

# Development, validation and application of in-silico methods to predict the macromolecular targets of small organic compounds

---

Neann Sarah Mathai

Thesis for the degree of Philosophiae Doctor (PhD)  
University of Bergen, Norway  
2021

UNIVERSITY OF BERGEN



# Development, validation and application of in-silico methods to predict the macromolecular targets of small organic compounds

Neann Sarah Mathai



Thesis for the degree of Philosophiae Doctor (PhD)  
at the University of Bergen

Date of defense: 10.12.2021

© Copyright Neann Sarah Mathai

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2021

Title: Development, validation and application of in-silico methods to predict the macromolecular targets of small organic compounds

Name: Neann Sarah Mathai

Print: Skipnes Kommunikasjon / University of Bergen

# Scientific environment

The work presented in the thesis was carried out at the Computational Biology Unit (CBU) hosted at the Department of Informatics and funded by Department of Chemistry, at the University of Bergen. I am also affiliated with the National Research School in Bioinformatics, Bio-statistics and Systems Biology (NORBIS) and the Norwegian Graduate School in Biocatalysis (BioCat).

This thesis was supervised by Associate Professor Johannes Kirchmair employed at the Department of Chemistry and CBU at the University of Bergen, Department of Pharmaceutical Sciences at the University of Vienna, and Centre for Bioinformatics at the University of Hamburg.







To my mother,  
my first chemistry teacher  
who taught me so much more.



# Acknowledgements

Thank you to Assoc. Professor Johannes Kirchmair for your supervision, time, guidance, and unceasing encouragement during this PhD. I am truly grateful to have been given the opportunity to be able to learn from you.

I am grateful to my co-supervisor Professor Nathalie Reuter for your guidance and support - thank you. I would like to thank the CBU, the Kjemisk Institutt, and Assoc. Professor Johannes Kirchmair for making this PhD possible. A big thank you to the HPC teams at the CBU and Sigma2 for your support.

To my teachers and mentors who stood by me during this very long and winding PhD journey and beyond, “thank you” is not enough. I will pay it forward.

To all of ACMD/Comp3D, I am grateful to have had you as my group. I am very thankful I met you in Hamburg and continued as a part of our remote lab. Thank you for the scientific collaboration and insights, the picnics and lab lunches, life talks and so much more. I am delighted for our friendship. Thank you to the CBU, my group in Bergen. You gave me a space to ask my questions and brainstorm, and to conduct my research. Most of all, I will forever be grateful for the support, guidance, and laughter over lunch. Thank you for being a big part of making Bergen a home for me. To my office-mates, the fun, perfecting the art of microwave popcorn, lunches on the fjord and so much laughter was glorious. Big thank you to friends who organized the pandemic-safe farewell.

To my village, in Bergen and beyond, thank you for all the dancing, singing, road trips to hunt waterfalls, hikes to spot trolls, barbecues and dinners, exercise classes, game nights, and more. Life is immeasurably richer because of you. Your prayers, love, encouragement, phone calls, visits, and late night “walk and talks” up the mountains and around fjords have been the biggest blessing and helped me run the race. You have kept me grounded and resting in Him. Thank you to my partner for making this past year exceptionally special. You are a fount of encouragement, support, joy and love. To my family, you made me, cloaked me in love and taught me to pray, to face my fears, and to value kindness, curiosity, learning, exploration and education. Thank you.



# Abstract

Computational methods to predict the macromolecular targets of small organic drugs and drug-like compounds play a key role in early drug discovery and drug repurposing efforts. These methods are developed by building predictive models that aim to learn the relationships between compounds and their targets in order to predict the bioactivity of the compounds.

In this thesis, we analyzed the strategies used to validate target prediction approaches and how current strategies leave crucial questions about performance unanswered. Namely, how does an approach perform on a compound of interest, with its structural specificities, as opposed to the average query compound in the test data? We constructed and present new guidelines on validation strategies to address these short-comings. We then present the development and validation of two ligand-based target prediction approaches: a similarity-based approach and a binary relevance random forest (machine learning) based approach, which have a wide coverage of the target space. Importantly, we applied a new validation protocol to benchmark the performance of these approaches. The approaches were tested under three scenarios: a standard testing scenario with external data, a standard time-split scenario, and a close-to-real-world test scenario. We disaggregated the performance based on the distance of the testing data to the reference knowledge base, giving a more nuanced view of the performance of the approaches. We showed that, surprisingly, the similarity-based approach generally performed better than the machine learning based approach under all testing scenarios, while also having a target coverage which was twice as large.

After validating two target prediction approaches, we present our work on a large-scale application of computational target prediction to curate optimized compound libraries. While screening large collections of compounds against biological targets is key to identifying new bioactivities, it is resource intensive and challenging. Small to medium-sized libraries, that have been optimized to have a higher chance of producing a true hit on an arbitrary target of interest are therefore valuable. We curated libraries of readily purchasable compounds by: i. utilizing property filters to ensure that the compounds have

key physicochemical properties and are not overly reactive, ii. applying a similarity-based target prediction method, with a wide target scope, to predict the bioactivities of compounds, and iii. employing a genetic algorithm to select compounds for the library to maximize the biological diversity in the predicted bioactivities. These enriched small to medium-sized compound libraries provide valuable tool compounds to support early drug development and target identification efforts, and have been made available to the community.

The distinctive contributions of this thesis include the development and benchmarking of two ligand-based target prediction approaches under novel validation scenarios, and the application of target prediction to enrich screening libraries with biologically diverse bioactive compounds. We hope that the insights presented in this thesis will help push data driven drug discovery forward.

# Scientific contributions

## Publications presented in this thesis:

- P1:** Mathai, N.; Chen, Y.; Kirchmair, J. Validation strategies for target prediction methods, *Briefings in Bioinformatics*, **2020**, 21(3), pp. 791–802. doi: 10.1093/bib/bbz026.
- P2:** Mathai, N.; Kirchmair, J. Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope, *International Journal of Molecular Sciences*, **2020**, 21(10), p. 3585. doi: 10.3390/ijms21103585.
- P3:** Mathai, N.; Stork, C.; Kirchmair, J. BonMOLière: Small-sized libraries of readily purchasable compounds, optimized to produce genuine hits in biological screens across the protein space, *International Journal of Molecular Sciences*, **2021**, 22(15), p. 7773. doi: 10.3390/ijms22157773.

*Reprints were made with permission from MDPI and Oxford University Press. All articles towards this thesis are Open Access and distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).*

## Additional publications not presented in this thesis:

- AP1:** Wald, J.; Pasin, M.; Richter, M.; Walther, C.; Mathai, N.; Kirchmair, J.; Makarov, V.A.; Goessweiner-Mohr, N; Marlovits., TC; Zanella, I.; Real-Hohn, A. Cryo-EM structure of pleconaril-resistant rhinovirus-B5 complexed to the antiviral OBR-5-340 reveals unexpected binding site, *Proceedings of the National Academy of Sciences of the United States of America*, **2019**, 116(38), pp. 19109–19115. doi: 10.1073/pnas.1904732116.



- AP2:** Chen, Y.; Mathai, N.; Kirchmair, J. Scope of 3D shape-based approaches in predicting the macromolecular targets of structurally complex small molecules including natural products and macrocyclic ligands, *Journal of Chemical Information and Modeling*, **2020** 60(6), pp. 2858–2875. doi: 10.1021/acs.jcim.0c00161.
- AP3:** Stork, C.; Hirte, S.; Mathai, N.; Kirchmair, J. Computational prediction of frequent hitters in target-based and cell-based assays, *Artificial Intelligence in the Life Sciences*, **2021** p. 100007. doi: 10.1016/j.ailsci.2021.100007.
- AP4:** Wilm, A.; Garcia de Lomana, M.; Stork, C.; Mathai, N.; Hirte, S.; Norinder, U.; Kühnl, J.; and Kirchmair, J. Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **2021**, 14(8), p. 790. doi: 10.3390/ph14080790.

## Conference presentations:

- CP1:** Validation strategies for target prediction methods: Scope and limitations, *Euro-QSAR 2018*, Sept. 16 - 20, 2018, Thessaloniki, Greece. (poster)
- CP2:** Maximizing the coverage of small-molecule target prediction methods, *5th annual NORBIS conference*, Sept. 30 - Oct. 2, 2019, Drøbak, Norway. (20 minute talk)
- CP3:** Maximizing the coverage of small-molecule target prediction methods using a combination of approaches based on molecular similarity and machine learning, *Bioinformatics in Bergen*, Oct. 16 - 17, 2019, Os, Norway. (3 minute flash talk)
- CP4:** Performance and scope of similarity-based and machine learning approaches for predicting the macromolecular targets of small molecules, *American Chemical Society Fall 2020 Conference*, Aug. 17-20, 2020, Virtual. (20 minute talk)
- CP5:** Performance and scope of a similarity-based and a random forest-based machine learning approach for small-molecule target prediction, *3rd RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry*, Sept. 28-29, 2020, Virtual. (2 minute flash talk and poster)

# Contents

Scientific environment	i
Acknowledgements	v
Abstract	vii
Scientific contributions	ix
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Machine learning . . . . .	3
1.3 The importance of target prediction . . . . .	4
1.4 Data for ligand-based target prediction . . . . .	6
1.5 Ligand-based target prediction methods . . . . .	9
1.6 Applying target prediction to generate screening libraries . . . . .	13
<b>2 Aims of this study</b>	<b>15</b>
<b>3 Methods</b>	<b>17</b>
3.1 Data sources for target prediction . . . . .	17
3.2 Chemical data processing and molecular descriptor calculation . . . . .	18

---

3.3	Calculation of molecular similarity . . . . .	19
3.4	Model development for target prediction . . . . .	19
3.4.1	Training the random forest classifiers for the ML approach . . . . .	21
3.5	Validation of the target prediction approaches . . . . .	22
3.6	Data source for compound library curation . . . . .	23
3.7	Rules-based filters for compound library curation . . . . .	24
3.8	Genetic algorithm for compound library curation . . . . .	25
<b>4</b>	<b>Strategies to validate target prediction methods</b>	<b>27</b>
<b>5</b>	<b>Development and validation of large-scale target prediction methods</b>	<b>41</b>
<b>6</b>	<b>Using target prediction to curate compound sets for screening libraries</b>	<b>59</b>
<b>7</b>	<b>Concluding discussions and future prospects</b>	<b>81</b>
7.1	Evaluating target prediction methods . . . . .	82
7.2	Development and validation of target prediction methods . . . . .	83
7.3	Applying target prediction to curate screening libraries . . . . .	85
7.4	Concluding remarks . . . . .	88
	<b>Bibliography</b>	<b>102</b>
	<b>Supporting information for P2</b>	<b>105</b>

# Chapter 1

## Introduction

### 1.1 Background

The drug discovery and development process is long and arduous, with high attrition rates at each step (Figure: 1.1). It is estimated that 80 - 90% of drug discovery projects fail in the discovery and pre-clinical stages, and a half of the remaining projects fail in Phase III trials [1]. The challenges of discovering a new drug are likely exacerbated by the fact that the early, “serendipitous” discoveries [2, 3] harvested the “low-hanging fruit” [4, 5]. The process of getting a compound to market as a drug is highly resource intensive, requiring a large amount of human expertise, funding and time [6]. Recent estimates suggest that the mean cost of bringing a drug to market range from \$1.3 billion [7] to \$2.8 billion [8, 9].

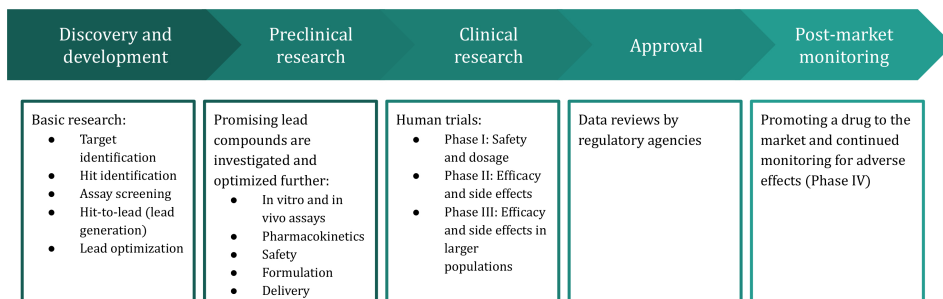


Figure 1.1: A general overview of the drug discovery and development process. Figure adapted from [10].

Computer-aided drug discovery (CADD) techniques are therefore routinely employed, both in industry and academic settings, during the early stages of drug discovery process to increase efficiency and reduce costs [6, 11]. A wide range of computational techniques are utilized to aid lead optimization. These include quantum mechanical and molecular mechanical calculations (to help understand the structure and dynamics of a molecular system) to virtual screening (to prioritize and reduce the number of compounds for a biological screening campaign) and quantitative structure-activity relationship (QSAR) techniques (to gain an understanding of the absorption, distribution, metabolism and elimination/excretion (ADME) properties of compounds) [6, 12, 13]. The increase in available computing resources and digital data [14, 15], has driven the increased use of computational techniques in drug discovery.

A ligand (which may also be known as a drug candidate or a drug, depending on how far along the investigative process the compound is) is typically a small molecule (a compound with a low molecular weight) which interacts with larger macromolecules, typically proteins, known as targets. The increased availability of bioactivity data, that is, data on the interactions between compounds and targets, has coincided with a paradigm shift in drug discovery from the “one drug one target” paradigm to “polypharmacology” [16].

Polypharmacology is driven by the fact that there are many proteins in nature, with over 20,000 proteins the human proteome alone, which have only about 1,000 characteristic ligand binding pockets (parts of the protein on which a small molecule may interact with) [17]. This means that a single compound, which binds to one of these pockets, is likely to interact with multiple proteins. It is estimated that on average drugs are active on six [18, 19] to twelve targets [20].

Therefore, for a drug to have the desired therapeutic effect, a set of targets need to be modulated to achieve efficacy, while avoiding others to reduce adverse side effects [21]. While the ability of a compound to interact with multiple targets poses challenges with respect to side effects, it also means that it is possible to repurpose or reposition drugs, which have already been through rigorous and expensive safety assessments, for other uses [22–24]. That is, a single drug may be used to treat multiple ailments depending on the targets and pathways it modulates. Repurposing drugs, approved and experimental, is a key strategy in accelerating drug development [1]. It is therefore vitally important that we gain an understanding of what targets a compound interacts with and may modulate.

## 1.2 Machine learning

There has been an increased use of artificial intelligence (AI) in a variety of fields, including drug discovery [25, 26]. AI systems attempt to replicate human intelligence and decision making. Machine learning (ML), a branch of AI, applies statistical and mathematical functions to input data to teach a computational machine (i.e. a model) to perform tasks such as predicting what the biomolecular targets of small molecules might be [27]. Broadly, ML approaches are divided into two paradigms, supervised or unsupervised learning, based on the type of input data used to train the machine to perform its task.

