universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Validation of the novel quantitative pharmacophore modeling algorithm QPhAR

verfasst von / submitted by
Matthias Schmid, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 606
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Drug Discovery and Development
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thierry Langer

Zusammenfassung

Quantitative-Struktur-Aktivitäts-Beziehungen (QSAR) sind eine wirksame Methode zur Untersuchung spezifischer Arzneimittel für eine Vielzahl von Wirkstoffzielen. QSAR nutzt die Korrelation zwischen der Molekularstruktur chemischer Verbindungen und ihrer biologischen Aktivität, um es Forschern zu ermöglichen, die Aktivität neuer potenzieller Liganden für ein Protein vorherzusagen. Ein neuer Ansatz bei QSAR-Studien ist die Verwendung von Pharmakophoren zum trainieren des Modells anstelle der kompletten Molekülstruktur. Diese Masterarbeit vergleicht den neuen QPhAR Algorithmus mit dem etablierten HypoGen Algorithmus von BioVia. Es wurde eine Literatursuche nach veröffentlichten Studien durchgeführt, in denen QSAR mit Pharmakophormodellen durchgeführt wurde, die mit dem HypoGen-Algorithmus erstellt worden waren. Mit denselben Datensätzen wurden Pharmakophormodelle mittels des QPhAR Algorithmus erstellt. Die Modelle wurden mit einem externen Testdatensatz validiert, um sie direkt miteinander vergleichen zu können. Die erhaltenen Ergebnisse zeigen, dass der QPhAR Algorithmus bei den meisten der untersuchten Systeme eine vergleichbare oder sogar bessere Leistung zeigt als der HypoGen Algorithmus.

Abstract

Quantitative-Structure-Activity-Relationship (QSAR) is an effective method to investigate the effect of different pharmaceuticals for a wide variety of drug targets. QSAR takes advantage of the correlation between chemical compounds' molecular structure and biological activity, enabling researchers to predict novel potential ligands' biological activity for a therapeutic target**.** A new approach in QSAR studies is to use only pharmacophores to train the model instead of whole molecular structures. The advantages of this approach are manifold, as for example, pharmacophores being less susceptible to spatial disturbances than molecular structures. This thesis compares the novel QPhAR algorithm to the established HypoGen algorithm from BioVia. A literature search was conducted for published studies in which QSAR was performed with pharmacophore models produced with the HypoGen Aagorithm. Multiple regression pharmacophore models were built using the QPhAR algorithm on the same data sets. The pharmacophore models of both algorithms were validated on the test set to compare them directly. The results obtained from this study have shown that the QPhAR algorithm performs on-par or even better than HypoGen on most of the evaluated protein targets.

Table of content

Acknowledgments

I want to express my sincere gratitude to Univ.-Prof. Mag. Dr. Thierry Langer for providing me with this master thesis project and supervising me during my Master's program.

Stefan M. Kohlbacher for his great support and help with the project, python scripts, and programming parts of my work. As well as his great patience and that he was always available for questions and discussions regarding my master thesis.

I want to thank Thomas Seidel for providing the chemoinformatics toolkit CDPKit and helping with the installation of programs on the university computers.

As well as the entire AG-Langer research group for the warm welcome and tremendous willingness to help with upcoming questions

Last but not least, I want to thank my Parents, Michaela Schmid and Rudolf Schmid, and my Girlfriend, Jeniffer Steinmassl, for the great support and for helping me complete this exciting and admirable chapter of my life

Introduction

Modern pharmacy and drug design see themselves confronted with more severe problems in the last decades than ever before. An aging population leads to an increased occurrence of age-related health problems such as diabetes, cancer, and dementia[1]. Furthermore, health care crises like the Ebola outbreak 2014 in west Africa[2] and the global COVID-19 pandemic show the importance of progress in pharmaceutical research. Fortunately, this progress never stagnates, and novel processes in the pharmaceutical industry as well as new insights generated by research pave the way for new and more effective methods accelerating drug design, hit identification, lead identification, and lead optimization.

The research on biologicals has gained more and more interest in the last decades. Biologicals are either directly synthesized in plants or modified versions of these molecules that show potential therapeutic use in different therapeutic approaches. Furthermore, drug repurposing is becoming an attractive proposition in drug design studies. This technique is based on the treatment of new therapeutic targets with known drug molecules that are already used in the treatment

of other diseases. Due to the involvement of de-risked molecules, drug repurposing can reduce the developmental costs and timeline in drug design studies[3]. Optimizing known drug molecules to improve druggability, biological activity, and bioavailability an important step in the drug development cycle. A powerful task for this task is bioisosterism[4]. A bioisostere is defined as "a molecule resulting from the exchange of an atom or of a group of atoms with an alternative, broadly similar, atom or group of atoms"[5].

All these mentioned approaches and methods represent powerful tools in the pharmaceutical processes and help to address the current and emerging hurdles for the healthcare systems. However, the medical care of the world population poses more significant challenges for pharmaceutical research and industry ever. Growing populations, changing lifestyles, and adapting pathogens[6,7] require an ability for quick adaption and a more efficient overall drug design process. Lengthy research and synthesis approaches of all potential therapeutics for a disease of which the majority will not succeed the clinical trials are therefor no longer an appropriate answer to the growing demands in current drug design.

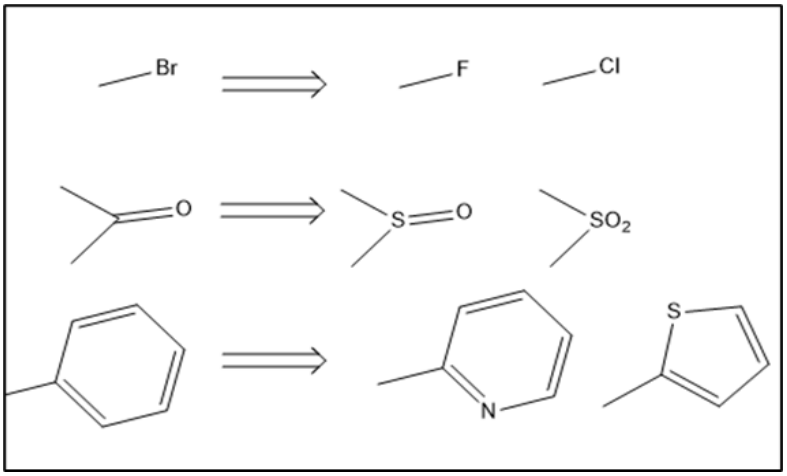

Figure 1: Examples of common bioisosteres. Each line consists of distinct functional groups that can be used to exchange each other while remaining the biological effect of chemical. The chemical structures are drawn with the ChemDraw® software. More examples are given in[71].

Powerful approaches to assist in the drug development pipeline are *in-silico* methods[8]. *In-silico* methods often utilize computer-based mathematical models for the simulation and prediction of pharmacological, physicochemical, and physiological processes and offer various advantages. Applied correctly, *in-silico* methods allow researchers to simulate and model biological systems like protein-ligand interactions and can closely mimic experimental setups in laboratories. Therefore, it is possible to save time, save money, and replace animal testing in early stage drug development by simulating such experiments on the computer. Furthermore, simulation algorithms allow

researchers to gain an insight into spatial movements of the target systems. For example, quantitative structure activity relationship studies can provide valuable insights regarding the impact of the physicochemical properties of different residues on the affinity of a lead compound to the target receptor. This information can then be used for conducting more directed structural modifications in the lead optimization phase which waste much less resources than undirected trial-and-error approach. Hence, *in-silico* methods help reduce the number of synthesized molecules in otherwise expensive and time-consuming experiments by filtering out the molecules with the best potential for new therapeutic agents. Another concept that has gained increased interest in the last decades is the use of pharmacophore based techniques. According to IUPAC, pharmacophores are defined as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [9]. Molecules are represented by the functional groups that are necessary for the interaction with the therapeutic binding site called pharmacophore features. Common pharmacophore features are Hydrophobic features (H), Hydrogen bond acceptors (HBA), Hydrogen bond donors (HBD), aromatic ring systems

(AR), positive ionizable groups (PI), and negative ionizable groups (NI) (*Figure2*).

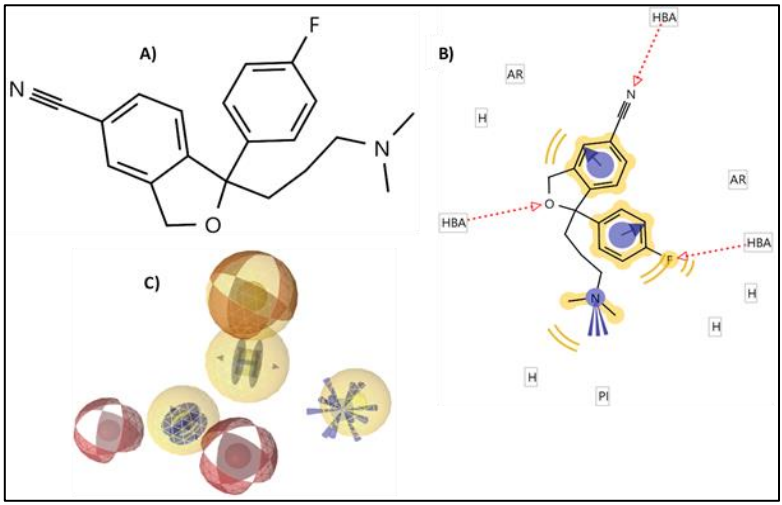

Figure2: 2-Dimensional structure of the drug molecule citalopram(A), a serotonin reuptake inhibitor, with the 2-Dimensional representation of the corresponding pharmacophore features(B) and the pharmacophore field of the molecule(C). The abbreviations of the pharmacophore features are H: Hydrophobic area, HBA: Hydrogen-Bond-Acceptor, PI: Positive ionizable, AR: Aromatic ring system. The molecule was drawn in LigandScout and the pharmacophore was generated by the implemented algorithm in LigandScout.

Pharmacophores are often used to replace the input molecules' complete steric and electrostatic interaction fields in *in-silico* studies. Using pharmacophores as input for experimental setups offers multiple benefits compared to

molecular structures. First, pharmacophores are more resistant to small conformational perturbations compared to molecule input data. Second, another is shown up when in terms of bioisosterism. As described earlier, bioisosteres are chemical groups that show similar biological effects on the receptor-binding site but can consist of totally different functional groups and substructures. A potential bias[10] for a model arises from the probability that a specific functional group or sub-structure occurs more often in the training dataset, thereby biasing the model towards the predominant bioisosteric form. However, by their nature, the different bioisosteric groups show the same interaction pattern at the receptor-binding site, and this fact can be exploited. Pharmacophores transform different functional groups with a similar interaction profile into an abstract chemical representation associated with a specific interaction type. The problem of a bias towards a dominant bioisosteric form in the dataset is negated. Besides the abstract representation of molecule structures, pharmacophores also abstract spatial information like the exact steric location and orientation of interactions by introducing tolerance ranges. Even though such a generalization is not desired all the time, for example in highly conserved protein binding sites, due to the loss of precise positional information of such interactions, it can help

to avoid overfitting a model on the training data set. Overfitting the model on the training set means it fits too closely or precisely against the training data. Thereby, the model cannot generalize to unseen data and will perform poorly on the test set. A simplified scheme about the difference between a generalized and an overfitted model is given in Figure 3. In this figure, The green circles represent the test set data points, and the blue circles represent the training set data points. The red straight line on the left side represents the generalized model. The black dashed lines represent the R-squared of the test set data points.

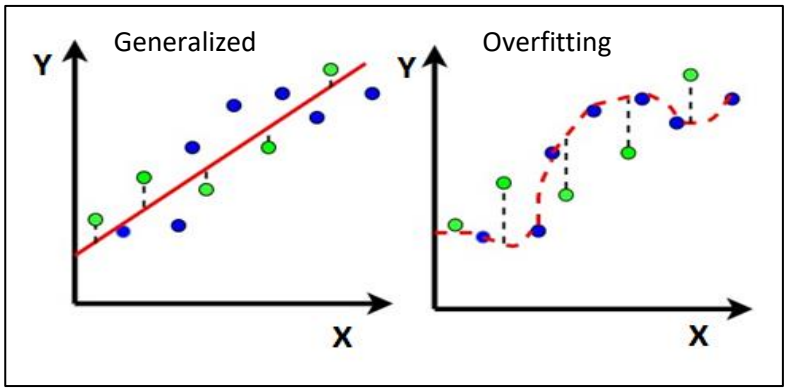

Figure 3: Scheme of the fitting line of a generalized model and an overfitting model. The red straight line on the left side represents the generalized model. The red dashed line on the right represents the overfitting model.

Considering all the advantages mentioned, it is no surprise that computer-based drug design[11] is one of the pivotal approaches in pre-clinical drug design nowadays. CADD can be roughly separated into two broad fields: structure-based drug design[12] and ligand-based drug design[13].

