THE EVOLUTION AND IMPACT OF QSAR MODELS IN DRUG DISCOVERY:

FROM LINEAR RELATIONSHIPS TO PERSONALIZED MEDICINE

Assist. Prof. Dr. Bilge ÖZLÜER BAŞER

ISBN: 978-625-8151-23-7

Ankara- 2024

THE EVOLUTION AND IMPACT OF QSAR MODELS IN DRUG DISCOVERY: FROM LINEAR RELATIONSHIPS TO PERSONALIZED MEDICINE

Assist. Prof. Dr. Bilge ÖZLÜER BAŞER

Mimar Sinan Fine Arts University, Faculty of Arts and Sciences, Statistics Department, İstanbul, Türkiye. bilge.baser@msgsu.edu.tr, ORCID ID:0000-0002-2400-6584

Editor Prof. Dr. Ayça ÇAKMAK PEHLİVANLI

DOI: https://doi.org/10.5281/zenodo.13475681

E mail: ubakyayinevi@gmail.com
www.ubakyayinevi.org
It is responsibility of the author to abide by the publishing ethics rules.
UBAK Publishing House – 2024©

ISBN: 978-625-8151-23-7
August / 2024
Ankara / Turkey

To Arya and Lara.

PREFACE

The trajectory of drug discovery over the past several decades has been profoundly influenced by advancements in computational methods, among which Quantitative Structure-Activity Relationship (QSAR) models hold a distinguished place. Initially conceived in the mid-20th century, QSAR models have evolved from simple linear correlations between molecular structures and biological activities into sophisticated tools that leverage machine learning, big data analytics, and systems biology to address the complexities of modern drug development. This transformation marks QSAR not only as a pivotal methodology in cheminformatics but as an essential framework for the rational design of new therapeutic agents.

As we navigate an era characterized by unprecedented technological advancements, it is essential to recognize the foundational role that statistics science plays in guiding these developments. In particular, the integration of machine learning and artificial intelligence into QSAR models underscores the necessity of robust statistical analyses that ensure the reliability, interpretability, and predictive accuracy of these models.

One of the challenges of writing a book in an interdisciplinary field is that no one is an expert in all aspects of the field at the same time. QSAR modeling requires collaboration across diverse disciplines. By fostering communication and knowledge sharing among statisticians, chemists, biologists, and pharmacologists, QSAR can accelerate the development

of safer and more effective medicines. This book bridges these disciplines by providing a comprehensive overview of QSAR's statistical foundations and modern applications. Thus, it helps researchers from different fields come together and communicate on a common ground and terminology.

Additionally, the book mentions QSAR’s emerging role in personalized medicine, where models predict individual responses to treatments based on genetic and molecular profiles. This paradigm shift towards precision therapeutics represents the future of drug discovery, where interventions are tailored to individual patients. Despite these advancements, challenges remain, including model interpretability, data quality, and generalizability across diverse chemical spaces.

Addressing these challenges is essential for the continued evolution of QSAR models and their application in drug discovery. This book is intended not only as a technical resource but also as a reflection on the broader implications of computational methods for shaping the future of pharmacology. I believe that "The Evolution and Impact of QSAR Models in Drug Discovery: From Linear Relationships to Personalized Medicine" will inspire further innovation and contribute to the advancement of efficient and personalized healthcare solutions.

Prof. Dr. Ayça ÇAKMAK PEHLİVANLI

TABLE OF CONTENTS

INTRODUCTION

Quantitative Structure-Activity Relationship (QSAR) models are computational techniques used to estimate the biological activity or properties of chemical compounds based on their chemical structure. These models are crucial tools in drug discovery and development, offering a more efficient and cost-effective approach compared to traditional experimental methods.

QSAR models have transformed the drug discovery landscape by enabling researchers to:

- Estimate the biological activity of novel compounds prior to their synthesis and experimental validation.
- Identify potential lead compounds from large chemical libraries quickly.
- Optimize the chemical structure of lead compounds to enhance their efficacy, selectivity, and safety.
- Reduce the cost and time for drug development by minimizing the need for extensive experimental testing.

This book aims to provide a comprehensive overview of QSAR models, their historical development, fundamental principles, technological advancements, modern approaches, applications in drug discovery, challenges, and future directions.

1. Historical Background

The origins of QSAR can be traced back to the late 19th and early 20th centuries, with foundational contributions from scientists like Hansch, Fujita, and Free-Wilson. They laid the groundwork for the development of mathematical models that correlate chemical structures with biological activities.

Over the years, QSAR methodologies have evolved significantly. The initial models were simple linear relationships between chemical structure and activity. However, with advancements in computational power and statistical methods, QSAR models have become more sophisticated, incorporating complex non-linear relationships and machine learning techniques.

Several key milestones have marked the evolution of QSAR models:

- **Early Linear Models (1960s-1970s):** Hansch and Fujita's pioneering work on the correlation of chemical structure with biological activity (Hansch & Fujita, 1964). Introduction of the Free-Wilson approach, which considered the additive effects of different chemical substituents (Free & Wilson, 1964).
- **Expansion to 3D-QSAR (1980s-1990s):** Development of 3D-QSAR models, incorporating spatial properties of molecules (Cramer et al., 1988).
- **Incorporation of Machine Learning and AI (2000s-Present):** Integration of machine learning and artificial intelligence techniques, leading to more accurate and robust QSAR models.

2. Fundamentals of QSAR Models

QSAR models operate on the principle that a molecule's biological activity can be quantitatively correlated with its chemical structure. This relationship is expressed through mathematical models that predict the biological activity based on various molecular descriptors. The basic concepts used in QSAR models are explained below.

2.1. Molecular Descriptors

These are numerical values that describe the chemical structure of a molecule. Descriptors are critical inputs for QSAR models and can be categorized into:

1D Descriptors: 1D (1-dimensional) descriptors are simple properties like molecular weight, number of atoms, number of bonds, number of hydrogen bond donors and acceptors, LogP, and topological polar surface area (Hansch & Leo ,1995). 1D descriptors play a crucial role in early-stage drug design because of the simplicity and computational efficiency. They often show significant correlation with biological activity in foundational level and they are useful for model interpretation and feature selection for statistical analysis. Despite all these advantages 1D descriptors derived from the basic molecular formula and disregard the detailed structure of the molecule. This can lead to significant information loss.

2D Descriptors: 2D (2-dimensional) descriptors provide a more detailed representation of a molecule compared to 1D descriptors. They capture various aspects of structural fragments, topological indices, bond and atom counts, connectivity indices, substructure, path and walk counts, molecular connectivity indices (Kubinyi, 1993). 2D descriptors have explicit chemical meanings to understand the factors influencing molecular behavior. They often lead to better predictive performance in QSAR models than 1D descriptors and have impact on a wide range of applications including biological activity, toxicity, solubility, and more. While 2D descriptors offer more detailed molecular representations compared to 1D descriptors, they also possess limitations that can impact their utility and accuracy in QSAR modeling. 2D descriptors are obtained from two-dimensional form of the molecule. However, molecules can adopt multiple conformations due to rotation around bonds and two-dimensional representations do not have this flexibility. In addition to this, 2D descriptors often failed to represent electronic properties, steric effects and the nuances of hydrogen bonding, van der Waals forces, and non-covalent interactions that play critical role in drug-receptor binding.