Models employing supervised learning learn to perform their task from training data which has been structured and labeled with the expected output value. That is, the training data have the correct answers to the task the model is being trained on. The model is fit to the training data and the trained model is then applied to new data to predict outcomes. For example, a target prediction model trained on data which consists of compound-target pairs that are labeled as interacting or non-interacting and would be used to label an unknown compound-target pair as interacting or non-interacting. In cases where the input data has been labeled with categorical variables (i.e. where a compound-protein pair is categorized as being either interacting or non-interacting, or perhaps as a strong binder, weak binder, or non-binder), the models are classification models. Classification models make predictions by classifying query data points into the output categories of training data. When the training data have been labeled with continuous values, the model is known as a regression model, e.g. predicting the binding affinity value of the compound-target pair. In comparison to classification models, regression models require higher quality training data, often limiting its use in large-scale bioactivity prediction.

Models employing unsupervised learning use input training data which has not been labeled with the output values. During the training, models find structure within the data to expose natural patterns among the data points. Models attempt to learn their task as the natural patterns in the training data emerge. For example, when the training data consists of a mixture of unlabeled interacting and non-interacting compound-target pairs, a goal of an unsupervised learning model could be to find patterns in the training data through which the training data would naturally be separable into interacting and non-interacting pairs (clustering). An unsupervised model may also highlight relationships between compounds of a particular type and targets of a particular type. Once trained, the model is applied to new pairs to make inferences on their interactions.

### 1.3 The importance of target prediction

In silico target prediction, using computational approaches to identify the possible macromolecular targets of small molecules, is a key tool in early drug discovery. Target prediction methods predict whether a query compound and macromolecule pair are interacting or non-interacting, thereby predicting the possible targets (interacting macromolecules) of a query compound. Target prediction is useful for a variety of tasks such as target deconvolution, elucidating the mode of action of compounds, drug repurposing, and predicting adverse effects of compounds. In-silico methods to predict the biomolecular targets of small compounds have a range of useful applications from drug discovery to cosmetics and agrochemicals. Across the general chemical industry, understanding the interactions with macromolecular targets aids assessing the safety and mode of action of compounds. Consequently, recent years have seen a growth in the development of in silico target prediction methods.

Target prediction methods have been developed using a number of different technology/model types, including similarity-based approaches [28], other ML-based approaches [29, 30], inverse/reverse-docking based approaches [31], and networks-based approaches [32]. Similarity-based approaches, also known as similarity-learning or nearest neighbor techniques, have a long history in CADD and use the similarity between a query and the knowledge base (also known as the reference/training data) to make predictions. Beyond utilizing similarity measures, target prediction has also been addressed using other ML models, such as random forests, support vector machines and neural networks, which use the models that were fitted on the training data to make predictions. Reverse docking approaches use docking scores of docked queries to make predictions, while network approaches build relationship networks to gain a systemic understanding of the data.

The data on which the method is developed is crucial to any artificial intelligence method. Depending on the types of data utilized, target prediction approaches may be categorized as [33-35]:

1. **Ligand-based approaches:** use molecular descriptors of the compounds to compare query compounds with compounds in the knowledge base and make predictions.
2. **Target-based approaches (also known as structure-based approaches):** use the structural information of the macromolecules to make predictions.
3. **Chemogenomic (or proteochemometric) based approaches:** use information from both the ligand and target sides to make predictions on interactions.

As ligand-based approaches only require the structural data of compounds and their bioactivity on macromolecules, their scope is wider than other types of target prediction approaches. Ligand-based approaches include methods that use similarity searches [28, 36–41] between a query compound and ligands in the knowledge base to infer possible targets for the query. Ligand-based ML models (such as linear regression [19], random forests [42, 43], support vector machines [43–45], neural networks [43, 46–50], etc.) trained on molecular descriptors of compounds, and chemical similarity networks [51–53] of compounds have also been used for target prediction.

Structure-based approaches use the three-dimensional (3D) structure of the macromolecules, often protein X-ray or nuclear magnetic resonance (NMR) structures, as the primary source of data to make predictions [54]. A common structure-based approach for target prediction is known as inverse docking, where the best orientation of query compounds complexed with a protein are scored to determine likely stable complexes. In contrast to ligand-based methods, docking approaches consider protein-ligand interactions as well as binding modes and some solvation effects [54]. Structure-based approaches are also better at determining the structural changes resulting in activity cliffs, which may not be visible in ligand descriptors and therefore missed by ligand-based methods [55]. However, docking multiple queries against multiple targets is computationally more expensive than ligand-based approaches and predictions are limited to targets with resolved 3D structures. Additionally, while the scoring functions used by docking methods rank the ligands and poses, correlating docking scores with binding affinity continues to pose a challenge and it is therefore difficult to rank targets using docking approaches [56]. Similarity-based approaches, predicting the bioactivity of compounds based on the structural similarity between protein targets, have also been utilized to make predictions [57, 58]. Other structure-based approaches, such as molecular dynamics simulations of a ligand interacting with a protein, can be used to elicit more information on binding. However, it is generally not computationally feasible to use these methods for large-scale predictions.

Chemogenomics-based approaches are defined as methods that use information from both ligands and targets to make their predictions [33, 59, 60]. Many types of modeling approaches, such as similarity-based approaches, heterogeneous networks, random forests, etc. [59] have been used for chemogenomic approaches. Ligands may be described using their physicochemical structures as well as other annotations including side effects, Anatomical Therapeutic Codes (ATC) and gene expression responses [59]. While on the other side, targets may be described using descriptors based on their 3D structures, genomic sequences, protein sequence, disease annotations etc. [59]. Additional cross-term descriptors, known as interaction fingerprints, that describe the interactions between a ligand and its target may also be used to develop a target prediction approach [33].



Interaction fingerprints are generated from crystal structures or docking experiments and codify the presence or absence of interactions such as hydrogen-bonds, hydrophobic and ionic interactions between a ligand and a target [23]. As with structure-based approaches, the scope of a chemogenomic-based approach is limited by the availability of target data for target descriptor calculation [33].

## 1.4 Data for ligand-based target prediction

Data are the foundation upon which a target prediction method is built [61]. Understanding the properties of the data is therefore key to understanding the scope and limitations of a developed method. Ligand-based target prediction methods use the structure of compounds (i.e. the ligands) and their bioactivity data (annotations of which proteins the compounds do or do not interact with) to make predictions.

### Data sources and limitations

A number of publicly available databases containing bioactivity information may be used as data sources [41, 61, 62], including databases such as ChEMBL, PubChem, BindingDB, PROMISCOUS and SuperTarget (Table 1.1). While the data in these sources varies due to differing foci, there is overlap of data points among the different data sources [61]. For example, ChEMBL database [63], one of the most popular data sources for target prediction as it balances size and quality, includes data from literature, as well as PubChem, bespoke Malaria screening data from different pharmaceutical companies, and BindingDB [64, 65].

While there is a large quantity of available data, which is increasing, there are still challenges associated with data availability. Data sets reflect just a tiny part of the chemical space and are highly imbalanced. This is because there is more information on particular ligands of explored chemical series [33, 58], and there is more information on well-studied targets compared with less explored targets [33, 66]. The data also often record a larger proportion of compounds which are known to be active on (i.e. interacting with) proteins compared to compounds confirmed to be inactive on proteins. [59, 67–69]. These biases in the available data limit the development and application of target prediction methods.

Some modeling approaches require data from both active and inactive classes, such as a ML classifier that classifies a compound as being active or inactive on a protein. Strategies to address the data biases during model development are often employed. These

Table 1.1: Examples of data sources for bioactivity data.

Name	Description	Data	URL
ChEMBL	Large bioactivity database with manually curated data from literature as well as PubChem and other data sources	2.1M compounds, 14k targets and 17.3M bioactivities	<a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a>
PubChem	Large open database of chemical information including small molecule screening data in PubChem BioAssay	110M compounds, 96k targets and 297M bioactivities	<a href="https://pubchem.ncbi.nlm.nih.gov">https://pubchem.ncbi.nlm.nih.gov</a>
BindingDB	Database of measured bioactivities of small molecules and druggable targets	982k compounds, 8.5k targets, and 2.3M binding data points	<a href="https://www.bindingdb.org/bind">https://www.bindingdb.org/bind</a>
DrugBank	Database of drugs and drug targets	14.5k compounds, 4.9k targets and 19k interactions	<a href="https://go.drugbank.com">https://go.drugbank.com</a>
PROMISCUOUS	Database of relationships between drugs, proteins, side-effects and indications	988k compounds, 4.9k targets and 19k interactions	<a href="https://bioinformatics.charite.de/promiscuous2">https://bioinformatics.charite.de/promiscuous2</a>
SuperTarget	Database with drug-target interactions	195k compounds, 6.2k targets and 332k interactions	<a href="https://bioinformatics.charite.de/supertarget">https://bioinformatics.charite.de/supertarget</a>

include enriching the inactive class by randomly selected compound-target pairs to represent inactive interactions [42], selecting compound-target pairs that are dissimilar to active compound-target pairs to represent inactive pairs [70], oversampling from the minority class [71] and under sampling or cluster sampling (sampling a selected number

interactions based on similarity) from the majority class [72]. In addition to accounting for the data bias in model development, a target prediction method needs to be thoroughly validated using validation strategies that account for these biases. In particular, it is important to understand how reliable a target prediction method is for a specific query.

## Representing molecular structures

There are multiple ways of representing molecular structures. One of most common representations of compounds in chemical information databases is a string format known as Simplified Molecular-Input Line-Entry System (SMILES) [73]. A SMILES string represents the two-dimensional (2D) or 3D graph structure of a molecule in a single line using ASCII characters to represent atoms, bonds and stereochemistry. Natural language processing techniques may be applied to the string representations of compounds to predict chemical properties and biological activity [74]. Traditionally however, quantitative encodings of compounds are utilized to make predictions [75].

One popular way to encode compounds for target prediction and virtual screening is the Extended Connectivity Fingerprint (ECFP) [76]. ECFPs are graph-based topological descriptions of a compound which encode the local properties of the compound. These fingerprints are generated by defining the radius of the local environment and the length of the bit vector of the fingerprint. To encode a compound, the algorithm begins by assigning a numerical identifier to the atoms of the compound (Figure 1.2 A). The algorithm starts at the first atom, encoding the substructures found in the environment around the atom. The radius of the environment is then increased around an atom it is encoding in an iterative fashion (Figure 1.2 B) until the predefined radius is reached. The algorithm then moves onto the next atom and repeats the encoding process until all atoms and their environments are encoded. The substructures identified on the compound as the algorithm traverses the molecular graph are then hashed into a single value. A binary bit-vector fingerprint of a set length is generated where each bit set to 1 represents the presence of a structural feature in the compound, while 0 represents the absence.

When choosing values for the radius of the environment and the length of the fingerprint, one must balance computational cost against the structural detail encapsulated by the fingerprint. While ECFP fingerprints are fast to generate [76], the larger the radius the more iterations per atom need to be conducted and more substructures may be identified. Longer fingerprints capture more detail but suffer from sparseness as majority of the bits are set to 0. Conversely, the shorter binary fingerprint is, more substructures are hashed

into a single position (hash collision) lowering the detail captured.

Once compounds have been encoded quantitatively, the encoding may be utilized to train predictive models. For similarity-based approaches, for example, one of multiple similarity coefficients [77] may be calculated to quantify the similarity between a pair of compounds which form the basis of the predictions made. Likewise, quantitative encodings are also used as input descriptions to fit more complex models such as random forest classification models.

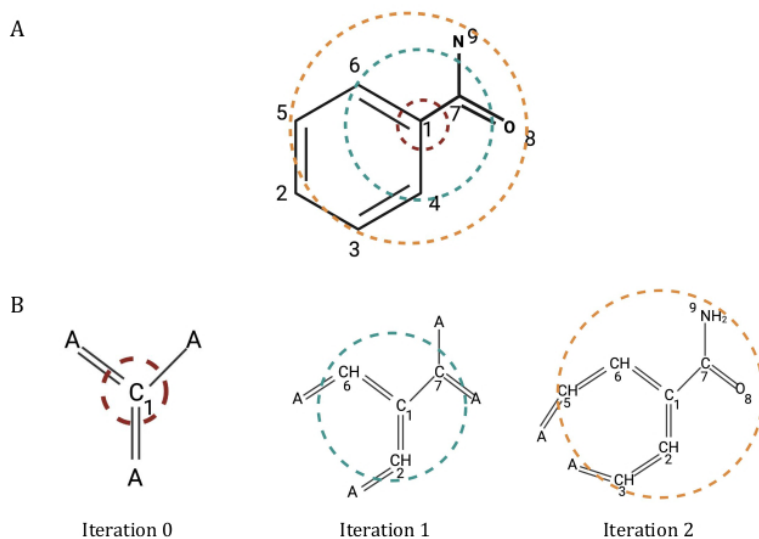


Figure 1.2: Example of local circular environments for Benzoic acid amide depicting the numbering of atoms (A) and the substructures found during the iterative identification process around atom 1 for two iterations (B). Atom “A” depicts a generic non-hydrogen atom. The red circle indicates the environment within a radius of 0 around the first atom, the teal circle indicates the environment within a radius of 1 (all atoms and bonds concerning the first atom away from the considered atom) and the orange circle indicates the environment within a radius of 2 (all atoms and bonds concerning the first two atom away from the considered atom). Adapted with permission from Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 2010,50, 742–754. ©2010 American Chemical Society.

## 1.5 Ligand-based target prediction methods

Target prediction is a rapidly developing field. There are a variety of ligand-based target prediction methods and many of these are available as free web services [24]. The types

of models used by the methods and the rigor with which their predictive performance has been evaluated varies.

The Similarity Ensemble Approach (SEA; <http://sea.bkslab.org>) is an early, and consequently one of the most widely used, target prediction method [78]. SEA uses minimum spanning trees to cluster the ligand sets for targets. The similarity between a query compound and the ligand sets are then assessed to make predictions. A raw score between a query and a target is calculated via a target's ligand set. The raw score is the sum of all the similarity coefficients between the query and all the ligands in a set. To reduce the bias caused by ligand set size the significance of the raw score is calculated based on the distribution of similarity scores between randomly curated ligand sets of different sizes. This significance, of the similarity between a query and the targets' ligand sets, is used to rank targets. SEA was originally built on ligand sets of 246 targets from the MDL Drug Data Report [78]. SEA has been tested through multiple rounds of prospective validation, where predictions which were previously not annotated were experimentally tested to see if they were true interactions [78, 79]. Notably, a large-scale validation study was carried out by Novartis where 656 drugs, which were approved for human use, were tested against 73 protein targets using annotations found in the ChEMBL database to build the minimum spanning tree [80]. SEA predicted 1241 interactions of which 348 interactions could be retrospectively verified in proprietary database [80]. Novartis tested 694 of the remaining unknown predictions, confirming that 48% of these predictions were true interactions, while 46% were disproved and about 6% of predicted interactions were ambiguous [80].

