UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Validation Strategies for Robust Assessment of QSAR Models

Filipa Alexandra da Silva de Noronha Almendra

Mestrado em Engenharia Informática

Dissertação orientada por:
Prof. Doutor André Osório e Cruz de Azerêdo Falcão

2024

Dedico a minha Dissertação em Engenharia Informática à minha querida avó, Rosa Gonçalves da Silva, que me acompanhou diariamente desde o meu nascimento até parte do presente ano.

Agradecimentos

Quero agradecer à minha mãe, Helena Maria Gonçalves da Silva Almendra, por me apoiar incondicionalmente e por me ter dado tudo para chegar onde estou hoje. É uma inspiração para mim, especialmente pela sua força, resiliência e bom coração.

Quero agradecer aos meus colegas e amigos da FCUL que me acompanharam desde o primeiro ano de faculdade e que não só me ajudaram a integrar na faculdade e a superar inúmeros desafios, como também me fizeram ter a melhor experiência académica possível. André Mendes, Rafaela Rodrigues, Miguel Costa, Fábio Estanqueiro e Sérgio Ferreira.

Quero agradecer aos meus amigos e família, que me acompanham e se preocupam comigo, tendo um impacto positivo na minha vida.

Quero agradecer ao meu professor e orientador, André Falcão, por me ter despertado a sua atenção nas aulas de mestrado pelo seu envolvimento em ciência de dados, aplicado a bioinformatics e cheminformatics, por me ter aceite para desenvolver este projeto muito interessante e que foi exatamente de encontro ao que eu procurava, e por me ter acompanhado no decorrer do mesmo.

Resumo

A relação estrutura-atividade quantitativa (QSAR) é um conceito importante para a descoberta e design de medicamentos. Os modelos QSAR são um bom exemplo de ferramentas tecnológicas que são aplicadas para melhorar os processos de descoberta de medicamentos. Estes modelos podem ser utilizados por exemplo para a classificação de moléculas em ativas ou não ativas, com base em descritores moleculares. O conceito fundamental por detrás da possibilidade de fazer previsões da atividade molecular com base nas suas propriedades baseia-se no princípio de que moléculas com estruturas semelhantes geralmente apresentam valor de atividade semelhantes. Efetivamente, o desenvolvimento e validação destes modelos são regulamentados e existem várias condições que têm de ser cumpridas para serem efetivamente implementados. No entanto, mesmo com os modelos a cumprir estes requisitos e a apresentar baixos erros de validação, muitos modelos ainda não são utilizados no atual desenvolvimento de medicamentos ou na pesquisa farmacológica. Neste trabalho são exploradas algumas possíveis causas para este acontecimento. Entre as quais, foi analisada a qualidade de fazer previsões para duas abordagens de validação diferentes: retrospetiva e prospetiva. A abordagem prospetiva é uma abordagem de validação mais realista, pois o conjunto de teste contém apenas os dados mais recentes, a partir de um certo ano, para um dado alvo, o que permite simular a tarefa de fazer previsões futuras. Além disso, uma ferramenta QSAR foi utilizada para criar modelos baseados no espaço métrico para multi-dimensões com o objetivo de explorar o impacto da semelhança estrutural das moléculas no desempenho e capacidade de fazer previsões dos valores de atividade das moléculas dos modelos QSAR.

14 alvos não-relacionados e vários modelos de aprendizagem automática foram escolhidos para avaliar a qualidade das previsões para os dois tipos de validação. As estratégias de validação requerem formas diferentes de dividir os conjuntos de dados, em conjuntos de treino e teste, que podem ser aleatórias (retrospetivas) ou baseadas no ano em que as moléculas foram documentadas (prospetivas). Na fase de pré-processamento dos alvos, uma tarefa crucial é a representação das moléculas, uma vez que permite obter informação sobre as suas características estruturais e de atividade. Uma forma de representar moléculas é através dos fingerprints moleculares, sendo que os Morgan fingerprints representam a estrutura das moléculas através de vetores de tamanho fixo binários, o que é ideal para servir como input para os modelos de aprendizagem automática. Além disso, para realizar a avaliação do desempenho dos modelos foram utilizadas métricas de classificação, tais como exatidão, precisão, recall, f1-score e Matthew's correlation score. O teste estatístico não-paramétrico, teste de Friedman, é também utilizado para comparar a performance dos vários modelos.

Neste projeto foi feita uma comparação em termos de desempenho entre a validação de modelos QSAR com uma abordagem retrospetiva e prospetiva. Esta fase do trabalho envolveu recolher dados, realizar o pré-processamento de dados, escolher modelos de aprendizagem automática adequados ao contexto e validar os modelos. Esta parte do estudo permite identificar uma grande discrepância ao testar os modelos com e sem considerar o ano em que as moléculas foram registadas. Após realizar estas tarefas, os resultados foram analisados e foi estudado o possível impacto da semelhança estrutural das moléculas. As abordagens de modelação nesta fase foram Support Vector Machines, Random Forests, Extreme Gradient Boosting e Redes Neuronais. Os resultados revelam, em média, uma diferença de 35% e a diferença mais significativa de 65% entre as duas abordagens de validação. Esta grande discrepância é preocupante, uma vez que a abordagem prospetiva corresponde a uma validação que se aproxima mais da ação de realizar previsões para o futuro. Os resultados apresentados ao longo deste estudo revelam em vários momentos que considerar apenas os dados mais recentes para teste tem um grande impacto na performance dos modelos.

Com base nos resultados obtidos para ambos os tipos de validação foi possível identificar um padrão, em que a validação retrospetiva revelou estar associada a uma menor diversidade de moléculas em termos de semelhança estrutural das moléculas, sendo que a distância medida entre as moléculas foi inferior em pelo menos três casos. O que sugere que para a validação retrospetiva as moléculas têm uma maior semelhança estrutural entre si, o que não se reflete na validação prospectiva, onde os modelos têm mais dificuldade em fazer previsões com moléculas novas ou diversas. Relativamente à análise do possível impacto da estrutura das moléculas foram criados conjuntos de dados, na qual apenas moléculas consideradas diferentes entre si foram adicionadas. A medição do nível de semelhança entre as moléculas foi feita através do cálculo da distância entre as moléculas do conjunto de treino e do conjunto de teste, utilizando o índice de jaccard. Moléculas com um índice de semelhança inferior a 60% são consideradas distintas. Após a criação dos conjuntos de treino e teste foi possível treinar e validar os modelos existentes com os mesmos.

No contexto de descoberta de medicamentos existe um princípio em que se assume que moleculas com estruturas semelhantes tem tendência a ter perfis de atividade semelhantes. Este princípio é utilizado como base em muitas ferramentas e técnicas computacionais para a previsão de bioatividade, na qual a semelhança estrutural intermolecular é um indicador chave para inferir o comportamento farmacológico. Para testar o impacto desta suposição foi utilizada uma ferramenta automática para modelação QSAR, uma ferramenta desenvolvida fora deste projeto. Este algoritmo completo para modelação QSAR foi desenhado para automatizar o processo de construção de modelos QSAR, desde recolher e selecionar dados até à avaliação do desempenho dos modelos. Esta ferramenta distingue-se dos modelos de aprendizagem automática frequentemente utilizados para problemas QSAR, na medida em que foi criada de raiz e aborda aspetos críticos da construção de modelos QSAR. Desta maneira, foi possível estudar o impacto de utilizar uma abordagem baseada na estrutura e semelhança intermolecular comparativamente a utilizar modelos de aprendizagem automática para modelação QSAR. Ao utilizar estes modelos com base

em espaços métricos foi explorado de que maneira a semelhança estrutural molecular pode afetar a qualidade dos modelos QSAR.

Após realizar as duas experiências para estudar o impacto da estrutura molecular na previsão de atividade dos alvos, através da criação de conjuntos de dados não redundantes e da implementação da ferramenta baseada no espaço métrico, foi observado que os resultados não são melhores que os obtidos com os avançados e gerais modelos de aprendizagem automática. Desta forma, é sugerido que outras características moleculares são necessárias para capturar completamente as nuances das ligações moleculares. Neste caso, o princípio de similaridade, embora importante, não revela ser suficiente para explicar a bioatividade, exigindo o desenvolvimento e implementação de uma abordagem mais detalhada em modelação QSAR. O sucesso dos modelos do estado da arte comparado aos de espaço métrico sugere que relações complexas entre características moleculares e atividade vão além da similaridade estrutural.

Estes resultados apresentam evidências de um problema, para o qual devem ser levantadas preocupações relativamente a incluir validação prospetiva, juntamente com medidas preventivas para o baixo desempenho dos modelos, na modelação QSAR. Este trabalho permitiu estudar a diferença entre uma abordagem de validação retrospetiva e prospetiva no contexto de modelação QSAR. Foi possível explorar o impacto da semelhança estrutural entre as moléculas dos vários alvos, utilizando uma ferramenta que permite a criação de modelos com base em espaços métricos para multi-dimensões e também através da criação de data sets não redundantes com base na distância intermolecular. Sem dúvida, este tema é atual e relevante, e este trabalho permitiu contribuir para a área de modelação QSAR.

Palavras-chave: Relação Estrutura-Atividade Quantitativa (QSAR), Aprendizagem automática, Validação, Prospectiva, Descoberta de medicamentos

Abstract

Quantitative Structure-Activity Relationship (QSAR) is vital in drug discovery and design. QSAR models are a good example of how technological tools are applied to elevate drug discovery processes, and they can classify molecules as active or non-active based on molecular descriptors. The development and validation of these models are regulated, and various measures must be met for their implementation. However, even with models meeting these conditions and presenting low validation errors, these models are still not widely used in current drug development or pharmacological research.

This work explores a possible cause for this event. More specifically, I will study the quality of predictions for two different validation approaches: retrospective and prospective. The prospective approach is a more realistic validation approach as the test set only contains the latest records documented for a particular target, simulating the task of making future predictions. In addition, the impact of structural similarity on QSAR modelling will be assessed with hyperdimensional metric space models. 14 diverse targets and various models were selected to assess the quality of predictions for the two types of validation. The modelling approaches include Support Vector Machines, Random Forests, Extreme Gradient Boosting and Neural Networks. The validation strategies require different ways of partitioning the datasets, which can be random (retrospective) or based on the year the molecules were documented (prospective).

Results show, on average, a difference of 35% and the most significant difference of 65% between the two validation approaches. When implementing a prospective approach, this big discrepancy can be problematic as this approach is a more realistic validation than the retrospective approach.

These results provide evidence that should raise concerns about including a prospective validation alongside preventive measures for the low performance of the models when performing this necessary type of validation in QSAR modelling.

Keywords: Quantitative Structure-Activity Relationship (QSAR), Machine Learning, Validation, Prospective, Drug Discovery

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Drug discovery is a multidisciplinary field that combines efforts and research to treat and prevent diseases. The methodologies used to investigate and develop effective drugs have evolved alongside the knowledge and resources available over the years. [68] Nowadays, computational techniques are largely integrated into many drug discovery processes, such as target identification or lead optimisation. [51, 62] Quantitative Structure-Activity Relationship (QSAR) models are a clear and good example of how technological tools are applied to support and elevate drug discovery processes. [18, 20] These flexible models can be constructed by combining different parts and methodologies addressing major issues in the area. In general, QSAR modelling allows a preliminary *In Silico* - experiment using a computer or through computer simulation - evaluation of the properties of molecules, in which savings in terms of money, time, and resources can be achieved. [8, 25, 51, 9]

Despite all the rigorous foundations for QSAR model fitting and validation, there is still a huge gap between produced models and their usage in the industry. Many models proclaim extremely low validation errors. [69, 54, 50, 71, 29] Which raises the question, why are they not widely accepted in the pharma-industry or why do most models, even though fully published and deployed in websites are not used in actual drug development or pharmacological research. [65, 19, 32, 11, 34] Besides the modelling approach or the quality of the existing data, it is necessary to have a realistic assessment of the quality of QSAR models, eventually by fitting models with old data and evaluating them using data collected afterwards, instead of already available data, which typically is biased since after an active compound has been found, many similar molecules are researched around it thus many times sharing similar activity properties [63]. To such an extent, in this work, I propose to compare, for 14 well-known and not related targets, the quality of predictions using two different validation approaches. Firstly, a Retrospective Validation, where data from the existing dataset is randomly selected for validation, and a Prospective Validation, where for a given target data published until a given date is used for training, and all data collected afterwards will be used for validation. As far as I know, this hasn't been fully assessed yet. Naturally, it is impossible to avoid biases completely, as new studies may be based on existing scaffolds and adapt from them. Nonetheless, results from prospective validation models are expected to be significantly worse than those from retrospective validation.

