From a small-size training set of 63 compounds, a potent mu-opioid receptor (MOR)

GPCR proteins agonist was found. One hundred seventy-six possible antimalarial hits were found by integrating QSAR and virtual screening [62,63].

5.1. Regression Analysis

Regression analysis is a statistical technique for simulating the relationship between a dependent variable and one or more independent variables. It seeks to identify the line with the best fit that minimizes the sum of the squared residuals. The relationship between variables can be inferred by estimating the regression equation coefficients. Early QSAR techniques like Hansch and Free Wilson analysis heavily utilized 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.

By fusing aspects of network analysis and piecewise linear regression, interpretable QSAR models have been created using network-based linear regression. In a study on inhibitors of polo-like kinase-1 and linear regression, to find prediction models of an extensive and structurally varied dataset of 530 chemicals, QSAR models were created. A discriminant–regression model (DIREM), a common discrete–continuous QSAR technique, is another form of regression model that combines discriminant and regression studies to investigate structure–activity connections for substances. The effectiveness of PLS-based QSAR models was assessed, and they were compared with the outcomes of multiple linear regression (MLR) and principal component regression (PCR) in a comparative analysis study on 5-nitrofuran-2-yl derivatives as inhibitors of Mycobacterium TB H37Rv. Compared to PCR, the findings of the PLS and MLR analyses demonstrated significantly higher predictive power and reliability, attesting to the dependability of these techniques [64–68]. Although numerous medication optimization studies have successfully used linear regression analysis and its derivates, there are still substantial drawbacks, including underlying linearity, overfitting, restricted interpretability, the necessity of high-quality data, and false vector space assumptions.

5.2. K-Nearest Neighbor

The k-nearest neighbors (kNN) algorithm represents labeled and unlabeled data nodes in a multidimensional feature space. The k-nearest neighbors (kNN) methodology is a straightforward distance-learning strategy in which an unknown member is categorized based on 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 [69].

Choosing the right number of nearest neighbors to utilize in the kNN algorithm can be challenging because doing so can lead to unfavorable false-positive or false-negative rates. This was addressed by introducing the similarity ensemble approach (SEA), which proved to be a more organized method for determining the right number of neighbors in kNN analysis. The SEA compares chemical similarity values to a randomized background score, similar to the BLAST sequence similarity search method [70].

A study developed a 3D QSAR model for 30 drugs with anti-HIV activity using a kNN model. According to most of the training set's nearest neighbors, this kNN model categorized the chemicals. The technique determined the essential structural characteristics that lead to the compounds' anti-HIV efficacy [69]. Another study developed QSAR models for 50 compounds with an anti-HIV activity using the kNN–molecular field analysis method. The results showed the importance of electrostatic and steric interactions in influencing the anti-HIV activity of the compounds [71]. Consensus kNN QSAR was used in a study and proved to be a practical method for quickly screening the estrogenic activity of organic compounds. It is a flexible method for predicting the estrogenic activity of organic compounds in silico [72]. Utilizing a deep neural network in conjunction with the kNN approach to creating QSAR models for a collection of 1000 chemicals having anticancer activity was also claimed to be helpful. According to the study, the main structural

characteristics contributing to the compounds' anti-cancer efficacy could be determined using the kNN approach [73].

5.3. Naïve Bayes

Naïve Bayes is a probabilistic classifier commonly assuming that features are independent, simplifying the modeling process. It determines 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 [74].

On 18 sizable, varied in-house QSAR datasets, a study examined the capacity of Pipeline Pilot Naïve Bayes (PLPNB) and random forest to produce precise predictions. According to the study, PLPNB could predict binary and multicategory activities with accuracy and was computationally efficient. Large-scale virtual screening for important pharmacological features, such as cytochrome P450 inhibition, human plasma protein binding, and animal model bioavailability, have demonstrated their effectiveness [75,76]. Another study showed that when used in QSAR modeling, the Naïve Bayes model delivers the lowest mean error when the data points are distributed uniformly. This study uses QSAR as an example to demonstrate the Naïve Bayes model's optimality [77]. 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 [78].