SwissTargetPrediction (<http://www.swisstargetprediction.ch>), another similarity-based approach, utilizes a logistic regression comprised of the 2D fingerprint similarity and the 3D molecular shape similarity between a query compound and knowledge base ligands [81]. The result of the logistic regression is then used to assign a probability of a query being active on the knowledge base compound's targets. The assigned probability is based on probabilities of a prediction being correct during internal validation tests run with the knowledge base data. The probability values are used to rank the targets for the query compound. SwissTargetPrediction covers over 2,600 protein targets using data from the ChEMBL database [19]. SwissTargetPrediction was evaluated using retrospective validation, that is, query compounds with known targets are selected from existing data and are used for testing. The average performance of the test compounds showed that SwissTargetPrediction performed the best when the test compounds were a random selection of the data, compared to test compounds that were selected based on their scaffolds, the assay where they were tested or the time (new compounds added to subsequent versions of ChEMBL) the compounds' bioactivity was recorded.

Chemical Similarity Network Analysis Pulldown (CSNAP; <https://services.mbi.ucla.edu/CSNAP/index.html>) is an example of a networks-based similarity-based approach [51]. CSNAP is built on data from the ChEMBL database and retrieves compounds from the knowledge base which are similar to the query compound. CSNAP uses a 2D similarity coefficient to retrieve compounds and calculates a Z-score to account for the significance of the similarity. The retrieved knowledge base compounds are then clustered into chemical similarity networks based on their pairwise 2D similarity. A network scoring function is then used to rank the targets of the first order neighbors of the query compound, returning a ranked list of targets for the query. CSNAP was validated using 206 external compounds that are active on 6 targets, retrieved from the Directory of Useful Decoys. Performance of the method was reported through the frequency of known targets ranked among the top-1, top-5 and top-10, with the method able to retrieve 94% of the targets for the queries among the top-10 ranked targets.

Some methods approach target prediction as a pair-input problem, where a compound-target pair is classified either as interacting or not-interacting [82]. When popular machine-learning classifiers (such as random forests, support vector machines, gradient boosting etc.) are applied to target prediction the target prediction problem is often approached as binary relevance problem. Binary relevance is an intuitive way of decomposing a multi-label problem (in the case of target prediction, what are the different targets a query interacts with) to a series of binary classification problems. That is, individual machines are trained for each target and a query is independently classified as interacting or non-interacting for each target.

An example of the binary relevance decomposition approach to target prediction is the work carried out by Bosc et al. [42]. For each target derived from the ChEMBL database, two types of random forest classification models were developed. The first is a standard random forest model (called the QSAR model in this work) and the second is a Mondrian conformal predictor (MCP) which used random forests as the underlying model type. The QSAR model calculates the probability of the query compound belonging to the active and inactive class of compounds for a target. The sum of these two probabilities is 1 and the query is labeled active or inactive depending on which class has the greater probability. With the MCP model, the prediction probabilities of a query are calibrated to determine the significance of the query’s predicting probability based on the prediction probabilities of compounds in a calibration set. The higher the significance of a prediction probability is for a class, the more likely a query belongs to the class. With an MCP model, a prediction is not just either active or inactive, but it could be both active and inactive or neither active nor inactive, based on the significance of the prediction probabilities. Bosc et al. analyzed the average classification performance of all the queries for each individual target model and showed that when “both” is considered a

valid prediction, the MCP models perform better than the standard QSAR models and when “both” is considered an incorrect prediction, the performance of the MCP models is comparable to the QSAR models. The authors argue that having information about the ambiguity of a prediction is useful to a user and the MCP models are currently used for target prediction on the ChEMBL website covering 352 targets [63].

Similarly Mayr et al. built classification models for over 1,300 ChEMBL assays [43] using different model types: k-nearest neighbors, naïve Bayes, random forests, support vector machines, a re-implementation of the SEA, feed-forward neural networks, convolution neural networks, and recurrent neural networks. Performance of the models was measured using classification metrics averaged over the query compounds. The authors showed that the deep learning feed-forward neural networks followed by the support vector machines performed the best at predicting activity. The study was repeated by Robinson et al. who argue that the deep learning model (feed-forward neural network) had a comparable performance to the support vector machines [83].

Among the ligand-based methods, a few combine a similarity-based approach with more complex ML approaches. This is primarily done through model stacking, where outputs from the first layer of models (typically different types of models) are used as inputs into a second layer of models, which then make a prediction.

HitPick (<http://mips.helmholtz-muenchen.de/hitpick/>) gets the most similar knowledge base compound to the query, and ranks the targets of this knowledge base compound using laplacian modified naïve Bayes models of each of the targets [84]. HitPick covers 1,375 human druggable targets. The performance was tested with about 20,000 interactions and presented as a function of the query’s similarity to the knowledge base [84]. With HitPickV2, the target coverage was updated and increased to cover 2,739 human targets. The method was also altered to select 10 distinct targets (not just the targets of the most similar knowledge base compound) from the most similar knowledge base compounds which are then ranked based first on their similarity (from most similar knowledge base compound) and then on target score from the naïve Bayesian model [85].

The Polypharmacology Browser 2 (PPB2; <https://ppb2.gdb.tools/>) method covers 1,720 target proteins and offers eight different modes for target prediction [86]. Three of the modes use different fingerprints (Molecular Quantum Numbers (MQN), Shape and Pharmacophore Fingerprint (Xfp), or ECFP) and a similarity-based approach to rank targets based on the most similar knowledge base compound associated with a target. Another three modes are combination modes (based on fingerprint type) which select the 2,000 most similar knowledge base compounds and then builds a Laplacian

modified multi-label ECFP-naïve Bayes model to rank the targets. A seventh mode is a global ECFP-naïve Bayes model and the final mode is a global ECFP-deep neural network model [86]. Based on the average performance measured by two different classification metrics, the similarity-based approach (nearest neighbor) and the combination approach (nearest neighbor combined with the naïve Bayes) performed the best using ECFP fingerprints with a radius of 4.

More recently, the STarFish study stacks different combinations of predictions from a k-nearest neighbor similarity search, random forests, and feed-forward artificial neural networks as inputs into logistic regression models to determine if a query is active or inactive on a target [87]. The k-nearest neighbor in the STarFish approach retrieves the targets of the 10 most similar knowledge base compounds and ranks the targets based on the number of times they are predicted by the retrieved knowledge base compounds. Random forest models using a binary relevance approach and a single multilayer perceptron feed-forward neural network were also built to predict targets of a query compound. Model stacking, using a logistic regression as the meta-model, was explored and a binary relevance construct was used for the stacking. That is a different logistic regression for each target, was used for the logistic regressions which took the prediction outputs from the other models as features to determine activity on a target. STarFish was trained on data from ChEMBL covering 1,907 unique targets. Performance metrics were averaged for the test compounds, however one of the test sets were natural product compounds which are fairly different to the knowledge base compounds compared to other test compounds. Stacking using the predictions from both the k-nearest neighbor and the random forests performed the best, followed closely by using only the k-nearest neighbor as input to the logistic regression. The latter was selected as the optimal model as it had a high performance with lower computational costs.

## 1.6 Applying target prediction to generate screening libraries

Compound libraries are a cornerstone of drug discovery [88]. Screening campaigns, testing a large number of compounds on desired targets, using biological and computational tools are routine in early drug discovery [88, 89]. Compound libraries for these campaigns are usually either compiled as focused libraries [90–93], where a library has been optimized with compounds to be screened against specific targets, or as general libraries, where a library has been optimized to increase the diversity of its compounds and increase the likelihood of finding bioactive compounds for any target [90–94]. With the



understanding that protein targets function as part of complex cellular systems with other protein targets (i.e., polypharmacology), there is a renewed interest in phenotypic screens [95]. Consequently, acquiring general purpose compound libraries have also gained interest. Commercially available compound libraries, however, are known to suffer from low hit-rates [96]. Designing compound libraries with a high likelihood of containing bioactive compounds is therefore important. In addition to increasing the likelihood of finding true hits, that is the library is composed of truly promiscuous compounds and not compounds which are likely to cause false readouts in assays [95, 97–99], a good library should also contain compounds with desirable physicochemical properties [95, 100]. A further consideration when designing screening libraries is keeping up with “novelty erosion” [101]. That is, libraries should contain compounds from newly explored areas of the chemical space [101] and show bioactivity on “new” and not just established targets.

Approaches to designing compound libraries include: utilizing property and substructure filters to ensure that the compounds have desired physicochemical properties and remove compounds with unwanted functional groups [98, 100, 102–104], exploiting the knowledge of in-house medicinal chemists to select key compounds and scaffolds [100], and employing selection algorithms [105], such as clustering [98, 106], ranking [102, 103], iterative selection [107] or evolutionary optimization algorithms [100], to select compounds for a library from a larger pool of compounds.

# Chapter 2

## Aims of this study

There are multiple approaches to build computational methods to predict the macro-molecular targets of small organic compounds. This thesis presents the development, validation and application of ligand-based target prediction methods. The thesis aims to answer the following questions:

Firstly, how would a medicinal chemist know which target prediction approach to use given the properties of the compound in question? This question emerges because the success of a target prediction approach is largely dependent on the knowledge base on which the approach is built. The reported performances of these approaches therefore depend on how the test data relates to the underlying knowledge base. Yet, most often the performance of a target prediction approach is only measured as an average of the tested queries. While these averaged metrics are useful in assessing the performance, they do not present a comprehensive picture as a user is likely to use the approach with queries that are different to the average of the tested data. Therefore, we begin this study by assessing the state-of-the-art target prediction methods and, importantly, the validation strategies used to evaluate the models (**Chapter 4**). To address shortcomings of the existing validation strategies of target prediction methods, we created strategies to account for underlying biases in the knowledge base and to help a user understand the applicability domain of the approach.

Secondly, how can we develop target prediction methods that have a wide chemical and biological scope and apply validation strategies to test these methods so that the performance results are interpretable to an end user given their query compounds? To address this question, we developed target prediction methods based on two popular approaches that utilize supervised learning: a similarity-based approach and a binary relevance random forest based approach (**Chapter 5**). In this work, the binary relevance random

forest based approach is called the ML approach as it harnesses more complex ML algorithms than basic similarity to make predictions. The objectives of this work were to develop target prediction methods which maximize the use of the existing bioactivity data, ensuring a wide target coverage, and to explore the performance and scope of the two approaches with respect to the similarity of the compounds in the test data to the knowledge base.

Finally, can target prediction be applied to curate libraries of compounds for screening decks which have a high likelihood of producing hits on any target of interest? Screening collections of compounds with experimental assays or virtual screens is a key step in identifying tool compounds for drug development. Increasing the possibility of finding true hits from a screening campaign is therefore valuable. Small to medium-sized screening libraries that are more likely to contain bioactive compounds on targets of interest at the early stages of drug discovery projects. One approach to increase the likelihood of true hits from screens is to optimize compound libraries with compounds that have drug like properties and predicted activity on a wide range of proteins. We developed a computational approach to select an optimal set of compounds for small to medium sized compound libraries (**Chapter 6**). To do this, we first identified purchasable compounds with drug-like physicochemical properties, and then predicted the bioactivities of these compounds. Finally, an evolutionary algorithm is used to select compounds for the optimized libraries, maximize the likelihood of selecting bioactive compounds and compounds with predicted activities on a diverse range of targets.

# Chapter 3

## Methods

To understand the value of large-scale target prediction methods, large volumes of data were retrieved and processed for method development and validation. This chapter summarizes the data sources, technologies and general procedures that were used to standardize the data and develop the models during the course of this work. Detailed information is provided in the individual Methods sections of the publications presented in this thesis.

### 3.1 Data sources for target prediction

The ChEMBL database [64] was used as the source of data to develop and validate the target prediction methods explored in this work (**P2**). The ChEMBL database is an open access database of manually curated bioactivity data that is primarily sourced from literature published in medicinal chemistry journals. ChEMBL also pulls data from other established databases such as PubChem BioAssay and BindingDB [64], as well as specialized data sets such as the natural product-like compounds from the University of Dundee, the Malaria Box Compound Set and contributions from commercial organizations such as GlaxoSmithKline and AstraZeneca [64, 65].

In this work, the PostgreSQL database dumps of the ChEMBL databases were retrieved from the ChEMBL website, restored locally and used to extract relevant data. Version 24 [108] of the ChEMBL database (ChEMBL24) was used to develop the similarity-based and machine learning target prediction approaches (**P2**). To validate the performance of the target prediction approaches (**P2**), test data was retrieved from ChEMBL24 and version 25 [109] of the ChEMBL database (ChEMBL25). A validated similarity-based

target prediction approach was applied to curate libraries of compounds for screening (**P3**) using the latest version of ChEMBL, version 27 (ChEMBL27) [110], available at the time. All the data retrieved from the various versions of the ChEMBL database were preprocessed and cleaned prior to the development and validation of the target prediction methods.

## 3.2 Chemical data processing and molecular descriptor calculation

Data must go through a rigorous cleaning and transformation process before it can be utilized for analysis. With chemical data, this includes standardization of the chemical structures, as there are multiple ways to represent a single compound and no widely accepted standard practices to structurally depict compounds [111].

All the compounds in the work presented in this thesis were retrieved from their respective sources in the Simplified Molecular Input Line Entry System (SMILES) format. SMILES strings are a ubiquitous compound representation format which is based on a defined set of grammar rules to express the graph (atoms as vertices and bonds as edges) formed by a molecular structure [73].

The compound structures were preprocessed and standardized using Python scripts and the RDKit toolkit [112] to remove salt and solvent components, neutralize any charges, and obtain the SMILES string of the canonical tautomer as the representative structure of the processed compound. Additional processing to identify compounds with unwanted atom types (atoms other than C, H, O, N, P, S, F, Cl, Br, and I) and compounds outside molecular weight limits was also conducted to remove compounds from further consideration.

Following the structural standardization, molecular descriptors were calculated for the compounds. In this work, the Morgan fingerprint of a compound, obtained by RDKit, provides the primary descriptors used for model development. Morgan fingerprints are RDKit’s implementation of the circular ECFP [76]. The fingerprint of a compound is generated by noting the presence or absence of specified substructures within the local environment of its atoms, defined by the radius of the fingerprint, and moving iteratively from one atom to the next of the compound. The resulting fingerprint is a series of bits of a defined length, with each bit set to 1 or 0 to denote the presence or absence of features within a compound. In this work, we use Morgan2 fingerprints (Morgan fingerprints with a radius of 2), at a length of 2048 bits, as this has been shown to be the best performing

descriptors for target prediction and virtual screening [20, 39, 86, 113].