3D Descriptors: 3D (3-dimensional) descriptors are derived from the three-dimensional coordinates of the atoms in a molecule. They can capture geometric and spatial properties of the molecule such as molecular volume, molecular surface areas, shape descriptors, electrostatic descriptors, pharmacophore descriptors

and 3D fingerprints (Cramer et al., 1988). 3D descriptors provide to understand how a molecule interacts with its biological target with well-defined binding sites. They are essential in molecular docking studies by providing to facilitate virtual screening of compound libraries to identify potential drug candidates. Although 3D descriptors are the most detailed, they have several limitations such as computational complexity, dependence on accurate 3D structures, sensitivity to small changes, lack of standardization, limited applicability for non-rigid molecules and integration with 2D descriptors. Addressing these shortages requires advanced computational techniques, rigorous validation, and the integration of complementary descriptors to develop robust and reliable QSAR models. Despite these challenges, the benefits of 3D descriptors in capturing the intricate details of molecular interactions make them invaluable tools in drug discovery and development.

2.2. Regression Methods

QSAR models use statistical techniques to derive the relationship between molecular descriptors and biological activity. Common regression methods include:

Linear Regression: Linear regression was among the initial techniques utilized in QSAR modeling as seen in the Hansch-Fujita analysis, which relates physicochemical properties to biological activity (Hansch & Fujita, 1964). Kubinyi analyzed the

combined effect of multiple descriptors on biological activity with Multiple Linear Regression (Kubinyi, 1993). Linear methods are easy to implement and effective for simple relationships with small datasets. However, they assume a linear relationship which may not always be accurate. Therefore, they have limited capacity to manage complex, non-linear relationships.

Non-linear Regression: Uses more complex functions to capture non-linear relationships between descriptors and biological activity. It is useful in cases where biological activity is influenced by complex interactions between molecular descriptors (Hansch & Leo, 1995). These models are capable of modeling complex relationships and more flexible than linear regression. On the other hand, computationally intensive and require accurate selection of the appropriate non-linear function.

Machine Learning Methods: Machine learning methods have gained popularity in QSAR modeling owing to their capability to handle complex, high-dimensional data. Machine learning methods, classified under supervised, unsupervised, and reinforcement learning paradigms, have significantly advanced QSAR modeling by providing robust tools for handling complex data and non-linear relationships. The choice of method depends on the specific requirements of the QSAR task, including data characteristics, model complexity, and computational resources available. Techniques like support vector machines, neural

networks, random forests, gradient boosting algorithms and Bayesian networks are increasingly used for QSAR modeling due to their ability to handle complex and non-linear data (Cherkasov et al.,2014).

3. Technological Advancements in QSAR Modeling

3.1. Impact of Computational Power and Software Development

The evolution of QSAR models has been significantly influenced by advancements in computational power and software development. The development of powerful processors and parallel computing has allowed researchers to handle larger datasets and more complex models. Various software tools and platforms have been developed specifically for QSAR modeling, making it more accessible and efficient for researchers such as *MOE (Molecular Operating Environment)* (Chemical Computing Group ULC, 2024), *Open Babel* (O'Boyle et al., 2011) and *AutoQSAR* (Tropsha & Golbraikh, 2007).

3.2. Integration of Big Data and High-Throughput Screening

The integration of big data and high-throughput screening (HTS) technologies has provided vast amounts of biological activity data, which are crucial for developing robust QSAR models.

HTS leverages automation, miniaturization, and large-scale data analysis to identify potential lead compounds in drug discovery. HTS allows the rapid testing of thousands to millions of compounds for biological activity, generating large datasets for QSAR modeling

(Macarron et al., 2011). HTS-driven QSAR models enhance the ability to predict the biological activity, toxicity, and therapeutic potential of chemical compounds. By leveraging HTS data, QSAR models can accelerate the lead identification and optimization process, facilitate drug repurposing, and improve safety assessments. As HTS technology continues to advance, its integration with QSAR modeling will play an increasingly crucial role in the efficient and effective discovery of new therapeutics.

Big Data Analytics involves the process of examining large and varied datasets, or "big data," to uncover hidden patterns, correlations, and insights. In the context of QSAR modeling, big data analytics can significantly enhance the predictive power and robustness of models by leveraging vast amounts of chemical and biological data. Techniques for managing and analyzing large datasets, including data mining and machine learning, have been integrated into QSAR modeling to handle the complexity and volume of HTS data (Chen et al., 2018).

4. Modern QSAR Approaches and Future Directions

Modern QSAR approaches leverage advanced computational techniques, machine learning, and integrative data analysis to improve the predictive power and applicability of QSAR models. These approaches have expanded the traditional boundaries of QSAR modeling, enabling more accurate predictions and broader applications in drug discovery and development.

4.1. Use of Deep Learning and Neural Networks in QSAR

Deep learning and neural networks represent advanced techniques that have transformed QSAR modeling by enabling the analysis of complex, high-dimensional data and capturing intricate non-linear relationships. Neural networks, particularly deep neural networks, consist of multiple layers of interconnected neurons that process data in a hierarchical manner. These networks consist of multiple layers of neurons, allowing them to learn hierarchical representations of data. Each layer extracts various levels of features from the input data, enabling the network to model complex relationships between molecular descriptors and biological activity (LeCun et al., 2015).

Deep Neural Networks (DNNs):

Deep learning and neural networks have been applied in various ways within QSAR modeling to predict biological activity, optimize drug candidates, and discover new therapeutic targets.
DNNs can model complex, non-linear relationships between molecular descriptors and biological activity. Each layer in the network learns a different level of abstraction, enabling the capture of intricate patterns.

LeCun et al. (2015) used DNNs to predict the inhibitory activity of compounds against specific enzymes, where traditional linear models fail to capture the underlying non-linear interactions. Ma et. al. (2015) demonstrated the application of deep neural networks in QSAR modeling, showing improved predictive performance over traditional methods by capturing non-linear relationships between molecular

descriptors and biological activity. Mayr et. al. (2016) applied deep learning techniques to predict the toxicity of chemical compounds, highlighting the ability of DNNs to model complex relationships and improve prediction accuracy. Yang et al. (2019) investigated the molecular representations learned by deep neural networks for property prediction, demonstrating their ability to model complex molecular properties. Altae-Tran et al. (2017) introduced one-shot learning in the context of drug discovery, using deep neural networks to make accurate predictions with limited data, highlighting the potential of DNNs in QSAR modeling.

Unlike traditional models that require hand-crafted features, DNNs can automatically learn relevant features from raw data through multiple layers of abstraction.

Combining deep learning with reinforcement learning allows the optimization of compounds by exploring chemical space and iteratively improving the properties of drug candidates (Mnih, et al., 2015).