The basis of structure-based drug design is information about the three-dimensional structure of target proteins and their binding site cavity. Structure-based methods are effectively fast and specific in lead identification and optimization approaches that can help understand the interaction on a molecular level. These applications are used for multiple assessments such as conformational changes of target binding sites upon binding a ligand, target-ligand interactions, and binding energetics. Commonly used methods in structure-based drug design are molecular docking[14], virtual screening (VS)[15] studies, and molecular dynamic (MD)[16] simulations.

Virtual screening allows the researcher to search a library of chemical compounds for new potential ligands for a therapeutic endpoint. It can be separated into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS). In structure-based virtual screening, the model is generated based on the ligand in its active

conformation in the receptors binding site, while for the ligand-based virtual screening, multiple conformations of the known ligands are calculated, and a merged or shared model of all calculations is created.

In molecular docking, potential ligands are scored by a mathematical function, the scoring function[17], to calculate the binding energy of each ligand in the binding site. The best scoring ligands can be used as a template to search for structurally similar molecules that are likely to show a similar activity on the therapeutic target. An example for a workflow in SBDD consists of the following steps: Identification and preparation of the target protein, identification of the therapeutic binding site, preparation of compound library, molecular docking and scoring functions, molecular dynamics simulation, and binding free energy calculations and is schemed in Figure 4.

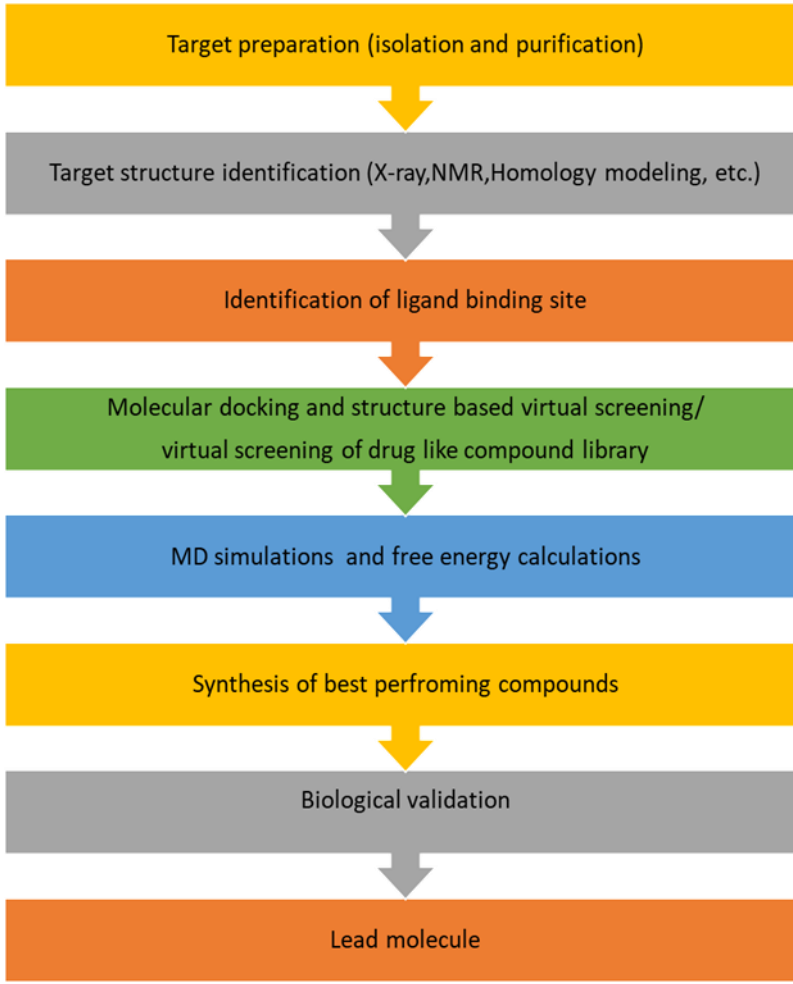

Figure 4: Schematic of a basic workflow in SBDD

On the other hand, ligand-based drug design is handy when there is no experimental information about the three-dimensional structure of the drug target available, but chemical compounds are known to bind to the therapeutic target under investigation. The correlation between the structure of the ligands and their physicochemical properties makes it possible to generate descriptive models. Many successful drug design studies have already been performed using ligand-based drug design, and due to technological progress, these methods are more often combined with modern artificial intelligence approaches such as machine learning and deep learning approaches. The most important methods in ligand-based drug design are pharmacophore modeling and 3D-quantitative-structure-relationship studies.

Pharmacophore modeling has been increasingly used in pre-clinical drug design studies in the last decades. Pharmacophore modeling structures can be either structure-based or ligand-based. In structure-based pharmacophore modeling approaches, common workflows produce a pharmacophore model of an active ligand bound to the receptor or the empty receptor binding site, called an apo-site pharmacophore model. While an apo-site pharmacophore model consists of every possible pharmacophore feature in

the binding cavity, a ligand bound to the receptor-binding site is most likely in its active biological conformation. Hence, the obtained model can be used directly for further drug design studies, such as compound library screening for other potential ligands and drug repurposing approaches.

However, in a lot of drug design studies, there is no structural information of the receptor available. In that case, it is possible to generate a ligand-based pharmacophore model from known active compounds (Figure 5). For the elucidation of a ligand-based pharmacophore model, two steps are required. First, multiple conformations of the ligands must be generated to adequately cover the conformational space of the ligands since the bioactive conformation of the ligands is unknown. The 3D structures of the conformers are aligned to each other, and a shared pharmacophore model is built on the shared chemical features of the molecule conformations. If all aligned conformers provide a functional group with equivalent physicochemical properties at a specific location, a pharmacophoric feature is placed at this position. To increase the quality of the model, inactive or less active compounds can be considered in the pharmacophore model generation process. Hence, it is possible to insert spatial constraints to the model occupied by the inactive molecules.

Furthermore, chemical features that are not observed in most of the active ligands should either be made optional or removed from the final model. By incorporating such optimizations, it is possible to increase the model's specificity and sensitivity to avoid false positives and false negatives, respectively. Finally, to confirm the quality of the model, it must be validated on an external test set. The test set consists of molecules that show activity on the therapeutic endpoint but were not used for the model's training, as well as inactive molecules. It is used to evaluate whether the derived pharmacophore models can distinguish active compounds from the inactive ones[21].

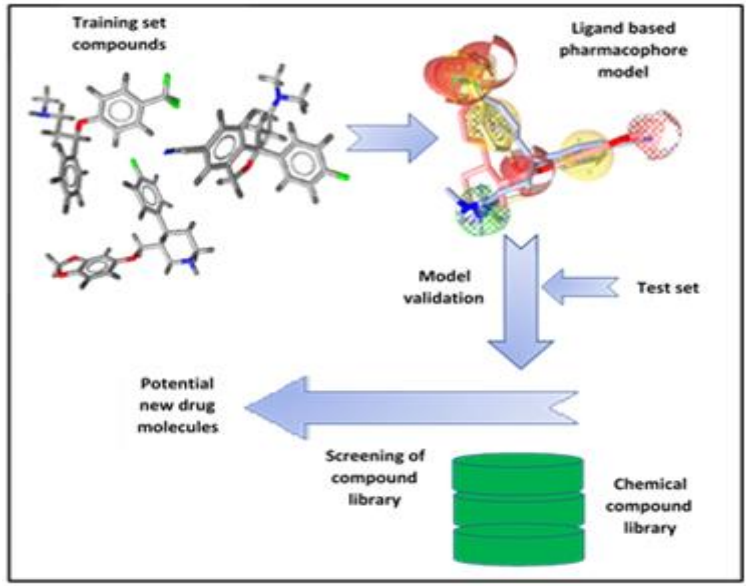

Figure 5: Schematic of a classic ligand-based pharmacophore modeling workflow. The 3-dimensional models of the molecules and the pharmacophores were generated using LigandScout.

Nowadays, several software packages for the generation of pharmacophore models from input molecules are available, for instance, PHASE by Schrödinger, HypoGen by BioVia, LigandScout, and GALAHAD, listed in table 1 below.

Table 1: Commonly used software in pharmacophore modeling and ligand-protein interaction studies

Software	Description	Reference
PHASE	Pharmacophore based tool using a tree-based partitioning algorithm to comprehensively map common spatial arrangement of functional groups in a set of bioactive ligands	[18]
LigandScout	Fully automated tool for generation of pharmacophore models which detects and classifies protein-ligand interactions. Additionally, has a virtual screening algorithm implemented for further studies with the produced models.	[19]
Catalyst **HypoGen**	Based on an algorithm that identifies the three-dimensional configuration of chemical features of a set of ligands.	[20]
GALAHAD	Performs flexible alignments of small molecules that bind to target proteins and share similar interaction patterns.	[21]

Another technique that is increasingly used in pre-clinical drug design over the last years is QSAR, first described by Hansch et al. in 1962[22]. In QSAR studies, a statistical model is created upon the correlation between biological activity and molecular structure. In consecutive steps, the new model can predict the biological properties of investigative ligand molecules. The objective of QSAR studies is to find a correlation between the molecular structure and the corresponding biological effect on the receptor of ligands. In its early days, simple biological and chemical parameters, such as the logP value or the Ki value, were correlated in QSAR studies. Nowadays, QSAR has evolved into a sophisticated method that can handle thousands of molecules and processes several thousand descriptors[23]. Molecular descriptors describe structural properties of molecules in numerical form for a processing by mathematical methods. Some of the commonly used descriptors in QSAR studies are listed below (*Table 2*).

Table 2: Molecular descriptors used in QSAR studies.

Descriptor type	Description	Example
Constitutional	Commonly used descriptors that reflect the chemical properties of a molecule without providing information about connectivity	Molecular Weight number of atoms in the molecule
Electronic	Describe electrostatic properties of the molecule or atomic bonds	HOMO and LUMO[24]
Topological	Represents the intramolecular connectivity of the atoms. Can be used for modeling pharmacokinetic, physicochemical, and biological properties	Zagreb connectivity indices[25]
Geometrical	Capture 3D information regarding molecular size, molecular shape, and atomic distribution. Geometrical descriptors are calculated from the 3D coordinates.	MoRSE descriptors[26]
Thermodynamic	Thermodynamic descriptors are used to relate chemical structure to observed chemical response	Molecular refractivity (molREF)[27]

There are different subtypes of descriptors to define different molecular characteristics. In QSAR studies, five different types of descriptors are used, namely, constitutional descriptors, electronic descriptors, geometrical descriptors, thermodynamic descriptors, and topological descriptors. The descriptors are described in more detail in table 2. Depending on the descriptors and how they are derived, QSAR can be separated into the following subtypes: (1) 1-D-QSAR, which studies the correlation of global molecular properties such as logP or pK_i value. (2) 2-D-QSAR, where the target property correlates with two-dimensional properties like 2-D pharmacophores or connectivity indices. (3) 3-D-QSAR studies the correlation between the molecules' activity and noncovalent interactions with their surroundings. (4) 4-D-QSAR, an extension of 3-D-QSAR that incorporates an ensemble of ligand configurations. (5) 5-D-QSAR, which is 4-D-QSAR that additionally takes different induced fit models in mind, and finally (6) 6-D-QSAR, an extension of 5-D-QSAR that incorporates multiple solvation models[28].

There are some critical requirements for QSAR studies. First, an appropriate data set is required, which means that it contains molecules with known molecular data such as biological activity. Second, the dataset must be appropriately

separated into training and test set. The training set should have a sufficient large range for the biological activity of the molecules in the data set, and the activity values should not cluster at a specific activity range. This is important to ensure that the model can learn from a diverse enough dataset and is not biased to a particular activity range. Third, the chemical properties of the compounds should not show autocorrelation to avoid overfitting the data. Autocorrelation[29] describes the correlation of values of the same variable in different observation points. For example, constitutional descriptors are chemical descriptors defined by a specific chemical structure. Suppose autocorrelation occurs between the values of this variable. In that case, it means that very similar structures occur several times in the dataset, posing a potential risk of a bias in model training. Finally, the retrieved model must be validated to confirm its applicability and predictive power[30].

QSAR is often used with complex machine learning models to handle the vast amount of today's available chemical data. Machine learning allows machines to learn from given input data without the need to program them to do so explicitly. These models can subsequently be used to make predictions based on the input data, calculate probabilities for certain

events, recognize data groups or clusters, and improve processes in the software domain. In QSAR, machine learning teaches algorithms to recognize certain structural characteristics of input molecules with specific biological properties on a given therapeutic target. The most commonly used machine learning algorithms that have been used in QSAR studies are linear regression[31,32], support vector machines[33], random forest[34], and deep learning[35].

QSAR techniques can be separated into two types, linear and non-linear approaches in general. For linear techniques, linear regression, multiple linear regression, and partial least squares are commonly used examples. Examples of non-linear QSAR methods are k-nearest-neighbors and artificial neural networks. Available programs that perform 3-D-QSAR studies are HypoGen by BioVia, PHASE by Schrödinger, and CoMFA.