### 3.3 Calculation of molecular similarity

A fundamental concept of cheminformatics is the similarity principle. This posits that the compounds with a similar structure are likely to have similar properties. That is, birds of a feather flock together. When bit-based fingerprints, such as the Morgan2 fingerprints, are used to describe compounds, coefficients to quantify the similarity or the distance (where distance is  $1 - \text{similarity}$ ) between a pair of compounds may be calculated [77]. In this work the Tanimoto coefficient (TC) (also known as the Jaccard index) is used to measure the similarity between two compounds. The TC of two compounds ( $m_1$  and  $m_2$ ) is calculated by dividing the number features that the compounds have in common by all unique features present within the pair (Equation 3.1). The TC is bound between 0 and 1, with 1 indicating a complete match and 0 indicating a complete mismatch between two compounds. Calculating the similarity between compounds underpins both the similarity-based approach for target prediction (which was developed for this work) and the validation of the target prediction approaches.

$$TC(m_1, m_2) = \frac{m_1 \cap m_2}{m_1 \cup m_2} \quad (3.1)$$

### 3.4 Model development for target prediction

In **P2** we investigate two types of target prediction approaches: a similarity-based approach and a binary relevance, random forest based (ML) approach (Figure: 3.1). With the similarity-based approach, targets are predicted for a query compound based on the similarity of the query to the compounds in knowledge base. By the similarity principle, the more similar a query compound is to a compound in the knowledge base, the more likely it is that the query compound will be bioactive on the knowledge base compound’s targets. As such, for a query, a ranked list of proteins is returned based on the similarity of the query and the proteins’ ligands (i.e., compounds found in the knowledge base). In **P2** the similarity between two compounds, as measured by the TC, was calculated using RDKit’s `DataStructs.FingerprintSimilarity` method [112]. In **P3**, where the similarity-based approach was applied to generate compound libraries, the highly optimized and significantly faster, `search.knearest_tanimoto_search_arena` method implemented in the chemfp toolkit [114] was used to calculate the TC between two compounds.

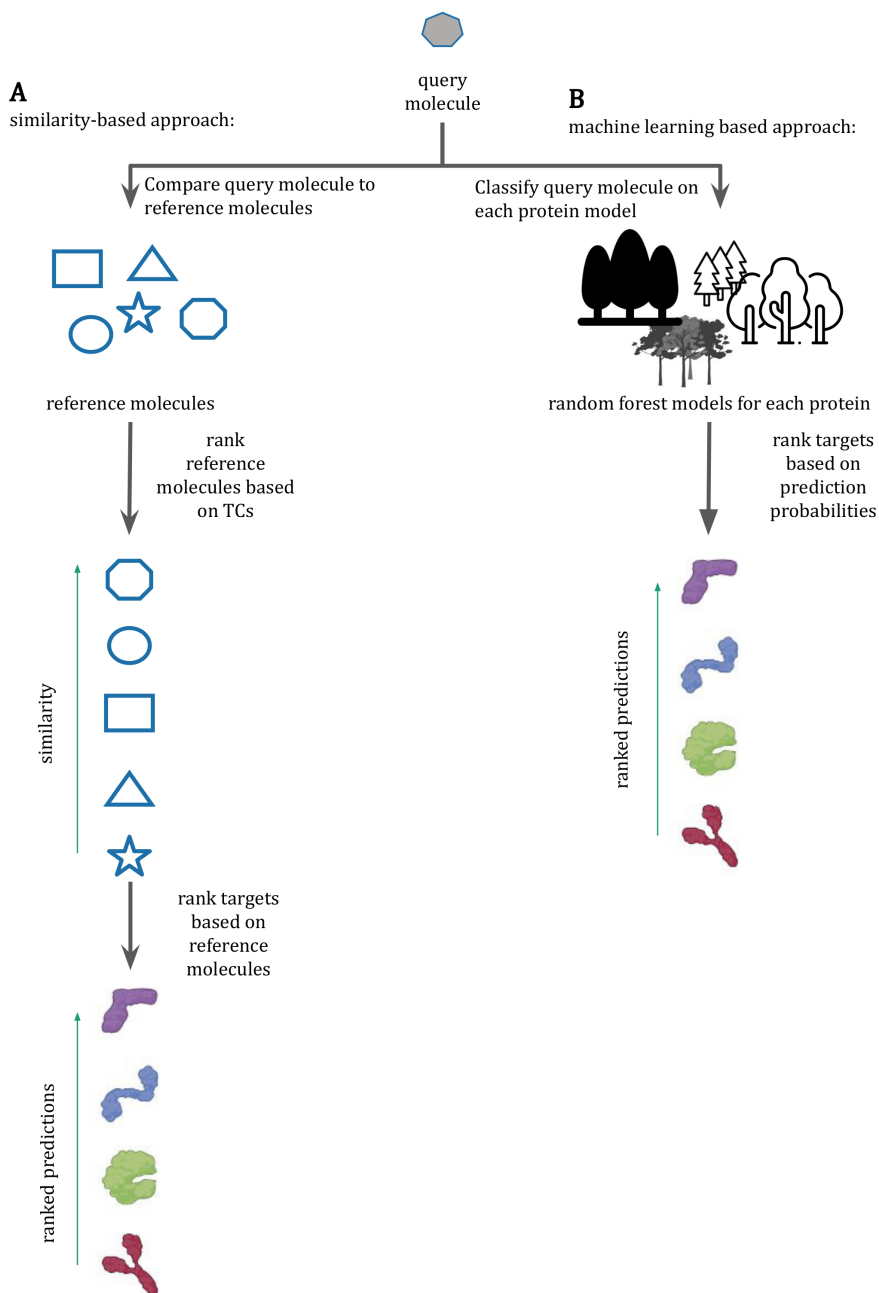


Figure 3.1: Workflows depicting the similarity-based approach (A) and the machine learning based approach (B) for target prediction

The ML approach, on the other hand, uses a binary relevance model with random forest classifiers to make predictions on the bioactivity of a query compound. For each target a random forest binary classifier was trained on data composed of at least 25 compounds which are known to be active on the target and ten times as many compounds which are known or assumed to be inactive. To rank the targets, a query is classified by each target model independently and the prediction probabilities of the query belonging to the active classes of all the classifiers are used to rank the proteins from most likely to be a target to least likely. The classifiers were implemented using the scikit-learn library [115].

### 3.4.1 Training the random forest classifiers for the ML approach

A random forest classifier is a model that is composed of multiple decision tree classifiers, with each tree in the forest casting a vote on which class a query belongs to. The aggregated vote is then used to classify a query. Each decision tree classifier could be thought of as a flowchart developed using a sub-sample that was sampled with replacement of the knowledge base. Every tree in the forest uses a different sub-sample of the data, generating different trees and therefore a better overall prediction. A single decision tree grows by using the most prominent features of the data first to decide on the class membership (Figure 3.2). In the case of target prediction, the substructures of the compound that have the highest correlation with a compound being active or inactive on a protein target are used first by the decision tree. Using multiple decision trees (i.e., a random forest) increases the variance among the trees, as they are each built on different sub-samples of the data, avoiding over-fitting the model and improves the predictions made.

Most hyperparameters for the classifiers were set to their default values. Values for the number of trees and the maximum depth of the trees (Table 3.1) were selected using a 10-fold cross-validation grid search protocol for each classifier. During the grid search, the performance of the classifiers was measured using the Matthews correlation coefficient (MCC) (Equation 3.2) metric. The best combination of the explored parameters (as measured by a high average MCC score) was selected and used to retrain the model on the full training data. MCC was selected as the performance measure as it is a robust measure, particularly with imbalanced data, because it considers the proportion of all the classes of the confusion matrix, namely true positive (TP), true negative (TN), false positive (FP) and false negative (FN).



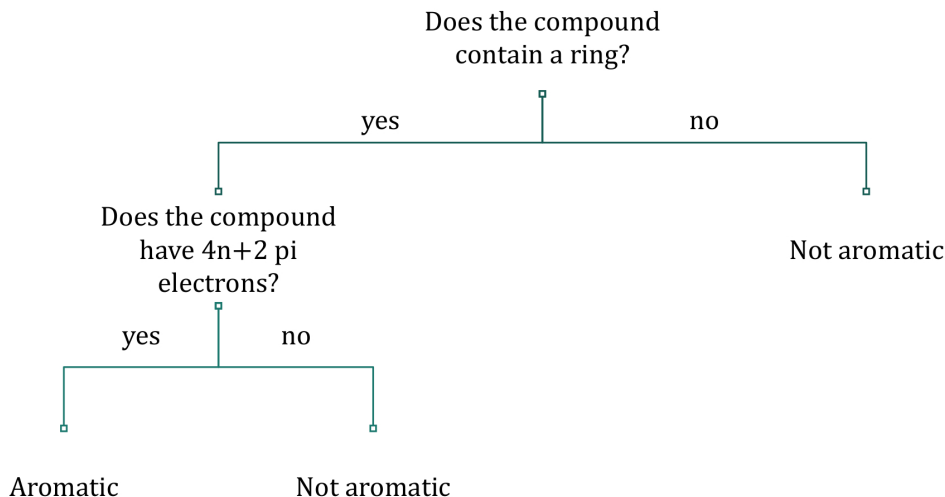


Figure 3.2: Toy example of a decision tree used to determine if a compound is aromatic or non-aromatic

Table 3.1: Hyperparameters explored during the grid search protocol to optimize the target classifiers

Hyperparameter	Values explored
n_estimators: number of trees	200, 500, 1000
max_depth: maximum depth of tree	25, 45, 50, 75, 100

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.2)$$

### 3.5 Validation of the target prediction approaches

In **P1** we make the case that the top-k metric, disaggregated by the similarity of queries or test set compounds to the reference/training set data, is a powerful metric to measure the performance of a target prediction approach. This metric conveys the average performance and provides a sense of how an approach would work on a specific compound given the compound’s relationship to the approach’s knowledge base. We expand on this in **P2** to thoroughly compare the similarity-based and the ML approaches to target prediction, defining two versions of the top-k metric: the success rate and the recovery rate:

1. **Success rate:** the percentage of queries with at least one known target present among the top-k predicted targets.
2. **Recovery rate:** the percentage of known bioactivities (ligand-target pairs) present among the top-k predicted targets.

The target prediction approaches were assessed using the two performance metrics (success and recovery rates) and under three testing scenarios (the Standard testing, Standard time-split and Close-to-real-world scenarios) in order to obtain a robust and realistic measure of performance:

1. **Standard testing scenario with an external test set:** over 44,000 query compounds with at least one known target represented in the knowledge base.
2. **Standard time-split validation scenario:** over 18,000 “new” compounds with at least one known target represented in the knowledge base.
3. **Close-to-real-world scenario:** over 20,000 “new” compounds with targets that may or may not be represented in the knowledge base.

## 3.6 Data source for compound library curation

The ZINC20 database [116, 117] was used as the source of purchasable compounds for the curation of the libraries. The ZINC20 database is a collection of purchasable compounds from 150 chemical vendors which is refreshed every 90 days. Nearly eight-million compounds, in SMILES format, that were listed as “in-stock”, annotated as “anodyne”, had a charge state of -1, 0, or +1, and a calculated logP value between 0 and 4, were downloaded from the ZINC20 web service [117]. Compounds listed as “anodyne” have been tested against and passed through a thorough set of reactivity filters and pan assay interference compounds (PAINS) patterns [118]. Compounds marked as “anodyne” are therefore not likely be “bad actors” or “nuisance compounds”. That is, they are less likely to be reactive or cause pan-assay interference [119]. The retrieved compounds were preprocessed as described in section Section 3.2.

### 3.7 Rules-based filters for compound library curation

To curate compounds for a screening library (**P2**), we first assembled a pool of candidate compounds (PCC). The purchasable, anodyne, compounds from the ZINC20 database were retrieved and standardized. To bias the PCC towards compounds with key “drug-like” physicochemical and structural properties, compounds matching the following were removed from further consideration:

1. Less than 18 or more than 30 heavy atoms (identified using RDKit’s Lipinski.HeavyAtomCount method)
2. Less than one or more than four rings (identified using RDKit’s CalcNumRings method)
3. Ring systems with more than three fused rings (identified using RDKit’s GetRingInfo and AtomRings methods to identify the ring systems and number of rings per system present a compound)
4. More than eight rotatable bonds (identified using RDKit’s rdMolDescriptors.CalcNumRotatableBonds method)
5. More than three hydrogen bond donors (identified using RDKit’s Lipinski.NumHDonors method)
6. More than seven hydrogen bond acceptors (identified using RDKit’s Lipinski.NumHAcceptors method)
7. Charged carbon atoms (identified using RDKit atom properties)
8. Without at least one oxygen or nitrogen atom (identified using RDKit atom properties)

Utilizing compounds which were flagged as “anodyne” from the ZINC20 database reduced the possibility of “bad actor” [97, 98] behavior within the library. To further reduce the possibility of “bad actor” behavior within the library, compounds that contained the substructures listed in the “remove” and “extreme caution” categories of the SMILES ARbitrary Target Specification (SMARTS) [120] patterns compiled by Chakravorty et al. [121] were also removed. The SMARTS patterns used were compiled to identify nuisance compounds and correctly identified 57% of noisy GSK compounds in

the study's validation [121]. An additional SMARTS pattern ("S(=O)(=O)O") was also used to filter out compounds with the tosyl esters, a highly reactive functional group not captured by the other patterns.

Compounds which pass through these filters and have predicted targets, using a similarity based target prediction approach, are then assembled as the PCC. Compounds for the libraries are selected and optimized from the compounds in the PCC.

### 3.8 Genetic algorithm for compound library curation

We developed an approach (**P2**) to generate a library of compounds which have a higher-than-average likelihood of being bioactive on an arbitrary target of interest. This approach includes applying a genetic algorithm to optimize the selection of compounds for the compound library. In our implementation, a gene is defined as a compound and an individual is a set of genes, i.e., a set of compounds. By extension then, a population is therefore defined as a set of individuals, i.e., a set of compound sets. A population, and the individuals within it, evolves over generations and the fittest individual is selected as the optimized library.

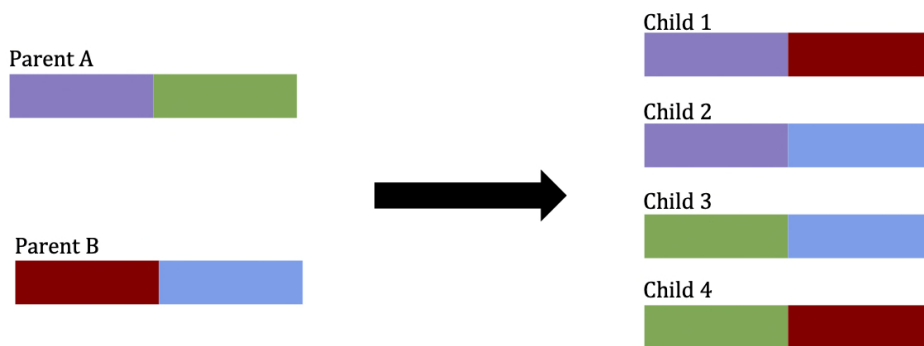


Figure 3.3: Generation of four child individuals (compound sets) from two parents using a single point crossover.