DNNs can be trained to predict the activity of compounds against multiple targets simultaneously, leveraging shared patterns across different biological activities (Ramsundar et al., 2015).

Convolutional Neural Networks (CNNs):

CNNs are particularly effective for data with spatial or grid-like topology, such as molecular graphs. They apply convolutional filters to extract local patterns. CNNs are widely used in QSAR modeling to

predict molecular properties and activities. They are effective in learning features directly from molecular images and graphs, improving the predictive accuracy and robustness of QSAR models. The examples provided illustrate the versatility and power of CNNs in various QSAR applications, from plasma protein binding prediction to end-to-end QSAR systems. Processing molecular graphs and images, extracting features that correlate with biological activity (Duvenaud et al., 2015). Goh et al. (2017) demonstrated the use of CNNs to predict the biological activity of molecules by encoding their 2D structures as pixel images. The CNNs learned spatial features that correlate with biological activities, outperforming traditional descriptor-based QSAR models. DeepChem (Ramsundar et al., 2019), an open-source deep learning framework, supports the use of CNNs for QSAR modeling by providing tools for molecular image processing and feature extraction.

Graph Neural Networks (GNNs):

GNNs extend the capability of neural networks to graph data structures, which are common in molecular representations. GNNs are revolutionizing QSAR modeling by enabling the direct use of molecular graphs, which capture the inherent structural and connectivity information of molecules. These examples illustrate the versatility and power of GNNs in predicting molecular properties, improving the accuracy and robustness of QSAR models, and enhancing their applicability in drug discovery.

Kearnes et al. (2016) discusses the advantages of molecular graph convolutions over traditional molecular fingerprints, highlighting their

ability to learn more detailed and nuanced representations of molecular structures. GNNs were applied to predict drug-target interactions, demonstrating improved accuracy and interpretability over fingerprint-based models.

Gilmer et al. (2017) employed GNNs to predict various molecular properties, demonstrating significant improvements in predictive performance over traditional QSAR models. The study uses GNNs to predict the activity of drug-like molecules, focusing on properties such as bioavailability and ADMET (absorption, distribution, metabolism, excretion, and toxicity).

Schütt et al. (2017) integrates quantum machine learning with graph networks to model chemical properties, showcasing the potential of GNNs in quantum chemistry. The study uses GNNs to predict electronic properties of molecules, demonstrating the capability of these networks to capture complex quantum interactions.

Ryu et al. (2019) applied Bayesian graph convolutional networks (GCNs) to predict molecular properties with uncertainty quantification, which is crucial for risk assessment in drug discovery. In the study GNNs were used to predict properties such as solubility and toxicity, with the Bayesian approach providing a measure of uncertainty in the predictions.

Hung & Gini (2021) demonstrated the application of GCNs for mutagenicity prediction. The GCNs processed molecular graphs to

predict the mutagenic potential of compounds, showing improved accuracy over traditional QSAR models.

Generative Models:

Generative models, including variational autoencoders, generative adversarial networks, conditional variational autoencoders, and recurrent neural networks are powerful tools in QSAR modeling for generating novel molecular structures with desired properties. These models have been applied successfully in various drug discovery contexts, demonstrating their potential to enhance the discovery and optimization of new drug candidates.

Variational Autoencoders (VAEs): Generate new compounds with desired properties by learning the distribution of known active compounds. VAEs are used to generate new compounds by learning the latent space of molecular structures. They encode molecules into a latent space and decode them back to molecular structures, allowing the generation of novel compounds with specific properties.

Kingma & Welling (2013) utilized VAEs to generate drug-like molecules that optimize specific activity profiles while maintaining drug-likeness. The VAEs learned the distribution of known active compounds and generated new molecules with improved properties.

Generative Adversarial Networks (GANs): Consist of a generator and a discriminator, where the generator creates new compounds,

and the discriminator evaluates their plausibility. This setup allows the generation of novel compounds that resemble real molecules (Goodfellow et al., 2014) and Popova et al. (2018) used GANs to design novel molecules with desired biological activities. The GANs generated compounds that were experimentally validated to have high binding affinity to specific targets.

Molecular Generative Model Based on Conditional VAEs*:* Conditional VAEs generate new molecules based on specified conditions or desired properties, allowing for more controlled generation of compounds.

Kang et al. (2018) was employed a conditional VAE to generate molecules with specific pharmacokinetic properties, such as solubility and permeability. The model successfully generated compounds that met the desired criteria.

Generative Models in the Open Molecule Generator (OMG)*:* The OMG framework uses generative models to explore chemical space and generate novel compounds with optimized properties. It integrates reinforcement learning to further refine the generated molecules. OMG was used to discover new inhibitors for a specific protein target. The generative model produced several promising candidates, which were validated through biological assays (Olivecrona et al., 2017).

Recurrent Neural Networks (RNNs) for Molecular Generation: RNNs are used to generate sequences of SMILES (Simplified Molecular-Input Line-Entry System) strings representing molecules. These models can learn the syntax and semantics of chemical structures, enabling the generation of valid and novel molecules. Segler et al. (2018) was trained an RNN on a large dataset of SMILES strings to generate new drug-like molecules. The generated molecules were found to have diverse and novel structures with potential biological activity.

GAN-Based Drug Generation with Reinforcement Learning: Combining GANs with reinforcement learning (RL) to optimize the generated molecules for specific drug-like properties. The RL component guides the generation process to produce compounds that meet predefined criteria.

You et al. (2018) used a GAN-RL hybrid model to generate molecules with high binding affinity and low toxicity. The generated compounds were then synthesized and tested, showing promising results.

4.2. Multi-task Learning and Transfer Learning Applications

Multi-task learning (MTL) and transfer learning (TL) are advanced machine learning techniques that have shown great promise in QSAR modeling. MTL involves training a single model on multiple related tasks simultaneously, allowing the model to leverage shared information across tasks. This approach can improve predictive

performance, especially when data for individual tasks are limited. By training on multiple biological activities or different datasets, MTL can improve the generalizability and robustness of QSAR models. MTL models are used to predict the biological activities of compounds against multiple targets, leveraging the shared structural information across different tasks. Ramsundar et al. (2015) used MTL to predict the activity of compounds against various cancer cell lines. The MTL approach improved predictive performance by sharing information across tasks.
MTL models can predict multiple ADMET properties simultaneously, providing a comprehensive assessment of a compound's pharmacokinetic profile (Unterthiner et al., 2014).

TL involves pre-training a model on a large dataset and then fine-tuning it on a smaller, task-specific dataset. This approach is particularly useful when data for the target task are scarce. Altae-Tran et al. (2017) applied TL to transfer knowledge from a model trained on the ZINC database to a specific project aimed at discovering new antibiotics. The TL approach significantly improved prediction accuracy. Xu et al. (2017) applied MTL and TL to integrate chemical structure data with high throughput screening data, predicting multiple biological endpoints with improved accuracy.

Chen et al. (2019) utilized a pre-trained model on a large dataset from ChEMBL (Gaulton et al., 2012) and fine-tuned it for specific QSAR tasks, such as predicting the activity of compounds against specific protein targets. Pre-trained models on large chemical datasets can be

fine-tuned for specific QSAR tasks, improving performance and reducing the need for large task-specific datasets.