This master's thesis aims to validate a new algorithm for quantitative pharmacophore modeling, namely Quantitative Pharmacophore Activity Relationship (QPhAR)[38], developed by Stefan M. Kohlbacher in the group of Univ. Prof. Mag. Dr. Thierry Langer is implemented on top of the Chemical Data Processing Toolkit (CDPKit) by Thomas Seidel[36]. For validation, this thesis compares the performance of the

produced pharmacophore models of the QPhAR and HypoGen algorithm using test sets of published datasets, with which QSAR studies were performed employing the HypoGen algorithm. For comparison of the performance, the statistical metrics R-squared (R^2), root-mean-square-error (RMSE), and standard error are considered.

Pertinent algorithms and statistics

HypoGen by BioVia

BioVia's Discovery studio[37] is an automated tool for pharmacophore pattern recognition from a collection of chemical compounds based on the correlation of chemical structures with biological activity. The models produced by BioVia, named hypothesis, consist of pharmacophore features representing certain interaction patterns of the ligands to the protein receptor. Features can be created for the following types of functional groups: positive and negative ionizable functional groups, hydrophobic areas, aromatic ring systems, and hydrogen bond acceptors and donors. The produced models can be used directly as a database search query within the Discovery Studio software. Many commonly used algorithms define the bioactive conformation of the

input molecules as the conformation that shows a local energy minimum. HypoGen, on the other hand, always tries to use a wide range of accessible conformations of the considered molecules within a user-defined energy threshold. HypoGen does not focus on the conformation with a local energy minimum because several studies have revealed that the binding conformation of a small molecule in the target binding pocket is not always the one with the local energy minimum.[38] The conformers' redundancy is one problem with the sampling of conformers and the consecutive search for the conformation with a local energy minimum. Many different conformations are created during sampling, and then an attempt is made to filter out representative families of conformers. The HypoGen[38] Algorithm considers all conformations of a biomolecule to avoid such problems. HypoGen[43] uses data from biological assays like the IC_{50} or K_i value to generate pharmacophore models that can quantitatively predict the activity of chemical compounds for a specific target. However, the molecules in the training set for the HypoGen Algorithm must possess the same binding mode.

The HypoGen Algorithm works with pharmacophores derived from the input molecules, and its workflow consists of three

consecutive steps: constructive, subtractive, and optimization phase. The input data is divided into active and inactive compounds in the constructive phase. Furthermore, a pharmacophore is generated on the two most active compounds. Next, only the generated hypotheses (pharmacophore models) that fit a minimum subset of chemical features from molecules in the training set are kept.

All hypotheses common to the inactive compounds in the training set are discarded in the subtractive phase. Compounds are considered inactive if they have an activity value of 3.5 log units (the user can specify this value) lower than the most active compounds. Finally, in the optimization phase, all produced hypotheses are scored based on the RMSE values of the predictions against the training set to improve the quality of the pharmacophore models. Minor changes are made to increase the predictive power of the models, and in the end, the ten simplest models with the best activity predictions are reported to the user.

A drawback of this method is that the hypotheses are produced on a subset of highly active compounds. Since the information obtained from less active compounds is discarded this leads to the presumption that predictions on pharmacophores of less active compounds will be worse

because of the lack of domain knowledge. Furthermore, the algorithm presents the top ten best hypotheses to the user instead of one model that shows the best results, adding some ambiguity about the model's quality.

QPhAR algorithm

The novel quantitative pharmacophore modeling algorithm QPhAR[39] developed by Stefan M. Kohlbacher relies on multiple consecutive steps. First, a template must be chosen. The template could either be one of the training samples, for example, the most rigid molecule or any molecule or pharmacophore that the user deemed relevant. Second, the pharmacophores or pharmacophores derived from the molecules in the training set are aligned to the template. Third, the features of the aligned pharmacophore models are clustered. A representative feature is chosen or generated in the next step for each cluster. In the following post-processing steps, clusters with non-conclusive information are discarded, and relevant clusters are kept in the quantitative pharmacophore model. Finally, the remaining features are used as input for a regression machine learning model in the subsequent modeling process. A schematic illustration of the QPhAR algorithm is shown in Figure 6.

Template selection

Template selection is one of the most crucial steps in pharmacophore model generation. A bad template leads to follow-up errors in further modeling steps and pharmacophore models of bad quality. The template is chosen by aligning a set of two molecules to each other via pharmacophore alignment, taking into consideration the conformational flexibility of the molecules. Additionally, directed pharmacophore features are translated into undirected spherical features to minimize the noise introduced by additional direction information. Afterwards, a pharmacophore is built upon the best alignment on the first molecule. As the first molecule, the most rigid molecule based on the number of intramolecular non-hydrogen bonds is chosen.

Alignment

After a template has been chosen, the remaining training samples are aligned against the template using pharmacophore alignment. The aligned pharmacophore features are stored in separate data structures, so-called containers. Each feature type is stored in a specific container,

resulting in six containers for hydrophobic features, aromatic features, positive/negative ionizable features, and hydrogen bond acceptor/donor features. Each of these pharmacophore features is then associated with the activity of the parent pharmacophore. Directed pharmacophore features are again translated into undirected spherical ones to minimize the introduction of noise.

Clustering

Next, the pharmacophore features are processed by a clustering. Hereby, a maximum distance hierarchical clustering algorithm is applied, and its cutoff is treated as a hyperparameter. The default cutoff is the radius of pharmacophore features (1.5 Å). A distance matrix containing the euclidean distances between all pharmacophore features in a container is used as the input for the clustering algorithm

Post-processing

As final steps after clustering, two post-processing steps are applied. First, representative features of each cluster are selected. Second, clusters with non-conclusive activity data

are removed and only high-impact pharmacophore features are kept for the final model.

In the ideal case, each cluster can be represented by one specific feature. If this is not possible, multiple features have to be chosen to represent a cluster. The representative feature can either be one of the existing features in the cluster or the product of merging all overlapping features. The merged feature inherits all activity values from the features it represents. Overlapping means that the distance between the features is smaller than the radius of the pharmacophore feature sphere.

After this step, the quantitative pharmacophore consists of multiple features, each representing a cluster of pharmacophore features from the training set. However, some of these features will not add information to the final model. A simple example is imagining two molecules with the same scaffold but different residues. One of them is biologically active and the other inactive. A merged pharmacophore model will contain features representing the scaffold and features representing the different residues. While the residues' features explain the molecule's activity, the features representing the scaffold do not provide any useful information to the final pharmacophore model.

Clusters containing features that do not contribute information are seen as non-conclusive and are discarded to increase the quality of the quantitative pharmacophore. Furthermore, features that are encountered only once are considered outliers and are removed from the final model due to the missing validation of the features' importance. After that, a merged and cleaned quantitative pharmacophore model is obtained for the subsequent machine learning process.

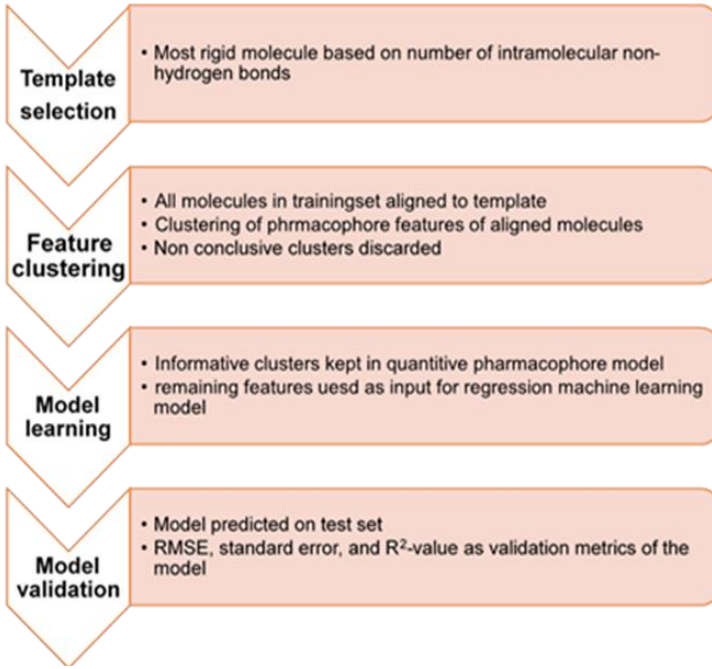

Figure 6: Schematic of the QPhAR workflow.

Statistical evaluations

Fort the validation of the pharmacophore models generated on behalf of this thesis, various statistical evaluations are carried out. These are discussed in detail in the following sections.

Grubbs-outlier test

First, a Grubbs-outlier test[40] is performed to identify and clean the dataset from outliers if necessary. The Grubbs-outlier test is described by equation 1 (Eq1). With X and Sn denoting the sample mean and standard error, respectively.

$$\text{Eq1:}\quad G = \frac{\max\left(Xi - \overline{X}\right)}{Sn}$$

Eq1: Grubbs-outlier-analysis

Paired T-test

Second, a paired t-test[41] on two related values is performed on the predictions of the QPhAR Algorithm. This t-test is done for dependent values, for example, two values calculated from the same sample. The null hypothesis of this two-sided t-test is that two related or repeated samples have identical

average (expected) values. As a result of this, the statistics of the t-test help us validate whether the obtained predictions are comparable or different between both algorithms. The threshold of the p-value for which the null hypothesis is rejected is 0.05. The calculation of the test statistics is depicted in equation2 (Eq2). X_D and S_D denote the mean and standard deviation of the difference between the pairs, respectively, and n denotes the number of pairs.

$$\text{Eq2: } t = \frac{\bar{X}_D}{\frac{Sn}{\sqrt{n}}}$$

Eq2: T-test equation

Wilcoxon signed-rank test

Third, additionally to the paired T-test, a Wilcoxon signed-rank test[42] is performed. It is used to check whether two related samples or values come from the same distribution.

Paired t-test and Wilcoxon signed-rank test are performed using python packages provided by Scipy[43] available at: https://www.scipy.org/.

Leave-one-out cross validation

Next, a retrospectively leave-one-out-cross-validation[44] is performed. In this method, a compound of the training set is omitted and then the RMSE value is calculated on the novel data set. These steps are repeated until every compound has been omitted once. This method is used to make sure that the calculated metrics do not depend on a single compound of the training set.

Y-scrambling analysis

A Y-scrambling[45] analysis is performed to make sure that the received RMSE values are not obtained just by chance. The fundamental steps of the Y-scrambling analysis are the following: First, the RMSE values are calculated for the original data. Second, the data pairs of the experimental and predicted activity of the dataset are shuffled and the RMSE values are calculated for the new incorrect paired dataset. This step will be repeated several times. The assumption is that for a robust model, the results of the shuffled, incorrect data pairs will be considerably worse compared to the original data. If this is not the case, the generated model is not robust on the given datasets. The shuffling of the data pairs was

accomplished using the random shuffling method provided by NumPy[33]. A total of 100 iterations was set for the shuffling step.

Tanimoto coefficient

Finally, the average Tanimoto coefficient[46] of the test set to the training set molecule fingerprints is calculated. A low Tanimoto coefficient indicates high dissimilarity between the training- and test-set molecules. As a consequence, this can consecutively lead to poor templates for pharmacophore model generation and a decreased ability of the model to generalize to unseen data. However, the Tanimoto coefficient is just a rough overview of the structural similarity between test and training set compounds and not an absolute validation metric for the quality of an obtained model. The Tanimoto coefficient is calculated using the KNIME analytics platform desktop app[47].

Methods

Target acquisition

To collect data for the study, literature research was performed to find publications with QSAR studies using quantitative pharmacophore models created with the HypoGen algorithm by BioVia. The publications were filtered for datasets containing the molecules' structure and the experimental and predicted activity data using the HypoGen algorithm. 17 publications were selected containing data sets with comprehensive information about the structure of the input compounds as well as the experimental and predicted biological activity values of the test set molecules. The data sets were extracted from the publication and saved locally. The target systems used in this thesis and its corresponding publications are summarized in *Table 3*.

Table 3: This table represents the targets and corresponding references of the publications used in this study.

Target	Reference
AuroraB kinase	48
Polo-like-kinase	49
EBP	50
σ-1	
ERG2	
Tubulin-casein binding domain	51
5-Lipoxygenase	52
HIV1-Integrase	53
Cyclin A/CDK2 binding domain	54
Endothelin A	55
hERG potassium channel	56
AngiotensinII receptor type 1	57 58
Prolyl-oligopeptidase	59
CDC25B	60
P450 19 aromatase	61
Kappa opioid receptor	62

Datasets

Three targets from Laggner et al.[50] were used: Vertebrate emopamil binding protein (EBP), its fungal counterpart ERG2, and the Sigma-1 receptor. The other targets, Polo-like-kinase 1 (PLK1), AuroraB kinase, Neuraminidase, Tubulin-casein binding domain, etc., that are mentioned in table 3 are taken from the corresponding publications. The molecules of the datasets were drawn with the software ChemDraw® from PerkinElmer informatics[63] to gather smiles codes for further steps in the modeling process. Furthermore, the data sets were split into training and test sets as described in the corresponding publications to ensure a fair comparison of the HypoGen and QPhAR models. Next, the data sets were transformed into a SDF file, containing the structural 3-dimensional coordinate information of the molecules based on the smiles codes. Conformations were calculated using the ICon[64] algorithm implemented in LigandScout[19]. A maximum of 25 conformations per molecule was specified and the remaining parameters of the ICon algorithm were set as default. The training sets were checked with respect to the requirements for QSAR studies. During these controlling steps, the data in the training set was checked whether it fulfills the following criteria:

- Do the activity values span at least 3 log units or 2 log units, not considering outliers?
- Are the activity data clustered or homogeneously distributed?