A PCC is assembled from purchasable compounds which have predicted targets and have passed through a set of rules-based filters. The initial population is first generated by randomly selecting compounds for each individual from the PCC. The evolution is carried out by selecting the fittest individuals as parents, producing four children from a pair of

parents using a single point crossover (Figure: 3.3). One-third of the population for the next generation is composed of the fittest parents from the current generation, while the remaining two-thirds of the population is composed of the children. To add variation to the population, 10% of the compounds in a child are replaced by new randomly selected compounds, ensuring that a compound is not repeated within an individual. Further details of the approach, including the development of the fitness function that was used to determine the fitness of the individual compound sets are described in the Methods section of **P2**.

# Chapter 4

## Strategies to validate target prediction methods

The validation of most published in-silico methods to predict the biomolecular targets of small compounds leave key questions about performance unanswered. Often, performance is reported as an averaged performance of the model(s). That is, performance metrics, such as the MCC, are calculated based on a confusion matrix composed of all the compounds in the test set. For example, Bosc et al. [42] compared binary relevance approaches using random forest classifiers and Mondrian conformal prediction models, and classification metrics such as average sensitivity, specificity and correct classification rates of the models were used to quantify performance. The Bosc et al. [42] study showed the value of using conformal prediction to calibrate prediction probabilities to classify the activity of a query on a protein. Similarly, Mayr et al. [43] used a binary relevance decomposition to train and compare the average performance of multiple types of classifiers (such as neural networks, support vector machines, random forest, etc.) for target prediction. They showed that, on average, the overall average classification of the feed-forward neural network classifiers outperformed the other classifiers. Reporting these averaged performance metrics is valuable as it gives an overview of performance. However, compounds are not average, and using averaged measures do not give a user a sense of how reliable a method may be for their specific query. Model performance is heavily affected by the quality of the data on which models were trained, and specifically how similar a query or queries are to the knowledge base.

This chapter outlines the different validation strategies that can be employed to estimate the performance of a target prediction method. We reviewed the different ways in which data can be partitioned into training and test sets, then discussed the merits of the

different performance metrics. Finally, building on existing strategies, we developed new strategies to obtain more realistic performance measures. We argued to desegregate performance measures by similarity to the training set. This is a simple, powerful and underutilized strategy which provides the user an understanding of how much to trust a method's prediction based on the distance of their query compound to the knowledge base.

## **P1: Validation strategies for target prediction methods**

Neann Mathai, Ya Chen and Johannes Kirchmair

*Briefings in Bioinformatics*, **21(3)**, 791–802 (2020).

<https://doi.org/10.1093/bib/bbz026>


### **Contributions:**

N. Mathai, Y. Chen and J. Kirchmair conceptualized the research. N. Mathai conducted the research and analysis, with contributions from Y. Chen. N. Mathai wrote the manuscript, with contributions from Y. Chen and J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted from: Mathai N.; Chen Y.; J. Kirchmair, Validation strategies for target prediction methods. *Brief. Bioinform.* **2020**, *21(3)*, 791–802.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).

# Validation strategies for target prediction methods

Neann Mathai, Ya Chen and Johannes Kirchmair 

Corresponding author: Johannes Kirchmair, Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway and Center for Bioinformatics (ZBH), Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, 20146, Germany. Tel.: +47-55-58-34-64; E-mail: johannes.kirchmair@uib.no

## Abstract

Computational methods for target prediction, based on molecular similarity and network-based approaches, machine learning, docking and others, have evolved as valuable and powerful tools to aid the challenging task of mode of action identification for bioactive small molecules such as drugs and drug-like compounds. Critical to discerning the scope and limitations of a target prediction method is understanding how its performance was evaluated and reported. Ideally, large-scale prospective experiments are conducted to validate the performance of a model; however, this expensive and time-consuming endeavor is often not feasible. Therefore, to estimate the predictive power of a method, statistical validation based on retrospective knowledge is commonly used. There are multiple statistical validation techniques that vary in rigor. In this review we discuss the validation strategies employed, highlighting the usefulness and constraints of the validation schemes and metrics that are employed to measure and describe performance. We address the limitations of measuring only generalized performance, given that the underlying bioactivity and structural data are biased towards certain small-molecule scaffolds and target families, and suggest additional aspects of performance to consider in order to produce more detailed and realistic estimates of predictive power. Finally, we describe the validation strategies that were employed by some of the most thoroughly validated and accessible target prediction methods.

**Key words:** target prediction; polypharmacology; model validation; data bias; classification; performance metrics

## Introduction

Fueled by the growing amount of chemical and biological data, the availability of powerful phenotypic screening technologies [1], and a shift in small-molecule drug discovery from the 'one drug one target' paradigm to 'polypharmacology' [2–5], *in silico* methods for the prediction of the biomacromolecular targets of small molecules have become one of the most intensely researched areas of cheminformatics in recent years. These methods are useful not only for the discovery of new medicines but also in the repositioning of existing approved drugs [6–9].

Target prediction methods are typically pair-input problems, in that they classify a query compound and a biomacromolecule pair as an interacting (positive) or a non-interacting (negative) pair. One categorization of target prediction methods, based on the types of data used, classifies methods into three overarching approaches: ligand-based, structure-based and chemogenomic approaches [10, 11]. Ligand-based approaches make predictions based on the similarity principle, which states that similar ligands (in the context of this review, small molecules) are likely to have similar targets. These methods typically make use of a variety of molecular descriptors to quantify and compare

Neann Mathai is a PhD student at the Department of Chemistry and the Computational Biology Unit (CBU) of the University of Bergen (UiB) and also affiliated with the University of Hamburg (UHH). Her research focuses on the development and application of computational methods for target prediction. Ya Chen is a PhD student at the Center for Bioinformatics (ZBH) of the UHH. Her research focuses on the development and application of computational methods for the prediction of the biomacromolecular targets of natural products.

Johannes Kirchmair is an associate professor in bioinformatics at the Department of Chemistry and the CBU of the UiB. He also is a group leader at the ZBH. His research activities focus on the development and application of *in silico* methods for the prediction of the biological activities, metabolic fate and toxicity of xenobiotics.

Submitted: 26 November 2018; Received (in revised form): 14 January 2019

© The Author(s) 2019. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



the physicochemical properties of small molecules. They do not rely on structural information on biomacromolecules. Their applicability domain is limited primarily by the available chemical and biological data. Structure-based approaches, such as ligand docking, use structural data on biomacromolecules as the main source of information to make predictions. They are generally more computationally expensive than ligand-based methods, and their primary limitations are defined by the availability of relevant target structures and accuracy of scoring functions. Chemogenomics approaches (or proteochemometric approaches) are defined here as methods that combine information from both ligands and targets to make their predictions [10–12].

There are several publications discussing techniques that can be used in validating target prediction models [13–20]. However, among the many recently published reviews on *in silico* target prediction, only few include a discussion of validation strategies [6, 10, 11, 21–26]. With this review we aim to provide a comprehensive reference of strategies for the validation of target prediction models. The review begins with a discussion of data partitioning schemes that are used to train and test models to measure their performance, highlighting their appropriateness and limitations. This is followed by an analysis of the metrics that are used to measure this performance and of established benchmark data sets. Building up on these components, we point out strategies to obtain more realistic estimates of the performance of target prediction models that account for the biases present in the underlying reference data. Finally, we describe the validation strategies that were employed by some of the most thoroughly validated and accessible target prediction methods.

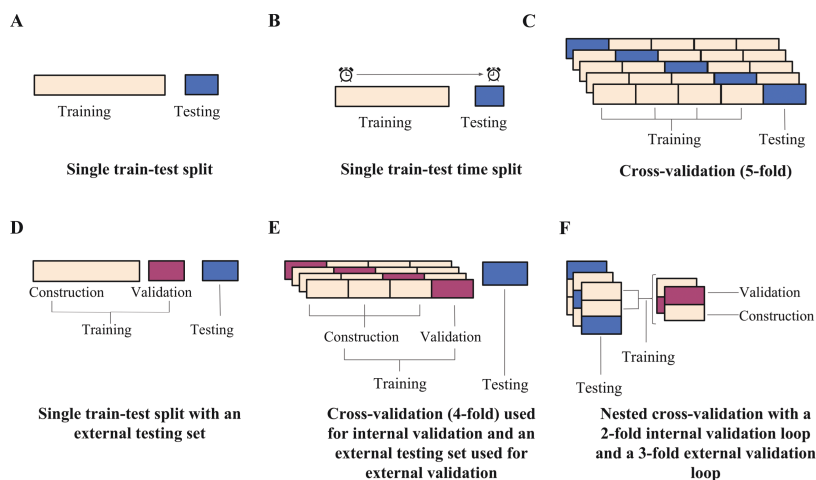
## Strategies for validating target prediction methods

Validation primarily serves two purposes: the selection of an optimal model and the evaluation of its generalized predictive performance [13, 14]. Model selection is commonly a result of an iterative model building process, during which models based on various algorithms and parameters are built on a training set and validated on a testing set. This validation procedure is generally referred to as internal validation. While often used as the sole means to report on the performance of models, internal validation is insufficient to determine the predictive performance as the iterative modeling procedure may introduce a bias toward the properties of the testing data and hence result in an overestimation of model performance. Data that are blinded to the model development process should therefore be used, in a process known as external validation, to obtain a more realistic representation of generalized performance [13]. As part of an external validation process, the training set may be further divided into a construction set (data used to train and parameterize the model) and a validation set (data used for the internal validation to optimize the model), while the testing set is held back for performance assessment [13]. With data in place to train and test the model, the metrics used to measure the performance during the testing need to be considered next. The choice of how a method was validated (that is the data partitioning schemes used for the validation) and how its performance was measured (the metrics used) are therefore essential in understanding the reported performance.

## Data-partitioning schemes

In the simplest case, models can be trained on one set of data and tested on another set created by random selection (Figure 1A). Such a single train–test split procedure is only effective if the training and testing sets are sufficiently large, diverse and representative of the parameter space [13, 14, 20]. However, as the limited amount of available data usually does not allow for large testing sets, the resulting test statistics may, to some extent, be an artifact of how the data were split and not an indicator of generalized performance [13, 14, 16, 18, 25]. Instead of random selection, a single split of the data into a training and a testing sets may alternatively be prepared using a time-split approach, where the model is trained on data compiled before a given date and tested on data generated later (Figure 1B). The time-split approach simulates a real-world scenario where a finalized model is put to use and new interactions are predicted [17]. Martin et al. [27] proposed a ‘realistic split’ approach, where compounds are clustered based on chemical similarity to mirror the exploration of new chemical scaffolds over time. In the realistic split approach, the larger compound clusters form the training set (~75% of the total number of compounds), while the remaining smaller clusters and singletons (~25%) are reserved for the testing set. The authors showed that when predicting activities of high throughput screens, a single 75:25 train–test split reported over-optimistic performance results when the split was created using a random sampling (as the compounds in the testing set were similar to the training set). In contrast, their sampling approach provided more realistic performance estimates.

To get a more robust estimate of how a model generalizes, cross-validation (CV) schemes have emerged, which partition the data in multiple ways to increase the variation in the training and testing data and to reduce the influence of how the data is split on the resulting testing statistics. A simple CV procedure is the *n*-fold CV, which involves randomly partitioning the data into *n* partitions and iteratively selecting each partition as the testing data while training the model on the remaining partitions (Figure 1C). The result is *n* models and *n* testing statistics, the latter of which are then averaged to give a more realistic estimate of a model's performance [15, 19]. When *n* is equal to the number of observations, the scheme is known as the leave-one-out CV (LOOCV), with each observation playing the role of the testing set once. LOOCV is known to produce over-optimistic estimates of performance in the current context as there is a high likelihood of finding similarity between the testing molecule and the training set [13]. Therefore, typically a 5- or 10-fold CV scheme is chosen where the observations are divided into 5 or 10 folds, respectively. The folds for an *n*-fold CV are often created through random sampling. Pair-input prediction methods however are known to perform better when the tested pairs contain small-molecule or target components that are present in the training data, as such randomly generated folds for validation may produce over-optimistic performance results [16, 18, 25]. Alternative sampling methods, like stratified sampling, aim to address this issue by constructing folds with desired representations. For stratified sampling, data are first divided into the different output strata (positive or negative interactions for example) and are then randomly selected from the strata so that the desired ratio of observations is represented in the folds [14]. The folds for a CV performance assessment may also be designed to ensure that all interaction pairs involving a particular compound, compound cluster (i.e. structurally related compounds) (Figure 2A), a target (Figure 2B) or even



**Figure 1.** Illustrations of example data partitioning schemes: (A) a single train-test split, (B) a single train-test split of chronological data, (C) a 5-fold CV scheme, (D) a single train-test split into construction and validation sets for internal validation and an external testing set for external validation, (E) a 4-fold CV scheme used for internal validation with a testing set reserved for external validation and (F) a nested CV scheme with a 2-fold loop for internal validation and a 3-fold loop for external validation.

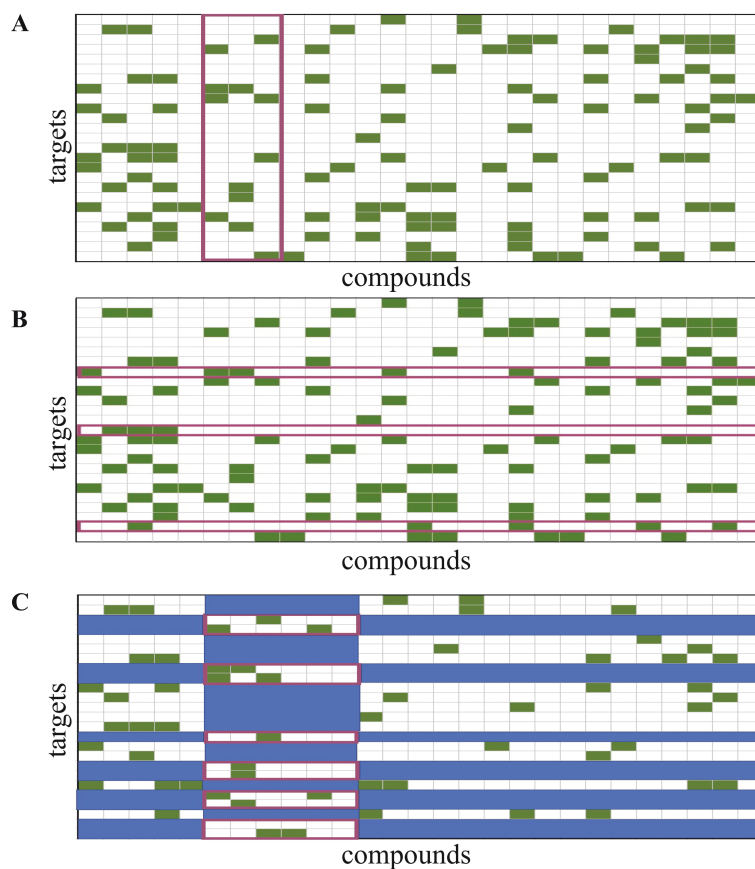
molecule–target pairs (Figure 2C) are assigned to the same fold. These types of schemes are useful to estimate the accuracy of a method with compounds or targets with limited prior knowledge [25]. As schemes with such designed folds are likely to have fewer or no similar components between the training and testing data, the performance will be lower than that measured with a standard  $n$ -fold CV [16, 18, 25]. In order to give a more thorough estimation of predictive performance, it is therefore recommended that the results obtained from standard  $n$ -fold CV are compared to those obtained from more challenging designed-fold testing scenarios [11, 18, 25].