4.3. Systems Biology Approaches

System biology approaches involve modeling the complex interactions within biological systems to understand how compounds affect these systems at a holistic level. Systems biology approaches can identify key pathways and networks influenced by compounds, enhancing the interpretability and relevance of QSAR predictions (Hood & Flores, 2012).

Systems biology approaches in QSAR involve integrating various types of biological data to create more comprehensive and accurate models. These methods include the integration of genomic and proteomic data, network-based modeling, pathway analysis, and multi-omics data integration. By considering the complex interactions within biological systems, these approaches enhance the predictive power and interpretability of QSAR models, making them invaluable tools in drug discovery and toxicology.

Integrating Genomic and Proteomic Data: Genomic and proteomic data provide insights into the molecular mechanisms affected by compounds, enhancing the predictive power of QSAR models.

Cheng et al. (2013) and Iorio et al. (2016) used genomic data to predict the response of cancer cell lines to various drugs. By integrating gene expression profiles with chemical descriptors, the QSAR models achieved higher accuracy in predicting drug efficacy.

Rix et al. (2007) and Geenen et al. (2021) integrated proteomic data with QSAR models to predict the off-target effects of drugs, leading to better predictions of adverse drug reactions.

Network-Based QSAR Models: Network-based approaches use biological networks to provide a holistic view of how compounds interact with various biological entities, such as proteins, genes, and metabolites.

Barabasi et al. (2011) and Gysi et. al. (2020) created a network-based QSAR model by integrating protein interaction networks with chemical data to predict drug-target interactions. This approach improved the prediction of compound efficacy and off-target effects.

Pathway-Based QSAR Models: Pathway-based approaches incorporate data on biological pathways to understand the mechanistic effects of compounds.

Judson et al. (2010) and Wang et al. (2019), integrated pathway analysis with QSAR modeling to predict the impact of environmental chemicals on human health. This approach provided insights into the pathways affected by different chemicals.

Multi-Omics Data Integration: Multi-omics approaches combine different types of omics data (genomics, transcriptomics, proteomics, metabolomics) to build comprehensive QSAR models. Integrating multi-omics data with QSAR models can enhance the understanding of the molecular mechanisms underlying drug action and toxicity, leading to more accurate predictions of biological activity (Cheng et al., 2013).

Hasin et al. (2017) and Nguyen et al. (2019) developed a multi-omics QSAR model that integrated genomic, transcriptomic, and proteomic data to predict the toxicity of environmental chemicals. This integrative approach provided a more comprehensive understanding of the mechanisms underlying toxicity.

4.4. Hybrid Approaches

Hybrid approaches combine QSAR with other computational methods, such as molecular docking, molecular dynamics, and pharmacophore modeling. These integrative models can provide a more comprehensive understanding of the interaction between drugs and their biological targets, leading to better predictions and more effective drug design (Sliwoski et al., 2014).

4.5. Cloud Computing

Cloud computing provides scalable computational resources, while big data analytics enables the handling and analysis of vast amounts of data. These technologies will facilitate the development and deployment of large-scale QSAR models, making it easier to process and analyze large datasets efficiently (Chen et al., 2018).

4.6. Personalized Medicine

Personalized medicine aims to tailor medical treatment to individual characteristics, such as genetic profiles and personal biomarkers.

QSAR models can be adapted to predict personalized drug responses, optimizing treatment efficacy and minimizing adverse effects for individual patients (Hood & Flores, 2012).

QSAR models have the potential to revolutionize personalized medicine by leveraging detailed molecular and genetic information to tailor treatments for individual patients. This can lead to more effective and safer therapies, optimized dosages, and reduced healthcare costs, ultimately improving patient outcomes and advancing the field of precision medicine. As computational techniques and data availability continue to evolve, the integration of QSAR models in personalized medicine will become increasingly impactful.

Potential for QSAR Models in Personalized Medicine

Personalized Drug Selection

QSAR models can predict the efficacy and toxicity of drugs based on individual genetic and molecular profiles, which is critical in personalized medicine. By analyzing a patient's specific genetic makeup, QSAR models can help identify the most effective drug with the least side effects for that individual. For instance, predicting how different patients with varying genetic backgrounds will respond to a particular chemotherapy drug can significantly enhance treatment outcomes (Cruz-Monteagudo et al., 2014).

Dose Optimization

QSAR models can determine the optimal drug dosage for individual patients, minimizing adverse effects while maximizing therapeutic benefits. Tailoring doses of anticoagulants like warfarin based on genetic markers that affect drug metabolism can prevent complications such as bleeding or clotting (Roden et al., 2019). Adjusting the dose of warfarin for patients with variations in the CYP2C9 and VKORC1 genes ensures safer and more effective anticoagulation therapy.

Predicting Drug-Drug Interactions

QSAR models can predict potential interactions between multiple drugs a patient is taking, which is crucial for patients on complex medication regimens. This is especially important for elderly patients or those with chronic conditions requiring multiple medications (Srinivasan et al., 2014). Identifying harmful interactions between statins and other common medications like certain antibiotics can prevent adverse drug reactions.

Assessing Drug Toxicity

QSAR models can help predict the toxicity of drugs in individuals based on their genetic and metabolic profiles. Screening out drugs that may cause severe side effects in specific patient populations during the drug development process can improve safety (Liu et al., 2017). Predicting liver toxicity risks in patients with certain genetic polymorphisms affecting drug metabolism can guide safer drug prescriptions.

Developing Companion Diagnostics

QSAR models can aid in the development of companion diagnostics that predict which patients will benefit from a particular drug. This approach is instrumental in creating tests that identify patients likely to respond to targeted cancer therapies based on molecular markers (Cruz-Monteagudo et al., 2014). For instance, developing diagnostic tests for HER2-positive breast cancer patients who will benefit from trastuzumab (Herceptin) improves treatment efficacy.

Advancing Pharmacogenomics

QSAR models can integrate pharmacogenomic data to predict drug responses and guide personalized treatment plans. Utilizing genetic information to select appropriate drugs and dosages for patients with cardiovascular diseases can enhance therapeutic outcomes (Roden et al., 2019). Applying pharmacogenomic data to adjust antihypertensive treatments based on individual genetic profiles ensures more effective blood pressure management.

Facilitating Drug Repurposing

QSAR models can identify existing drugs that might be effective for new indications based on similarities in molecular profiles. This approach can expedite the discovery of new therapeutic applications for approved drugs (Ekins et al., 2013). Discovering new therapeutic applications for approved drugs in treating rare genetic disorders accelerates drug development.

Enhancing Clinical Trials

QSAR models can be used to stratify patients in clinical trials based on predicted responses to treatment. This approach can improve the efficiency and success rates of clinical trials by selecting participants more likely to benefit from the treatment (Srinivasan et al., 2014). Stratifying patients in oncology trials based on predicted responses to experimental therapies ensures more targeted and effective treatments.