Although the provided data were used in QSAR studies using HypoGen, controlling the training data is important since it is possible that the molecules perform very well in the published study but not well in general.

Machine learning modeling and hyperparameter selection

The underlying machine learning model type is a hyperparameter of the quantitative pharmacophore algorithm. To avoid overfitting, simple machine learning algorithms are recommended. The following six regression algorithms have been tested in this study: random forest regression[34], principal component analysis[65] (PCA) combined with linear regression[31,32], ridge regression[66], lasso regression[67], PCA + ridge regression, and partial least squares regression[68] (PLS). The parameters' n_estimators' and 'max_depth' of the random forest were optimized too, with the values being [10,15,20] and [2,3], respectively. The

remaining parameters were kept at default. For ridge and lasso regression and their respective combinations with PCA the parameter 'fit intercept' was set to both 'True' and 'False'. The parameters of the other machine learning model algorithms were kept at their default parameters. The following additional hyperparameters of the quantitative pharmacophore algorithm were optimized:

- Weight type: [distance, nrOfFeatures, None]
- Threshold: [1, 1.25, 1.5, 1.75, 2, 2.5, 3]

For the QPhAR algorithm, the machine learning models were trained using the scikit-learn Python package[69].

Regression machine learning model validation

Different metrics for model validation were calculated, whereas the most important values for validations were the root-mean-square-error (RMSE), standard error, and R^2-value. Simultaneously, the same metrics were calculated for the predictions of the HypoGen model on the test set given in the corresponding publications. Finally, the best model based on the metric values was chosen as a representative QPhAR model for the target.

Creation of plots

The activity scatter plots and error bar plots for the visualization of the validation metrics and statistical evaluations of the pharmacophore models were created using the matplotlib package[70] available for Python.

Results and discussion

In the results section, the received validation metrics of the pharmacophore models produced by the QPhAR and HypoGen algorithm on the test set are discussed. The R-squared value (R^2), Root-mean-square-error (RMSE), and standard error (StdError) of the quantitative models for each target are given below (

Table *4*). the QPhAR algorithm performed head-to-head or better than the HypoGen algorithm on multiple targets. Considering the validation metrics, the QPhAR models show better results for the Emopamil binding protein, ERG2, Sigma1 receptor, and the hERG potassium channel compared to the HypoGen models. Considering the R^2 values of these target systems, it is clear that for each target, the QPhAR model performs better on the test set compared to the HypoGen algorithm. The Models of the ERG2 and

Emopamil binding protein result in an R-squared value of 0.578 and 0.522, and 0.115 and 0.221 for the QPhAR and HypoGen model, respectively. Hence, for the Sigma1 and hERG K^+-channel target systems, the statistics show that the QPhAR models have R-squared values of 0.668 and 0.367, whereas the HypoGen model gives R-squared values of -1.156 and -0.563 (Figure 7).

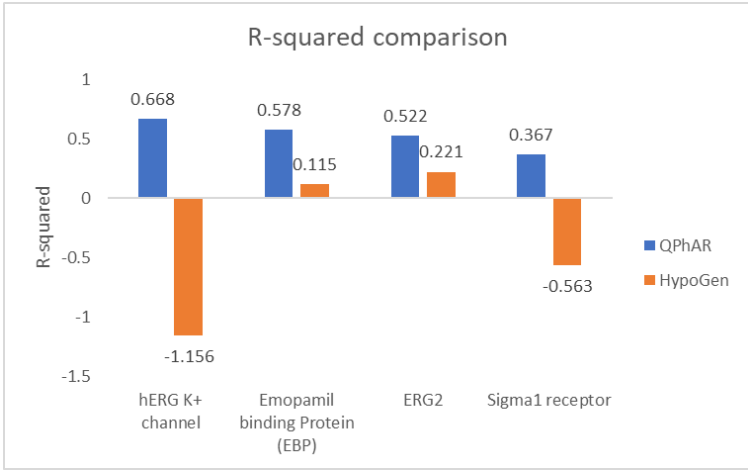

Figure 7: Bar plot of the received R-squared values for the hERG K^+-potassium channel, EBP, ERG2, and Sigma1-receptor target systems.

Moreover, the QPhAR algorithm performs on-par with HypoGen for the CyclinA/CDK2 binding domain, kappa opioid receptor, P450 19 aromatase, and the Angiotensin II receptor I target systems. Regarding the R-squared values, the QPhAR models of the CyclinA/CDK2, Kappa opioid receptor, and the Angiotensin II receptor I target systems

show R^2 values of 0.876 and 0.894, 0.504 and 0.699, 0.195 and 0.371, 0.343 and 0.301, and 0.398 and 0.401 for the QPhAR and HypoGen models, respectively (Figure 6). Even though a better performance of 0.19 and 0.17 regarding the R^2 value, both the pharmacophore model of the QPhAR and HypoGen algorithm perform mediocre on the test set, even if such a difference in the R^2 value could be an indication for a better performance of the HypoGen models, the RMSE value and the standard error show that both models perform similarly on the test set.

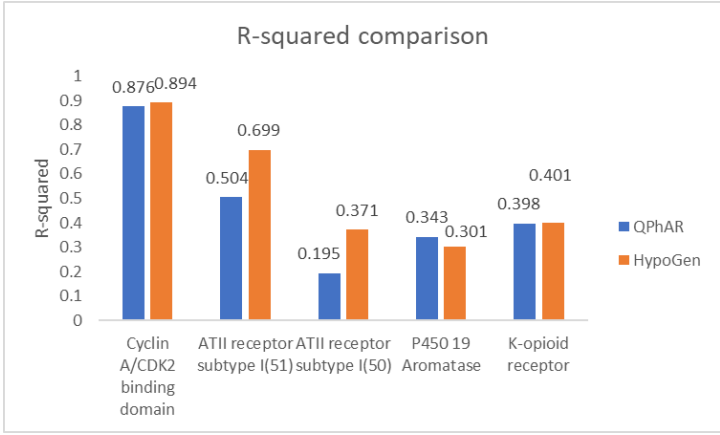

Figure 8: Bar plot of the received R-squared values for the Cyclin A/CDK2, Angiotensin II receptor, P450 19 Aromatase, and Kappa-opioid receptor target systems.

The target systems Tubulin, CDC25B, and Aurora-B kinase result in a low to mediocre performance regarding the R^2 values. The HypoGen models of these targets are interesting because they show a very high R^2 value paired with a very low RMSE value, which could indicate overfitting of the models on the test set. Hence, further analysis of the HypoGen models are necessary to check whether the models overfit the data or not. For the targets, Endothelin A, Polo-like kinase, 5-lipoxygenase, and HIV-1 integrase, the HypoGen algorithm show R^2 values of 0.154 and 0.640, 0.304 and 0.814, 0.450 and 0.870, 0.580 and 0.874, 0.242 and 0.927, 0.282 and 0.783, 0.562 and 0.838, and 0.083 and 0.470 for the QPhAR and HypoGen model, respectively (Figure 9).

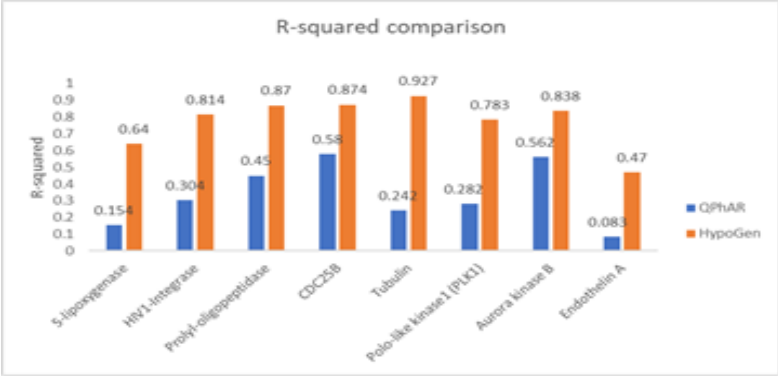

Figure 9: Bar plot of the received R-squared values for the 5-lipoxygenase, HIV-1 Integrase, Prolyl-oligopeptidase, CDC25B,Tubulin, PLK1, Aurora-kinase B and Endothelin A target systems.

Table 4: This table represents validation metrics of the HypoGen and QPHAR pharmacophore models for each target system.

Target	Algorithm	R^2	RMSE	StError
5-lipoxygenase	QPhAR	0.154	0.782	0.444
	HypoGen	0.64	0.51	0.284
HIV1-Integrase	QPhAR	0.304	0.877	0.504
	HypoGen	0.814	0.454	0.277
Cyclin A/CDK2 binding domain	QPhAR	0.876	0.414	0.214
	HypoGen	0.894	0.383	0.199
hERG K+ channel	QPhAR	0.668	0.554	0.396
	HypoGen	-1.156	1.413	0.809
ATII receptor subtype I (58)	QPhAR	0.504	0.704	0.369
	HypoGen	0.699	0.549	0.334
Prolyl-oligopeptidase	QPhAR	0.45	1.022	0.732
	HypoGen	0.87	0.496	0.342
Emopamil binding Protein (EBP)	QPhAR	0.578	1.135	0.673
	HypoGen	0.115	1.643	1.118
ERG2	QPhAR	0.522	1.319	0.68
	HypoGen	0.221	1.683	1.045

Continue Table 4

Target	Algorithm	R^2	RMSE	StError
Sigma1 receptor	QPhAR	0.367	0.934	0.576
	HypoGen	-0.563	1.468	0.941
CDC25B	QPhAR	0.58	0.44	0.256
	HypoGen	0.874	0.241	0.146
Tubulin	QPhAR	0.242	0.798	0.465
	HypoGen	0.927	0.249	0.163
ATII receptor subtype I (57)	QPhAR	0.195	0.609	0.349
	HypoGen	0.371	0.538	0.291
P450 19 Aromatase	QPhAR	0.343	0.749	0.467
	HypoGen	0.301	0.773	0.46
K-opioid receptor	QPhAR	0.398	0.987	0.604
	HypoGen	0.401	0.984	0.664
Polo-like kinase1 (PLK1)	QPhAR	0.282	0.872	0.479
	HypoGen	0.783	0.48	0.232
Aurora kinase B	QPhAR	0.562	0.458	0.248
	HypoGen	0.838	0.279	0.148
Endothelin A	QPhAR	0.083	1.209	0.650
	HypoGen	0.467	0.921	0.492

The low performance of the QPhAR algorithm on these targets could have several reasons. First, the HypoGen algorithm mainly considers the high activity ligands and discards low activity ligands., The considered ligands may perform well on the given test set, but inferior on other data splits. Due to missing cross-validation studies of the HypoGen algorithm, we were not able to test this hypothesis. In contrast, the QPhAR algorithm considers all ligands in its calculations. Second, it is possible that the randomly generated conformations do not include correct conformations for the model generated during the QPhAR workflow. In simple words, if none of the generated conformations come close to the generated model, then the model cannot perform well on the test set. Finally, the structural difference between the training set and test set molecules is a possible explanation for the low performance of some of the targets. Hence, a low Tanimoto coefficient between the fingerprints of the training and test set molecules indicates a high structural difference between those chemicals that could lead to a bad template for the model training. On the other side, one of the biggest advantages of the QPhAR algorithm is that it works with pharmacophores as input data, and thereby, structural differences between the

training and test set molecules do not necessarily affect the quality of the models.