5.4. 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 framework known as "ML-QSAR" was established in a study in which machine learning methods were used for QSAR modeling. SVM was discovered to be one of the most often used machine learning algorithms in QSAR modeling. The framework was created to facilitate the selection of appropriate strategies among existing algorithms according to the application area requirements and to help develop and improve current approaches [79].

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 accurately compared to other models [80]. Another study examined SVM's effectiveness and prognostication power in HEPT derivative QSAR modeling. This investigation showed that SVM outperformed different approaches, including artificial neural networks, in terms of prediction [81]. SVM was also used to simulate phenethylamines' structure–activity relationships (SAR). To categorize antagonists and agonists and forecast their effects, the study used SVM, which it discovered to be a reliable method in the SAR/QSAR field [82].

Another study evaluated the effectiveness of 16 machine learning algorithms, including SVM, on 14 QSAR datasets and concluded that various ML algorithms offered different QSAR modeling approaches to uncover the connections between compound structures and properties [83]. 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 [84]. 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 [85]. To determine the structural elements that give aminopyrimidine-5-carbaldehyde oxime derivatives a potent vascular endothelial growth factor (VEGF)-2 inhibitory action, 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 [86].

5.5. 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 topologies and ensemble methods have been extensively used in modeling QSAR/QSPR features of small compounds and conducting pharmacokinetic and pharmacodynamic studies, alongside work in other fields of chemoinformatics. In particular, CNN's unmatched capacity for image analysis made it possible to visualize protein structures as '3D images' with four separate atom-type channels. These 3D-CNNs were used to compare the microenvironments of amino acids and predict how mutations might affect the structure of proteins [87]. A Transformer–CNN architecture was suggested in a study for QSAR modeling and interpretation. Convolutional and element-wise feed-forward layers were used in place of all recurrent units in the design, and it was discovered that the Transformer-CNN architecture produced good results for small datasets and converged quickly for QSAR tasks [88].

Recurrent neural networks (RNNs), also known as long short-term memory (LSTM) networks, are built to recognize both short-term and long-term dependencies in sequential input. For applications like de novo drug design, where they learn the structural patterns and rules from SMILES strings to produce novel molecules, LSTM networks have been used in the context of QSAR. Deep reinforcement learning, variational autoencoders, and generative adversarial networks (GANs) are other cutting-edge methods used to generate compounds with precise molecular features while learning latent representations of molecules. These methods aid in the discovery of novel medication candidates and the exploration of new chemical territory [27,89–93]. 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 [94]. To perform QSAR analysis utilizing three-dimensional photographs of chemical structures, a brand-new DL-based method dubbed DeepSnap was created. Without extracting descriptors, this method may also forecast the potential toxicity of many compounds to different receptors. To perform QSAR analysis utilizing three-dimensional photographs of chemical structures, a brand-new DL-based method dubbed DeepSnap was created. Without extracting descriptors, this method may also forecast the potential toxicity of many compounds to different receptors [95]. CNN, RNN, and deep-learning-based methods have also shown promising results in QSAR modeling.

6. 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 primary method for evaluating the accuracy of generated models for the activity prediction of compounds that have not yet been synthesized. Understanding the variables that control molecular characteristics and creating new compounds with advantageous features depend on QSAR models, which provide information on the association between activities and structure-based molecular descriptors [96,97].

Even though 3D-QSAR techniques like CoMFA take structural conformation into account, they are computationally intensive and can introduce errors related to conformation prediction, ligand orientation, and structural alignment. Consequently, 2D-QSAR models can provide a viable alternative and sometimes even outperform 3D-QSAR strategies [38]. The creation of verified models for accurate and precise prediction of a compound's biological actions is the ultimate goal of QSAR analysis. When creating QSAR models, metrics like R2 and QCV2 are generally optimized. The performances of the final models are assessed using comparable metrics computed on external datasets [98].

A comparative study on 5-nitrofuran-2-yl derivatives as inhibitors of Mycobacterium tuberculosis H37Rv used statistical parameters, including squared correlation coefficients, cross-validated correlation coefficients, and Fischer's value for statistical importance, to assess the quality of the generated QSAR models. Another study examined several statistical parameters of 44 published QSAR models for biologically active substances that were externally validated and presented in academic journals. They concluded that using the coefficient of determination (R2) alone was insufficient to determine if a QSAR model was viable. There are benefits and drawbacks to these defined criteria for external validation that should be considered in QSAR investigations [99].