Most computational approaches require parametrization (e.g. the value of  $k$  in a  $k$ -nearest neighbour model) via iterative optimization, during which different values of the parameters are explored so as to minimize the prediction error. The repeated use of the identical training and testing sets from a single train-test split for this optimization procedure is likely to result in selection bias. That is, the optimized models may be biased towards the properties of the specific testing data [13, 14]. In cases where CV is used not only to estimate the performance of a model but also to determine the best parameters for the final model, the CV is first repeated over the different values of the parameters so as to minimize the CV error, and the parameters with the lowest validation error rates are selected for the final optimal model [14, 15]. Due to the limitations of data utilized for the development of target prediction models (such as implicit biases, data imbalance and incomplete interaction knowledge), the performance of a model determined through internal  $n$ -fold CV is often over-optimistic because of selection bias [18, 25]. Therefore, the performance results of this internal validation should not be considered as a rigorous estimate of the performance of the selected model. Instead, external validation should be used to evaluate the performance of the method once the model has been selected [14]. However, using a single testing set reserved for external validation (Figure 1D and E) may still produce performance statistics that are not reflective of the

generalized performance but are an artifact of the testing and training split and requires the testing set to be withheld from the model [13].

Nested CV has consequently emerged as a scheme to perform external CV and better estimate unbiased performance (Figure 1F) [13–15]. In nested CV, two CV loops are run: an inner ‘internal validation’ CV loop is used for model selection and parameter optimization, and an outer ‘external validation’ loop is used for model evaluation. In the inner loop, models are trained using construction data and tested using validation data over all unique parameter values. The parameters that produced the lowest internal CV error are then used to build models for the external CV loop, where models are trained on the training set and tested on the testing set. As the testing set has remained independent of the parameter selection process, the external CV errors, often presented as an average error, are a more realistic estimate of the generalized error of the model [13–15]. It is important to note that with each iteration of the outer loop, the combination of parameters may be different due to the nature of the data in the internal loop that was used to optimize them. Nested CV does however provide the best estimate of performance [11, 14].

Often, as is the case with all the validation schemes described, even when using the data in the testing set for external validation, a final model, with parameters unchanged, is trained on the full data. The performance measures therefore do not evaluate this final model but the process of building the model. These measurements are dependent on how the data are split into the training and testing sets [13–15]. Repeated CV and repeated nested CV, to allow for data variance by resampling the folds over each repetition, have thus been recommended as a means of converging on true performance [14]. Repeated validation, commonly known as bootstrapping, is resampling the training and testing sets and repeatedly calculating performance metrics many times over. This iterative process allows for the calculation of the variation and confidence intervals of the



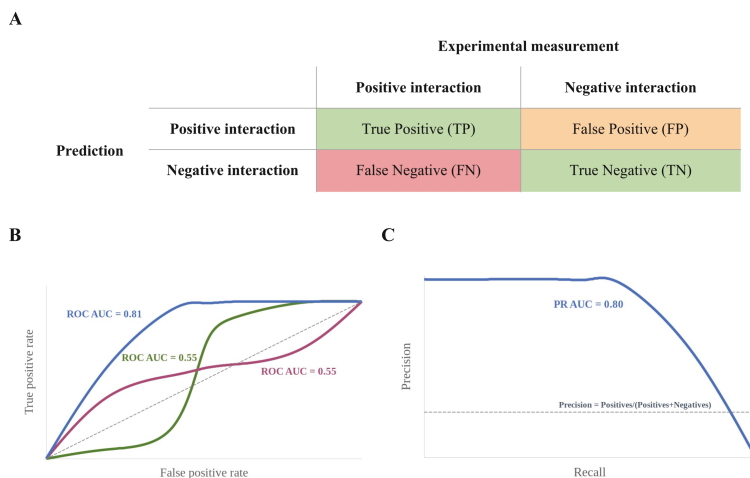
**Figure 2.** Examples of CV-testing folds designed to have (A) all data points involving specific queries within 1-fold (points inside the purple box), (B) all data points involving specific targets within 1-fold (points inside the purple box) and (C) all data points involving the components of query compounds-target pairs within one testing fold (points inside the purple boxes). The data points covered by the blue boxes are omitted from both training and testing data during the CV round involving the purple boxed data as the testing set, and the remaining data points are used as the training set. Interacting pairs are shown in green while (putative) non-interacting pairs are shown in white (adapted from Pahikkala et al. [25]).

performance metrics. Krstajic et al. [14] propose a repeated nested CV scheme, where the internal and external validation loops each have 50 repetitions, and the lowest and highest error metric, in addition to the average error metric, are reported to show the variance in the method's performance. They recommend using random  $n$ -fold CV for the internal loop and stratified CV for the external loop when using repeated nested CV to develop and evaluate a model [14].

In addition to reporting statistical metrics generated from the above validation schemes, illustrative case studies are also often reported to highlight the performance of a method. However, reporting on just a few case studies is not a sufficiently rigorous approach to determine a model's performance [26]. Ideally, large-scale experimental studies would need to be conducted that allow not only thorough validation but also a demonstration of a method's potential impact. However, due to cost, such large-scale studies are generally not carried out.

### Performance metrics

In its most basic form, target prediction can be regarded as a binary classification problem: a small molecule either interacts with a biomacromolecule (a positive interaction) or it does not (a negative interaction). Based on this premise, a common evaluation technique is to complete the confusion matrix. The confusion matrix shows how the predictions made by a method on a testing data set (in the current context, data on small molecules) compare to the known recorded interactions of these compounds. A two-class confusion matrix consists of a set of four tallies of the prediction results: the number of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) predictions (Figure 3A). Metrics to describe the performance of a method are then calculated using these entries. Importantly, the FP predictions may in fact include undiscovered or unreported interactions and may therefore be more precisely referred to as assumed FP predictions. Performance metrics generally do



**Figure 3.** (A) A binary classification confusion matrix with the four categories of prediction (FPs may include putative false positives); (B) ROC curves: the closer the curves are to the top left-hand corner, the better. AUC values alone may be deceptive as a lack of correct early predictions may be offset by an increased number of correct predictions later, leading to high AUC values. This scenario is shown by the green and purple curves. (C) Precision-recall curve: the closer the curve is to the top right corner, the better the model's performance.

not account for this kind of missing data, and it is therefore more appropriate to consider this component as potential FP predictions.

Two simple measures calculated from the confusion matrix are the model's sensitivity (SE) and specificity (SP). SE (also recall or TP rate) quantifies the model's ability to detect positive interactions and is the fraction of how many of the known positive interactions are identified by the target prediction method

$$SE = \frac{TP}{TP + FN} \quad (1)$$

SP, or TN rate, quantifies the model's ability to detect negative interactions and is the fraction of how many known, or assumed, negative interactions are identified by the prediction method

$$SP = \frac{TN}{TN + FP} \quad (2)$$

Precision (PR), or positive predictive value, quantifies how many of the predicted interactions are known interactions for a compound or a set of compounds

$$PR = \frac{TP}{TP + FP} \quad (3)$$

Accuracy (ACC) is a basic metric of the overall performance of binary classifiers that quantifies the proportion of correct predictions

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

A limitation of this metric is that it does not account for data set imbalance, which is a ubiquitous issue in target prediction, where data are often made up of a small number of recorded ligand-target interactions (positive class) and a large number of observed or assumed non-interactions (negative class). In this context, a target prediction method that correctly predicts most

non-interactions but fails to identify known positive interactions would obtain high ACC values, despite its inability to correctly identify the targets of small molecules [28].

A metric that does consider the proportion of all classes in the confusion matrix and therefore addresses the issue of imbalanced data is the Matthews Correlation Coefficient (MCC). The MCC quantifies the correlation between the predictions and their true value

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (5)$$

MCC values range from  $-1$  to  $+1$ , with  $+1$  indicating perfect prediction,  $0$  a prediction as good as random and  $-1$  a prediction that is in total disagreement with the measured data. Although the MCC is regarded as one of the most robust measures of the quality of binary classification, it is rarely used in target prediction. In the special case when a model predicts very few FPs and very few TPs at the same time, the MCC value will be deceptively high [29].

Other correlation metrics, such as Cohen's kappa ( $\kappa$ ) are sometimes used to measure the performance of a classifier. Cohen's kappa measures the similarity between two sets of classifications (in this case, the predicted classes and the known classes for interactions). Kappa quantifies how much better or worse a classifier is compared to random chance [30–32].

All metrics discussed so far aim at quantifying the ability of classifiers to discriminate interacting from non-interacting pairs of small molecules and biomacromolecules. However, rather than only predicting categories, most target prediction models return a score or probability that is used to rank predicted (non-) interactions. The ability of a target prediction method to recognize interacting pairs of ligands and targets and to rank them early in the hit list ('early recognition') is a key parameter for the goodness and value of such models. A straightforward and often used measure of early recognition is the top- $k$  metric,

which quantifies the percentage of compounds for which a defined number of known interactions is ranked among the top- $k$  positions. Statements such as ‘for  $X\%$  of all tested molecules, at least one known target was ranked among the top  $k$  targets’ are used to report performance. Note that the top- $k$  metric obviously depends on an arbitrary cut-off (the  $k$  value) and the number of targets considered for ranking, and it does not account for the statistical likelihood of random pick [33].

The receiver operating characteristic (ROC) curve is used to determine early enrichment, without an earliness cut-off. The ROC curve is an easily interpretable plot of the TP rate (SE) on the y-axis versus the FP rate (1-SP) on the x-axis, and it is drawn by calculating the cumulative positives and negatives as one moves down a rank-ordered list (Figure 3B) [34]. The closer a ROC curve approaches the top left corner of the graph, the better the rank-ordered list is, since TPs are identified early on, achieving early enrichment. A ROC curve that approaches the diagonal represents the random classification of small molecule and target pairs. Parts of the ROC curve located below the diagonal indicate a performance that is worse than random ranking.

The ROC curve considers both the correctly classified positive values (SE on the y-axis has TP in the numerator) and the correctly classified negative values (1-SP has TN in the denominator) and is therefore a good measure for balanced data sets [28, 35]. In contrast, the precision-recall curve plots PR (which has TP in the numerator) on the y-axis versus recall (which also has TP in the denominator) on the x-axis and is therefore ideal at visualizing how well positives appear at the top of the ranking, particularly when the data set has an imbalanced distribution between positives and negatives (Figure 3C) [28]. Unlike the ROC curve, the closer the precision-recall curve is to the top right edge, the better. The random classification of small molecule and target pairs results in a precision-recall curve that approaches the straight line, where PR is equal to the fraction of positives in the data set. Parts of the curve located below this line indicate a performance that is worse than random ranking.

The goodness of a classifier, as reflected by ROC and precision-recall curves (and others), can, in part, be quantified by the area under the curve (AUC). AUC values are bound between 1, for ideal models, and 0, for models that make predictions that are entirely the opposite of the recorded results. To draw conclusions about a model's early recognition ability, both AUC values and the original curve need to be considered, as models that perform differently with respect to early enrichment may have the same AUC since a lack of early recognitions may be offset by later recognitions (Figure 3B) [36, 37].

As the AUC metrics are not sensitive to early recognition, the robust initial enhancement (RIE) was developed as a single parameterized metric based on the enrichment factor (which is the factor by which known interactions are ranked more often within the top- $k$  predictions compared to random selection of  $k$  predictions)

$$RIE(\alpha) = \frac{\sum_{i=1}^n e^{-\alpha r_i/N}}{\left(\sum_{i=1}^n e^{-\alpha r_i/N}\right)_{random}} \quad (6)$$

The RIE uses a decreasing exponential weight to calculate how much better a ranked list of interactions is compared with the list with random distribution of the positive and negative targets [38, 39]. The RIE value is dependent on the early cut-off exponential parameter ( $\alpha$ ) and the ratio of positive interactions in the list, the product of which is the exponent component of the metric. RIE values therefore cannot be compared, unless the same cut-off and proportion of actives are present, making it harder to compare different methods [34, 39].

The Boltzmann-enhanced discrimination of ROC (BEDROC) metric, developed by Truchon et al. [34] for easier comparison, is the RIE metric scaled between 0 and 1, with 1 implying perfect prediction

$$BEDROC(\alpha) = \frac{RIE(\alpha) - RIE_{min}(\alpha)}{RIE_{max}(\alpha) - RIE_{min}(\alpha)} \quad (7)$$

A BEDROC value of 0.5 is when the observed cumulative distribution function (the cumulative number of actives versus the number of predictions in a rank-ordered list) has the same shape as the cumulative distribution function exponentially parameterized by the  $\alpha$  parameter. This allows BEDROC scores with the same  $\alpha$  parameter to be compared. The BEDROC metric is therefore more useful in discriminating a method's early recognition capabilities than an AUC due to the exponential weights and allows for easier comparison than the RIE metric [34, 39].

### Benchmark data sets for target prediction

Benchmark data sets can be useful for the comparative assessment of target prediction approaches. However, due to the complexities involved in compiling high-quality representative data sets, only few have been reported to date. One of the more widely used [22, 40, 41] benchmark data sets for target prediction is the Yamanashi data set [42], which was compiled from different sources and comprises 5127 drug-target interactions of 932 drugs and 989 targets for G protein-coupled receptors (GPCRs), ion channels, enzymes and nuclear receptors. Koutsoukas et al. [43] published a benchmark data set consisting of ~100 k compounds compiled from the ChEMBL database [44] used to compare the performance of different machine-learning algorithms [43]. Peón et al. [45] compiled two benchmarking data sets for their comparative study of ligand-centric methods for target prediction, one with 183 k active compounds with activities (EC50, Ki, Kd or IC50) below 10  $\mu$ M and one with 147 k active compounds with activities below 1  $\mu$ M. The data set used for externally testing SwissTargetPrediction has been made available for use as a benchmark [46]. Most recently, Wang and Kurgan [47] compiled and curated a data set from several different databases, consisting of 449 compounds, 1469 targets and 34 k interactions. One of a very few sources offering a complete data matrix of compounds tested against an array of different proteins is the kinase data set published by Davis et al. [48], which comprises 72 diverse kinase inhibitors measured against 442 kinases and was suggested by Pahikkala et al. [25] as a high-quality data set for testing target prediction methods. Two benchmark data sets specifically designed for testing structure-based methods have also been reported [49].