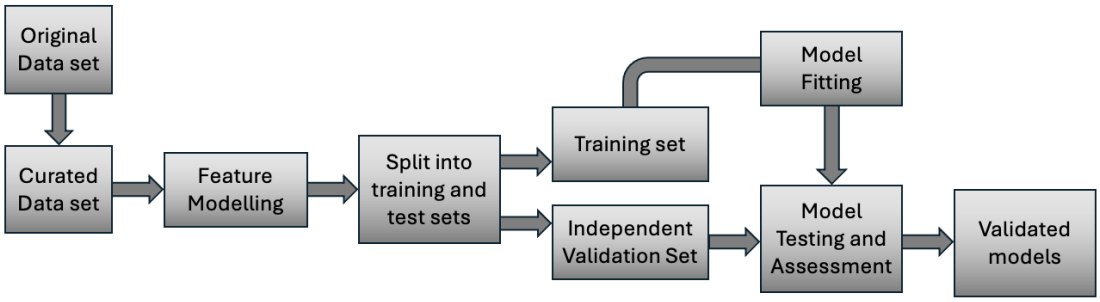


Figure 1.1: QSAR modelling workflow.

To further remove any model bias, it is a goal to test and use several state-of-the-art modelling approaches for each dataset and select the best model for comparison so that any possible modelling bias will be removed. It is further important to highlight that this study does not aim to identify a better-performing model but to explore the impact of different validation types by training existing machine learning (ML) QSAR models with data extracted from an online database and different data set partitions.

This work aims to present insights into the impact of the different validation types in the context of QSAR models by selecting different approaches to work with targets chosen for this study. Considering a common QSAR modelling workflow 1.1, this study will explore an additional step between feature modelling and splitting the curated data sets into train and test sets to help answer the question: How can models that were validated with data generated after the training phase, make accurate predictions for new molecules?

Several steps were performed in this work to address the abovementioned issues, aiming to complement experimental drug discovery processes with the advantages of data, machine learning techniques and available computational power. Firstly, be up-to-date with the state of the art and challenges addressed or to address in the field. Secondly, gather relevant and representative data sets from an online database. Next, data must be pre-processed to ensure the sets do not contain irregular or incomplete entries. Choose different machine learning models that are adequate for the type of problem in question. Following this step, the data sets can be partitioned according to the two approaches being studied and used for training and testing the models selected. Finally, the results can be analysed, considering the research question and hypothesis.

In drug discovery, it is often assumed that molecules with similar structures tend to exhibit identical bioactivity profiles. This principle underlies many computational strategies for predicting the bioactivity of compounds, where structural similarity serves as a key indicator for inferring pharmacological behaviour. To test this assumption more robustly, a metric space (MS) tool [38] was employed, with a further extension from the original two-dimensional framework to N dimensions. This extension allows for a more detailed and nuanced representation of molecular structures, enabling the exploration of higher-dimensional relationships between structural features and bioactivity.

The algorithm maps complex molecular features into a lower-dimensional metric space while retaining crucial information related to pharmacological properties. This tool, tailor-made for the context of QSAR modelling, allows the classification of molecules as active or non-active using probabilistic models. This open-source QSAR modelling tool presents some advantages compared to related work. The main advantage of this framework is that it automates many steps of QSAR modelling and includes valuable additions for enhanced efficiency and robustness. Additionally, it provides visual support, specifically a molecular space visualisation that may ultimately be used for visual structure-activity relationships (SAR) analysis. It should also increase the understanding of SAR in the drug discovery field. In addition, the models' workflow design considers the possible lack of machine learning/programming expertise from users by presenting a tool that does not require advanced parameterisation and does not depend on users' decisions.

This algorithm will be tested considering the nature of the available data, which is crucial when developing tools that aim to be implemented in the real world, as it can define whether the tool will be able to present accurate results according to reality or not and that experts from various areas will be able to use confidently.

1.1 Structure of the document

This document is organised as follows:

- Chapter 2 - Background
- Chapter 3 - Related Work
- Chapter 4 - Methodology
- Chapter 5 - Data
- Chapter 6 - Retrospective vs Prospective validation results
- Chapter 7 - Impact of structural similarity on QSAR modelling
- Chapter 8 - Conclusion

Chapter 2

Background

The following chapter addresses relevant concepts and definitions needed to comprehend and conduct the present work.

2.1 QSAR modelling

Classification problems are a fundamental aspect of machine learning, where the goal is to assign data points to predefined categories or classes based on input features. In classification, the model learns from labeled training data to identify patterns and relationships between features and class labels, ultimately predicting the correct category for new, unseen data. Challenges in classification include imbalanced datasets, where some classes are underrepresented, and overfitting. [3]

In QSAR modelling, classification problems often involve predicting whether a chemical compound will have a particular biological activity or fall into a specific toxicity category. Here, the input features are molecular descriptors that capture the chemical properties of the compounds, and the target variable is typically binary or categorical, such as ”active” or ”inactive”. The QSAR model classifies compounds based on their structural properties, aiming to predict their biological behavior. The goal is to build models that can reliably predict the activity or toxicity of novel compounds. [51, 12]

There are several types of QSAR, such as 1D-QSAR, 2D-QSAR, 3D-QSAR, 4D-QSAR, and string-based and chemical descriptor-based.[51] String-based QSAR relies on activity prediction using SMILES (Simplified Molecular Input Line Entry System) strings, which allow for a rigorous structure specification of chemical species using short ASCII representations. [67] This notation offers a simple and efficient way to work with molecules programming-wise. In contrast, chemical descriptor-based QSAR relies on numerical descriptors that quantify different aspects of molecular structure, such as molecular weight, polarity, or 3D shape. These descriptors are derived from the molecular structure and are used to build predictive models. Moreover, combining string-based and chemical descriptor-based approaches by using SMILES to generate descriptors for prediction is possible.

2.2 Target-Ligand Interactions

In the context of drug discovery, a pharmacological target is typically a biological molecule, such as a protein, enzyme, or receptor associated with a specific disease or biological pathway. The goal of drug design is to identify compounds (ligands) that can bind to and modulate the activity of a target to achieve a desired outcome in treating a disease. In other words, a drug (ligand) can physically attach to a target molecule, as properties of the ligand complement specific areas of the target (bind). Once the ligand is bound with the target, modulation corresponds to how the drug influences the target's activity by either increasing (activating) or decreasing (inhibiting) the activity of the target molecule. After this process, a therapeutic effect can be achieved. [40, 23] These interactions, target-ligand-based interactions, describe how a ligand (small molecule or drug candidate) binds to the target. The strength and nature of these interactions determine the ligand's biological activity. Understanding these interactions is key to predicting whether a ligand will successfully influence the target's function, such as inhibiting a protein or activating a receptor. QSAR connects a ligand's structure to its potential interactions with a biological target, allowing for predictions of target-ligand interactions before experimental testing. This accelerates drug discovery by identifying promising candidates likely to bind to a target effectively.

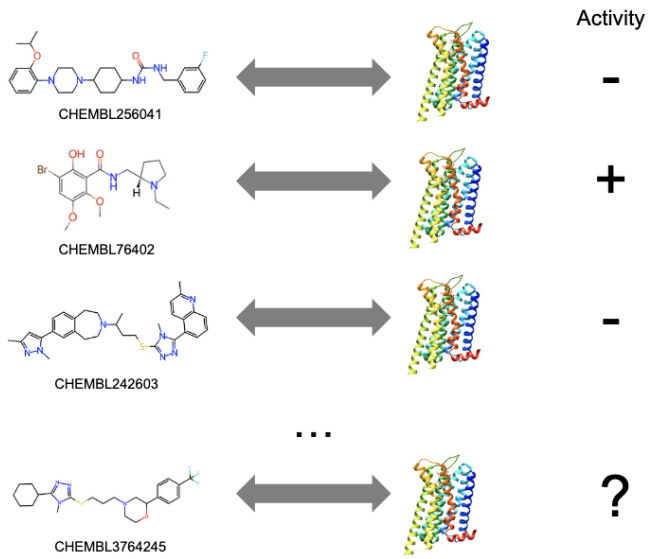
Furthermore, the bioactivity value of a drug reflects its potency and binding affinity to a specific receptor. A lower bioactivity value indicates higher potency and stronger binding affinity, requiring a lower concentration to achieve its effect. Binding affinity shows the strength of the drug-receptor interaction. QSAR models are essential in predicting new drug candidates' bioactivity and binding affinity. These can be based on machine learning models, which can be used for regression and classification purposes, where the predictors consist of physicochemical properties. 2.1The fundamental concept of predicting molecular activity based on their properties relies on the principle that molecules with similar structures generally exhibit similar activities.[41] In QSAR studies, researchers have explored different ways to represent molecular structures, aiming to predict the quantitative relationships between biological activity and structural features. [39, 12]

2.3 Representation of molecules

In QSAR studies, the representation of molecules is essential for understanding the relationship between structural features and biological activity and for assessing molecular similarity.

Molecular descriptors/fingerprints are the numerical representations of a molecular structure, playing an important role in QSAR as they allow for the assessment of molecular similarity. Several fingerprint types include Morgan, FeatMorgan, AtomPair, Torsion, RDKit, Avalon, Layered, MACCS and Pattern. [52, 38]

Among these descriptors, Morgan Fingerprints can be represented in binary format, producing fixed-length vectors, which are well-suited for input into machine learning algorithms. Generating Morgan Fingerprints is also computationally efficient, which is beneficial when dealing with potentially large data sets. They are widely used in cheminformatics and have been proven to



CHEMBL256041	0	1	0	...	0	0	N
CHEMBL76402	0	0	0	...	0	0	P
CHEMBL242603	0	1	0	...	0	0	N

Figure 2.1: QSAR approach, where the descriptors of the molecules with known activities are used to train a model that can predict the activities of untested molecules.

perform well in various QSAR studies. [13, 17]

2.4 Modelling approaches

Among the currently explored methods for classification in QSAR, using technological approaches, existing machine learning models, such as Logistic Regression (LR) [45, 2, 46], K-Nearest Neighbours (KNN) [1, 36], Support Vector Machine (SVM) [14, 39, 70, 31, 4], Random Forest (RF) [61, 70, 14, 39] and Extreme Gradient Boosting (XGBoost) [4, 70] are widely used. For instance, the SVM, RF, and XGBoost models were used in a study to discriminate between tuberculosis inhibitors and tuberculosis noninhibitors. [70] In addition, deep learning algorithms have also been included in the most recent studies with many variants of Neural Networks (NN). [27, 26] In the field of drug discovery, these models can be used for different applications and combined with other techniques to contribute to more efficient and complete processes, taking advantage of the data available. For instance, the classification of compounds (as active or inactive), the virtual screening of database compounds, and investigating the impact of molecular representations are some tasks performed in the area while resourcing these machine learning models. [14]

2.4.1 Support Vector Machine

The SVM model is a supervised learning algorithm commonly used for classification problems and is particularly good for binary classification. The key idea is to find the optimal hyperplane to separate data points in a high-dimensional space, separating data points from different classes. It maximises the margin, or the distance, between this hyperplane and the nearest data points from each class, called support vectors. This maximised margin helps the model generalise new, unseen data better.

For cases where the data is not linearly separable, SVM uses kernel functions to map the data into a higher-dimensional space, making it easier to find a separating hyperplane, such as Linear, Radial Basis function, Polynomial and Sigmoid kernels. Two additional parameters that allow balance bias and variance preventing underfitting or overfitting are C and *gamma*. C is a regularisation parameter that controls the trade-off between maximizing the decision boundary margin and minimizing misclassification errors. A smaller C allows a wider margin, reducing the chance of overfitting and penalises less the misclassified points, while a larger C narrows the margin, potentially increasing accuracy on the training set but risking overfitting. The *gamma* parameter affects the shape of the decision boundary by determining the flexibility of the model and the level of overfitting or underfitting of the training data. A larger *gamma* leads to more complex decision boundaries, while a smaller *gamma* results in smoother ones.

By identifying the most important molecular features and creating a clear decision boundary, SVM helps predict whether a molecule is likely to be active or inactive against a given target. [68] In QSAR, SVM can help predict molecular activity by finding boundaries that separate active and inactive compounds. It has been used in past studies for screen radiation protection, gene interaction, anti/non-anticancer molecule classification, and kinase mutation activation. [14]

2.4.2 Random Forest

Random forest is an ensemble model that combines predictions from multiple trees to enhance model performance. It operates by constructing multiple decision trees during training, where each tree is built using a random subset of the data and a random selection of features. A decision tree is a model that splits data into branches based on feature decisions to make predictions. The random selection of features involves selecting a random subset of features from the full set of input features for each decision tree or model, ensuring each tree in the forest is built on different features, reducing the likelihood of overfitting. The final prediction of the RF is made by aggregating the outputs of these individual trees by taking the majority vote for classification tasks. Combining predictions of multiple decision trees can create a more robust and accurate model.