7. Interpretability and Explainability of ML-QSAR Models

The creation of verified models for accurate and precise prediction of a compound's biological actions is the ultimate goal of QSAR analysis. The interpretability and explainability of ML-QSAR models promote transparency, reproducibility, and trust in the models' predictions, allowing researchers and stakeholders to make informed decisions regarding drug discovery and development. Six artificial datasets of varying degrees of complexity were produced as part of a study to compare various QSAR model interpretation techniques. These datasets were used in the study's investigation of a wide range of descriptor and algorithm pairings and the Structure–Property Correlation Index (SPCI) method of universal interpretation. The study showed that predictivity might decline more quickly than interpretation performance and that even models with good predictivity may occasionally have subpar 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. 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 new way to visualize QSAR models is described in a publication that streamlines analysis by adding a new measure of model similarity. The method relies on projecting models into a two-dimensional plane, where the separation between two models is proportional to the variation in their expected activities [103]. Another study creates predicted QSAR models that may be projected onto the atoms of a molecule by combining direct kernel-based PLS with Canvas 2D fingerprints. The work offers a model visualization that can be used to determine which atoms are most important for forecasting activity [104].

Being unable to explain why a neural network generates a prediction is a significant impediment to the application of AI models due to the 'black box' approach. In addition to discouraging chemists from utilizing deep learning predictions, this has caused neural networks to pick up undetectable bogus correlations. Counterfactual interpretation is a

technique for reading ML models that can be used to comprehend why a model generates a specific prediction. Counterfactuals are local interpretations that can disclose the contributions of atoms or fragments within particular molecules to identify the most beneficial or detrimental motifs to consider for future alterations in the context of QSAR models. Because they resemble counterfactuals, instance-based techniques have been claimed to provide 'natural' model interpretations for researchers. Various approaches to interpretation have been established; however, there are no appropriate standards to assess how well they apply to the interpretation of QSAR models. An approach known as STONED (Structure–Topology Optimization for Novel Explanatory Discoveries) is suggested in a study; it produces molecular counterfactuals for any model. These molecular counterfactuals offer skeletal, molecular structure-based explanations. All molecules produced by STONED are legitimate substances, so the method does not require training a counterfactual generator. This simplifies the procedure and eliminates the necessity of a generative counterfactual creator [100,105,106].

8. 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, enabling efficient exploration of chemical space and predicting biological activities. Multiple algorithms built for QSAR modeling significantly highlight 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 to accelerate the drug discovery process, reduce costs, and increase 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.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Small Molecule Drug Discovery Market Size, Report by 2032. Available online: https://www.precedenceresearch.com/small-molecule-drug-discovery-market (accessed on 24 May 2023).
2. Brown, F.K. Chapter 35—Chemoinformatics: What is it and How does it Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*; Bristol, J.A., Ed.; Academic Press: New York, NY, USA, 1998; Volume 33, pp. 375–384. [CrossRef]
3. Polanski, J. 4.26-Chemoinformatics. In *Comprehensive Chemometrics*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 2020; pp. 635–676. [CrossRef]
4. Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*, 151. [CrossRef]
5. Polanski, J. 4.14-Chemoinformatics. In *Comprehensive Chemometrics*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 459–506. [CrossRef]
6. Gasteiger, J. *Handbook of Chemoinformatics*; Wiley: New York, NY, USA, 2003. [CrossRef]
7. Varnek, A.; Baskin, I.I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inform.* **2011**, *30*, 20–32. [CrossRef]