Reducing Healthcare Costs

By personalizing treatments, QSAR models can reduce the trial-and-error approach in prescribing, leading to more cost-effective healthcare. This approach can reduce hospitalizations and adverse events by tailoring treatments to individual patient profiles (Cruz-Monteagudo et al., 2014). Cost savings from reduced adverse drug reactions and hospital readmissions through personalized medication plans improve healthcare efficiency.

Improving Patient Outcomes

Personalized medicine guided by QSAR models can lead to more effective treatments with fewer side effects, improving overall patient outcomes. Enhanced management of chronic conditions through tailored therapeutic strategies can significantly improve patients' quality of life (Roden et al., 2019). Improved management of chronic pain by personalizing opioid prescriptions to minimize addiction risks and maximize pain relief enhances patient well-being.

5. Developing a QSAR Model from Scratch

The process of developing a QSAR model involves a series of methodical steps, ranging from data collection to model validation, each contributing to the robustness and predictive power of the final model.

5.1. Data Collection and Preparation

Data Collection

The initial step in QSAR modeling involves the acquisition of a comprehensive dataset comprising chemical compounds with known biological activities. Public databases such as ChEMBL and PubChem (Kim et al., 2019), as well as proprietary datasets, serve as valuable sources for obtaining such data.

Data Cleaning

Data cleaning is a crucial step in the development of QSAR models, as the quality and reliability of the data directly influence the accuracy and predictive power of the resulting models. The presence of erroneous, inconsistent, or redundant data can lead to misleading conclusions and suboptimal model performance. Therefore, rigorous data cleaning is essential to ensure the integrity of the dataset used for QSAR modeling.

Steps in Data Cleaning

Removal of Duplicates

Identification of Duplicates:

- Duplicates can arise from multiple sources, such as merging datasets from different sources or repeated entries within the same dataset.
- Duplicate entries should be identified based on unique identifiers like compound names, chemical structures (e.g., SMILES strings or InChI (IUPAC International Chemical Identifier) keys), and biological activity measurements.

Elimination Process:

- Upon identification, duplicate entries should be carefully examined to retain the most reliable and complete data point.
- Redundant duplicates are then removed to prevent bias and ensure data integrity.

Correction of Erroneous Data

Detection of Errors:

- Erroneous data can result from transcription errors, instrument malfunctions, or inconsistencies during data entry.

- Common errors include impossible or implausible values for molecular descriptors (e.g., negative molecular weights) and biological activities outside the expected range.

***Correction Strategies**:*

- Verification against original sources or experimental records is essential for correcting errors.
- If verification is not possible, such data points should be flagged and potentially removed from the dataset.

Handling Missing Data

***Identification of Missing Values**:*

- Missing data can significantly impact the quality of the QSAR model, leading to biased or invalid predictions.
- It is important to identify missing values in both molecular descriptors and biological activity measurements.

***Imputation Methods**:*

- Imputation techniques can be employed to estimate missing values, such as mean imputation, median imputation, or more sophisticated methods like k-nearest neighbors imputation or multiple imputation by chained equations.
- The choice of imputation method depends on the nature of the data and the extent of missing values.

Outlier Detection and Handling

***Identification of Outliers**:*

- Outliers are data points that deviate significantly from the rest of the dataset and can result from experimental errors or genuine variations in the data.
- Statistical methods such as z-scores, Interquartile Range (IQR) analysis, and visualization techniques like box plots can be used to identify outliers.

***Decision on Outliers**:*

- Outliers should be carefully examined to determine whether they result from experimental errors or represent true biological variability.
- Depending on the context, outliers may be retained, corrected, or removed to enhance model robustness.

Normalization and Standardization

Normalization and standardization are essential preprocessing steps in QSAR modeling. These techniques adjust the scales of the data to ensure that no single descriptor disproportionately influences the model, thereby enhancing the accuracy and reliability of predictions.

***Normalization**:*

Normalization of the selected descriptors is crucial to ensure they are on a comparable scale, thereby preventing any single descriptor from disproportionately influencing the model. Normalization involves scaling the data to a specific range, typically [0, 1], which ensures that no single descriptor dominates the model due to its scale. Min-max normalization, decimal normalization, quantile normalization or log transformation are common techniques used for this purpose.

***Standardization**:*

Standardization, or z-score normalization, involves rescaling the data to have a mean of 0 and a standard deviation of 1. Similar to z-score normalization, mean normalization scales the data around the mean without dividing by the standard deviation. Robust scaling uses the median and the IQR to scale the data, making it robust to outliers. Unit vector scaling uses each data point such that the norm (magnitude) of the vector representing the data point is 1. This is useful in some machine learning contexts where the direction of the vector is important but not its length. MaxAbs Scaling uses each feature by its maximum absolute value. This method is useful when the data is already centered at zero and you want to preserve sparsity. These are commonly used methods for standardization.

Consistency Checks

Ensuring Consistent Units:

- It is essential to ensure that all descriptors and biological activities are reported in consistent units. For example, molecular weights should be consistently reported in Daltons, and concentrations in molarity.
- Inconsistencies in units can lead to incorrect interpretations and predictions.

Chemical Structure Verification:

- Chemical structures should be verified to ensure they are correctly represented. Tools such as Open Babel or alternative cheminformatics software can be used to standardize representations and remove any inconsistencies in chemical structures.

Descriptor Calculation

Molecular descriptors, which quantitatively represent various properties of chemical compounds, are integral to QSAR modeling. Tools such as MOE, Open Babel, and RDKit (Landrum, 2016) facilitate the calculation of these descriptors, encompassing one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) properties.

5.2. Descriptor (Feature) Selection

To enhance the model's predictive accuracy, it is essential to select relevant descriptors from the pool of calculated descriptors. Effective feature selection can improve model performance, reduce overfitting, and enhance interpretability.

Importance of Descriptor Selection

1. **Improves Model Performance**: Selecting relevant features can enhance the predictive power of the model.
2. **Reduces Overfitting**: Eliminating irrelevant or noisy features helps in building a more generalizable model.
3. **Enhances Interpretability**: Models with fewer, more relevant features are easier to interpret and understand.
4. **Reduces Computational Cost**: Fewer features mean reduced computational requirements, making the modeling process faster and more efficient.

Statistical filter methods such as correlation analysis, and principal component analysis, alongside machine learning techniques like recursive feature elimination, sequential feature selection or embedded methods like LASSO ((Least Absolute Shrinkage and Selection Operator) or tree-based models are employed to identify and retain the most pertinent features.

5.3. Model Building

Algorithm Selection

The choice of the algorithm is a critical determinant of the QSAR model's performance. Commonly employed algorithms include linear regression, random forests, support vector machines, and neural networks, each with distinct strengths and applicability depending on the complexity of the data and the nature of the relationship between descriptors and biological activity. Each algorithm has its strengths and limitations, making it suitable for different types of data and modeling tasks. Researchers should consider the nature of their dataset, the complexity of the relationships between descriptors and biological activities, and the specific goals of their study when selecting an algorithm.