Table 5: This table lists average Tanimoto coefficients of the fingerprints of training to test molecules for each target system. Abbreviations: TC: Average Tanimoto

Target-system	**TC**
5-lipoxygenase	0.252
HIV1-Integrase	0.384
Cyclin A/CDK2 binding domain	0.551
hERG K+ channel	0.299
ATII receptor subtype I(51)	0.633
Prolyl-oligopeptidase	0.379
EBP	0.295
ERG2	0.298
Sigma1 receptor	0.309
CDC25B	0.406
Tubulin	0.369
ATII receptor subtype I(50)	0.616
P450 19 Aromatase	0.374
K-opioid receptor	0.468

Continue Table 5

Target-system	TC
Polo-like kinase1 (PLK1)	0.338
Aurora kinase B	0.506
Endothelin A	0.439

The calculations result in average Tanimoto coefficients between 0.252 and 633 for the different target systems (Table 5). For example, QPhAR models that perform well on the test set show low and higher Tanimoto coefficients, like the hERG K^+-channel with a Tanimoto coefficient of 0.299 and the CDK2 target system with a coefficient of 0.551. On the other side, QPhAR models that perform mediocre or bad on the test set have comparable Tanimoto coefficients. For example, the Endothelin-A target system with a coefficient of 0.439 or the 5-Lipoxygenase target with a coefficient of 0.252. These examples point out that QPhAR models that perform bad and good have similar Tanimoto coefficients and confirm that the QPhAR algorithm can perform well on a target, even though the training set is structurally different from the test set. In conclusion, the evaluations show that each pharmacophore model generated by the QPhAR algorithm have a comparable Tanimoto coefficient for the

fingerprints of the training to test set molecules. The given values indicate that the low performance of the pharmacophore models of the Endothelin A, Prolyl-oligopeptidase, PLK1, 5-lipoxygenase, and HIV-1 integrase target systems do not rely on the structural differences of the training to the test set molecules. In the following, the targets and their results will be analyzed in more detail.

hERG K+ channel

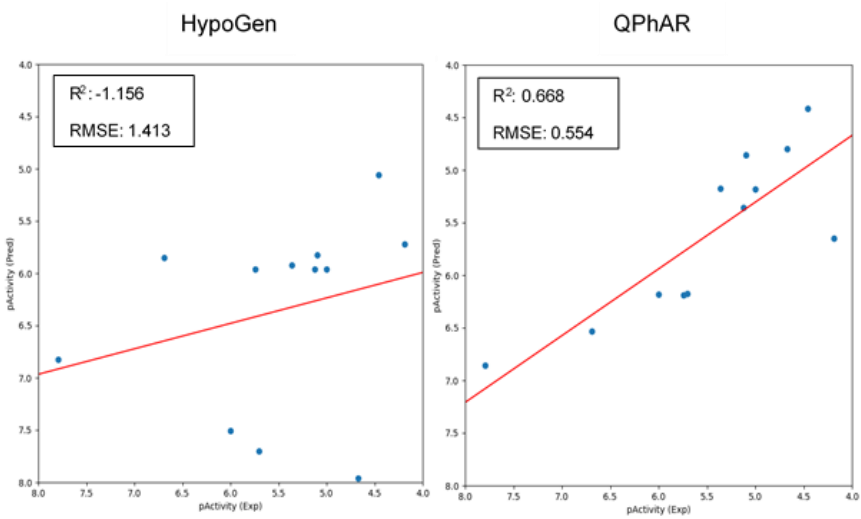

Figure 10: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPhAR (right) algorithm

The pharmacophore models for the hERG potassium channel target systems show R^2 values of -1.156 and 0.668, an RMSE value of 1.413 and 0.558, and standard errors of 0.809 and 0.396 for the HypoGen and QPhAR model, respectively. The scatter plots (*Figure 10*, right) of the QPhAR predictions against the experimental activity values show an almost perfect regression line. On the other hand, the scatter plot of the predicted activity values of the HypoGen (Figure 10, left) model indicates hardly any relationship between the predicted and experimental activity values.

Statistics

Table 6: Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.06
Wilcoxon-signed test	0.099
Average Tanimoto-coefficient	0.299

The p-Value of the two-sided t-test and Wilcoxon-signed test confirm that the average expected mean of the QPhAR predictions and HypoGen predictions are not the same. Hence, the predictions of both algorithms are significantly different from each other. The average Tanimoto coefficient of the training molecules to the test set molecules is 0.299. This is interesting because it confirms that the QPhAR algorithm can perform well on a structural different input dataset, which is one of its strengths.

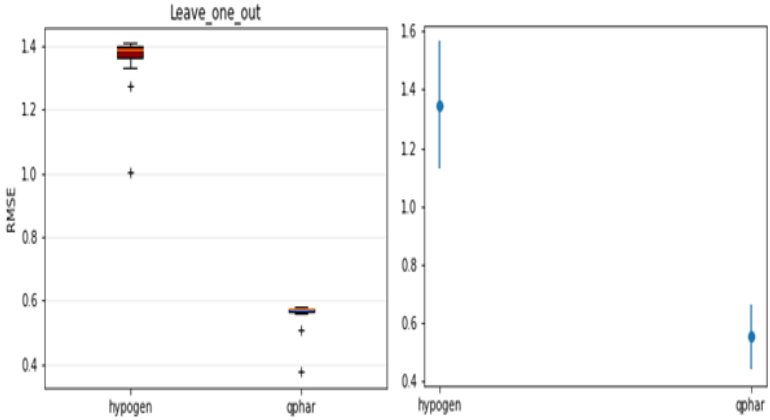

Figure 11: his figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the average RMSE values of the HypoGen and QPhAR models as a box plot. The right figure shows the error bars plot of the two models with the error bars at 95% confidence interval.

The RMSE values obtained by the leave-one-out validation fit the values received from the models in the regression

analysis on the test set. These results indicate that the obtained RMSE values do not depend on a single compound in the training set.

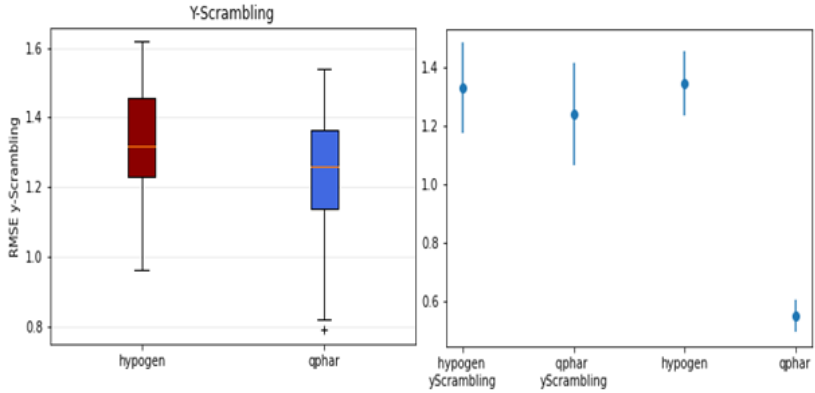

Figure 12: This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPhAR model (blue) on the left side and the error bar plot of the RMSE values on the right side with a 95% confidence interval.

The Box plot and error bar plot of the Y-scrambling analysis (*Figure 12*) of the HypoGen and QPHAR model show interesting results. For the QPHAR model, the RMSE value of the Y-scrambling approach differs significantly from the RMSE value of the regression analysis. On the other hand, the scrambled RMSE value of the HypoGen model does not differ from the actual RMSE of the HypoGen model. This shows that the QPhAR model and its strong performance were not obtained by chance from a certain data split but

rather indicates successful training of the model. Furthermore, the QPhAR model is expected to generalize well to unseen data with similarly low error, which is also confirmed by the test set performance presented above. In contrast, the HypoGen model seems to be rather specific to the given data split due to the results and RMSE values obtained from the Y-scrambling analysis. These results are also in strong agreement with the negative R^2 value, the high RMSE value, and the high standard error of the HypoGen model on the test set.

Emopamil binding protein (EBP)

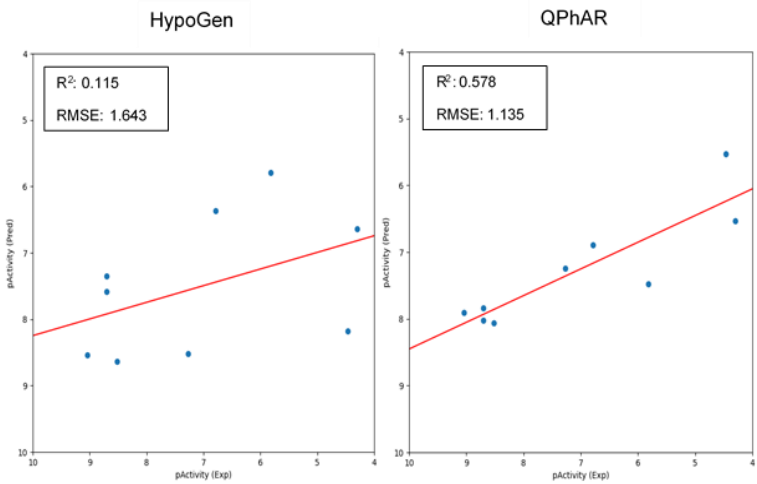

Figure 13: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPhAR (right) algorithm.

The test set performance of the HypoGen and the QPHAR models are 0.227 and 0.578 for the R2, 1.643 and 1.135 for the RMSE, and 1.119 and 0.683 for the standard error, respectively. The plots of the HypoGen (*Figure 13*, left) and QPhAR (*Figure 13*, right) pharmacophore models show that the QPhAR model produces a nice regression line for the predicted and experimental activity values, whereas the HypoGen model does not show any relationship between the activity values. Comparing the validation metrics, the QPhAR model shows better performance for the R^2, the RMSE, and the standard error, even though there is room for improvement in both cases.

Statistics

Table 7: Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.694
Wilcoxon-signed test	0.652
Average Tanimoto-coefficient	0.295

The results of the t-test and Wilcoxon-signed test show a p-value of 0.694 and 0.652, respectively. This confirms that the expected mean of the predictions of both the QPhAR and HypoGen algorithms are similar to each other. In combination with the activity scatterplots (*Figure 13*), however, it is recognizable that the QPhAR algorithm predicts much more accurately than the HypoGen algorithm for this target.

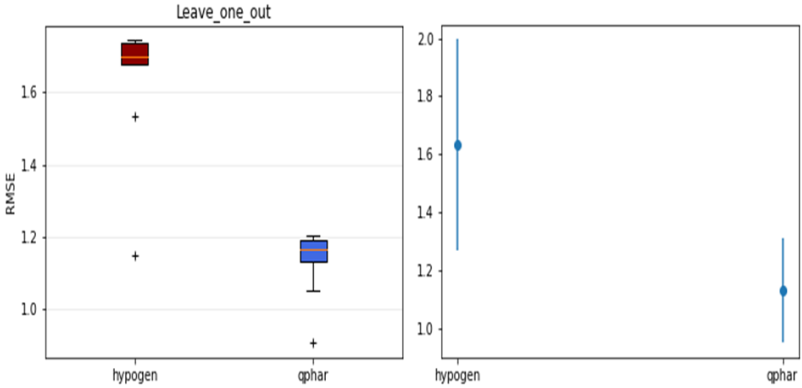

Figure 14: This figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the average RMSE values of the HypoGen and QPhAR models as a box plot. The right figure shows the error bars plot of the two models with the error bars at 95% confidence interval

The plots of the leave-one-out analysis (*Figure 14*) show that the average RMSE value from the QPhAR and the HypoGen models agree with the values obtained in the regression analysis. This indicates that the RMSE values are not dependent on a single molecule in the training set. The error bars within a 95% confidence interval slightly overlap. This suggests that the RMSE values are not significantly different from each other, which is consistent with the t-test and Wilcoxon-signed test analysis.

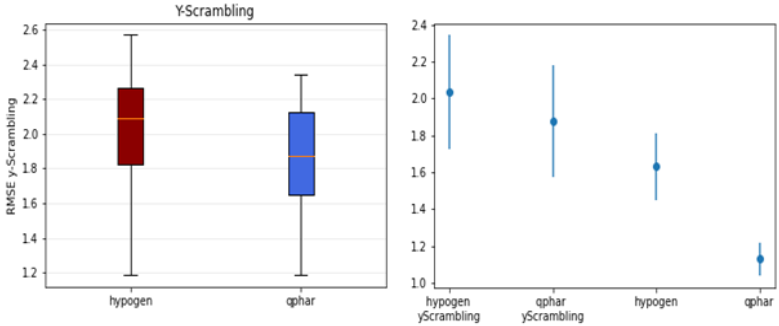

Figure 15 This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPhAR model (blue) on the left side and the error bar plot of the RMSE values on the right side with a 95% confidence interval.

The results of the Y-scrambling analysis (*Figure 15*) show that the RMSE value of the QPhAR model differs from the Y-scrambling RMSE value even with the error bars at a 95% confidence interval. However, the error bars of the HypoGen

model heavily overlap with the error bars of the Y-scrambling analysis. This indicates that the model for this target system is not robust and strongly influenced by the data split.

Considering all this information, it is clear the QPhAR model performs better for the Emopamil binding protein target system than the HypoGen pharmacophore model. Although the QPhAR model does not show a very good RMSE value and a relatively high standard error, it shows a good R^2 value and good results in the visual analysis of the predicted activity value. As mentioned before, a reason for the average performance of both algorithms could be that non-optimal conformations of the test set molecules were chosen in the data preparation step. This might result in a sub-optimal pharmacophore template for the model training and consecutively to a model with a mediocre to low quality. In summary, the performance and statistical evaluations indicate successful model building for the QPhAR algorithm, which clearly shows better results than the HypoGen algorithm.