### Strategies for obtaining more realistic estimates of model performance

Rigorous validation schemes, involving external validation, in combination with information-rich performance metrics, quantify how well a method has generalized. However, the data employed for target prediction models are usually heavily biased. In opposition to reality, for example, chemical databases commonly have an overrepresentation of known actives compared to known inactives [10, 24, 26]. Established drug targets are much better represented by the available chemical, structural and biological data than other biomacromolecules [11, 50]. Additionally, the synthesizability of compounds and the fact that medicinal chemistry tends to generate congeneric

series of compounds lead to significant biases in the represented scaffolds [11, 51]. These biases are a natural result of the drug-development environment and lead to concentrations of information on certain targets and scaffolds.

Some targets are more challenging to predict than others due to the specific properties of individual targets or the structural and functional relationships between the biomacromolecules covered by a target prediction model. For example, due to its large and malleable ligand-binding site and no clear pharmacophoric requirements, cytochrome P450 (CYP) 3A4 binds to a broad variety of ligands [52, 53]. These properties mean that, despite the availability of a substantial body of structural, chemical and biological data, CYP3A4 is a particularly challenging target to address for both ligand and structure-based methods [54]. It is also much more difficult for target prediction methods to discriminate small-molecule activity among structurally and/or functionally related biomacromolecules. That is, it will be more challenging to correctly predict a protein kinase inhibitor's selectivity profile for kinases than it is to understand whether the compound will also bind to a certain GPCR. For all these reasons, the number of biologically tested compounds or the number of crystal structures by which a target is represented in the reference data is not the only factor that determines how difficult it is for a model to make predictions for a specific molecule or target.

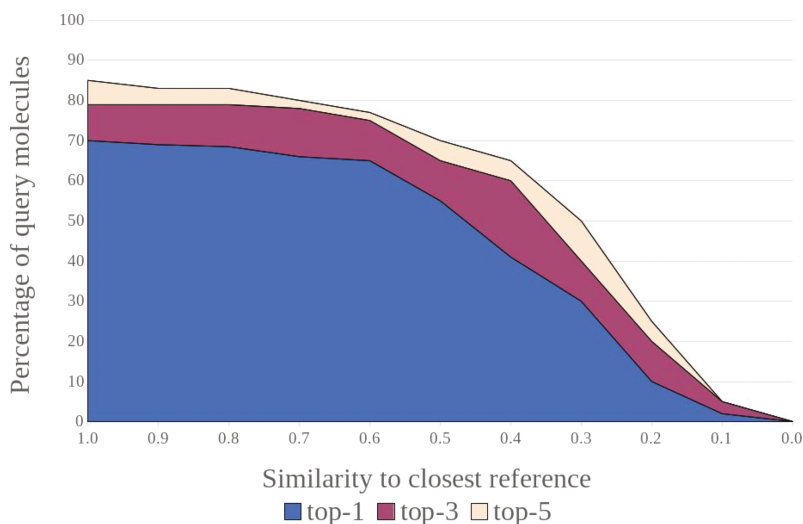
Given these data biases and challenges, it is clear that averaged performance metrics have limited significance as they obfuscate the predictive power of a method across queries and target classes. In fact, the individual characteristics of the targets and molecules covered by a target prediction model and by the testing set will determine the measured performance of a model. It is therefore generally not possible to directly compare results on model performance obtained from different studies as these usually use different data for model training and testing.

To obtain a more realistic representation of the performance of a target prediction model, a number of measures may be carried out to ameliorate the impact of the data and model biases:

- (i) A combination of metrics and methods that are more robust against the imbalance [10, 11, 55, 56] between known actives and inactives in the data set (e.g. precision-recall curve, PR AUC and the MCC) should be used for model testing. It is also useful to present the confusion matrices of the performance tests, so that further metrics may be calculated and used to compare methods.
- (ii) For any averaged performance metrics, their minima, maxima and distributions of values should be reported. A repeated validation scheme to calculate ROC curves would be useful in evaluating performance, as an average ROC curve with its confidence interval can be shown for assessment.
- (iii) Stratified sampling may be applied to construct more realistic data sets that mimic the real world, for training and testing. Caution must be exercised to ensure that oversampling of a class does not result in a model that is overfit.
- (iv) External data should be used for the evaluation of model performance.
- (v) In addition to a standard CV or nested CV, the performance of a model should also be evaluated using the various designed folds to establish performance estimates under conditions where there is no knowledge of the query molecule or target (Figure 2) in the training data.

- (vi) From a ligand perspective, building on established concepts in applicability domain research [45, 57–61], a weighted performance metric should be derived that is an improvement on the averaged metrics that quantify generalized performance. Such a metric would account for the difficulty of predicting the targets of individual query molecules as a function of the structural similarity between the query and the training instances (in the case of structure-based approaches, the similarity to the closest bound ligand may be used). Graphical approaches can be powerful tools to visualize such relationships, as shown by the example in Figure 4. These strategies can provide a better understanding of a method's capacity for inter- and extrapolation and help with the definition of the applicability domain.
- (vii) Performance metrics could also take into account the complexity of the (known) bioactive chemical space for the individual targets (in particular, in terms of size and diversity) as it is indicative of the number of ligand-binding pockets and subpockets, their size, shape, flexibility and specificity (in terms of pharmacophoric requirements).
- (viii) From a target perspective, a weighted performance metric could be used that takes into account the coverage and complexity of the conformational phase space relevant to ligand binding. Parameterizing such a performance metric is a non-trivial task, as in most cases the relevant conformational phase space remains unknown to a large extent. As an approximation, tools such as SIENA [62] may be used to automatically align protein-binding sites and quantify structural deviations among them.
- (ix) The druggability of a target, which is the likelihood of being able to modulate a target's activity with a small molecule [63, 64], may also be an indicator of how difficult it is, in particular for a docking algorithm, to make predictions for a specific target. Buried ligand-binding sites featuring hydrogen bond donors and acceptors are, for example, typically less challenging to address with small molecules than shallow hydrophobic interfaces on the protein surface (as often observed for protein-protein interaction interfaces) [65]. Docking algorithms show similar trends; ligand-binding sites that lack directed interactions or are solvent exposed are more challenging, for example.
- (x) The structural and functional relationships between the individual targets covered by a model should also be taken into account. TP predictions of targets that are related and therefore more challenging to discriminate should be assigned a higher weight than correct predictions for targets that are distinct. Likewise, a putative FP prediction of a target that is in agreement with activity recorded for a related target should be assigned a lower weight. Putative FP predictions are cases where compounds are predicted as active on a particular target, but no bioactivity data are available to confirm or refute this prediction. Given the low likelihood of a compound being active on a random biomacromolecule, for the purpose of evaluation, the general assumption made is that the compound is indeed inactive on that target. However, in the case of closely related targets there is a good chance that a compound confirmed to be active on one target is also active on the other. Ideally, the structural similarity of targets would be assessed based on the comparison of 3D structures of the ligand-binding sites. Given the complexities involved





**Figure 4.** Success rates for a target prediction model (e.g. percentage of compounds for which at least one known target was ranked among the top 1, top 3 and top 5 positions) versus the maximum similarity between the individual query compounds and their closest related compounds in the reference data. Such plots are powerful tools to visualize a method's capacity for inter- and extrapolation and help with the definition of the applicability domain.

in such comparisons, this is generally not feasible on a large scale. Instead, the sequence similarity of the protein domains involved in ligand binding may be used as a rough indication of the structural similarity of targets as perceived from a ligand's perspective.

- (xi) While there is no universal gold standard data set, evaluating a model's performance on benchmarking data sets will allow for easier comparison among methods.
- (xii) In addition to the many strategies involving statistical means, a critical discussion of representative examples can be very useful to better understand the scope and limitations of target prediction models. This could include comparing the performance of a model for well-represented versus underrepresented targets or highlighting the ability of a model to discriminate targets of a group of related biomacromolecules versus a group of distinct targets.

### Examples of how popular target prediction methods have been validated

Today, a large number of target prediction models are accessible via (mostly free) web services [2, 21, 50, 66–69]. The rigor applied in the evaluation of these methods varies greatly. For some models, their predictive power has been demonstrated by a small number of case studies (e.g. ChemMapper [70], Mantra [71, 72] and TarFisDock [73]). A substantial proportion of models have been evaluated on larger sets of data (e.g. ChemProt [74], CSNAP [75], DR. PRODIS [41], HitPick [76], Semantic Link Association Prediction (SLAP) [77], SuperPred [78] and TargetHunter [79]). Others have undergone systematic statistical validation by CV (e.g. SPiDER [80] and SwissTargetPrediction [81]). In one case, namely Similarity Ensemble Approach (SEA) [82], large-scale experimental evaluations have been reported. We describe four examples of popular target prediction models that have

undergone some of the most thorough validation experiments reported so far.

**SEA** (<http://sea.bkslab.org>) is an early ligand-based method that predicts the targets of small molecules based on their similarity to ligand sets of a reference database [82]. SEA has been tested through multiple rounds of prospective validation [82, 83]. The largest study reported so far is by Novartis and included the analysis of 1241 predicted interactions for 656 approved drugs. Of the predicted interactions, 348 were retrospectively verified. Further 694 predictions were experimentally tested, of which 48% were confirmed and 46% were disproved [84]. A number of studies have since used SEA [85–87] to identify, for example, the targets of the small molecule ogerin as the adenosine  $A_{2A}$  receptor and of SLV 320, an adenosine  $A_1$  antagonist, as an inhibitor of GPCR68 [88]. SEA has undoubtedly had the largest impact and use of all target prediction methods, and this can be attributed to its early development and the large-scale experimental testing by Novartis that is not typically feasible.

**SwissTargetPrediction** (<http://www.swisstargetprediction.ch>) is a ligand-based similarity method that uses both 2D fingerprints and 3D shape, combined in a logistic regression, to predict the likely targets of small molecules [81]. SwissTargetPrediction covers more than 2600 targets from five organisms (human, mouse, rat, cow and horse) and is arguably one of the most thoroughly statistically validated target prediction methods in existence [46]. The method also suggests the orthologs and paralogs of the predicted biomacromolecules as potential targets. SwissTargetPrediction was evaluated by a standard and two designed 10-fold CV runs. For the 1st designed CV run, molecules with similar scaffolds were incorporated into the same CV fold to estimate the performance of the method when the method is used with structurally distinct ligands [81]. This experiment was repeated using an additional 2nd filter to group molecules that were tested in the same assay within the same fold, thus reducing

the probability of a comparison of ligands from the same series [81]. For all the CV experiments, the folds were created to have 10 times as many negative interactions as positive ones, with the number of negative interactions supplemented by randomly pairing ligands and targets with no known positive interactions together. As expected, the performance of the method was lower for the designed CV runs (distinct scaffolds ROC AUC 0.979; distinct scaffolds and assays ROC AUC 0.932) than it was for the standard CV (ROC AUC 0.994). The effects of ligand properties (e.g. number of heavy atoms and lipophilicity) on the prediction accuracy were also investigated. In order to estimate the performance on new molecules, a 2nd external testing set that was composed of 213 molecules with 346 positive and 278 new interactions recorded in the consecutive version of the ChEMBL database. The testing set was expanded with randomly assigned ligands and targets to ensure that there were five times as many negative interactions than positive interactions in the testing set. On these data, the model obtained a ROC AUC of 0.87.

**SPiDER** (<http://modlab.cadd.ethz.ch/software/spider/>) [80] is a ligand-based method that utilizes self-organizing maps in combination with 'fuzzy' CATS pharmacophore descriptors [89] and Molecular Operating Environment (MOE) descriptors [90]. Validation of the method was carried out through a stratified 10-fold CV during which a prediction was considered successful if all known targets of a query were predicted within a defined significance threshold. The results from the CV were combined to calculate the ROC curve and ROC AUC value of 0.92 [80]. The capacity of SPiDER to predict the biomolecular targets of small molecules was demonstrated by a number of studies involving synthetic molecules [80, 91–94] as well as natural products [92, 95]

**SLAP** (<http://cheminfov.informatics.indiana.edu:8080/slap/>) is a network-based method that uses data from 17 sources and a semantic network linking the diverse and related data types (chemical compound, substructure, side effect, chemical ontology, target, disease, gene family, tissue, pathway and gene ontology) [77]. A chemical compound and a target are considered to be associated based on the defined path patterns, which include characteristics such as the length and the type of nodes involved in the paths between them. To evaluate the model's performance, four testing sets were compiled with known drug–target pairs from DrugBank and random drug–target pairs (serving as negative interactions), such that the ratio of positive and negative interactions was 1:1, 1:4, 1:8 and 1:12. The ROC AUCs (about 0.92 for all sets) and the precision–recall curves were reported for these tests, along with the performance measures by target class. SLAP was also evaluated on 23 confirmed drug–target pairs that were identified with SEA, and it was found that the method is not capable of identifying cross-boundary targets. In addition, SLAP was evaluated on 444 drug–target pairs recorded in MATA-DOR [96] (and not represented in the network) and successfully identified 170 of these interactions with high confidence.

## Conclusions

A plethora of *in silico* models have become available in recent years and are increasingly utilized to guide efforts to identify the biomacromolecular targets of small molecules. While the modeling approaches have come of age, there is room for further improvement in the validation of the methods. Ideally, target prediction methods would be tested in large-scale, prospective studies, but high expenses in terms of costs and time are, in general, prohibitive to such efforts. Therefore, developers and

users rely on robust retrospective (statistical) analyses. One of the most elaborate efforts of retrospective validation was published for SwissTargetPrediction, where a standard CV, two CVs with designed folds and a time-split approach were executed and analyzed in combination.

One of the most obvious deficits of current approaches to retrospective validation is their limitation to the global assessment of model performance, which can vary substantially for individual query molecules and targets as they are represented in the reference data to different extents. Here, the development of weighted scoring functions that account for the challenges involved in predicting the interaction of specific pairs of small molecules and biomacromolecules is desirable and urgently needed. A 2nd major limitation of current retrospective studies is their lack of comparability, which is a result of a lack of established, high quality, benchmark data sets and the complexities involved in the validation of target prediction models. It will take time for both of these issues to be resolved, but there are several immediate steps that can be taken to obtain more realistic estimates of model performance. As a minimum requirement, any target prediction method should undergo a systematic statistical validation. In particular, it is important for parameterized models to undergo external validation, and the results obtained from this test should be discussed with respect to the results obtained from internal validation. The discussion of representative test cases is desirable, e.g. the ability of a model to discriminate bioactivities of small molecules on structurally distinct targets in contrast to structurally related targets.