RF is valued in QSAR for its robustness and capability to handle complex, non-linear relationships between molecular descriptors and biological activity.

2.4.3 Extreme Gradient Boosting

The Extreme Gradient Boosting is also an ensemble model, particularly well-suited for classifying molecular activity, such as predicting whether a molecule is active or inactive against a biological target. Similarly to RF, the XGBoost is based on decision trees, but it differs significantly in how these trees are constructed and combined to make predictions. Unlike the bagging method used in Random Forests, where decision trees are built independently and their predictions are averaged, XGBoost uses a boosting technique. This approach constructs decision trees sequentially, with each new tree addressing the errors made by the previous ones. XGBoost employs gradient descent at each stage to reduce classification errors, progressively enhancing the accuracy of the model's predictions. While Random Forests provide a more straightforward and robust approach, XGBoost allows for more targeted and refined predictions, making it highly effective in molecular classification tasks. Additionally, it incorporates a regularization technique to prevent models from getting too complex, by penalising them.

In QSAR, XGBoost is preferred for its high performance and ability to capture detailed patterns in data.

2.4.4 Neural Network

A simple neural network is considered to have a different approach than machine learning models because it can learn from experience and previous errors with more autonomy when making predictions. A feedforward neural network (FNN) is the simplest type, where information flows from the input to the output layer to make predictions without loops - sending the output back into the network and using it as input. A FNN can have several configurations, such as Single Layer Perceptron (SLP) and Multilayer Perceptron (MLP). An MLP is a more advanced version of the SLP, with one or more hidden layers between the input and output. A neural network is a computational model with layers of interconnected neurons. Each neuron computes outputs by applying weights to inputs, adding a bias, and passing the result through an activation function. The network learns by adjusting these weights through backpropagation to minimize errors, enabling it to recognize patterns and make predictions.

In this context, the input is the molecular fingerprints, which are processed through the hidden layers before predicting whether a molecule is active or inactive. NNs in QSAR can model complex non-linear relationships and interactions between molecular features, enhancing predictive power. These MLPs allow for efficient training while still providing valuable insights into molecular classification.

2.4.5 K-Nearest Neighbors

The K-Nearest Neighbours model can be used for regression and classification problems, although the latter is more common. This simple supervised machine learning model is largely implemented and can handle both categorical and numerical data. This algorithm's strategy is to make predictions based on the distance of a new point to the K nearest observations. The distance of a query

point to the others can be measured in several ways: Euclidean, Manhattan or Minkowski distance. After calculating the distances to all points, the K nearest neighbours are sorted and selected, and for classification problems, the majority vote among neighbours is used. However, this model is computationally expensive as it calculates the distance to every point in the training set, and it is also intensive in terms of memory, as it stores all data.

When constructing neighbourhoods in KNN, both the number of neighbours and the weighting of points within each neighbourhood can be adjusted to suit different problem requirements. The weight function used in predictions can be set to either a uniform or distance-based approach. A uniform weight function assigns equal importance to all points within the neighbourhood. In contrast, a distance-based function assigns weights inversely proportional to the distance from the query point, giving more influence to closer neighbours. Alternatively, there is the option to use a user-defined function.

In the context of QSAR, this model can be used either to classify if a molecule is active or not active or for regression to predict the activity value, using the molecular descriptors to train the KNN.

2.4.6 Bagging classifier with a Neural Network estimator

It was previously presented how a simple feedforward neural network, more specifically a multilayer perceptron, could provide efficient training of the data sets and be able to model complex non-linear relationships while being scalable to large data sets. Similar to what is implemented in a Random Forest, a bagging technique can enhance the robustness of machine learning models by combining predictions from multiple models trained on different subsets of the data. Bagging indicates a dataset is divided into multiple bootstrap samples - random subsets with replacement. Sampling with replacement means that samples are a subset of observations randomly selected, in which an observation can be used more than once. This ensures each model learns different aspects of the dataset, contributing to a more complete and diverse algorithm. The bagging classifier can receive a model as a base estimator and the number of estimators to use. For classification tasks, once all models have been trained, the final prediction is defined by the majority vote between all models.

Overall, having a neural network as a base estimator for a bagging classifier can help reduce overfitting by allowing a broader perspective of the data with the bootstrapping samples and reducing the impact of outliers/noise. In QSAR, these aspects are crucial to learning valuable insights from data while ensuring robustness, generalisation and scalability.

2.4.7 Logistic Regression

Logistic Regression is a supervised machine learning model used for binary classification problems. This model does not stand out as other approaches, although it can be effective and simple for problems that are not overly complex and are linearly separable. LR uses a logistic (sigmoid) function to map the input features into a probability value ranging between 0 and 1, where the goal

is to predict the probability that a given input belongs to one of two classes. The LR uses log-odds to facilitate the task of distinguishing the ratio of something happening to not happening, odds, and achieve symmetry in the problem by taking the logarithm of the odds, which helps create a clear linear separation. After calculating the log-odds, it is possible to map the values into the logistic function, assigning each point a direct probability. In the context of QSAR, this is applied to the ratio of the probability of a molecule being active to the probability of a molecule being inactive.

In QSAR, LR can be used to model the relationship between molecular descriptors and their activity values or to investigate which molecular features contribute to the desired or adverse biological activity.

2.5 Overfitting and underfitting

Furthermore, bias and variance errors are two terms related to statistical inference that are very relevant to understanding a model's behaviour and allow the refinement or development of better machine learning models. The concept of bias corresponds to the difference between the predicted and actual values, while variance is the variability of the predictions for different subsets of data. When building machine learning models, an optimum fitting is the right trade between not over-simplifying or overlearning during the training phase. A characteristic of a good machine learning model is that it can generalise unseen data well, allowing predictions to be made for future data. There can be two main situations when the model is not performing well during the training and testing phase: underfitting and overfitting. Suppose a model is not performing well neither in the training phase nor in the testing phase. In this case, the model is likely oversimplifying the learning process and not paying attention to detail - underfitting -, which means the model has high bias and low variance. A different case can be when a model performs well in the training phase but does not make accurate predictions on testing data; this suggests the model is incapable of generalising and is likely to be learning from noise and too many details, even though it predicts training data accurately - overfitting -, which suggests the model has low bias and high variance. Ideally, the model should have low bias and low variance to perform well on testing data.

2.6 Generation of hyperdimensional metric space models

As mentioned previously, a metric space tool will be employed to assess the impact of structural similarity when predicting the bioactivity of compounds and compare the performance of metric space-based models with machine learning models used in QSAR modelling. The workflow of this tool can be deconstructed in data access and processing, descriptors calculation, data set modelability estimation, feature selection, model building and validation. While steps such as data access and processing, feature selection, model building, and validation are familiar from traditional machine learning, the calculation of descriptors and estimation of data set modelability are crucial and transformative for this workflow. This section will address these topics and further

explain how the hyperdimensional metric models are generated.

2.6.1 Data access, processing and descriptors calculation

There are three possibilities to access data. Firstly, it is possible to fetch specific target compounds directly from the ChEMBL database, retrieving processed data with calculated descriptors and fingerprints ready to be used by the machine learning models. Secondly, it is also possible to have CSV files as input, whose descriptors and fingerprints are then calculated. Finally, the algorithm can also read previously prepared data, which contemplates activity values and descriptors/fingerprints.

One of the critical issues in QSAR modelling is dealing with the high dimensionality of data problems, which can be addressed by identifying and selecting a subset of descriptors able to predict the desired biochemical property. This way, the workflow was designed to calculate molecular descriptors using RDKit, and it can even compute nine fingerprint types. [38]

2.6.2 Data set modelability

The data set modelability feature, performed before time-consuming trials, consists of a preliminary assessment of the feasibility of constructing robust QSAR models by employing a designated descriptor space for a data set of molecules. To perform this, the modelability index (MODI) is calculated by resourcing to a k-nearest neighbours approach-based criteria [28], with k either 3 or 5. This way, modellers can get an estimation of the predictability of the computed models, where a low MODI index might indicate the data set is not recommended for model building.

2.6.3 Model building

Several steps are performed in model building, like feature selection, modelling algorithms, and handling overfitting, among others. The workflow is implemented with and without feature selection. The latter's predictive performance is assessed to verify the impact of the feature selection methods. Regarding the model built with feature selection, the method chosen uses a Random Forest to identify and rank the features according to their importance. Regarding machine learning approaches to build QSAR models, the SVM and RF were considered to be present in the QSAR modelling workflow, considering their capacity to deal with non-linear and complex problems and high dimensional data from QSAR problems, posing a lower chance of overfitting compared to other models. Optionally, a fingerprint filtering feature can be used to optimise making inferences by using a Random Forest to extract the N most important bits of the fingerprints. This parameter defines the percentage of best bits to select. Finally, to assess model predictability, the data was divided into training and test sets, and a stepwise estimation model was built and validated by gradually adding the ranked variables into new fitted models to find the best set of features with minimum predictive error. The model, internally validated using the best features, is then subjected to external validation.

2.6.4 Model validation

Regarding model validation, there are two important phases: a phase to perform an internal validation using an internal test to assess the quality of the model and a phase using an external test set (IVS) to verify the predictivity of the model for parts of the data sets that were not involved in the training process, allowing the retrieval of additional performance evaluations on the quality of the model, after the model is built and selected. The internal validation step involves using a K-fold cross-validation, where the data is divided into K folds, leaving a different subset for the validation set in each fold, ensuring the model learns from different parts of the data set and providing a robust estimate of the model's performance on unseen data. The proportion of variance explained (PVE) and root mean squared error (RMSE) are used for external evaluation of the models. PVE assesses how much of the variance in the observed data is captured by the model, where the higher the PVE, the stronger the predictive capability of the model. RMSE is the measure of the differences between the predicted and actual values, and it aims to have the lowest RMSE possible.

2.6.5 Molecular space visualisation

This tool offers a molecular space visualisation that is presented by generating a topographical map, where molecules are grouped depending on their structural similarity. To produce the visualisation of a metric space, it is recommended to perform a dimensionality reduction, in which the criterion for distance preservation asserts that any complex geometric structure inherent in the data manifold can be condensed into a reduced number of dimensions. Afterwards, the quality of this transformation is measured by the difference between the original distances and the distances in the projected space. [37] This way, addressing issues imposed by high-dimensional problems and presenting robust visual support for molecular visualisation is possible.

In other words, the algorithm includes a feature for calculating a probabilistic surface of molecular activity (PSMA), using molecular similarity as input, where a similarity matrix is transformed into 2-Dimensions (2D) using a dimension reduction algorithm, transforming similarities into distances and projecting them into a 2D space. This is followed by applying a kernel density estimation function that will compute the probability of activity for each coordinate in the new projected space.

Chapter 3

Related Work

To effectively harness the vast quantity of chemical data and the array of analytical tools available, it is essential to focus on the quality of the data as well as the methodologies applied. While advanced computational methods, such as QSAR models, offer significant potential for predicting chemical behaviour, their accuracy and reliability depend heavily on the integrity of the underlying data.