Sigma-1 receptor

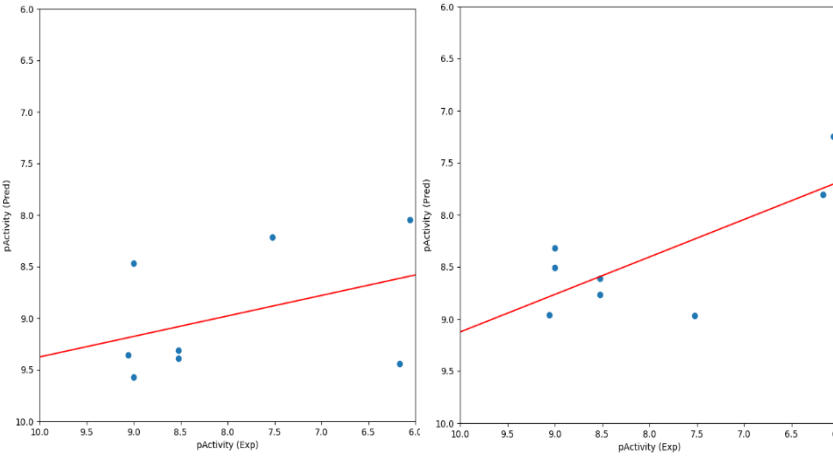

Figure 16: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPhAR (right) algorithm.

The QPhAR and HypoGen pharmacophore model performance metrics on the test set are 0.367 and -0.568 for the R^2 value, 0.935 and 1.468 for the RMSE, and 0.567 and 0.941 for the standard error, respectively. The scatterplot of the experimental and predicted activity values of the QPhAR model (*Figure 16*, right) shows a mediocre regression line and a recognizable relationship between the predictions. However, the plot of the HypoGen model (*Figure 16*, left) does not suggest a relationship between the predicted and experimental activity values.

Statistics

Table 8: Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.003
Wilcoxon-signed test	0.017
Average Tanimoto-coefficient	0.309

The two-sided t-test and Wilcoxon signed test result in a p-value of 0.003 and 0.017, respectively. This indicates that the mean of the predicted activity values is significantly different between the QPhAR and HypoGen models. This is consistent with the remarkably different performances on the test set between the pharmacophore models. The low Tanimoto coefficient indicates structurally different training and test set molecules and is a good example for the fact that

the QPhAR can create good models on data with structural perturbations.

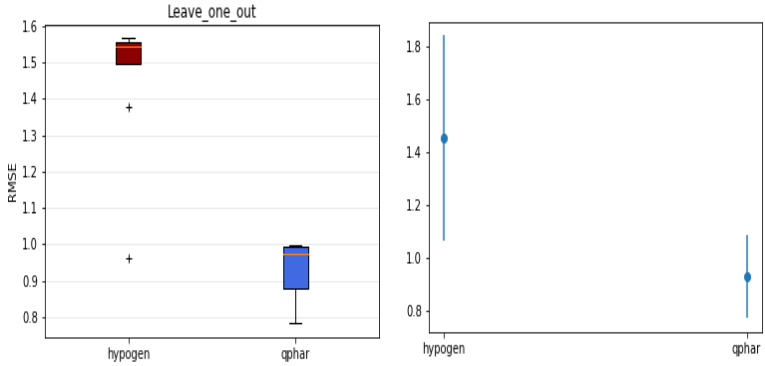

Figure 17: This figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the average RMSE values of the HypoGen and QPHAR models as a box plot. The right figure shows the error bars plot of the two models with the error bars at 95% confidence interval

The bar plot (*Figure 17*, left) and error bar plot (*Figure 18,* right) of the leave-one-out analysis shows that the RMSE values fit the values from the regression analysis, which indicates the received RMSE values are not dependent on a single molecule in the training set.

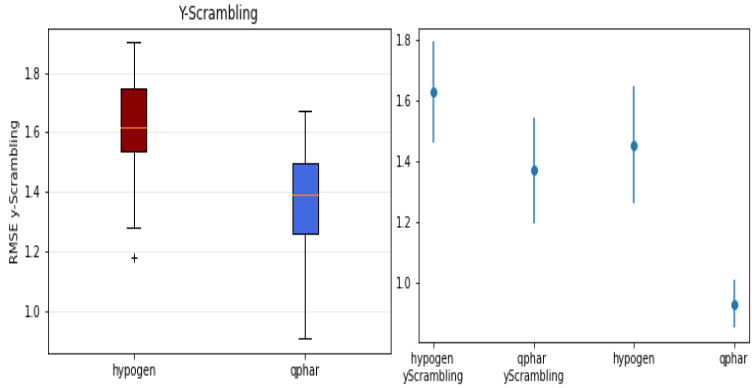

Figure 18: This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPHAR model (blue) on the left side and the error bar plot of the RMSE values on the right side with a 95% confidence interval.

The plots of the Y-scrambling analysis (*Figure 18)* show that the RMSE value of the Y-scrambling approach differs significantly from the RMSE value received from the regression analysis, and also the error bars with a 95% confidence interval do not overlap in case of the QPhAR model. This indicates that the QPhAR model is not influenced by the data split. On the other hand, the HypoGen model shows the exact opposite. The error bars with a 95% confidence interval overlap between the RMSE values of the quantitative model and the Y-scrambling analysis, which suggests that the HypoGen model is not robust.

Considering the statistical evaluations regarding the Sigma-1 receptor target system, it points out that the QPhAR model performs better on this target than the HypoGen model. It shows a good correlation in the activity scatter plot with relatively good performance metrics even though it has a slightly high standard error. The statistical evaluations confirm that the QPhAR model is robust and performs well on the test set.

CDK2

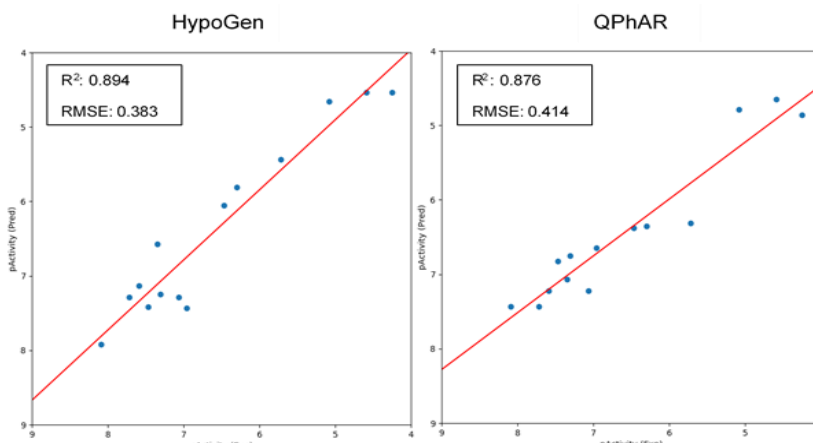

Figure 19: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPHAR (right) algorithm

The model validation metrics for the HypoGen and the QPhAR models are 0.894 and 0.876 for the R^2, 0.384 and 0.414 for the RMSE, and 0.199 and 0.212 for the standard

error, respectively. The validation metrics indicate that the QPhAR model performs on-par with the HypoGen model on the given test set. This statement can be confirmed when considering the activity plots of the HypoGen (Figure 19, left) and QPhAR (Figure 19, right) model. Both models result in a very nice regression line and show a recognizable relationship between the predicted and experimental activity values.

Statistics

Table 9: Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.923
Wilcoxon-signed test	0.925
Average Tanimoto-coefficient	0.551

The p-values of the two-sided t-test and Wilcoxon-signed test indicate that the mean of the predictions of both, the HypoGen and QPhAR model is similar. This supports the results from the activity plots and supports the statement that

both models perform on-par on the test-set for the CDK2 target system.

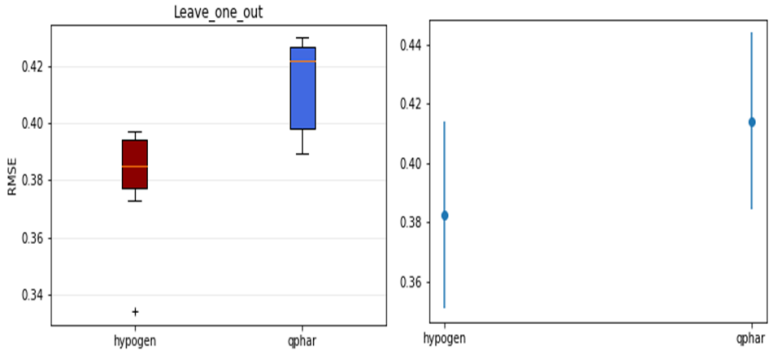

Figure 20: This figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the average RMSE values of the HypoGen and QPHAR models as a box plot. The right figure shows the error bars plot of the two models with the error bars at 95% confidence

The boxplot and error bar plot of the Leave-one-out analysis (*Figure 20*) show that both models' mean RMSE values correspond with the RMSE values received by the regression analysis. This indicates that the obtained RMSE values do not depend on a single compound in the training set.

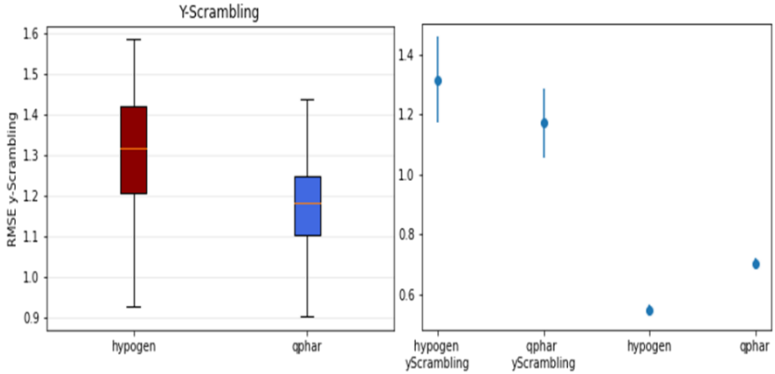

Figure 21: This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPHAR model (blue) on the left side and the error bar plot of the RMSE values on the right side with a 95% confidence interval.

The Y-scrambling analysis (*Figure 21*) of both models' models shows that the RMSE values of the Y-scrambling approach are significantly different from the RMSE values received by the regression analysis. This is an important indicator that the received model does not just randomly generate the received metrics and thus further increases the trust in our model.

The collected information for the model of the CDK2 target show that the QPhAR model performs on-par with the HypoGen model on the test set. The statistical analysis confirms this statement. The t-test and Wilcoxon-signed test

analysis show that the predicted activity values of both models come from the same distribution and thus ensure similar validation metrics of the regression analysis. The leave-one-out analysis results in RMSE values comparable to the analysis values and confirms that a specific compound in the training set does not bias the results. Furthermore, the Y-scrambling analysis confirms that the models are not only obtained by chance and, therefore, increase the confidence in the results.

K-opioid receptor

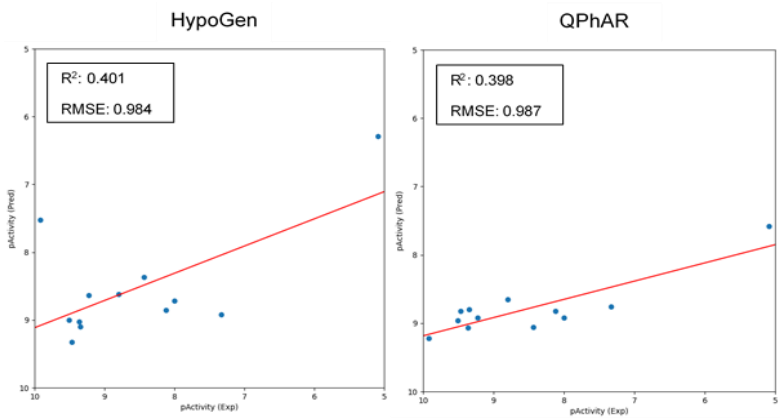

Figure 22: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPhAR (right) algorithm.

The model validation metrics for the HypoGen and the QPhAR models are 0.401 and 0.398 for the R2, 0.984 and

0.987 for the RMSE, and 0.664 and 0.604 for the standard error, respectively. The activity scatterplots of the HypoGen model (*Figure 22*,left) and the QPhAR model (*Figure 22*, right) show both comparable regression lines and a recognizable relationship between the predicted and experimental activity values. However, the predicted activity values have a narrower distribution for the QPhAR model compared to the HypoGen model.

Statistics

Table 10: Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.386
Wilcoxon-signed test	0.470
Average Tanimoto-coefficient	0.468

The t-test and Wilcoxon-signed test results show a p-value of 0.386 and 0.470, respectively. This indicates that the predictions of both, the QPhAR and HypoGen model, have

the same expected mean. This is confirmed by the received model validation metrics that are very similar between both models. Furthermore, the plots of both models show an equal distribution of the predicted activity values.