Data Partitioning

To facilitate model training and validation, the dataset is typically partitioned into a training set and a test set, with a conventional split of such as 70% for training and 30% for testing which is holdout method. This partitioning is simple and quick to implement but it is performance sensitive to the specific data split. k-fold cross-validation divides the data into k subsets; each subset is used as a test set while the rest serve as the training set. It is a technique primarily used for model validation, but it inherently involves data partitioning as part of its process. In general, typical k values are 5 or 10. The performance metric is

averaged across all k iterations to provide a comprehensive evaluation of the model. This method reduces bias, provides robust performance estimates but it is computationally intensive. Another technique for partitioning is leave-one-out cross-validation which uses each data point as a single test case; the model is trained on the remaining data. It utilizes all data for training, so provides less biased partitioning. However, it has the limitation that extremely computationally intensive. For imbalanced data stratified k-fold cross-validation is more reliable despite of its computationally limitations. It ensures each fold has the same class distribution as the whole dataset. Another technique is train-validation-test split. This technique splits data into three sets for training, validation, and testing like 60-20-20 or 70-15-15. It allows for hyperparameter tuning and model selection. On the other hand, it requires more data to allocate to all three sets.

The choice of partitioning method depends on dataset size, computational resources, and specific modeling needs. Proper partitioning techniques ensure robust model evaluation and enhance generalizability.

Model Training

The training set is utilized to fit the chosen algorithm, effectively establishing the relationship between molecular descriptors and biological activity. This process is supported by software tools such as scikit-learn (Pedregosa et al., 2011) for Python, tidymodels package

(Kuhn & Wickham, 2020) for R, AutoQSAR, or KNIME, which offer robust frameworks for model development.

5.4. Model Validation

Internal Validation

Internal validation techniques, including k-fold cross-validation, are employed to evaluate the model's robustness and generalizability.

External Validation

External validation entails assessing the model's predictive performance on the test set, which was not used during training. This step provides an unbiased estimate of the model's accuracy and its potential applicability to new, unseen data. The methods that are commonly used holdout method and external dataset. The external dataset method validates the model using an independent external dataset not used in the model development process. This provides an unbiased assessment of model performance.

Y-Randomization (Permutation Test)

Randomly shuffle the biological activity labels and rebuild the model. If the model’s performance significantly drops, it indicates that the original model’s performance is not due to chance.

Applicability Domain (AD)

Defining the AD of the model is essential to ensure that predictions are reliable. The AD can be established using methods such as the leverage approach or distance-based methods, which delineate the chemical space within which the model's predictions are considered valid.

- **Define AD**: Establish the chemical space where the QSAR model makes reliable predictions. Methods like the leverage approach or distance-based methods can be used.
- **Assess Predictions**: Ensure that new compounds fall within the applicability domain of the model before making predictions.

Statistical Metrics

Statistical metrics are crucial for evaluating the performance of QSAR models. They help in assessing how well a model predicts the biological activity of compounds based on their chemical structure. Several statistical metrics commonly used in QSAR model validation are R^2 (Coefficient of Determination), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), Q^2 (Predictive Squared Correlation Coefficient), RMSEP (Root Mean Square Error of Prediction), Bias and Variance, RSS (Residual Sum of Squares), F1 Score, AUC-ROC (Area Under the ROC Curve), and MCC (Matthews Correlation Coefficient).

These statistical metrics are critical for assessing the performance and robustness of QSAR models. Using a combination of these metrics

provides a comprehensive evaluation, helping to ensure the models are accurate, reliable, and generalizable.

Visual Validation

- **Residual Plots**: Plot the residuals (difference between predicted and actual values) to check for patterns. Random distribution of residuals indicates a good model fit.
- **Parity Plots**: Plot predicted vs. actual values to assess how well the model predicts across the range of data.

5.5. Model Optimization

Hyperparameter Tuning

To enhance the model's performance, hyperparameters (i.e., parameters that govern the learning process) are tuned using techniques such as grid search, random search, Bayesian optimization, gradient-based optimization, hyperband and successive halving. These methods systematically explore different combinations of hyperparameters to identify the optimal settings.

Ensemble Methods

Incorporating ensemble methods, which combine multiple models to form a single predictive model, can further improve accuracy and mitigate overfitting. Common ensemble techniques include bagging,

boosting, and stacking, each offering distinct advantages in model performance enhancement.

5.6. Model Interpretation and Analysis

Result Interpretation

Understanding the relationships between descriptors and biological activity is crucial for deriving actionable insights from the QSAR model. Techniques such as feature importance analysis and partial dependence plots can elucidate these relationships, offering transparency and interpretability.

Visualization

Visualization tools play a pivotal role in communicating the model's performance and insights. Graphical representations such as residual plots, parity plots, and feature importance charts aid in illustrating the model's accuracy and the significance of individual descriptors.

5.7. Deployment and Application

Model Deployment

Once validated, the QSAR model can be deployed within a software tool or platform for routine use in drug discovery projects. This involves integrating the model into the workflow of the research team, ensuring it is accessible and user-friendly.

Continuous Monitoring

Ongoing monitoring and periodic updating of the QSAR model are imperative to maintain its accuracy and relevance. As new data becomes available, the model should be retrained and revalidated to incorporate the latest information and maintain its predictive power.

6. Challenges and Limitations in QSAR Modeling

6.1. Data Quality and Availability

Issue: QSAR models heavily rely on the quality and quantity of available data. Inconsistent, sparse, or erroneous data can lead to inaccurate models.

Solution: Ensuring high-quality, standardized data through rigorous experimental procedures and data curation practices. Public databases like ChEMBL and PubChem can be valuable resources for high-quality datasets (Gaulton et al., 2012).

- ***Standardization****:* Implementing standardized protocols for data collection and curation to ensure consistency and reliability.
- ***Data Augmentation****:* Using data augmentation techniques to artificially increase the size of datasets, thereby improving model training (Shorten & Khoshgoftaar, 2019).

6.2. Model Interpretability

Issue: Complex models, especially those using machine learning and deep learning, can be difficult to interpret, making it challenging to understand the relationships between molecular descriptors and biological activity.

Solution: Developing interpretable machine learning models and incorporating feature importance analyses to elucidate the key descriptors driving model predictions (Ribeiro et al., 2016).

- ***Explainable AI (XAI)**:* Developing and applying XAI techniques to make machine learning models more interpretable. Techniques like SHAP (SHapley Additive exPlanations) values can help explain individual predictions by attributing them to specific features (Lundberg & Lee, 2017).
- ***Visualization Tools**:* Utilizing visualization tools to illustrate the relationships between molecular descriptors and predicted activities, making it easier to interpret and communicate model results.

6.3. Generalizability

Issue: QSAR models trained on specific datasets may not always generalize well to new, diverse chemical spaces. This limitation can reduce the model's applicability to novel compounds.

Solution: Using diverse training datasets and cross-validation techniques to improve model generalizability. Transfer learning approaches can also help adapt models to new data (Chen et al., 2019).

6.4. Overfitting

Issue: Overfitting happens when a model is overly complex and captures noise in the training data instead of the underlying patterns, resulting in poor performance on new data.