8. Bajorath, J.; Bajorath, J. (Eds.) Chemoinformatics and Computational Chemical Biology. In *Methods in Molecular Biology*; Springer Science+Business Media: Humana Totowa, NJ, USA, 2011. [CrossRef]
9. Kapetanovic, I.M. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chem.-Biol. Interact.* **2008**, *171*, 165–176. [CrossRef]
10. Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J.G.; Stephan, R.; Page, R.; Vondrášek, J.; et al. The LOTUS initiative for open natural products research: Knowledge management through Wikidata. *bioRxiv* **2021**. [CrossRef]
11. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20. [CrossRef] [PubMed]
12. Banerjee, P.; Erehman, J.; Gohlke, B.O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II—A database of natural products. *Nucleic Acids Res.* **2015**, *43*, D935–D939. [CrossRef]
13. Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **2018**, *46*, D1217–D1222. [CrossRef] [PubMed]
14. Wu, Y.; Zhang, F.; Yang, K.; Fang, S.; Bu, D.; Li, H.; Sun, L.; Hu, H.; Gao, K.; Wang, W.; et al. SymMap: An integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Res.* **2019**, *47*, D1110–D1117. [CrossRef] [PubMed]
15. Ru, J.; Li, P.; Wang, J.; Zhou, W.; Li, B.; Huang, C.; Li, P.; Guo, Z.; Tao, W.; Yang, Y.; et al. TCMSP: A database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform.* **2014**, *6*, 13. [CrossRef] [PubMed]
16. Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res.* **2012**, *41*, D1089–D1095. [CrossRef]
17. Krenn, M.; Aspuru-Guzik, A.; Nigam, A.; Friederich, P. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *arXiv* **2020**, *1*, 045024. [CrossRef]
18. Engel, T.; Gasteiger, J. (Eds.) *Chemoinformatics: Basic Concepts and Methods*; Wiley: New York, NY, USA, 2018. Available online: https://www.wiley.com/en-dk/Chemoinformatics:+Basic+Concepts+and+Methods-p-9783527331093 (accessed on 7 May 2023).
19. Xue, H.; Stanley-Baker, M.; Kong, A.W.K.; Li, H.; Goh, W.W.B. Data considerations for predictive modeling applied to the discovery of bioactive natural products. *Drug Discov. Today* **2022**, *27*, 2235–2243. [CrossRef] [PubMed]
20. Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity—A Review. *Qsar Comb. Sci.* **2003**, *22*, 1006–1026. [CrossRef]
21. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940. [CrossRef]
22. Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045–D1053. [CrossRef]
23. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]
24. Siramshetty, V.B.; Grishagin, I.; Nguyễn, Đ.T.; Peryea, T.; Skovpen, Y.; Stroganov, O.; Katzel, D.; Sheils, T.; Jadhav, A.; Mathé, E.A.; et al. NCATS Inxight Drugs: A comprehensive and curated portal for translational research. *Nucleic Acids Res.* **2022**, *50*, D1307–D1316. [CrossRef]
25. Sussman, J.L.; Lin, D.; Jiang, J.; Manning, N.O.; Prilusky, J.; Ritter, O.; Abola, E.E. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1998**, *54*, 1078–1084. [CrossRef]
26. Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2020**, *2*, 171–180. [CrossRef]
27. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Publ.* **2017**, *4*, 120–131. [CrossRef] [PubMed]
28. Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, *6*, 1527–1542. [CrossRef] [PubMed]
29. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: A review and practical guide. *J. Cheminform.* **2020**, *12*, 56. [CrossRef] [PubMed]
30. Rahman, R.; Dhruba, S.R.; Ghosh, S.; Pal, R. Functional random forest with applications in dose-response predictions. *Sci. Rep.* **2019**, *9*, 1628. [CrossRef] [PubMed]
31. 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. [CrossRef] [PubMed]
32. Feinberg, E.N.; Joshi, E.; Pande, V.S.; Cheng, A. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848. [CrossRef]
33. 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. [CrossRef] [PubMed]
34. Xiong, J.; Xiong, Z.; Chen, K.; Jiang, H.; Zheng, M. Graph neural networks for automated de novo drug design. *Drug Discov. Today* **2021**, *26*, 1382–1393. [CrossRef]
35. Kubinyi, H. Evolutionary variable selection in regression and PLS analyses. *J. Chemom.* **1996**, *10*, 119–133. [CrossRef]

36. Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.T.D.; McDowell, R.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375. [CrossRef]
37. Dehmer, M.; Varmuza, K.; Bonchev, D. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2012. [CrossRef]
38. Lo, Y.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [CrossRef]
39. Chandrasekaran, B.; Abed, S.N.; Al-Attraqchi, O.; Kuche, K.; Tekade, R.K. *Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 731–755. [CrossRef]
40. Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277. [CrossRef]
41. Ash, J.; Fourches, D. Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *J. Chem. Inf. Model.* **2017**, *57*, 1286–1299. [CrossRef] [PubMed]
42. Concepts and Experimental Protocols of Modelling and Informatics in Drug Design. ScienceDirect. Available online: https://www.sciencedirect.com/book/9780128205464/concepts-and-experimental-protocols-of-modelling-and-informatics-in-drug-design (accessed on 24 May 2023).
43. Machine Learning Descriptors for Molecules. ChemIntelligence. 5 January 2021. Available online: https://chemintelligence.com/blog/machine-learning-descriptors-molecules (accessed on 14 May 2023).
44. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opin. Drug Discov.* **2021**, *16*, 949–959. [CrossRef] [PubMed]
45. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80–93. [CrossRef]
46. Priya, S.; Kumar, A.; Singh, D.B.; Jain, P.; Tripathi, G. Machine learning approaches and their applications in drug discovery and design. *Chem. Biol. Drug Des.* **2022**, *100*, 136–153. [CrossRef]
47. Bajorath, J. Molecular Similarity Concepts for Informatics Applications. *Methods Mol. Biol.* **2017**, *1526*, 231–245. [CrossRef]
48. Sun, H.; Tawa, G.J.; Wallqvist, A. Classification of scaffold-hopping approaches. *Drug Discov. Today* **2012**, *17*, 310–324. [CrossRef]
49. Zheng, S.; Lei, Z.; Haitao, A.; Chen, H.; Deng, D.; Yang, Y. Deep scaffold hopping with multimodal transformer neural networks. *J. Cheminform.* **2021**, *13*, 87. [CrossRef] [PubMed]
50. Jenkins, J.L.; Glick, M.; Davies, J. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144–6159. [CrossRef]
51. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J.A.; Tagliabue, S.G.; Todeschini, R.; Schneider, G. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **2018**, *1*, 44. [CrossRef]
52. Bhattacharjee, H.; Burns, J.; Vlachos, D.G. AIMSim: An accessible cheminformatics platform for similarity operations on chemicals datasets. *Comput. Phys. Commun.* **2023**, *283*, 108579. [CrossRef]
53. Luo, M.; Wang, X.S.; Tropsha, A. Comparative Analysis of QSAR-based vs. Chemical Similarity Based Predictors of GPCRs Binding Affinity. *Mol. Inform.* **2016**, *35*, 36–41. [CrossRef] [PubMed]
54. Dong, J.; Yao, Z.; Zhu, M.; Wang, N.; Lu, B.; Chen, A.F.; Lu, A.; Miao, H.; Zeng, W.; Cao, D. ChemSAR: An online pipelining platform for molecular SAR modeling. *J. Cheminform.* **2017**, *9*, 27. [CrossRef]
55. Yoshimori, A.; Bajorath, J. The SAR Matrix Method and an Artificially Intelligent Variant for the Identification and Structural Organization of Analog Series, SAR Analysis, and Compound Design. *Mol. Inform.* **2020**, *39*, 2000045. [CrossRef] [PubMed]
56. Hu, H.; Bajorath, J. Systematic assessment of structure-promiscuity relationships between different types of kinase inhibitors. *Bioorganic. Med. Chem.* **2021**, *41*, 116226. [CrossRef] [PubMed]
57. Yoshimori, A.; Hu, H.; Bajorath, J. Adapting the DeepSARM approach for dual-target ligand design. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 587–600. [CrossRef] [PubMed]
58. Chakravarti, S.K.; Alla, S.R.M. Descriptor Free QSAR Modeling Using Deep Learning with Long Short-Term Memory Neural Networks. *Front. Artif. Intell.* **2019**, *2*, 17. [CrossRef]
59. Ponzoni, I.; Sebastián-Pérez, V.; Requena-Triguero, C.; Roca, C.P.; Martínez, M.J.; Cravero, F.; Díaz MP, M.; Páez, J.A.; Arrayás, R.G.; Adrio, J.; et al. Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Sci. Rep.* **2017**, *7*, 2403. [CrossRef]
60. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488. [CrossRef]
61. Kumar, N.; Acharya, V. Machine intelligence-driven framework for optimized hit selection in virtual screening. *J. Cheminform.* **2022**, *14*, 48. [CrossRef]
62. Tsou, L.K.; Yeh, S.H.; Ueng, S.; Chang, C.; Song, J.; Wu, M.; Chang, H.T.; Chen, S.; Shih, C.; Chen, C.; et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Sci. Rep.* **2020**, *10*, 16771. [CrossRef] [PubMed]
63. Neves, B.M.; Braga, R.C.; Melo-Filho, C.C.; Moreira-Filho, J.T.; Muratov, E.N.; Andrade, C.H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **2018**, *9*, 1275. [CrossRef] [PubMed]
64. Duchowicz, P.R. Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors. *Cells* **2018**, *7*, 13. [CrossRef] [PubMed]
65. Cardoso-Silva, J.; Papageorgiou, L.G.; Tsoka, S. Network-based piecewise linear regression for QSAR modelling. *J. Comput.-Aided Mol. Des. Vol.* **2019**, *33*, 831–844. [CrossRef]