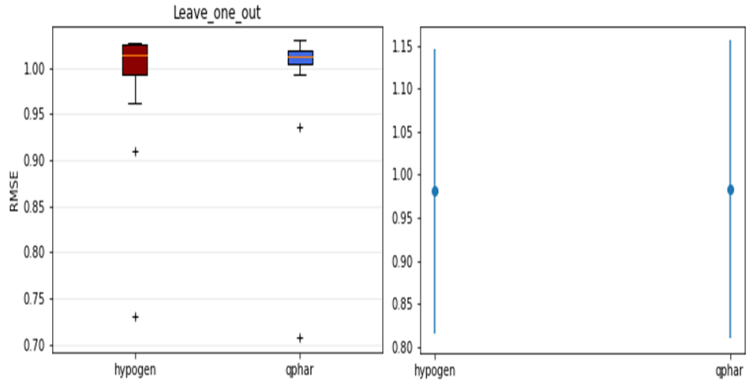

Figure 23: This figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the average RMSE values of the HypoGen and QPhAR models as a box plot. The right figure shows the error bars plot of the two models with the error bars at 95% confidence interval

The average RMSE values of the leave-one-out analysis reflect the values of the regression analysis for both models, which means that the models are not biased towards a single compound of the training data set. The error bar plot (*Figure 23,* right) shows that the RMSE values within a 95% confidence interval are approximately identical between the QPhAR and HypoGen models.

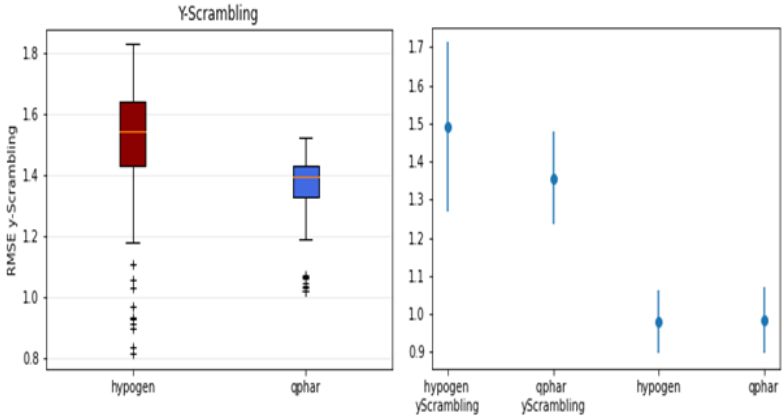

Figure 24: This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPhAR model (blue) on the left side and the error bar plot of the RMSE values on the right side.

The results of the Y-scrambling analysis (*Figure 24*) show that for both the HypoGen and QPhAR models, the y-scrambled RMSE value differs significantly. This fact increases the confidence in both models. Considering the activity plots (*Figure 22*) and evaluation metrics, the QPhAR model performs on-par with the HypoGen model for this target system. The R2 value, the RMSE value, and standard error are very similar between the pharmacophore models. The visual analysis confirms this statement with nice regression lines for both models. Furthermore, the statistical evaluations justify the equivalent performance of both

pharmacophore models on the test set of the target system. The T-test and Wilcoxon-sign test statistics show that the predicted activity values of both models are similar and the predictions of the QPhAR algorithms are in the 95% confidence interval of the HypoGen algorithm. In addition, the leave-one-out analysis confirms that both models are not based on a specific compound from the training set. Furthermore, The Y-scrambling analysis confirms that both models do not just perform well on the test set by chance.

Endothelin A

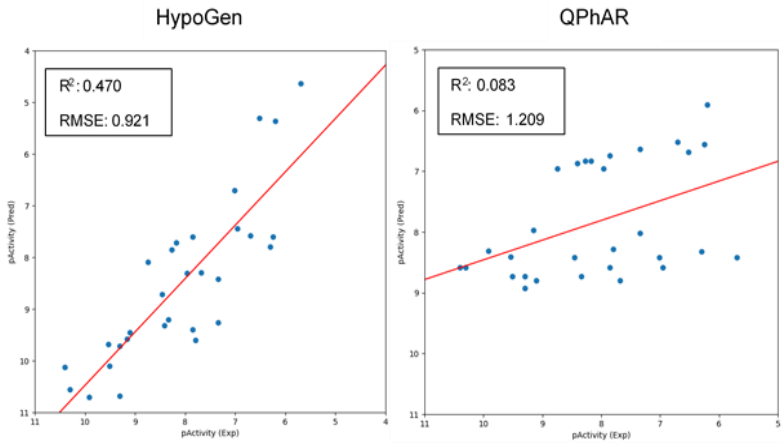

Figure 25: This Figure shows the activity scatter plots of the predicted against the true activities of the quantitative pharmacophore model on the test set of the HypoGen (left) and the QPhAR (right) algorithm.

The scatter plots of the predicted activity values show that the values of the HypoGen (*Figure 25*, left) model result in a

nice regression line and a relationship between the predicted and experimental activity values. In comparison, the predicted activity values of the QPhAR model (*Figure 25*, right) do not show a relationship between the predicted and experimental activity values. The model metrics are R^2 values of 0.470 and 0.083, an RMSE value of 0.921 and 1.209, and a standard error of 0.492 and 0.650 for the HypoGen and QPhAR pharmacophore model, respectively. The scatter plots of the predicted activity values show that the QPhAR model performs worse on the test set than the HypoGen model for this target system.

Statistics

Table 11 Summary of t-test and Wilcoxon's sign test statistics. The given statistic values are the t-statistic for the t-test and the minimum sum of ranks above or below zero for Wilcoxon's sign test, respectively.

	p-value
Two-sided t-test	0.021
Wilcoxon-signed test	0.018
Average Tanimoto-coefficient	0.439

The t-test and Wilcoxon-signed test p-values of 0.021 and 0.018, respectively, confirm that the null hypothesis is rejected. Hence, the predicted activity values of the QPhAR pharmacophore model differ from the ones received by the HypoGen pharmacophore model.

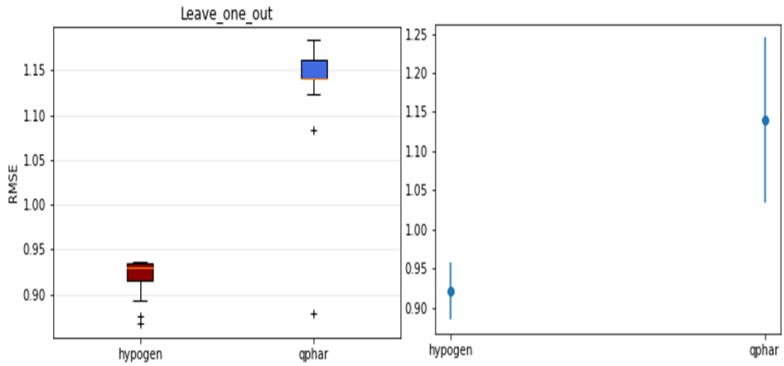

Figure 26: This figure shows the RMSE plots of the leave-one-out analysis. The left figure shows the boxplot of the average RMSE values of the HypoGen and QPhAR models. The right figure shows the error bars plot of the two models with the error bars at 95% confidence interval

The mean RMSE value of the leave-one-out method (*Figure 26*) is similar to RMSE of the model's performance on the test set. The equal RMSE values indicate that both models do not depend on a certain molecule in the training set. Furthermore, the error bars of the calculated RMSE values do not overlap between the HypoGen and QPhAR model.

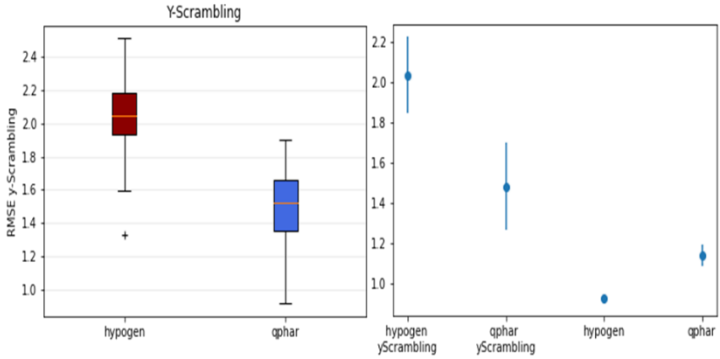

Figure 27: This figure shows the box plot of the Y-scrambling analysis of the RMSE Values of the HypoGen model (red) and the QPhAR model (blue) on the left side and the error bar plot of the RMSE values on the right side.

The RMSE value of the Y-scrambling approach differs from the regression analysis RMSE value for both models (*Figure 27*, right). If we collect all information for validation, we can see that the QPhAR model for the Endothelin A target performs worse on the test set compared to the HypoGen model. The QPhAR model shows a low R2 value, with a high RMSE value and standard error. The visual analysis also shows no relationship between the datapoints for the QPhAR model. One possible explanation for the bad performance is that a bad template was selected. As already mentioned, selecting the template for the QPhAR algorithm is one of the most important steps. However, the control of the training set

and the statistical analysis of the QPhAR model exclude a training set that is not suitable for QSAR and a bias of the model towards a certain training set molecule as the reason for the bad performance on the Endothelin A target system.

Final overview

When taking a closer look at the results of the target systems CyclinA/CDK2 binding domain, kappa-opioid receptor, P450 19 aromatase, and Angiotensin II receptor subtype I, it is possible to say with high confidence that the QPhAR algorithm performs on-par with the HypoGen algorithm. This statement is confirmed by the activity scatter plots with very similar distribution and linear correlation of the predicted activity values of the HypoGen and QPhAR pharmacophore model. Furthermore, confidence in these models is increased by the statistical evaluations of the models. The test statistics and p-values of the T-test and Wilcoxon's signed-rank test show that the predicted activity values of both, the HypoGen and QPhAR algorithm, are modeled with the same distribution. That means both models retrieve predictions from a distribution that cannot be distinguished from each other. Hence, this indicates that the algorithms give comparable predicted activity values on the test set. The

leave-one-out method shows that the calculated RMSE values of the models are valid and that the quantitative models do not depend on a certain compound within the training set. Additionally, the Y-scrambling analysis shows that the models for these target systems are not obtained by pure chance and the model has learned sound relationships between the input data and the predicted activity values. This is indicated by the significant difference between the RMSE value of the scrambled activity values from the RMSE values of the regression analysis.

Considering the results of the hERG potassium channel, EBP, ERG2, and Sigma-1 receptor target systems, the QPhAR algorithm performs better than the HypoGen algorithm based on the validation metrics and the visual validation of the activity plots. For all of these target systems, the predicted activity values of the HypoGen model show extremely low R^2 values and high RMSE values. In contrast, the QPhAR models show good validation metrics for the EBP, ERG2, and sigma-1 receptors.

The Aurora kinase B is an interesting target system in this study. The validation metrics of the two models show that the HypoGen model gives a better R^2 metric than the QPhAR model, and the RMSE value of the HypoGen model is

conspicuously low with a value of 0.279. The QPhAR model, on the other side, shows a slightly higher RMSE value of 0.458. The outstanding RMSE value paired with the excellent R^2 value of 0.838 could indicate possible overfitting of the quantitative pharmacophore model on the test set. The same could be true for the CDC25B target which shows a particularly high R^2 of 0.927 and a low RMSE value of 0.241 compared to an R^2 of 0.580 and an RMSE of 0.440 for the QPhAR model. The Aurora kinase B is a representative example for these targets in the results section above.

Taking a closer look at the Endothelin A, HIV1 integrase, Tubulin/colchicine binding domain, and polo-like kinase target systems, the regression analysis and visual analysis of the activity plots show that the HypoGen algorithm performs better than QPhAR. There are multiple possible reasons why the QPhAR algorithm performs worse on these target systems. One reason was eliminated since the training datasets were checked for the requirements for QSAR at the beginning. The Tanimoto coefficient analysis could explain a second possible reason. For each of these targets, the average Tanimoto coefficient of the training set molecules on the test set molecules ranges from 0.2 to 0.44. A low Tanimoto coefficient shows that the molecular structures of

the training set are highly dissimilar to the molecules of the test set. This might result in differing pharmacophores that cannot be aligned. Therefore, the QPhAR model cannot generalize from the training to the test set. However, the evaluations of the Tanimoto coefficients of the target systems show that the QPhAR algorithm can perform very well on structurally different input data and that the low performance is not a result of inappropriate input.

Conclusion

In conclusion, results obtained in this thesis show that the QPhAR algorithm performs better or at least on par with the HypoGen algorithm on the evaluated target systems. The models were validated on the test set, and the best models for each target were confirmed by regression analysis and multiple statistical evaluations to increase the confidence in the models' results. Two targets are interesting for further investigations regarding the HypoGen model because they show outstanding R^2 metrics and RMSE values, which could indicate an overfitting of the HypoGen model. The QPhAR algorithm performs moderately on these targets based on the R^2 values but shows better RMSE values than the HypoGen models. The QPhAR model shows inferior performance to

the HypoGen model on four targets. In general, this thesis was show that the novel QPhAR algorithm is a reasonable alternative to the established HypoGen algorithm for quantitative pharmacophore modeling.

Regarding future perspectives, it would be interesting to compare the QPhAR algorithm with other quantitative pharmacophore algorithms like PHASE by Schroedinger[23,71]. Unfortunately, hardly any data is available for quantitative models generated by PHASE. An attempt was made to find appropriate literature, but unfortunately no publications with QSAR studies that fulfill all necessary requirements could be found.