Solution: Implementing regularization techniques, cross-validation, and pruning methods to prevent overfitting and ensure model robustness (Hawkins, 2004).

6.5. Descriptor Selection

Issue: Choosing suitable molecular descriptors is essential for effective QSAR modeling. Poor descriptor choice can lead to inaccurate predictions.

Solution: Utilizing feature selection methods and domain expertise to identify relevant descriptors that capture the essential chemical and biological properties (Tropsha & Golbraikh, 2007).

7. Comparative Analysis: Traditional vs. Modern QSAR Models

7.1. Traditional QSAR Approaches

Advantages

a. Simplicity

- **Implementation**: Traditional QSAR methods like Linear Regression and Multiple Linear Regression are straightforward to implement using basic statistical software.
- **Understanding**: The mathematical foundation of these models is simple, making them accessible to a wide range of researchers.

b. Interpretability

- **Coefficient Analysis**: The coefficients in linear models provide clear insights into how each molecular descriptor influences the biological activity.
- **Transparency**: The models are transparent, allowing researchers to understand the decision-making process.

c. Computational Efficiency

- **Speed:** Traditional models require less computational power and can be run on standard hardware.
- **Resource Usage:** They are suitable for small to medium-sized datasets, ensuring efficient use of computational resources.

d. Established Techniques

- **Reliability:** Traditional QSAR methods have been used for decades, providing a reliable framework for certain types of data.

- **Reproducibility:** These methods are well-documented and widely understood, facilitating reproducibility of results.

Limitations

a. Linearity Assumption

- **Model Limitation:** Assumes a linear relationship between descriptors and biological activity, which may not hold for complex biological systems.
- **Performance:** May fail to capture the true nature of the relationship, leading to suboptimal performance.

b. Limited Flexibility

- **Non-linearity:** Traditional methods struggle with capturing non-linear relationships and interactions between descriptors.
- **Feature Interaction:** They may not account for complex interactions among multiple features.

c. Overfitting

- **High-Dimensional Data:** With many descriptors relative to the number of data points, these models can overfit the training data.
- **Generalizability**: Overfitting reduces the model's ability to generalize to new, unseen data.

d. Manual Descriptor Selection

- **Feature Engineering**: Requires significant manual effort to select and engineer relevant descriptors.
- **Bias:** Manual selection can introduce bias and may miss important features.

7.2. Modern QSAR Approaches

Advantages

a. **Handles Non-Linearity**
   - **Complex Relationships**: Machine learning techniques like Random Forests, Support Vector Machines, and Neural Networks can model complex, non-linear relationships.
   - **Accuracy**: Typically achieve higher predictive accuracy on diverse and complex datasets.

b. **Higher Predictive Power**
   - **Performance**: Modern models generally outperform traditional models in terms of predictive accuracy and robustness.
   - **Flexibility**: Capable of handling a wide variety of data types and structures.

c. **Automatic Feature Selection**
   - **Embedded Methods**: Techniques like Random Forests inherently perform feature selection, identifying the most relevant descriptors automatically.
   - **Efficiency**: Reduces the need for manual feature engineering and selection.

d. **Advanced Techniques**
   - **Deep Learning**: Advanced neural networks can learn hierarchical representations of data, capturing intricate patterns and dependencies.
   - **Graph-Based Models**: Techniques like Graph Neural Networks can directly model molecular structures.

Limitations

a. **Complexity**
   - **Interpretability**: Modern QSAR models are often seen as "black boxes," making it difficult to interpret and understand the decision-making process.
   - **Transparency**: Lack of transparency can hinder the understanding of how predictions are made.

b. **Computational Intensity**
   - **Resource Requirements**: These models require significant computational power and time, especially for training on large datasets.
   - **Infrastructure**: May necessitate specialized hardware like GPUs and high-performance computing resources.

c. **Risk of Overfitting**
   - **Model Complexity**: High-capacity models like deep learning can overfit the training data if not properly regularized.
   - **Validation**: Requires rigorous validation techniques to ensure generalizability.

d. **Implementation Complexity**
   - **Technical Expertise**: Building and tuning modern QSAR models require advanced knowledge in machine learning and computational techniques.
   - **Tooling**: Utilizes advanced software libraries and frameworks, which may have a steep learning curve.

The comparison made above are summarized in the *Table 1*.

Table 1. Comparative Analysis

Feature	Traditional QSAR Models	Modern QSAR Models
Simplicity	Easy to implement and understand	Complex implementation requiring advanced expertise
Interpretability	Transparent and easy to interpret	Often seen as "black boxes"
Computational Efficiency	Low computational requirements	High computational demands
Handling Non-Linearity	Limited to linear relationships	Capable of modeling complex, non-linear relationships
Predictive Power	Moderate	Generally higher
Feature Selection	Requires manual selection	Often includes automatic selection
Risk of Overfitting	Prone with high-dimensional data	Can overfit if not properly regularized
Manual Effort	Significant in feature engineering	Reduced due to automated methods
Infrastructure	Requires standard hardware	May need specialized hardware like GPUs

Established Techniques	Long-standing and reliable	Cutting-edge but evolving

The choice between traditional and modern QSAR models depends on the specific needs of the research. Traditional models are suitable for simpler, linear relationships and are valued for their simplicity and interpretability. Modern QSAR models, though more complex and computationally intensive, offer superior performance and flexibility, making them ideal for handling complex datasets with non-linear relationships. Understanding the strengths and limitations of each approach allows researchers to select the most appropriate modeling technique for their specific applications.

8. Conclusion

QSAR models have transformed the drug discovery landscape by providing a powerful computational tool to predict the biological activity of compounds. This transformation has led to significant reductions in time and cost associated with drug development. As the field continues to evolve, QSAR models are expected to become even more integral to the drug discovery process, contributing to the identification of new therapeutics and the optimization of existing ones.

The future of QSAR modeling is bright, with numerous opportunities for further innovation and impact. Continued advancements in computational methods, data integration, and interdisciplinary collaboration will drive the development of more accurate, robust, and versatile QSAR models. These models will play a pivotal role in

advancing personalized medicine, improving drug safety and efficacy, and ultimately enhancing patient outcomes.

As the field progresses, it is imperative for researchers and practitioners to remain informed about the latest advancements in QSAR modeling and to engage actively in addressing ongoing challenges. Such proactive engagement will be crucial for harnessing the full potential of QSAR models.

QSAR models have profoundly impacted the drug discovery and development landscape. Over the decades, these models have evolved from simple linear relationships to complex, machine-learning-driven algorithms that integrate vast amounts of chemical and biological data. QSAR models offer significant advantages in terms of efficiency, cost reduction, and predictive power, making them indispensable tools in modern medicinal chemistry.

In summary, QSAR models have revolutionized the drug discovery process, offering powerful tools for predicting biological activity, optimizing lead compounds, and ensuring drug safety. With continued innovation and collaboration, the future holds great promise for further advancements in this dynamic and impactful field.

REFERENCES

Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS Central Science, 3(4), 283-293.