High-quality data is critical for constructing robust models and generating reliable predictions. This underscores the importance of rigorously assessing the quality of the QSAR methods and the data utilised to train these models. Data quality issues, such as incorrect representation of molecular structures or inaccurate activity information in public databases, can severely undermine the performance of QSAR models. These inaccuracies often stem from experimental errors and can propagate to model predictions, leading to suboptimal or misleading results. [74] Addressing these challenges requires a multi-faceted approach. One effective strategy is carefully selecting targets and compounds, ensuring the data used is as accurate and relevant as possible. Another crucial step is the manual curation of databases, which involves verifying and correcting data to minimise errors. Data splitting and cross-validation are also techniques widely implemented that allow the creation of diverse sets for training and testing phases that will enable testing the models and ensure the model can generalize well to unseen data. [64, 70, 29, 45] These sets are created randomly from the original data sets to ensure the splits represent the overall sets and focus on the relationship between chemical structures and their biological activities. The temporal aspect of the data sets is not typically used in traditional QSAR modelling for various reasons, like lack of time-dependent trends, random nature of data collection, or to focus on chemical diversity. However, for predicting the future activity of compounds, it can be relevant to perform a temporal validation that can reflect the fact that data is collected over the years and that when a researcher is developing new drug candidates based on existing knowledge, the models will be tested against a completely new and recent batch of compounds. This will ensure the model is forward-looking and can generalize well to new chemical structures rather than simply reproducing patterns in historical data.[24]

Moreover, advanced techniques such as feature selection and feature learning play a pivotal role in mitigating the impact of experimental errors. For example, feature selection reduces the

number of molecular descriptors, simplifies the model, and lowers computational costs while preventing overfitting. Focusing on the most relevant descriptors makes the models more efficient and enhances their interpretability, allowing researchers to understand how specific molecular features influence biological activity. This refined selection of features often leads to improved model accuracy, as targeted models generally outperform those overloaded with irrelevant or redundant descriptors. Although feature learning is generally less common than feature selection, it is a valuable alternative when computational efficiency is crucial. Feature learning automatically extracts a reduced set of new features directly from the chemical structure of compounds, eliminating the need for traditional molecular descriptors. While this approach can complicate model interpretability, it enables the efficient capture of essential structural patterns that traditional descriptors might miss. Both techniques have been extensively studied, independently and in combination, and research indicates that hybridizing feature selection and feature learning can significantly enhance model accuracy, particularly when they provide complementary information. This suggests that a hybrid approach, leveraging the strengths of both techniques, may offer a powerful solution for improving the quality and reliability of QSAR models, thereby enhancing the predictability and utility of chemical data analysis. [58]

Aspects related to the design and validation of QSAR models can directly impact their applicability. [35, 47] Past studies identify and set fundamental steps when designing and developing a QSAR model, and, effectively, there are models that present satisfactory results for different applicability domains and targets. [34, 20, 24] Strategies to make models more reliable and applicable for external prediction and regulatory purposes can be implemented in different phases and assess different topics, such as data, validation, models, and performance. Regarding validation, both internal and external validation are important for model evaluation, but external validation is less commonly applied. External validation tests the model on an independent dataset not used during model development and is essential for assessing its ability to generalize to new, unseen compounds. In contrast, internal validation evaluates the model's performance on the training data, offering insights into its fit but not guaranteeing accuracy for external data. Though internal validation refines the model, external validation is crucial for confirming its predictive accuracy in real-world applications, yet it is often limited by data availability or time constraints. [30, 65]

Clear and well-defined procedures ensure coherent and accurate results when screening chemical databases and virtual libraries. Valid QSAR models should comply to key principles, such as employing an unambiguous algorithm, defining the domain of applicability, and using appropriate metrics for goodness-of-fit, robustness, and predictivity. A crucial aspect of QSAR modelling is the applicability domain (AD), which defines the chemical space where a model can reliably make predictions. A well-defined AD is necessary to avoid extrapolating to compounds structurally different from those in the training set, as this could lead to unreliable predictions. To establish the AD, models must fulfil specific statistical criteria—such as correlation coefficients and coefficients of determination—ensuring their predictive reliability. Moreover, scientists should evaluate the relationship between the AD's size, the coverage of the virtual screening library, and the ac-

curacy of predictions. This balance is vital for identifying potential hits without compromising prediction reliability. External validation then assesses the model's predictive performance within the defined AD. Once a model demonstrates strong predictive power, it can be confidently applied to the virtual screening of chemical databases. Ultimately, rigorously defining and validating the applicability domain is key to enhancing the reliability and effectiveness of QSAR models in predicting chemical activities. These established principles enable the use of QSAR models in the regulatory assessment of chemical safety, providing model development and validation standards. [64, 43]

Chapter 4

Methodology

4.1 Validation Strategies

Two approaches were implemented to explore the impact of the type of validation in the QSAR models, as represented in Figure 4.1. On the one hand, a random data partition is implemented by splitting 20% of the whole dataset for the test set, which resembles a retrospective validation. This way, the models were trained and tested with data known beforehand and shuffled in terms of year of documentation. On the other hand, to simulate a prospective validation, the molecules for every target were ordered by the year of its documentation so that the most recent records would not be used for training. In this aspect, defining which records should be considered recent is necessary. Considering targets have a significant discrepancy in the year of documentation of their molecule records, it would not be viable to define one year for all targets for selecting the records for the test set. This way, it is more appropriate to select approximately 20% of the size of each dataset, ordered by the year of the documentation of molecules, for the test set. In table 6.1, it is possible to check the years of partition for each data set.

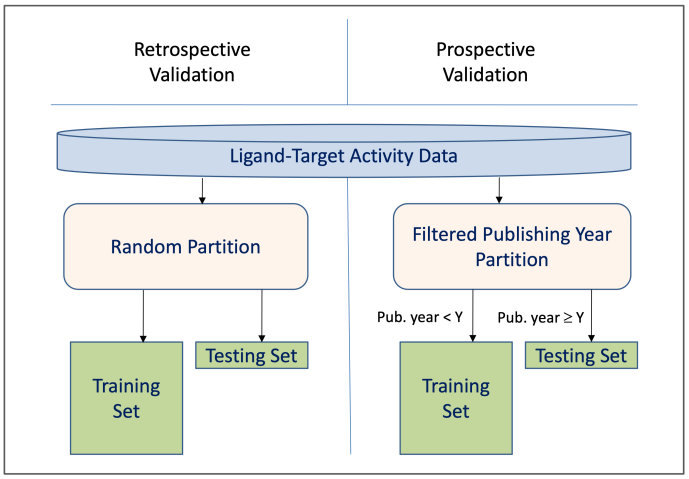


Figure 4.1: Visual representation of the two data partition approaches.

4.2 Dimensionality Reduction in Molecular Metric Space

Metric spaces are spaces where instances can only be defined relative to each other by distance or similarity function. [60] In this context, a metric space is an M X M dimensional distance matrix where M compounds are represented by M inter-molecular distances.

Dimensionality reduction in the context of molecular metric space refers to techniques used to simplify the high-dimensional data typically encountered in cheminformatics and QSAR modelling, where a large number of molecular descriptors represent each molecule. These descriptors, which quantify compounds' chemical and structural properties, often create a vast and complex feature space that can be difficult to manage due to redundancy, noise, and computational inefficiency.

By applying dimensionality reduction, the goal is to transform this high-dimensional molecular metric space into a lower-dimensional one while preserving the most important information. This makes the data more manageable and helps improve the performance of machine learning models by reducing noise, minimizing overfitting, and speeding up computations. The Principal Component Analysis (PCA) and feature selection are examples of techniques that can be used for dimensionality reduction. [38, 10, 33]

4.3 Assessment of structure similarity

As a first step to assess the impact of structural similarity, a new data set was created by removing structural redundancy. In other words, molecules that were considered too similar to others already present in the data set were not added, ensuring the remaining molecules were structurally diverse. This process included iterating over the molecules and measuring the distance between two molecules at a time by calculating the Jaccard index using their molecular fingerprints, where molecules are considered distinct if the index of similarity is 60% or inferior. After constructing the non-redundant data set, it is divided into training and validation sets, similar to the retrospective approach, and fitted using the machine learning models employed in the retrospective and prospective experiments. By comparing the performance of these models on the new data set, it will be possible to observe whether the absence of structural redundancy affects their ability to generalise and predict molecular bioactivity.

In parallel with the experiment of validating the new data set stripped of structural redundancy, hyperdimensional metric space models will be generated for the same targets. The models embed molecules into a high-dimensional space based on a predefined molecular distance metric (such as Tanimoto similarity). The goal is to construct a probability surface map that captures regions of molecular space where activity is more likely based on the similarity principle. The idea behind these models is to assess how well molecular similarity correlates with biological activity in a continuous, probabilistic framework. If the similarity principle is indeed the most influential factor, the regions of the hyperdimensional map associated with a higher probability of molecular activity are expected to align closely with actual areas of molecular activity. The performance of these

metric space models will be compared to those of the previous across all data sets. If the metric space models outperform the more sophisticated machine learning methods, this would further prove that structural similarity is the dominant driver of bioactivity, as the metric space models rely almost exclusively on the distance between molecules in structural space. In contrast, if the machine learning models outperform the metric space models, it would suggest that particular molecular features play a significant role in determining bioactivity.

4.4 Model evaluation

Following model fitting and testing, evaluating the models' performance is possible. For a classification problem, statistical metrics like Accuracy, Precision, Recall, F1 score, and Matthews Correlation Coefficient (MCC) offer a comprehensive quantitative evaluation of the classifier's performance. 4.1 Precision and Recall are calculated similarly; the first one measures the accuracy of positive predictions by focusing on how many of the predicted positives are actually correct, while Recall measures how effectively the model captures all actual positive cases from the dataset. Accuracy, Precision, Recall and F1 Score range from 0 to 1, and values closer to 1 can indicate better model performance. However, the relevance of each metric depends on the specific context and task. For example, accuracy is defined as the ratio of correct predictions to the total number of predictions; a higher accuracy value generally indicates more accurate classification, but it can be misleading in imbalanced datasets, where minority classes might be underrepresented.

Table 4.1: Classification Metrics for Model Evaluation

Metric	Expression	Description
Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$	The percentage of correct predictions out of all predictions.
Precision	$\frac{TP}{TP+FP}$	Measures the accuracy of positive predictions.
Recall	$\frac{TP}{TP+FN}$	Measures the ability to find all positive instances.
F1 Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of Precision and Recall, balancing the two.
Matthews Correlation Coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Balanced measure considering all four confusion matrix values.

The F1 score combines Precision and Recall into a single metric, calculated as the harmonic mean of the two, offering a balanced measure that is particularly useful in cases where there is a trade-off between Precision and Recall. This makes the F1 score valuable when false positives (FP) and false negatives (FN) carry similar consequences, as it penalizes extreme values in either Precision or Recall, aiming for a balance between the two. The MCC, on the other hand, takes into account all four components of the confusion matrix (true positives (TP), true negatives (TN), false positives, and false negatives) to provide a balanced measure that can handle imbalanced

datasets effectively. The MCC ranges from -1 to 1, where a value of 1 indicates perfect prediction, 0 indicates random prediction, and -1 signifies total disagreement between prediction and observation.

Together, these metrics provide a multi-dimensional view of model performance, enabling a deeper understanding of each model's strengths and limitations, as well as insights into the trade-offs inherent in classification tasks.

4.5 Model comparison

The Friedman test is well-suited for QSAR studies due to its simplicity, non-parametric nature, and focus on rank data. It provides a clear method for comparing multiple models across multiple data sets without making strong assumptions about the data distribution, making it a robust choice in this field. [16, 59, 39] A model is considered a treatment evaluated by its results for the different data sets. Each model is then ranked separately for each data set according to its performance, where the best models have a lower rank and vice versa. If the result is statistically significant, p-value < 0.05, post hoc tests may be performed to explore which specific groups or treatments differ from each other.

Models were ranked before implementing the Friedman test according to the performance, specifically, the MCC values. This way, for every target, models were ranked from 1 to 4, where 1 corresponds to the highest MCC value, highest performance, and 4 corresponds to the smallest MCC value, lowest performance. Two tables were created: one for the retrospective approach and another for the prospective approach. Each table contains 4 columns representing the possible rank values for the 4 models tested and 14 rows corresponding to the number of targets. These tables were given as input for the Friedman test function chosen from the R's agricolae package.

Chapter 5

Data

To conduct the study, it was necessary to choose a data source from the various options of online chemical databases. Due to many bioactive compounds and biological targets, the ChEMBL [73] database was an appropriate choice for gathering diverse pharmacological data. Also, this database is manually curated and is composed of molecules with drug-like properties, which is important to guarantee quality data for this study. The chosen targets corresponding to human proteins were selected because they had a large number of recorded ligands, which reflects the importance of that target as an object of study. After selecting the targets, data can be collected inside the Target Report Card, containing data sets for different biological activities, and the one with more records should be chosen.

Fourteen data sets were gathered, and different bioactivity measures were contemplated (Ki, IC50 or EC50). The goal is to measure the interaction between a ligand and its target, whether it involves activating or inhibiting the target. Therefore, any recorded activity value is relevant to the hypothesis. The chosen data sets can be consulted in Table 5.1, with the original data set lengths corresponding to the N-retrieved parcel.