66. Dudek, A.Z.; Arodz, T.; Galvez, J. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screen.* **2006**, *9*, 213–228. [CrossRef]
67. Raevsky, O.A.; Sapegin, A.; Zefirov, N.S. The QSAR Discriminant-Regression Model. *Quant. Struct.-Act. Relatsh.* **1994**, *13*, 412–418. [CrossRef]
68. Doreswamy; Vastrad, B. Predictive Comparative Qsar Analysis of as 5-Nitrofuran-2-Yl Derivatives Myco Bacterium Tuberculosis H37RV Inhibitors. *Healthc. Inform. Int. J.* **2013**, *2*, 47–62. [CrossRef]
69. Ajmani, S.; Jadhav, K.M.; Kulkarni, S.A. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* **2006**, *46*, 24–31. [CrossRef]
70. Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206. [CrossRef]
71. Raj, N.; Jain, S. 3d QSAR Studies in Conjunction With k-Nearest Neighbor Molecular Field Analysis (k-NN-MFA) on a Series of ResearchGate. 2011. Available online: 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 (accessed on 24 May 2023).
72. Asikainen, A.H.; Ruuskanen, J.; Tuppurainen, K.A. 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. *Environ. Sci. Technol.* **2004**, *38*, 6724–6729. [CrossRef]
73. Nigsch, F.; Bender, A.; Van Buuren, B.N.; Tissen, J.; Nigsch, E.A.; Mitchell, J.C. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422. [CrossRef] [PubMed]
74. Poroikov, V.V.; Filimonov, D.A.; Borodina, Y.V.; Lagunin, A.A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349–1355. [CrossRef]
75. Chen, B.; Sheridan, R.P.; Hornak, V.; Voigt, J.H. Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792–803. [CrossRef] [PubMed]
76. Kupervasser, O. *Quantitative Structure-Activity Relationship Modeling and Bayesian Networks: Optimality of Naive Bayes Model*; IntechOpen: Rijeka, Croatia, 2019. [CrossRef]
77. Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 837–843. [CrossRef]
78. Bender, A.; Jenkins, J.L.; Glick, M.; Deng, Z.; Nettles, J.H.; Davies, J.W. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456. [CrossRef] [PubMed]
79. Keyvanpour, M.R.; Shirzad, M.B. An Analysis of QSAR Research Based on Machine Learning Concepts. *Curr. Drug Discov. Technol.* **2021**, *18*, 17–30. [CrossRef] [PubMed]
80. Bugeac, C.A.; Ancuceanu, R.; Dinu, M. QSAR Models for Active Substances against Pseudomonas aeruginosa Using Disk-Diffusion Test Data. *Molecules* **2021**, *26*, 1734. [CrossRef]
81. Darnag, R.; Schmitzer, A.R.; Belmiloud, Y.; Villemin, D.; Jarid, A.; Chait, A.; Seyagh, M.; Cherqaoui, D. QSAR Studies of HEPT Derivatives Using Support Vector Machines. *Qsar Comb. Sci.* **2009**, *28*, 709–718. [CrossRef]
82. Niu, B.; Lu, W.; Yang, S.; Cai, Y.; Li, G. Support vector machine for SAR/QSAR of phenethyl-amines. *Acta Pharmacol. Sin.* **2007**, *28*, 1075–1086. [CrossRef]
83. Wu, Z.; Zhu, M.; Kang, Y.; Leung, E.L.; Lei, T.; Shen, C.; Jiang, D.; Wang, Z.; Cao, D.; Hou, T. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Brief. Bioinform.* **2021**, *22*, bbaa321. [CrossRef]
84. Alvarsson, J.; Lampa, S.; Schaal, W.; Andersson, C.; Wikberg, J.E.S.; Spjuth, O. Large-scale ligand-based predictive modelling using support vector machines. *J. Cheminform.* **2016**, *8*, 39. [CrossRef]
85. Liu, H.X.; Zhang, R.S.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. 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. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288–1296. [CrossRef] [PubMed]
86. Nekoei, M.; Mohammadhosseini, M.; Pourbasheer, E. 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. *Med. Chem. Res.* **2015**, *24*, 3037–3046. [CrossRef]
87. Torng, W.; Altman, R.B. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinform.* **2017**, *18*, 302. [CrossRef] [PubMed]
88. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48. [CrossRef]
89. Graves, A.; Mohamed, A.; Hinton, G.E. Speech Recognition with Deep Recurrent Neural Networks. arXiv (Cornell University). *arxiv* **2013**, arXiv:1303.5778. [CrossRef]
90. Kingma, D.P. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
91. Goodfellow, I.J. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
92. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.F.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]