Barabási, A. L., Gulbahce, N. & Loscalzo, J. (2011). Network-based strategies to understand the pharmacology of complex diseases. Nature Reviews Genetics, 12(1), 56-68.

Chemical Computing Group ULC. (2024). MOE (Molecular Operating Environment). Retrieved from https://www.chemcomp.com/.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250. Doi: 10.1016/j.drudis.2018.01.039

Chen, X., Kelly, J. M., Wang, L., et al. (2019). Chemprop: Toward interpretable graph neural networks for predicting chemical properties. arXiv preprint arXiv:1904.01561.

Cheng, F., Zhou, Y., Li, W., Liu, G., & Tang, Y. (2013). Network-based approach to prediction and population-based validation of in silico drug repurposing. Nature Communications, 4, 2636.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to? Journal of Medicinal Chemistry, 57(12), 4977-5010.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. Journal of the American Chemical Society, 110(18), 5959-5967. doi:10.1021/ja00226a005

Cruz-Monteagudo, M., Medina-Franco, J. L., & Pérez-Castillo, Y. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?. Drug Discovery Today, 19(8), 1069-1080.

Duvenaud, D. K., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems, 2224-2232.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., ... & Clark, A. M. (2013). Exploiting machine learning for end-to-end drug discovery and development. Nature Materials, 12(8), 702-708.

Free, S. M., & Wilson, J. W. (1964). A Mathematical Contribution to Structure-Activity Studies. Journal of Medicinal Chemistry, 7(4), 395-399. doi:10.1021/jm00334a001

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., ... Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research, 40(D1), D1100-D1107. doi:10.1093/nar/gkr777

Geenen, S., de Mooij, T., Bakker, P., ... Wessels, L. F. A. (2021). Integrating proteomics and phosphoproteomics improves drug response prediction for targeted therapy. Cell Reports Medicine, 2(1), 100210.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning, PMLR, 70, 1263-1272.

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. Journal of Computational Chemistry, 38(16), 1291-1307.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 2672-2680.

Gysi, D. M., Valle, Í. D., Ramirez, D., ... Zitnik, M. (2020). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proceedings of the National Academy of Sciences, 117(40), 24644-24650.

Hansch, C., & Fujita, T. (1964). *p*-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. Journal of the American Chemical Society, 86(8), 1616-1626. doi:10.1021/ja01062a035.

Hansch, C., & Leo, A. (1995). Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. American Chemical Society. ISBN: 978-0841220667.

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18, 83.

Hawkins, D. M. (2004). The problem of overfitting. Journal of Chemical Information and Computer Sciences, 44(1), 1-12.

Hood, L., & Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotechnology, 29(6), 613-624. doi:10.1016/j.nbt.2012.03.004

Hung, C. & Gini, G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. Mol Divers 25, 1283–1299 (2021). https://doi.org/10.1007/s11030-021-10250-2.

Iorio, F., Knijnenburg, T. A., Vis, D. J., ... Saez-Rodriguez, J. (2016). A landscape of pharmacogenomic interactions in cancer. Cell, 166(3), 740-754.

Judson, R. S., Martin, M. T., Reif, D. M., ... Kavlock, R. J. (2010). Integrating pathway-based toxicology with QSAR modeling. Toxicological Sciences, 118(1), 20-30.

Kang, S., Mulholland, J., Hirao, H., & Chai, C. L. L. (2018). Conditional molecular design with deep generative models. Journal of Chemical Information and Modeling, 58(3), 574-582.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design, 30(8), 595-608.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research, 47*(D1), D1102-D1109.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

Kubinyi, H. (1993). Quantitative structure-activity relationships in drug design. In J. Dearden, C. S. Patterson, & J. R. Boicelli (Eds.), Rational Approaches to Structure, Activity, and Ecotoxicology of Agrochemicals. Springer.

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Retrieved from https://www.tidymodels.org

Landrum, G. (2016). RDKit: Open-source cheminformatics. *Journal of Cheminformatics, 8*(1), 56. https://www.rdkit.org

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539

Liu, R., Wu, C., & Shen, X. (2017). Bioactivity and toxicity prediction using machine learning: recent advances and perspectives. Bioinformatics, 33(14), 2266-2275.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765-4774.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. Journal of Chemical Information and Modeling, 55(2), 263-274.

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., ... Sittampalam, G. S. (2011). Impact of high-throughput screening in biomedical research. Nature Reviews Drug Discovery, 10(3), 188-195. doi:10.1038/nrd3368.

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science, 3, 80.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

Nguyen, L., Wicaksono, A., Vincent, T., Draghici, S., & Nguyen, T. (2019). Integration of multi-omics data for predicting drug-induced liver injury. Bioinformatics, 35(19), 3541-3548.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3(1), 33. doi:10.1186/1758-2946-3-33.

Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de novo design through deep reinforcement learning. Journal of Cheminformatics, 9, 48.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Science Advances, 4, eaap7885.

Ramsundar, B., Eastman, P., Walters, P., Pande, V., & Leswing, K. (2019). DeepChem: An Open-Source Toolkit for Deep Learning in Drug Discovery. *Journal of Chemical Information and Modeling, 59*(10), 3999-4009.

Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. (2015). Massively Multitask Networks for Drug Discovery. arXiv preprint arXiv:1502.02072.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Rix, U., Superti-Furga, G., ... More Authors, ... & Müller, A. C. (2007). Chemical proteomic profiles of the anticancer drugs imatinib and dasatinib: An integrated approach for kinase inhibitor selectivity profiling. Proceedings of the National Academy of Sciences, 104(51), 20552-20557.

Roden, D. M., Altman, R. B., & Pinto, Y. M. (2019). Pharmacogenomics 2019: An update. Clinical Pharmacology & Therapeutics, 106(3), 467-473.

Ryu, S., Kwon, Y., & Kim, S. (2019). A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. Chemical Science, 10, 8438-8446.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. Nature Communications, 8, 13890.

Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science, 4(1), 120-131.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48. doi:10.1186/s40537-019-0197-0.

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacological Reviews, 66(1), 334-395.

Srinivasan, B., Dror, Y., & McHardy, A. C. (2014). Next-generation sequencing and bioinformatics in drug development: challenges and opportunities. Drug Discovery Today, 19(9), 1362-1372.

Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Current Pharmaceutical Design, 13(34), 3494-3504. doi:10.2174/138161207782794072

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., ... Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. Advances in Neural Information Processing Systems, 27, 1-9.

Wang, Z., Nichols, R. G., Maldonado-Valderrama, J., Yu, L., & Patterson, A. D. (2019). A pathway-based approach for predictive toxicology using gene expression data and machine learning. Environmental Science & Technology, 53(17), 10308-10318.

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., & Lai, L. (2017). Deep learning for drug-induced liver injury. Journal of Chemical Information and Modeling, 55(10), 2085-2093.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing

learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370-3388.

You, J., Liu B., Ying R., Vijay P. & Leskovec J. (2018). Graph convolutional policy network for goal-directed molecular graph generation. Advances in Neural Information Processing Systems, 6410-6421.

ISBN: 978-6-25815-123-7