Table 5.1: Description of selected targets

Gene	Target Protein Name	Associated Bio-activity (Y)	Total number of records (N-retrieved)	Total number of records (N-processed)	Number of Negatives	Number of Positives
EGFR	Epidermal growth factor receptor erbB1	IC50	16715	9505	6007	3498
DRD2	Dopamine D2 receptor	Ki	12011	8215	5381	2834
ERBB2	Receptor protein-tyrosine kinase erbB-2	IC50	4072	2703	1660	1043
DAT	Dopamine transporter	Ki	3106	1820	1200	620
HRH1	Histamine H1 receptor	Ki	2764	1613	1024	589
IGF1R	Insulin-like growth factor I receptor	IC50	4424	2965	1597	1368
IKBKB	Inhibitor of nuclear factor kappa B kinase beta subunit	IC50	2175	1726	1296	430
TNF	TNF-alpha	IC50	1139	950	863	87
CDK4	Cyclin-dependent kinase 4	IC50	1178	1089	418	671
HTR1B	Serotonin 1b (5-HT1b) receptor	Ki	1269	1053	572	481
MAOA	Monoamine oxidase A	Ki	1526	587	489	98
ADRB1	Beta-1 adrenergic receptor	IC50	1540	661	566	95
PPARG	Peroxisome proliferator-activated receptor gamma	EC50	5127	3855	2959	896
ABL1	Tyrosine-protein kinase ABL	IC50	3119	1966	1109	857

Each data set provides information regarding the molecules' measurements for each target. In

this context, the most relevant features are the ChEMBL ID, Smiles string, information about the molecule's activity and the year of its documentation. After gathering the desired targets, data processing can occur, ensuring that missing data are removed from the training and testing sets and that for duplicates, only the most recent record is kept. The number of records is now reduced after the pre-processing, corresponding to the N-processed column in Table 5.1.

In general, every molecule has a standard type, relation, value and unit. If the type of relation was '>' the value of activity would be set to 0 and in the case of being '<' the molecule would be discarded, as it was risky to consider the latter without having precise information about how smaller the presented value is compared to the actual one.

To have a better perception of the frequency of the possible values for the dependent variable, *y*, normalised activity values, a histogram is presented in Figure 5.1, showing the distribution of activity values ranging from 0 to 1, divided into five equal intervals. The bars represent the absolute frequency of occurrences within each interval, illustrating how often activity values fall into each specific range. It is possible to observe that the interval between 0 and 0.2 has a more significant concentration of values for almost every data set. In contrast, the interval of values between 0.8 and 1 generally appears to have more minor frequencies. This histogram also allows the identification of the data sets with more records, quickly, which are D(2) dopamine receptor (*DRD2*), Epidermal growth factor receptor (*EGFR*) and Peroxisome proliferator-activated gamma receptor (*PPARG*).

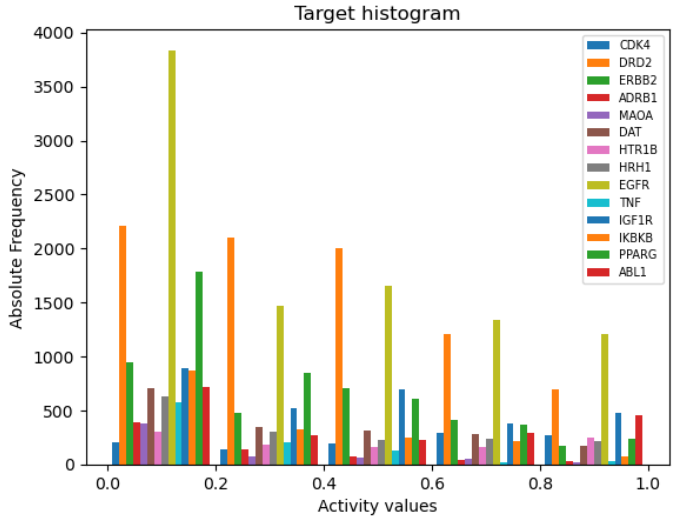


Figure 5.1: Frequency of Activity values of molecules for every data set

The ultimate goal is to perform the classification of active and non-active molecules with QSAR models based on similarities across sub-parts of molecules and their activity values. This way, these activity values correspond to the dependent variable for each target, which is what I want to predict. However, in this case, the data collected has continuous values instead of categorical values, so it is required to transform the regression data into classification data. To do this, there is the parameter regression threshold, *rthres*, that defines the threshold delimiting which

molecules should be considered active or inactive, receiving a value of 1 or 0, respectively. The threshold is defined according to the fact that in a human body, there are more inactive than active molecules. The threshold can be redefined, but the baseline ensures that having a *rthes* equal to 0.5 corresponds to considering molecules active with activity values inferior to 100 nm. This way reduces the asymmetry between the two classes.

Chapter 6

Retrospective vs Prospective validation results

6.1 Implementation

In terms of implementation, Python was the chosen programming language, as it is commonly used for data science and tailor-made for carrying out repetitive tasks and data manipulation. Also, using Python, it is possible to access packages such as Scikit-learn, a machine learning library for Python, which offers access to many models with functions ready to use and Pandas is a flexible and easy-to-use data analysis and manipulation tool built on top of Python, both implemented in the work. Secondly, as this research study involves working with molecules, the RDkit [44] environment, an open-source cheminformatics software available for Python, was used to perform tasks like fingerprint generation and the construction of molecules using SMILES.

This study uses a regression threshold of 100 nm in activity value to determine whether a molecule is classified as active or inactive, corresponding to a *rthres* value of 0.5. For the structural representation of molecules, circular morgan fingerprints were computed using the RDKit [44] toolkit. The radius was set to 3 and the fingerprint length to 2048. For each target, fingerprints are calculated for each documented molecule, which are the independent variables for training the models.

Regarding the machine learning models used, SVM was the only model to go through a cross-validated grid-search, aiming to find the best hyperparameters in general. From this, it was concluded that the predominant hyperparameters are C = 10 and gamma = 0,01, with a radial basis function kernel, so these were the parameters implemented for the training and testing of the SVM model. The RF and SVM classifiers were implemented using the Scikit-learn library. The RF's parameters are the default ones defined in Scikit-learn. The XGBoost model has its library, and the parameters were not modified. Regarding the NN, PyTorch modules were used to create a feed-forward fully connected neural network from root, being composed of two hidden layers, the first one with 512 neurons and the second one with 64 neurons, receiving an input with the size of the fingerprint chosen and with two as the size of the output, corresponding to the possibility of a molecule being either active or non-active. For the training process of the NN, a cross-entropy

loss criterion and an Adam optimizer with a learning rate of 0.001 were used. This configuration was chosen among other combinations tested on the NN.

All targets 5.1 were used to train every model described in 2.4.

Regarding the Friedman test, groups were created after the results were generated. There were 14 groups, each with a list of MCC values for the various classification models. The Friedman Test ranks the algorithms for each data set separately. If the result is not statistically significant, equivalent to having a significance level inferior to 0.05, a post hoc analysis should be done. The chosen implementation for treatment comparison with the Friedman test and the post hoc analysis was R's agricolae package. [53]

6.2 Comparison Retrospective and Prospective validation

6.2.1 Retrospective validation results

The results with the retrospective validation summarised for the best models in each target in Table 6.1, were satisfactory, with an MCC value minimum of 0.39 and a maximum of 0.77, in which MCC results were on average 0.64. Targets like TNF, HTR1B, HRH1, MAOA and DAT exhibit high MCC values in retrospective validation (above 0.7). The other targets present moderate MCC values, and the lowest MCC value is 0.5872 for the IKBKB target.

Table 6.1: Best models for each target with retrospective validation

Data sets	Model	Total number of observations (N-processed)		Metrics				
		Training set	Test set	Accuracy	Precision	Recall	F1_score	MCC
TNF	Support Vector Machine	760	190	0.9579	0.7647	0.7647	0.7647	0.7416
DRD2	XGBoost	6572	1643	0.8430	0.8125	0.7296	0.7688	0.6526
ABL1	Random Forest	1572	394	0.8299	0.7961	0.7707	0.7832	0.6436
HTR1B	Support Vector Machine	842	211	0.8768	0.9341	0.8095	0.8673	0.7601
PPARG	Support Vector Machine	3084	771	0.8729	0.7640	0.6721	0.7151	0.6359
IKBKB	Random Forest	1380	346	0.8526	0.7761	0.5909	0.6710	0.5872
HRH1	XGBoost	1290	323	0.8947	0.8917	0.8359	0.8629	0.7787
CDK4	Random Forest	871	218	0.8440	0.8561	0.8947	0.8750	0.6691
ABRB1	Support Vector Machine	528	133	0.9098	0.6957	0.7619	0.7273	0.6744
EGFR	Random Forest	7604	1901	0.8264	0.7659	0.7340	0.7496	0.6172
IGF1R	XGBoost	2372	593	0.8263	0.8246	0.8160	0.8202	0.6523
MAOA	Random Forest	469	118	0.9322	0.8824	0.7143	0.7895	0.7555
DAT	Random Forest	1456	364	0.9011	0.8512	0.8512	0.8512	0.7772
ERBB2	Random Forest	2162	541	0.8152	0.7773	0.7703	0.7738	0.6175

6.2.2 Prospective validation results

The results from the prospective validation summarised for the best models in each target in Table 6.2, were significantly reduced compared to the ones with the random data split. The average MCC value was 0.22, ranging from -0.08 and 0.72. This shows a clear drop in the performance of the models by having selected the 20% most recent records for the test set. None of the targets achieve high MCC values in prospective validation. The highest values are seen for DRD2, HTR1B,

IKBKB, and ERBB2, but these values are still noticeably lower than those of their retrospective counterparts. On the contrary, IGF1R presents the lowest MCC value, with a Recall close to 0. Generally, Precision is not balanced with Recall, which could mean the models cannot identify true positives without many false positives.

Table 6.2: Best models for each target with prospective validation

Data sets	**Model**	**Year for 20% data partition**	**Total number of observations (N-processed)**		**Metrics**				
			Training set	**Test set**	**Accuracy**	**Precision**	**Recall**	**F1_score**	**MCC**
TNF	Random Forest	2016	759	191	0.7435	0.000	0.000	0.000	0.000
DRD2	Random Forest	2016	6085	2130	0.8117	0.7632	0.5821	0.6605	0.5428
ABL1	Random Forest	2016	1007	959	0.4599	0.8421	0.1123	0.1981	0.1480
HTR1B	Random Forest	2010	775	278	0.8777	0.8043	0.8222	0.8132	0.7224
PPARG	Random Forest	2017	3048	807	0.8451	0.8571	0.0462	0.0876	0.1771
IKBKB	Support Vector Machine	2013	1281	445	0.8247	0.9672	0.4370	0.6020	0.5755
HRH1	Neural Network	2016	1221	392	0.7449	0.4262	0.2857	0.3421	0.1973
CDK4	Random Forest	2015	817	272	0.4853	0.9079	0.3416	0.4964	0.2354
ABRB1	XGBoost	2011	478	183	0.6667	0.5660	0.4412	0.4959	0.2569
EGFR	XGBoost	2017	7261	2244	0.6225	0.6846	0.3220	0.4380	0.2393
IGF1R	Random Forest	2014	2030	935	0.3326	0.6901	0.0753	0.1357	-0.0038
MAOA	Random Forest	2016	459	128	0.6484	1.000	0.0816	0.1509	0.2281
DAT	XGBoost	2012	1355	465	0.7032	0.7101	0.5000	0.5868	0.3797
ERBB2	XGBoost	2014	2147	556	0.7554	0.6731	0.6731	0.6731	0.4777

6.2.3 Assessment of differences

The plot (6.1) indicates that while the models perform well in retrospective validation, their performance drops considerably in prospective validation. Retrospective validation (blue bars) yields higher MCC values compared to prospective validation (orange bars) for all targets and the average MCC for retrospective validation is significantly higher than for prospective validation. However, DRD2, HTR1B and IKBKB present a similar performance for both validation approaches, in which the three targets have a size of at least 1000 entries. The difference between retrospective and prospective MCC values is quite pronounced for most targets, suggesting that models may be overfitting to the training data or significant differences in the data distributions or challenges in prospective scenarios. The target TNF has an MCC value of zero, and IGF1R has a negative MCC value near zero for prospective validation, indicating almost no predictive power. The biggest difference between the two approaches is 0.7416 for the TNF target and the smallest is 0.0117 for IKBKB. The last two columns correspond to the weighted average of every target for both approaches, reflecting the discrepancy between the retrospective and prospective types of validation, with a value of 0.6550 and 0.3095, respectively. This suggests a potential issue with the generalisability of the models to new, unseen data. Improving prospective validation performance would be crucial for ensuring that the models are robust and reliable in real-world applications.