93. Kusner, M.J. Grammar Variational Autoencoder. *arXiv* **2017**, arXiv:1703.01925.
94. Matsuzaka, Y.; Uesawa, Y. 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. *Front. Bioeng. Biotechnol.* **2019**, *7*, 65. [CrossRef] [PubMed]
95. Karpov, P.; Godin, G.; Tetko, I.V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminform.* **2020**, *12*, 17. [CrossRef]
96. Xu, Y. Development and Evaluation of Conformal Prediction Methods for QSAR. *arXiv* **2023**, arXiv:2304.00970.
97. Shayanfar, S.; Shayanfar, A. Comparison of various methods for validity evaluation of QSAR models. *BMC Chem.* **2022**, *16*, 63. [CrossRef]
98. Golbraikh, A.; Wang, X.; Zhu, H.; Tropsha, A. *Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment*; Springer: Cham, Switzerland, 2017; pp. 2303–2340. [CrossRef]
99. Spiegel, J.; Senderowitz, H. Evaluation of QSAR Equations for Virtual Screening. *Int. J. Mol. Sci.* **2020**, *21*, 7828. [CrossRef]
100. Matveieva, M.; Polishchuk, P.G. Benchmarks for interpretation of QSAR models. *J. Cheminform.* **2021**, *13*, 41. [CrossRef] [PubMed]
101. C3.ai. LIME: Local Interpretable Model-Agnostic Explanations. 2022. Available online: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/#:~:text=What%20is%20Local%20Interpretable%20Model,to%20explain%20each%20individual%20prediction (accessed on 24 May 2023).
102. Molnar, C. 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning. 2 March 2023. Available online: https://christophm.github.io/interpretable-ml-book/shap.html (accessed on 24 May 2023).
103. Izrailev, S.; Agrafiotis, D. A method for quantifying and visualizing the diversity of QSAR models. *J. Mol. Graph. Model.* **2004**, *22*, 275–284. [CrossRef]
104. An, Y.; Sherman, W.; Dixon, S.L. Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. *J. Chem. Inf. Model.* **2013**, *53*, 2312–2321. [CrossRef]
105. Wellawatte, G.P.; Seshadri, A.; White, A.J.P. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.* **2022**, *13*, 3697–3705. [CrossRef]
106. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. [CrossRef]