References

1. Aging population and the effects on health care. https://www.pharmacytimes.com/view/the-aging-population-the-increasing-effects-on-health-care.

2. RKI Ebola outbreak 2014-2015. https://www.rki.de/EN/Content/infections/epidemiology/outbreaks/Ebola_virus_disease/EVD_situation_summary.html;jsessionid=D09083F93130830E2B7EF6A31523BC5B.internet112.

3. Parvathaneni V, Kulkarni NS, Muth A, Gupta V. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug Discov Today*. 2019;24(10):2076-2085. doi:10.1016/j.drudis.2019.06.014

4. Lima L, Barreiro E. Bioisosterism: A Useful Strategy for Molecular Modification and Drug Design. *Curr Med Chem*. 2012;12(1):23-49. doi:10.2174/0929867053363540

5. Bioisosteres. https://www.cambridgemedchemconsulting.com/resources/bioisosteres/.

6. Chaguza C. Bacterial survival: evolve and adapt or perish. *Nat Rev Microbiol*. 2020;18(1):5. doi:10.1038/s41579-019-0303-5

7. Centers for Disease Control and. Antibiotic resistance threats in the United States. *US Dep Heal Hum Serv*. 2019:1-113. https://www.cdc.gov/drugresistance/biggest_threats.html.

8. Macalino SJY, Billones JB, Organo VG, Carrillo MCO. In silico strategies in tuberculosis drug discovery. *Molecules*. 2020;25(3). doi:10.3390/molecules25030665

9. Wermuth CG, Ganellin CR, Lindberg P, Mitscher L a. Glossary for Chemists of Terms Used in Medicinal Chemistry. *Pure Appl Chem*. 1998;70(5):1129-1143.

10. Bacelar M. Monitoring bias and fairness in machine learning models: A review. *Sci Prepr*. 2021;(May):0-2. doi:10.14293/s2199-1006.1.sor-.pp59wrh.v1

11. Gurung AB, Ali MA, Lee J, Farah MA, Al-Anazi KM. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. *Biomed Res Int*. 2021;2021. doi:10.1155/2021/8853056

12. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. *Int J Mol Sci*. 2019;20(11). doi:10.3390/ijms20112783

13. Badalà F, Nouri-mahdavi K, Raoof DA. Recent Advances in Ligand-Based Drug Design: Relevance and Utility. *Computer (Long Beach Calif)*. 2008;144(5):724-732.

14. Huang SY, Zou X. Advances and challenges in Protein-ligand docking. *Int J Mol Sci*. 2010;11(8):3016-3034. doi:10.3390/ijms11083016

15. Lavecchia A, Giovanni C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr Med Chem*. 2013;20(23):2839-2860. doi:10.2174/09298673113209990001

16. Lazim R, Suh D, Choi S. Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. *Int J Mol Sci*. 2020;21(17):1-20. doi:10.3390/ijms21176339

17. Romano T. Kroemer. Structure-Based Drug Design: Docking and Scoring. *Curr Protein Pept Sci*. 2007;8(4):312-328. doi:10.2174/138920307781369382

18. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA. PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des*. 2006;20(10-11):647-671. doi:10.1007/s10822-006-9087-6

19. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model*. 2005;45(1):160-169. doi:10.1021/ci049885e

20. Catalyst. Pharmacophore Identification Using Catalyst A pharmacophore is a representation of generalized. 2001.

21. Richmond NJ, Abrams CA, Wolohan PRN, Abrahamian E, Willett P, Clark RD. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J Comput Aided Mol Des*. 2006;20(9):567-587. doi:10.1007/s10822-006-9082-y

22. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature 194:178–180. https:// doi. org/ 10. 1038/ 19417 8b0.

23. Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today*. 2016;21(8):1291-1302. doi:10.1016/j.drudis.2016.06.013

24. HOMO and LUMO. https://en.wikipedia.org/wiki/HOMO_and_LUMO.

25. Gao Y, Imran M, Farahani MR, Muhammad H, Siddiqui A. Some connectivity indices and Zagreb Index of honeycomb graphs [327]. 2018;9(May). doi:10.13040/IJPSR.0975-9048.9(5).1000-04

26. Devinyak O, Havrylyuk D, Lesyk R. 3D-MoRSE descriptors explained. *J Mol Graph Model*. 2014;54:194-203. doi:10.1016/j.jmgm.2014.10.006

27. Le Fèvre RJW. Molecular Refractivity and Polarizability. In: Gold VBT-A in POC, ed. Vol 3. Academic Press; 1965:1-90. doi:https://doi.org/10.1016/S0065-3160(08)60298-1

28. Patel HM, Noolvi MN, Sharma P, et al. Quantitative structure-activity relationship (QSAR) studies as strategic approach in drug discovery. *Med Chem Res*. 2014;23(12):4991-5007. doi:10.1007/s00044-014-1072-3

29. Autocorrleation. https://en.wikipedia.org/wiki/Autocorrelation.

30. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: Where have you been? Where are you going to? *J Med Chem*. 2014;57(12):4977-5010. doi:10.1021/jm4004285

31. Yu C, Yao W. Robust linear regression: A review and comparison. *Commun Stat Simul Comput*. 2017;46(8):6261-6282. doi:10.1080/03610918.2016.1202271

32. Powers DA, Xie Y. Review of Linear Regression Models. *Stat Methods Categ Data Anal*. 2000:15-39. doi:10.1016/b978-012563736-7/50002-1

33. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

34. Jin Z, Shang J, Zhu Q, Ling C, Xie W, Qiang B. RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2020;12343 LNCS:503-515. doi:10.1007/978-3-030-62008-0_35

35. Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An Introductory Review of Deep Learning for Prediction Models With Big Data. *Front Artif Intell*. 2020;3(February):1-23. doi:10.3389/frai.2020.00004

36. T.Seidel. Seidel T (2021) Chemical data processing toolkit, GitHub repository [Internet]. [cited 2021 Mar 19] Available from: https:// github. com/ aglan ger/ CDPKit.

37. BioVia Discovery Studio. https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/biovia-discovery-studio/.

38. Li H, Sutter J, Hoffman R. Hypogen: An Automated System for Generating 3D Predictive Pharmacophore Models. In: Pharmacophore Perception, Development and Use in Drug Design; Guner, O, Ed; International University Line: La Jolla, CA,. 2000.

39. Kohlbacher SM, Langer T, Seidel T. QPHAR: quantitative pharmacophore activity relationship: method and validation. *J Cheminform*. 2021;13(1):1-14. doi:10.1186/s13321-021-00537-9

40. Aslam M. Introducing Grubbs's test for detecting outliers under neutrosophic statistics – An application to medical data. *J*

King Saud Univ - Sci. 2020;32(6):2696-2700. doi:10.1016/j.jksus.2020.06.003

41. Potochnik A, Colombo M, Wright C. Statistics and Probability. *Recipes Sci*. 2018;(Table 2):167-206. doi:10.4324/9781315686875-6

42. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bull*. 1945;1(6):80-83. doi:10.2307/3001968

43. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

44. Webb G, Sammut C, Perlich C, et al. Leave-One-Out Cross-Validation. In: ; 2010. doi:10.1007/978-0-387-30164-8_469

45. Y-scrambling. https://www.geeksforgeeks.org/y-scrambling-for-model-validation/.

46. Kumar A. Chemical Similarity Methods - A Tutorial Review. *Chem Educ*. 2011;16(August):46-50.

47. Berthold MR, Cebron N, Dill F, et al. KNIME - the Konstanz information miner. In: Preisach CBerhardt HSchmidt-Theime Let al. ., eds. Data Analysis,. *Mach Learn Appl Ger Springer*. 2009;11(1):26. http://portal.acm.org/citation.cfm?doid=1656274.1656280.

48. Wang HY, Li LL, Cao ZX, Luo SD, Wei YQ, Yang SY. A specific pharmacophore model of Aurora B kinase inhibitors and virtual screening studies based on it. *Chem Biol Drug Des*. 2009;73(1):115-126. doi:10.1111/j.1747-0285.2008.00751.x

49. Wang HY, Cao ZX, Li LL, et al. Pharmacophore modeling and virtual screening for designing potential PLK1

inhibitors. *Bioorganic Med Chem Lett*. 2008;18(18):4972-4977. doi:10.1016/j.bmcl.2008.08.033

50. Laggner C, Schieferer C, Fiechtner B, et al. Discovery of high-affinity ligands of σ1 receptor, ERG2, and emopamil binding protein by pharmacophore modeling and virtual screening. *J Med Chem*. 2005;48(15):4754-4764. doi:10.1021/jm049073+

51. Niu MM, Qin JY, Tian CP, et al. Tubulin inhibitors: Pharmacophore modeling, virtual screening and molecular docking. *Acta Pharmacol Sin*. 2014;35(7):967-979. doi:10.1038/aps.2014.34

52. Aparoy P, Kumar Reddy K, Kalangi SK, Chandramohan Reddy T, Reddanna P. Pharmacophore modeling and virtual screening for designing potential 5-Lipoxygenase inhibitors. *Bioorganic Med Chem Lett*. 2010;20(3):1013-1018. doi:10.1016/j.bmcl.2009.12.047

53. Barreca ML, Ferro S, Rao A, Luca L De, Zappala M. Pharmacophore-Based Design of HIV-1 Integrase Strand-Transfer Inhibitors. 2005:17-20.

54. Ece A, Sevin F. The discovery of potential cyclin A/CDK2 inhibitors: A combination of 3D QSAR pharmacophore modeling, virtual screening, and molecular docking studies. *Med Chem Res*. 2013;22(12):5832-5843. doi:10.1007/s00044-013-0571-y

55. Funk OF, Kettmann V, Drimal J, Langer T. Chemical function based pharmacophore generation of endothelin-A selective receptor antagonists. *J Med Chem*. 2004;47(11):2750-2760. doi:10.1021/jm031041j

56. Garg D, Gandhi T, Gopi Mohan C. Exploring QSTR and toxicophore of hERG K+ channel blockers using GFA and

HypoGen techniques. *J Mol Graph Model*. 2008;26(6):966-976. doi:10.1016/j.jmgm.2007.08.002

57. Paliwal S, Pal M, Yadav D, Singh S, Yadav R. Ligand-based drug design studies using predictive pharmacophore model generation on 4H-1,2,4-triazoles as AT 1 receptor antagonists. *Med Chem Res*. 2012;21(9):2307-2315. doi:10.1007/s00044-011-9756-4

58. Krovat EM, Langer T. Non-peptide angiotensin II receptor antagonists: Chemical feature based pharmacophore identification. *J Med Chem*. 2003;46(5):716-726. doi:10.1021/jm021032v

59. Kumar R, Parameswaran S, Bavi R, et al. Investigation of novel chemical scaffolds targeting prolyl oligopeptidase for neurological therapeutics. *J Mol Graph Model*. 2019;88:92-103. doi:10.1016/j.jmgm.2018.12.006

60. Ma Y, Li HL, Chen XB, et al. 3D QSAR Pharmacophore Based Virtual Screening for Identification of Potential Inhibitors for CDC25B. *Comput Biol Chem*. 2018;73:1-12. doi:10.1016/j.compbiolchem.2018.01.005

61. Schuster D, Laggner C, Steindl TM, Palusczak A, Hartmann RW, Langer T. Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors. *J Chem Inf Model*. 2006;46(3):1301-1311. doi:10.1021/ci050237k

62. Singh N, Nolan TL, McCurdy CR. Chemical function-based pharmacophore development for novel, selective kappa opioid receptor agonists. *J Mol Graph Model*. 2008;27(2):131-139. doi:10.1016/j.jmgm.2008.03.007

63. PerkinElmerInformatics. https://perkinelmerinformatics.com/.

64. Poli G, Seidel T, Langer T. Conformational sampling of small molecules with iCon: Performance assessment in comparison with OMEGA. *Front Chem*. 2018;6(June). doi:10.3389/fchem.2018.00229

65. Jolliffe IT, Cadima J, Cadima J. Principal component analysis : a review and recent developments Subject Areas. *PhilTransRSocA*. 2016;374(2065):1-16.

66. Fearn T. Ridge Regression. *NIR news*. 2013;24(3):18-19. doi:10.1255/nirn.1365

67. Kwon S, Han S, Lee S. A small review and further studies on the LASSO. *J Korean Data Inf Sci Soc*. 2013;24(5):1077-1088. doi:10.7465/jkdi.2013.24.5.1077

68. Boulesteix AL, Strimmer K. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*. 2007;8(1):32-44. doi:10.1093/bib/bbl016

69. Li H, Phung D. Journal of Machine Learning Research: Preface. *J Mach Learn Res*. 2014;39(2014):i-ii.

70. Matplotlib packages. https://matplotlib.org/.

71. Leach AR, Gillet VJ. *An Introduction to Chemoinformatics*.; 2007. doi:10.1007/978-1-4020-6291-9

Appendix

The molecule files containing training and test set molecules along with the corresponding conformations and the obtained validation metrics and statistics for each target can be found at: https://github.com/Matthiasschmid1995/MasterThesis-Appendix.git