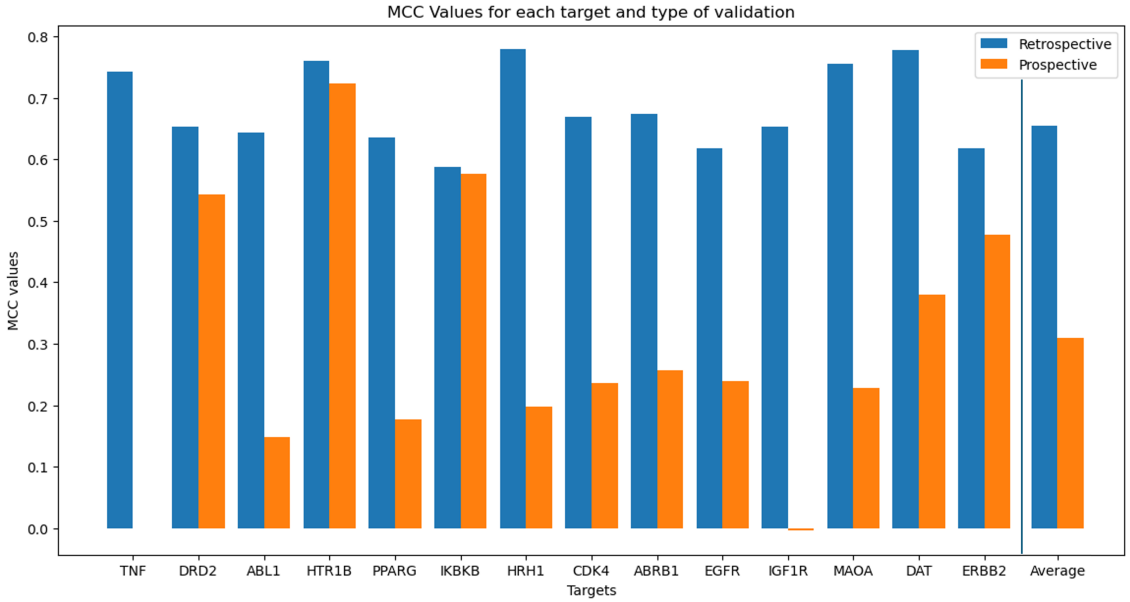


Figure 6.1: Bar plot comparing the retrospective and prospective approaches, considering the best model for each target.

6.2.4 Assessment of modelling approaches

It is possible to observe in Table 6.1 and Table 6.2 that there is no model that outperformed globally in both types of validation. For further model comparison and performance analysis, a Friedman test was conducted for both validation approaches.

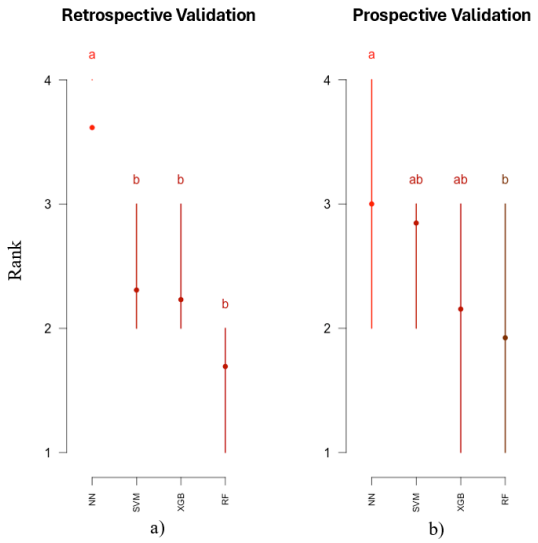


Figure 6.2: Friedman test results and interquartile ranges of the tested models.

Regarding the Friedman test for the retrospective approach, results showed significant differences between treatments, with a p-value from the Chi-square test of 1.79×10^{-3} with 13 degrees of freedom, which strongly suggests that there are statistically significant differences between the

different modelling approaches. Figure 6.2a) suggests that the neural network, compared to the rest, is significantly different, presenting a lower performance. Also, it is possible to observe how the ranks varied for each model, where even though the RF is in the same group as SVM and XGBoost, it shows higher rank values (better performance).

In the case of the prospective validation, the p-value of the Chi-squared test was 9.19×10^{-2} with 13 degrees of freedom, suggesting no statistically significant differences between the different models. By looking at the results of the Friedman test in image 6.2b), the NN is the model with a significant difference compared to the RF and has the worst ranking. Next, the XGBoost and SVM models belong to both groups, not being significantly different from either RF or NN, presenting a similar performance. It is also possible to see a higher variance of ranks for the NN, RF and XGBoost models with only the SVM being more consistent between ranks 2 and 3.

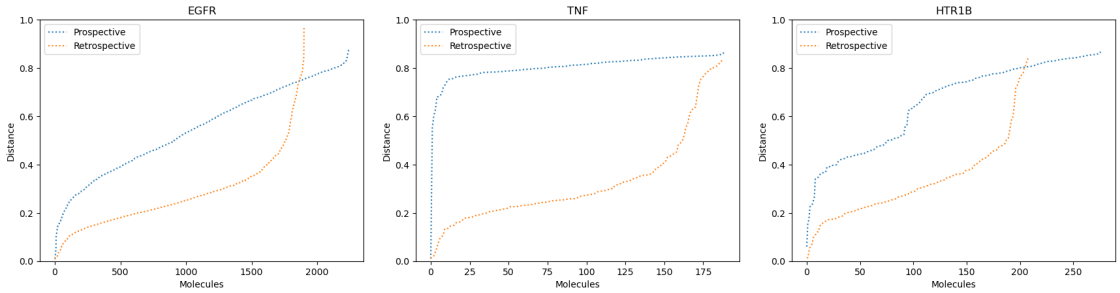
6.3 Discussion

It is important to acknowledge that these two problems are independent as the different conditions and data partitions generate different data sets between them. Overall, the prospective approach appears to be a more difficult problem than the retrospective approach. However, machine learning models must generalise well to unseen data, and these results indicate a lack of predictive power when performing temporal validation. Furthermore, no single model consistently outperforms the others overall, nor is there a model that excels under specific conditions. There appears to be a slight trend where XGBoost is preferred for larger data sets, while RF and SVM are more commonly used for smaller data sets.

Initially, the impact of the different types of validation was questioned, and what would happen when performing a prospective validation approach was discussed. The results presented in Tables 6.1 and 6.2 suggest that considering the year of documentation of the molecules has a significant impact on the models' performance. A possible reason to explain this event is that the data registered in the ChEMBL online database might have been retrieved differently over the years. This would lead to molecule records in the training set being very distinct from the testing set and, consequently, the models having less predictive power. Another aspect that can affect the models' performance is related to the level of structure similarity among the molecules in each target, which was presented in 6.2.3.

Results suggest that considering the year of documentation of the molecules significantly impacts the models' performance. A possible reason the models' performance might be affected is related to the composition and diversity of the target data sets, precisely, the level of structure similarity among the molecules in each target. The performance of the QSAR models may be impacted by the redundancy caused by having a large number of similar molecules in each target. This event was further studied 6.3, and three targets were chosen for structural comparison. EGFR, as it presents a high value for the retrospective approach and a value lower than the average for the prospective approach, HTR1B for having very high performance for both validation types and TNF did not show any result for the prospective approach based on 6.1. In 6.3, the closest

distance of each molecule in the test set to each molecule in the training set was kept and ordered by proximity. A very open curve suggests that many molecules in the training set are very distant from the test set. These plots reflect the pattern identified initially with the performance analysis in 6.1 with 6.3a and 6.3c having a distinct configuration from 6.3b, in which the first two show less open curves between the two approaches. Generally, the distance between the molecules from the train and test sets for the prospective approach is bigger than the retrospective approach. Measuring the distance between molecules can indicate the level of their structural similarity, where the bigger the distance between molecules, the less similar they are. It is possible to observe 6.3b a very distinct behaviour between the two types of validation, which is clearly reflected in the performance of this target.



(a) (b) (c)

Figure 6.3: Distance between the molecules in the train set and test set, for both validation approaches

These patterns suggest that QSAR models are highly susceptible to molecular structural patterns and have difficulty providing robust predictions for molecules outside the known domains. The optimistic curves 6.3 for retrospective validation show that most molecules in the test set are, in general, quite similar to at least another one in the train set. In contrast, this is not so much the case with prospective validation data sets, and the models struggle to make reliable predictions with more diverse or novel molecules.

Chapter 7

Impact of structural similarity on QSAR modelling

7.1 Non-redundant data sets

Following the findings in 6.3, a retrospective approach with structural separation was implemented to investigate the impact of molecular structure similarity further. This approach involved removing redundant data from the initial data sets (see Section 4.3) to isolate the effect of structural similarity on model performance and help address the question: How much can model predictions degrade when intermolecular structural redundancies are removed from the data sets?

7.1.1 Retrospective validation with Structural separation results

The results of the retrospective validation with the structural separation (Table 7.1) are comparable to the ones presented in the retrospective validation (6.2.1). However, the average MCC values were reduced by one-tenth compared to the retrospective approach, ranging from 0.20 to 0.78.

Table 7.1: Best models for each dataset with structural separation

Data sets	Model	Total number of observations (N-processed)		Metrics				
		Training set	Test set	Accuracy	Precision	Recall	F1_score	MCC
TNF	Random Forest	300	75	0.9733	1.0000	0.3333	0.5000	0.5695
DRD2	Support Vector Machine	2477	620	0.8500	0.7736	0.6833	0.7257	0.6253
ABL1	Support Vector Machine	628	158	0.8797	0.8788	0.6591	0.7532	0.6881
HTR1B	Random Forest	340	86	0.8256	0.7143	0.4762	0.5714	0.4825
PPARG	Support Vector Machine	1128	283	0.8834	0.7188	0.4894	0.5823	0.5303
IKBKB	Support Vector Machine	424	106	0.9434	0.8000	0.6667	0.7273	0.6996
HRH1	XGBoost	596	149	0.8523	0.7609	0.7609	0.7609	0.6541
CDK4	Support Vector Machine	296	74	0.7973	0.7273	0.6400	0.6809	0.5356
ABRB1	Random Forest	136	34	0.9412	1.0000	0.6667	0.8000	0.7888
EGFR	XGBoost	2450	613	0.8891	0.7317	0.7200	0.7258	0.6563
IGF1R	Support Vector Machine	699	175	0.8857	0.8611	0.6739	0.7561	0.6917
MAOA	Support Vector Machine	234	59	0.9492	1.0000	0.4000	0.5714	0.6156
DAT	XGBoost	689	173	0.8671	0.8182	0.6136	0.7013	0.6286
ERBB2	Support Vector Machine	737	185	0.8432	0.6471	0.5641	0.6027	0.5075

The targets ABRB1, IKBKB and IGF1R have the best MCC values, whereas HTR1B, ERBB2 and PPARG presented the lowest MCC values. It is possible to verify that models that performed

better in the previous experiment with the retrospective approach presented average results for the non-redundant data sets, with the exception of IKBKB, which had a better performance.

Regarding the Friedman test, results presented a Chi-squared p-value of 2.75×10^{-3} with 13 degrees of freedom, which strongly suggests that there are statistically significant differences between the different models. Three distinct groups are identified: the neural network with the highest median ranking (worst model), the XGBoost with a performance between ranks 2 and 3, and the SVM with the best performance. In this case, the RF is only significantly different from the SVM.

7.1.2 Discussion

Considering the retrospective approach using the non-redundant data sets 7.1.1, it is possible to observe a considerable reduction of the training and test sets compared to the sets without removing similar molecules. On average, the data sets were reduced by 38%, possibly making it more difficult to train and test the models for smaller targets. However, the decrease in terms of performance when using the non-redundant compared to the retrospective approach when using the redundant data sets was not significant.

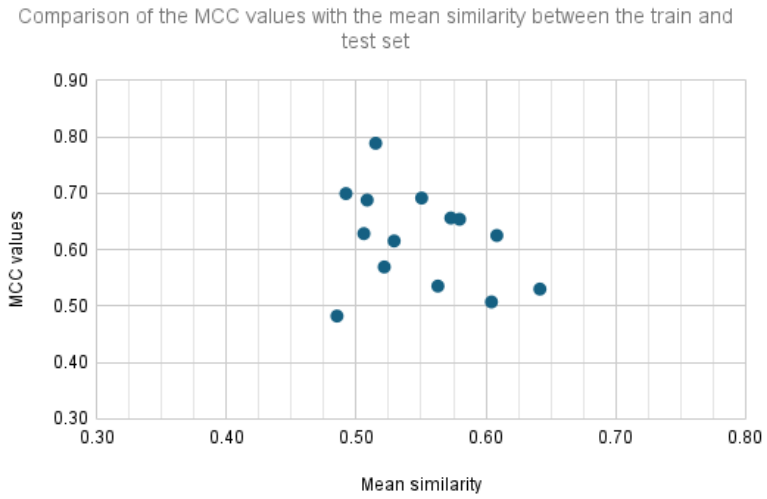


Figure 7.1: Scatter plot comparing MCC values with mean similarity.

The scatter plot presented in Figure 7.1 compares the best MCC values for each target with the mean similarity measured between the training and test set for each model and target. MCC values for the best models in each target range from 0.48 and 0.78. The mean similarity values range from 0.48 and 0.64. There is no strong linear correlation between mean similarity and the models' performance, as indicated by the scattered points. However, higher MCC values (>0.55) are mostly found within mean similarity values of 0.5 to 0.6, suggesting a potential optimal range for better performance. This graph could guide how to split the data into training and test sets, in which, if a specific range of similarity values consistently leads to better performance, it could inform the data preparation process. On the other hand, it is unclear from this plot that the higher

the mean similarity, the higher the MCC value. Whereas in the past experiment for the retrospective approach 6.3 the bigger the similarity the better the performance, here the results do not show the same dependency for the similarity, presenting similar MCC values for higher similarities.

Overall, using non-redundant data sets did not affect the ability of the models to generalize and predict molecular bioactivity.

7.2 Hyperdimensional metric space models

In parallel, a tool based on an N-dimensional metric space was used to provide deeper insights into the importance and impact of molecular structure on the quality of the models. This approach also aimed to verify whether performance would degrade when using structure-reliant models compared to common machine learning models.

7.2.1 Implementation

A metric space-based QSAR tool (see Section 2.6) was implemented to assess the impact of structural similarity on the quality of models.

Similarly to the previous experiments, while testing machine learning models, the implementation of the metric space-based models follows a consistent approach in terms of data handling, model evaluation, and model comparison. I developed several configurations by combining different fingerprint filtering values and dimensions, allowing for dynamic and controlled testing. The fingerprint filtering step minimises noise by focusing on critical structural features, which provides more targeted insights from the remaining parameters, simplifying the original data sets. Moreover, in a different approach, various reduced dimensions were tested to explore how varying the number of dimensions from a high-dimensional space of 2048 binary variables affects the model's ability to capture essential information while improving performance and manageability. Initially, I tested dimensions ranging from 5 to 45; however, evidence of improvement in model performance was only observed for 5 and 10 dimensions. By reducing the problem to 5 and 10 dimensions and employing fingerprint filtering values of 0.05, 0.1, and 1, I aimed to assess the effects of dimensional reduction on model efficiency and accuracy. Testing these two variables was important because dimensionality reduction helps simplify the data while preserving essential information, ultimately improving model performance by reducing noise and computational demands. In this context, d represents the number of dimensions, while ff denotes the fingerprint filtering feature.

7.2.2 Metric-space based models results

The results for assessing the importance of strict structure similarity for inferring binding activity were generally high, suggesting it is a significant component of QSAR models. Targets such as HRH1, ABL1, and HTR1B exhibited strong correlations with structural similarity, achieving

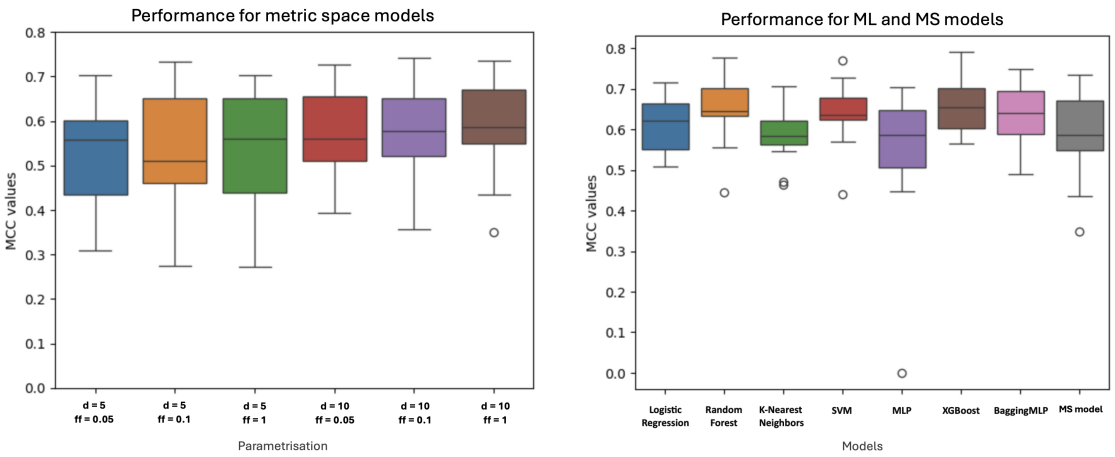
high MCC values above 0.7262. In contrast, targets like MAOA, TNF, and DRD2 showed lower correlations, with overall average MCC values across all targets of 0.6122.

Table 7.2: Best models for each target using hyperdimensional metric space models

Data sets	**Total number of observations (N-processed)**		***d***	***ff***	**Metrics**			
	Training set	**Validation set**			**Precision**	**Recall**	**F1_score**	**MCC**
HTR1B	723	241	10	0.05	0.8376	0.8750	0.8559	0.7262
ABL1	1177	393	10	1.0	0.8412	0.8563	0.8487	0.7352
ADRB1	474	159	5	0.1	0.6250	0.8696	0.7273	0.6855
CDK4	741	247	5	1.0	0.8768	0.8403	0.8582	0.6705
DAT	1290	430	10	0.05	0.6561	0.9118	0.7631	0.6471
DRD2	5082	1694	10	0.05	0.6196	0.8042	0.6999	0.5291
EGFR	5378	1793	10	0.1	0.6663	0.8529	0.7481	0.5926
ERBB2	1633	545	10	1.0	0.6449	0.8404	0.7298	0.5702
HRH1	955	319	10	0.1	0.7800	0.9286	0.8478	0.7420
IGF1R	2030	677	10	1.0	0.8684	0.8000	0.8328	0.6881
IKBKB	1173	392	10	1.0	0.6000	0.7582	0.6699	0.5614
MAOA	371	124	10	1.0	0.2813	0.9000	0.4286	0.4345
PPARG	2335	779	10	1.0	0.5395	0.8497	0.6600	0.5406
TNF	617	206	5	0.1	0.3333	0.8750	0.4828	0.4834

The comparison between the metric space-based models' results and the machine learning models reveals some key performance differences across classification metrics. Generally, machine learning models demonstrate higher precision, with values ranging from 0.6957 to 0.9341, indicating they are more effective at identifying true positives without as many false positives compared to the ones obtained with the metric space-based models, whose precision ranges from 0.2813 to 0.8768. Regarding recall, these models often achieved higher values, ranging from 0.7582 to 0.9286, suggesting they may be better at capturing all true positives, even at lower precision. Machine learning models show more balanced F1-scores, from 0.6710 to 0.8750, highlighting their ability to maintain a good trade-off between precision and recall; in contrast, the metric space models' F1-scores range from 0.4286 to 0.8582. When examining the Matthews Correlation Coefficient (MCC), machine learning models again tend to outperform, suggesting a stronger overall correlation between predicted and actual classifications. In summary, while the metric space-based models perform well in recall, capturing more active compounds, machine learning models generally offer better precision, F1-score, and MCC, making them more balanced and reliable for accurate classifications across these datasets.

Figure 7.2 presents a visual comparison of the performance of the various approaches based on the MCC scores. Firstly, Figure 7.2a allows the analysis of the parameterisation of the metric space-based algorithm, in which the features associated with an overall better performance of the model with 10 as the number of dimensions and the fingerprint filtering parameter set as 1. In this case, having the latter with a value of 1 associated with the model with the best performance indicates the filtering feature did not impact the model performance as expected. Secondly, it is possible to compare it with the other implemented models, represented in 7.2b. As it is possible to



(a) Parametrisation of the metric space-based tool (b) Performance of ML models and MS model

Figure 7.2: Comparison of models in terms of performance based on the MCC scores

observe, the models show similar ranges, and the metric space model seems to perform similarly to the other models. The approaches appear to be very similar, although the results of the Friedman test indicate there could be differences: statistic = 53.0125, p-value = 1.8931×10^{-6}. The p-value is smaller than the 0.05 established, which could indicate the algorithms are significantly different.

7.2.3 Discussion

The impact of the structure similarity of the molecules on the models can vary depending on several factors, like the number of molecules recorded for a target, experimental errors, or the ability of the chosen models to make predictions for every target. In this case, the metric space-based workflow involves building models by processing molecules projected into a multi-dimensional metric space and then transforming them into a lower-dimensional space that better preserves the relative distances between the molecules. By leveraging metrics that quantify the similarities and differences among molecular structures, this approach can provide insights into how new compounds are likely to behave based on known data, highlighting the impact of structural similarity on the performance of QSAR models.

As it is possible to observe in 7.2b, the metric space models appear to behave similarly to the machine learning models, although the latter presents a lower variability of the MCC values. On the one hand, for the targets selected, the MS models present good overall results, suggesting hyperdimensional metric space models could be possible candidates for QSAR modelling. On the other hand, the machine learning models generally offered superior classification metrics, making them more balanced and reliable for accurate predictions across the tested targets, suggesting that structure-reliant models do not enhance performance in predicting compound activity.

Overall, these results are aligned with the ones obtained using non-redundant data sets. The results of these experiments (7.1 and 7.2) suggest the machine learning models outperform the metric space models and, consequently, that there is no strong evidence that structural similarity

is the dominant driver of bioactivity. This implies that other molecular characteristics, such as stereochemistry, specific functional groups, or even complex quantum mechanical properties, are necessary to fully capture the nuances of molecular binding. Thus, while the similarity principle is important, it is insufficient to explain bioactivity, necessitating a more nuanced approach to QSAR modelling. Moreover, the success of the state-of-the-art models over metric space models could indicate that complex, non-linear relationships between molecular features and activity cannot be fully captured by relying on structural similarity alone. This would emphasize the need for more sophisticated feature engineering and model architectures in bioactivity prediction, particularly when dealing with highly diverse chemical libraries.

Chapter 8

Conclusion

In conclusion, the drug discovery field has been majorly impacted by the growing use of computational resources that allow automating processes and taking advantage of the large amounts of data made available over the years. QSAR modelling is one of these advances, allowing the prediction of new drug candidates' activity values based on molecular descriptors. However, several underlying issues are affecting machine learning models' predictive power and preventing them from achieving their full potential.

In this project, a comparison in terms of performance between validating QSAR models with a retrospective and a prospective approach was conducted using machine learning models and various targets contemplating diverse pharmacological data. This objective involved collecting data, pre-processing data, choosing machine learning models adequate for the context, and model validation. This part of the study provided insights into the big discrepancy between testing models with and without considering the year when the molecules were recorded, prospective and retrospective approaches. Following these tasks, results were analysed, and a study of the possible impact of the structure similarity of molecules was conducted, which involved two parts. Firstly, non-redundant data sets obtained from the original ones were used for testing the common machine learning models. Secondly, hyperdimensional metric space models were tested and compared with previous results obtained with the machine learning models.

The findings, when exploring the impact of structural similarity with the metric space-based models for multiple dimensions for retrospective and prospective approaches, led to a turning point in my work, where studying the impact of the type of validation used was considered to be a priority as it is a more general issue that could have a bigger impact on the QSAR modelling field, described from chapter 4 to chapter 6.3.

In terms of future work, this study already provided solid evidence on a relevant issue that is very likely affecting the predictive power of machine learning models in QSAR. However, conducting a more detailed assessment of the two approaches would be convenient for analysing the impact of every variable and conducting more tests to make precise comparisons. Regarding testing of the automated framework for QSAR modelling, it would be relevant to continue the testing by considering more parameters, like testing the regression threshold parameter, scalability, and computational complexity, and considering the findings on the impact of the retrospective and

prospective validation approaches.

Overall, I believe this study was essential to gaining knowledge on one of the most crucial issues in QSAR modelling and machine learning in general.

Bibliography

[1] Subhash Ajmani, Kamalakar Jadhav, and Sudhir A. Kulkarni. Three-dimensional qsar using the k-nearest neighbor method and its interpretation. *Journal of Chemical Information and Modeling*, 46(1):24–31, 2006. PMID: 16426036.

[2] Zakariya Yahya Algamal, Muhammad Hisyam Lee, Abdo M. Al-Fakih, and Madzlan Aziz. High-dimensional qsar classification model for anti-hepatitis c virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty. *Journal of Chemometrics*, 31(6):e2889, 2017. e2889 CEM-16-0087.R2.

[3] Amer Alnuaimi and Tasnim Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97:00133, 04 2024.

[4] Ismail Babajide Mustapha and Faisal Saeed. Bioactive molecule prediction using extreme gradient boosting. *Molecules*, 21(8):983, July 2016.

[5] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.*, 7(1):20, May 2015.

[6] Andreas Bender and Robert C. Glen. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, 2:3204–3218, 2004.

[7] Andreas Bender, Jeremy L. Jenkins, Qingliang Li, Sam E. Adams, Edward O. Cannon, and Robert C. Glen. Chapter 9 molecular similarity: Advances in methods, applications and validations in virtual screening and qsar. In David C. Spellmeyer, editor, *Annual Reports in Computational Chemistry*, volume 2 of *Annual Reports in Computational Chemistry*, pages 141–168. Elsevier, Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, United Kingdom CB2 1EW, 2006.

[8] Grant Beyleveld, Kris M White, Juan Ayllon, and Megan L Shaw. New-generation screening assays for the detection of anti-influenza compounds targeting viral and host functions. *Antiviral Res.*, 100(1):120–132, October 2013.

[9] Alexandre Blanco-González, Alfonso Cabezón, Alejandro Seco-González, Daniel Conde-Torres, Paula Antelo-Riveiro, Ángel Piñeiro, and Rebeca Garcia-Fandino. The role of AI in

drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals (Basel)*, 16(6), June 2023.

[10] L.J. Cao, K.S. Chua, W.K. Chong, H.P. Lee, and Q.M. Gu. A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1):321–336, 2003. Support Vector Machines.

[11] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, Viviana Consonni, Victor E Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: where have you been? where are you going to? *J. Med. Chem.*, 57(12):4977–5010, June 2014.

[12] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. Qsar modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014. PMID: 24351051.

[13] Isidro Cortés-Ciriano, Ctibor Škuta, Andreas Bender, and Daniel Svozil. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *J. Cheminform.*, 12(1):41, June 2020.

[14] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, Ch Madhu Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: A review. *Artif. Intell. Rev.*, 55(3):1947–1999, 2022.

[15] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, 43(W1):W612–W620, July 2015.

[16] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[17] Yi Ding, Minchun Chen, Chao Guo, Peng Zhang, and Jingwen Wang. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. *J. Mol. Liq.*, 326(115212):115212, March 2021.

[18] Dimitar Dobchev, Girinath Pillai, and Mati Karelson. In silico machine learning methods in drug development. *Curr. Top. Med. Chem.*, 14(16):1913–1922, October 2014.

[19] Arthur M Doweyko. QSAR: dead or alive? *J. Comput. Aided Mol. Des.*, 22(2):81–89, February 2008.

[20] Arkadiusz Dudek, Tomasz Arodz, and Jorge Galvez. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb. Chem. High Throughput Screen.*, 9(3):213–228, March 2006.

[21] Hanna Eckert and Jürgen Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, 12(5):225–233, 2007.

[22] Lennart Eriksson, Joanna Jaworska, Andrew P Worth, Mark T D Cronin, Robert M McDowell, and Paola Gramatica. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based qsars. *Environmental Health Perspectives*, 111(10):1361–1375, 2003.

[23] Steffen W. Ernst, Richard Knight, Jenny Royle, and Laura Stephenson. Chapter 22 - pharmaceutical toxicology. In Martin Wehling, editor, *Principles of Translational Science in Medicine (Third Edition)*, pages 265–279. Academic Press, Boston, third edition edition, 2021.

[24] Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023. PMID: 37216672.

[25] Johann Gasteiger. Chemoinformatics: Achievements and challenges, a personal view. *Molecules*, 21(2):151, 2016.

[26] Fahimeh Ghasemi, Alireza Mehridehnavi, Alfonso Pérez-Garrido, and Horacio Pérez-Sánchez. Neural network and deep-learning algorithms used in qsar studies: merits and drawbacks. *Drug Discovery Today*, 23(10):1784–1790, 2018.

[27] G. Gini, F. Zanoli, A. Gamba, G. Raitano, and E. Benfenati. Could deep learning in neural networks improve the qsar models? *SAR and QSAR in Environmental Research*, 30(9):617–642, 2019. PMID: 31460798.

[28] Alexander Golbraikh, Denis Fourches, Alexander Sedykh, Eugene Muratov, Inta Liepina, and Alexander Tropsha. *Modelability Criteria: Statistical Characteristics Estimating Feasibility to Build Predictive QSAR Models for a Dataset*, pages 187–230. Springer US, Boston, MA, 2014.

[29] Alexander Golbraikh and Alexander Tropsha. Predictive qsar modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des.*, 16(5/6):357–369, 2002.

[30] Paola Gramatica. Principles of qsar models validation: internal and external. *QSAR and Combinatorial Science*, 26(5):694–701, 2007.

[31] Kathrin Heikamp and Jürgen Bajorath. Support vector machines for drug discovery. *Expert Opinion on Drug Discovery*, 9(1):93–104, 2014. PMID: 24304044.

[32] Jianping Huang and Xiaohui Fan. Why QSAR fails: an empirical evaluation using conventional computational approach. *Mol. Pharm.*, 8(2):600–608, April 2011.

[33] Gordana Ivosev, Lyle Burton, and Ron Bonner. Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry*, 80(13):4933–4944, 2008. PMID: 18537272.

[34] M.T.D. Cronin J.C. Dearden and K.L.E. Kaiser. How not to develop a quantitative structure–activity or structure–property relationship (qsar/qspr). *SAR and QSAR in Environmental Research*, 20(3-4):241–266, 2009. PMID: 19544191.

[35] William L. Jorgensen. Qsar/qspr and proprietary data. *Journal of Chemical Information and Modeling*, 46(3):937–937, 2006.

[36] Gregory W. Kauffman and Peter C. Jurs. Qsar and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *Journal of Chemical Information and Computer Sciences*, 41(6):1553–1560, 2001. PMID: 11749582.

[37] S. Kausar and A.O. Falcao. A visual approach for analysis and inference of molecular activity spaces. *Journal of Cheminformatics volume 11*, 2019.

[38] Samina Kausar and Andre O Falcao. An automated framework for QSAR model building. *J. Cheminform.*, 10(1), December 2018.

[39] Samina Kausar and Andre O. Falcao. Analysis and comparison of vector space and metric space representations in qsar modeling. *Molecules*, 24(9):1698, 2019.

[40] Terry P. Kenakin. Chapter 1 - pharmacology: The chemical control of physiology. In Terry P. Kenakin, editor, *Pharmacology in Drug Discovery*, pages 1–18. Academic Press, Boston, 2012.

[41] Gilles Klopmand. Concepts and applications of molecular similarity, by mark a. johnson and gerald m. maggiora, eds., john wiley & sons, new york, 1990, 393 pp. price: $65.00. *Journal of Computational Chemistry*, 13(4):539–540, 1992.

[42] Gilles Klopmand. Concepts and applications of molecular similarity, by mark a. johnson and gerald m. maggiora, eds., john wiley & sons, new york, 1990, 393 pp. price: $65.00. *Journal of Computational Chemistry*, 13(4):539–540, 1992.

[43] Evgueni Kolossov and R Stanforth. The quality of qsar models: Problems and solutions. *SAR and QSAR in environmental research*, 18:89–100, 01 2007.

[44] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[45] Yi Li, Dahua Pan, Jianzhong Liu, Petra S. Kern, G. Frank Gerberick, Anton J. Hopfinger, and Yufeng J. Tseng. Categorical QSAR Models for Skin Sensitization based upon Local Lymph Node Assay Classification Measures Part 2: 4D-Fingerprint Three-State and Two-2-State Logistic Regression Models. *Toxicological Sciences*, 99(2):532–544, 08 2007.

[46] Yi Li, Yufeng J. Tseng, Dahua Pan, Jianzhong Liu, Petra S. Kern, G. Frank Gerberick, and Anton J. Hopfinger. 4d-fingerprint categorical qsar models for skin sensitization based on the classification of local lymph node assay measures. *Chemical Research in Toxicology*, 20(1):114–128, 2007. PMID: 17226934.

[47] Man Luo, Xiang Simon Wang, Bryan L. Roth, Alexander Golbraikh, and Alexander Tropsha. Application of quantitative structure–activity relationship models of 5-ht1a receptor binding to virtual screening identifies novel and potent 5-ht1a ligands. *Journal of Chemical Information and Modeling*, 54(2):634–647, 2014. PMID: 24410373.

[48] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry. *J. Med. Chem.*, 57(8):3186–3204, April 2014.

[49] Vinicius Gonçalves Maltarollo, Thales Kronenberger, Gabriel Zarzana Espinoza, Patricia Rufino Oliveira, and Kathia Maria Honorio. Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov.*, 14(1):23–33, January 2019.

[50] Kamel Mansouri, José T Moreira-Filho, Charles N Lowe, Nathaniel Charest, Todd Martin, Valery Tkachenko, Richard Judson, Mike Conway, Nicole C Kleinstreuer, and Antony J Williams. Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. *J. Cheminform.*, 16(1):19, February 2024.

[51] Jiashun Mao, Javed Akhtar, Xiao Zhang, Liang Sun, Shenghui Guan, Xinyu Li, Guangming Chen, Jiaxin Liu, Hyeon-Nae Jeon, Min Sung Kim, Kyoung Tai No, and Guanyu Wang. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience*, 24(9):103052, 2021.

[52] Andrea Mauri, Viviana Consonni, and Roberto Todeschini. Molecular descriptors. In *Handbook of Computational Chemistry*, pages 2065–2093. Springer International Publishing, Cham, 2017.

[53] Felipe Mendiburu. Agricolae: Statistical procedures for agricultural research. *R package version*, 1:1–8, 01 2010.

[54] Lewis Mervin, Alexey Voronov, Mikhail Kabeshov, and Ola Engkvist. QSARtuna: An automated QSAR modeling platform for molecular property prediction in drug design. *J. Chem. Inf. Model.*, 64(14):5365–5374, July 2024.

[55] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2):137–148, 2016. PMID: 26558489.

[56] Nina Nikolova and Joanna Jaworska. Approaches to measure chemical similarity – a review. *QSAR & Combinatorial Science*, 22(9-10):1006–1026, 2003.

[57] Nina Nikolova and Joanna Jaworska. Approaches to measure chemical similarity – a review. *QSAR & Combinatorial Science*, 22(9-10):1006–1026, 2003.

[58] I. Ponzoni, V. Sebastián-Pérez, and C. Requena-Triguero. Hybridizing feature selection and feature learning approaches in qsar modeling for drug discovery. *Sci Rep 7, 2403*, 2017.

[59] Edward E. Rigdon. Using the friedman method of ranks for model comparison in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(3):219–232, 1999.

[60] Mícheál Searcóid. *Metric Spaces*. Springer London, 2006.

[61] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. PMID: 14632445.

[62] Jian-Ping Lin Tao Wang, Mian-Bin Wu and Li-Rong Yang. Quantitative structure–activity relationship: promising advances in drug discovery platforms. *Expert Opinion on Drug Discovery*, 10(12):1283–1300, 2015. PMID: 26358617.

[63] Ana L. Teixeira and Andre O. Falcao. Structural similarity based kriging for quantitative structure activity and property relationship modeling. *Journal of Chemical Information and Modeling*, 54(7):1833–1849, 2014. PMID: 24897621.

[64] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010.

[65] Alexander Tropsha, Paola Gramatica, and Vijay K Gombar. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.*, 22(1):69–77, April 2003.

[66] Ravichandran Veerasamy, Harish Rajak, Avijeet Jain, Shalini Sivadasan, Parayil Varghese Christapher, and Ram Agrawal. Validation of qsar models - strategies and importance. *Int J Drug Design and Discov*, 2:511–519, 07 2011.

[67] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[68] Zhenxing Wu, Minfeng Zhu, Yu Kang, Elaine Lai-Han Leung, Tailong Lei, Chao Shen, Dejun Jiang, Zhe Wang, Dongsheng Cao, and Tingjun Hou. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings in Bioinformatics*, 22(4):bbaa321, 12 2020.

[69] Qisong Xu, Pui Shan Chow, Erte Xi, Randy Marsh, Shikar Gupta, and Krishna M Gupta. Evaluation of polymer-preservative interactions for preservation efficacy: molecular dynamics simulation and QSAR approaches. *Nanoscale*, -:–, August 2024.

[70] Qing Ye, Xin Chai, Dejun Jiang, Liu Yang, Chao Shen, Xujun Zhang, Dan Li, Dongsheng Cao, and Tingjun Hou. Identification of active molecules against Mycobacterium tuberculosis through machine learning. *Briefings in Bioinformatics*, 22(5):bbab068, 04 2021.

[71] Fumitaka Yoshida and John G Topliss. QSAR model for drug human oral bioavailability. *J. Med. Chem.*, 43(13):2575–2585, June 2000.

[72] Saeed Yousefinejad and Bahram Hemmateenejad. Chemometrics tools in qsar/qspr studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149:177–204, 2015.

[73] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.*, 52(D1):D1180–D1192, January 2024.

[74] L. Zhao, W. Wang, A. Sedykh, and H. Zhu. Experimental errors in qsar modeling sets: What we can do and what we cannot do. *ACS Omega*, 2017.