We submit that current reports on the performance of models often miss to convey the implications of the outcomes of statistical tests on the usefulness of target prediction methods under real-life conditions. In contrast to the common assumption made during model validation, investigators will most likely have prior knowledge of some biological properties of a compound. Armed with their expert knowledge they will often be able to identify false predictions. For the same reason, FP predictions on targets structurally related to the real target of a small molecule (e.g. predictions of activity on CYP1A2, whereas the compound actually is an inhibitor of CYP3A4 and not CYP1A2) can be useful as they may point researchers into the right direction, even though current validation approaches would commonly consider these predictions as false. It is also likely that investigators will have knowledge of several structurally related compounds exhibiting the same kind of biological activity rather than a singleton. By using multiple structurally related compounds as queries the signal-to-noise ratio can be improved. On the downside, in a real-life scenario, compounds of interest are likely to be more distant to the training data than the average compound of the testing set, which makes observing the applicability domain of a model an important issue.

Overall, we believe, and the recent reports in the literature show, that *in silico* models have become powerful tools to aid the identification of the mode of action of small molecules. We should not expect target prediction methods to generally be able to correctly rank the targets of a compound of interest among the top 1 or top 3 out of several hundreds or thousands of biomacromolecules. However, we are on a good track of developing models that are able to provide valuable guidance to experimentalists in their efforts to confirm the relevant targets of small molecules and to point out if a compound of interest is outside of the applicability domain of a model. This is a qualitative improvement to the challenging task of mode of action identification, and the increasing availability of chemical and biological data will lead to a further boost of theoretical methods for target prediction.



### Key Points

- *In silico* models have become important and powerful tools to efforts to identify the biomacromolecular targets of small molecules.
- Commonly followed strategies in assessing the performance of target prediction approaches do not adequately account for the heavy biases present in the chemical and biological data utilized for training and testing.
- A number of immediate measures can be taken to obtain more realistic estimates of the performance of target prediction models.
- New metrics that weigh the difficulty of individual predictions are urgently needed, as are benchmark data sets enabling the comparative performance analysis of target prediction methods.

### Acknowledgements

Dr Christoph Bauer from the University of Bergen is thanked for fruitful discussions and for proofreading the manuscript.

### Funding

Bergen Research Foundation (BFS2017TMT01 to N.M. and J.K); China Scholarship Council (201606010345 to Y.C.).

### References

- Moffat JG, Vincent F, Lee JA, et al. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;16:531–43.
- Chaudhari R, Tan Z, Huang B, et al. Computational polypharmacology: a new paradigm for drug discovery. *Expert Opin Drug Discov* 2017;12:279–91.
- Reddy AS, Zhang S. Polypharmacology: drug discovery for the future. *Expert Rev Clin Pharmacol* 2013;6:41–7.
- Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. *J Med Chem* 2014;57:7874–87.
- Proschak E, Stark H, Merk D. Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. *J Med Chem* 2019;62:420–44.
- Vanhaelen Q, Mamoshina P, Aliper AM, et al. Design of efficient computational workflows for *in silico* drug repurposing. *Drug Discov Today* 2017;22:210–22.
- March-Vila E, Pinzi L, Sturm N, et al. On the integration of *in silico* drug design methods for drug repurposing. *Front Pharmacol* 2017;8:298.
- Hodos RA, Kidd BA, Shameer K, et al. *In silico* methods for drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med* 2016;8:186–210.
- Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;18:41–58.
- Ezzat A, Wu M, Li X-L, et al. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 2018;2018:bby002.
- Cortés-Ciriano I, Ain QU, Subramanian V, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* 2015;6:24–50.
- Reker D, Schneider P, Schneider G, et al. Active learning for computational chemogenomics. *Future Med Chem* 2017;9:381–402.
- Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* 2014;6:47.
- Krstajic D, Buturovic LJ, Leahy DE, et al. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform* 2014;6:10.
- Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91.
- Guney E. Revisiting cross-validation of drug similarity based classifiers using paired data. *Genomics Comput Biol* 2017;4:e100047.
- Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J Chem Inf Model* 2013;53:783–90.
- Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;9:1134–6.
- Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv* 2010;4:40–79.
- Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, et al. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct Chem* 2011;22:795–804.
- Cereto-Massagué A, Ojeda MJ, Valls C, et al. Tools for *in silico* target fishing. *Methods* 2015;71:98–103.
- Hao M, Bryant SH, Wang Y. Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief Bioinform* 2018;2018:bby010.
- Li J, Zheng S, Chen B, et al. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;17:2–12.
- Chen X, Yan CC, Zhang X, et al. Drug–target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016;17:696–712.
- Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug–target interaction predictions. *Brief Bioinform* 2015;16:325–37.
- Brown AS, Patel CJ. A review of validation strategies for computational drug repositioning. *Brief Bioinform* 2018;19:174–7.
- Martin EJ, Polyakov VR, Tian L, et al. Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration ICs for realistically novel compounds. *J Chem Inf Model* 2017;57:2077–88.
- Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min* 2017;10:35.
- Baldi P, Brunak S, Chauvin Y, et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;16:412–24.
- Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257–68.
- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159.
- Zaretski J, Bergeron C, Rydberg P, et al. RS-predictor: a new tool for predicting sites of cytochrome P450-

- mediated metabolism applied to CYP 3A4. *J Chem Inf Model* 2011;**51**:1667–89.
34. Truchon J-F, Bayly CL. Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem. *J Chem Inf Model* 2007;**47**:488–508.
  35. Prati RC, Gustavo EAP, Monard MC. A survey on graphical methods for classification predictive performance evaluation. *IEEE Trans Knowl Data Eng* 2011;**23**:1601–18.
  36. Zhao W, Hevener KE, White SW, et al. A statistical framework to evaluate virtual screening. *BMC Bioinformatics* 2009;**10**:225.
  37. Kirchmair J, Markt P, Distinto S, et al. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? *J Comput Aided Mol Des* 2008;**22**:213–28.
  38. Sheridan RP, Singh SB, Fluder EM, et al. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J Chem Inf Comput Sci* 2001;**41**:1395–406.
  39. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform* 2013;**5**:26.
  40. Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform* 2014;**15**:734–47.
  41. Zhou H, Gao M, Skolnick J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci Rep* 2015;**5**:11090.
  42. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**:i232–40.
  43. Koutsoukas A, Lowe R, Kalantarmotamedi Y, et al. In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. *J Chem Inf Model* 2013;**53**:1957–66.
  44. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res* 2017;**45**:D945–54.
  45. Peón A, Dang CC, Ballester PJ. How reliable are ligand-centric methods for target fishing? *Front Chem* 2016;**4**:15.
  46. Gfeller D, Grosdidier A, Wirth M, et al. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res* 2014;**42**:W32–8.
  47. Wang C, Kurgan L. Review and comparative assessment of similarity-based methods for prediction of drug-protein interactions in the druggable human proteome. *Brief Bioinform* 2018;**2018**:bby069.
  48. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51.
  49. Schomburg KT, Rarey M. Benchmark data sets for structure-based computational target prediction. *J Chem Inf Model* 2014;**54**:2261–74.
  50. Lavecchia A, Cerchia C. In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov Today* 2016;**21**:288–98.
  51. Katsila T, Spyroulias GA, Patrinos GP, et al. Computational approaches in target identification and drug discovery. *Comput Struct Biotechnol J* 2016;**14**:177–84.
  52. Kirchmair J, Göller AH, Lang D, et al. Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov* 2015;**14**:387–404.
  53. Mustafa G, Yu X, Wade RC. Structure and dynamics of human drug-metabolizing cytochrome P450 enzymes. In: Kirchmair J (ed). *Drug Metabolism Prediction*, Vol. 38. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2014, 75–102.
  54. Leach AG, Kidley NJ. Cytochrome P450 substrate recognition and binding. In: Kirchmair J (ed). *Drug Metabolism Prediction*, Vol. 39. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2014, 103–32.
  55. Mousavian Z, Masoudi-Nejad A. Drug-target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin Drug Metab Toxicol* 2014;**10**:1273–87.
  56. Cheng T, Hao M, Takeda T, et al. Large-scale prediction of drug-target interaction: a data-centric review. *AAPS J* 2017;**19**:1264–75.
  57. Roy K, Ambure P, Kar S. How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* 2018;**3**:11392–406.
  58. Liu R, Glover KP, Feasel MG, et al. General approach to estimate error bars for quantitative structure-activity relationship predictions of molecular activity. *J Chem Inf Model* 2018;**58**:1561–75.
  59. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern Lab Anim* 2005;**33**:445–59.
  60. Netzeva TI, Worth A, Aldenberg T, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern Lab Anim* 2005;**33**:155–73.
  61. Mathea M, Klingspohn W, Baumann K. Chemoinformatic classification methods and their applicability domain. *Mol Inform* 2016;**35**:160–80.
  62. Bietz S, Rarey M. SIENA: efficient compilation of selective protein binding site ensembles. *J Chem Inf Model* 2016;**56**:248–59.
  63. Owens J. Determining druggability. *Nat Rev Drug Discov* 2007;**6**:187–7.
  64. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov* 2002;**1**:727–30.
  65. Guo Z, Li B, Cheng L-T, et al. Identification of protein-ligand binding sites by the level-set variational implicit-solvent approach. *J Chem Theory Comput* 2015;**11**:753–65.
  66. Koutsoukas A, Simms B, Kirchmair J, et al. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J Proteomics* 2011;**74**:2554–74.
  67. Kurgan L, Wang C. Survey of similarity-based prediction of drug-protein interactions. *Curr Med Chem* 2018. DOI: 10.2174/0929867325666181101115314.
  68. Sam E, Athri P. Web-based drug repurposing tools: a survey. *Brief Bioinform* 2017;**2017**:bbx125.
  69. Murtazaliev KA, Druzhilovskiy DS, Goel RK, et al. How good are publicly available web services that predict bioactivity profiles for drug repurposing? *SAR QSAR Environ Res* 2017;**28**:843–62.
  70. Gong J, Cai C, Liu X, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics* 2013;**29**:1827–9.
  71. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A* 2010;**107**:14621–6.

72. Carrella D, Napolitano F, Rispoli R, et al. Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics* 2014;**30**: 1787–8.
73. Li H, Gao Z, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* 2006;**34**:W219–24.
74. Kringelum J, Kjaerulff SK, Brunak S, et al. ChemProt-3.0: a global chemical biology diseases mapping. *Database* 2016;**2016**:bav123.
75. Lo Y-C, Senese S, Li C-M, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comput Biol* 2015;**11**: e1004153.
76. Liu X, Vogt I, Haque T, et al. HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics* 2013;**29**:1910–2.
77. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. *PLoS Comput Biol* 2012;**8**:e1002574.
78. Nickel J, Gohlke B-O, Erehman J, et al. SuperPred: update on drug classification and target prediction. *Nucleic Acids Res* 2014;**42**:W26–31.
79. Wang L, Ma C, Wipf P, et al. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J* 2013;**15**:395–406.
80. Reker D, Rodrigues T, Schneider P, et al. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A* 2014;**111**:4067–72.
81. Gfeller D, Michielin O, Zoete V. Shaping the interaction landscape of bioactive molecules. *Bioinformatics* 2013;**29**:3073–9.
82. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**:197–206.
83. Keiser MJ, Setola V, Irwin JJ, et al. Predicting new molecular targets for known drugs. *Nature* 2009;**462**:175–81.
84. Lounkine E, Keiser MJ, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 2012;**486**:361–7.
85. Mugumbate G, Abrahams KA, Cox JAG, et al. Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. *PLoS One* 2015;**10**:e0121492.
86. Yee SW, Lin L, Merski M, et al. Prediction and validation of enzyme and transporter off-targets for metformin. *J Pharmacokinetic Pharmacodyn* 2015;**42**:463–75.
87. Laggner C, Kokel D, Setola V, et al. Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat Chem Biol* 2011;**8**:144–6.
88. Huang X-P, Karpiak J, Kroeze WK, et al. Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65. *Nature* 2015;**527**:477–83.
89. Schneider G, Neidhart W, Giller T, et al. 'Scaffold-Hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed Engl* 1999;**38**: 2894–6.
90. Molecular Operating Environment. [https://www.chemcomp.com/MOE-Molecular\\_Operating\\_Environment.html](https://www.chemcomp.com/MOE-Molecular_Operating_Environment.html) (14 November 2018, date last accessed)
91. Brand S, Roy S, Schröder P, et al. Combined proteomic and in silico target identification reveal a role for 5-lipoxygenase in developmental signaling pathways. *Cell Chem Biol* 2018;**25**:1095–106 e23.
92. Merk D, Grisoni F, Friedrich L, et al. Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J Med Chem* 2018;**61**: 5442–7.
93. Kremer L, Schultz-Fademrecht C, Baumann M, et al. Discovery of a novel inhibitor of the hedgehog signaling pathway through cell-based compound discovery and target prediction. *Angew Chem Int Ed Engl* 2017;**56**:13021–5.
94. Merk D, Grisoni F, Friedrich L, et al. Scaffold hopping from synthetic RXR modulators by virtual screening and design. *MedChemComm* 2018;**9**:1289–92.
95. Rodrigues T, Sieglitz F, Somovilla VJ, et al. Unveiling (–)-englerin A as a modulator of L-type calcium channels. *Angew Chem Int Ed Engl* 2016;**55**:11077–81.
96. Günther S, Kuhn M, Dunkel M, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;**36**:D919–22.

# Chapter 5

## Development and validation of large-scale target prediction methods

A plethora of target prediction methods exist to predict the activity of compounds on macromolecular target. Target prediction is a key strategy in early drug discovery. The predictions made by a target prediction method help guide experimental research and ameliorate risks in the drug discovery process. The performance of target prediction methods are often reported as averages over all test queries. This leaves questions unanswered about the value of the different approaches to target prediction.

To understand the value of ligand-based target prediction methods, we developed two target prediction approaches with a wide target coverage: first, a similarity-based approach and second, a ML approach. Our ML approach formulates target prediction as a binary relevance problem where independent random forest classifiers are trained for each target. The similarity-based and ML approaches were selected for investigation as the data requirements for these approaches allowed for a large coverage of the target space, which is important in a general-purpose target prediction method. Both approaches were built and tested using data from the ChEMBL database (versions 24 and 25). Morgan fingerprints with a radius of 2 and a length of 2,048 bits were used as the descriptors of the compounds. The hyperparameters of the individual random forest classifiers were tuned using internal cross-validation and a grid search protocol with the MCC metric measure of performance used to select the optimal classifier.

To measure performance, we benchmarked the approaches using the top-k metric disaggregated by the similarity of the test data to the knowledge base of the approach (as

laid out in Chapter 4). The performance of the approaches was measured under three different testing scenarios. In the first testing scenario, the standard testing scenario, the approaches were tested on a large external data set. Under the second testing scenario, the standard time-split testing scenario, the approaches were tested on external data that was generated a year after the data used to train the models. In both these testing scenarios, all the test compounds had targets represented by the approaches. Finally, in the third testing scenario, called the close-to-real-world testing scenario, all the new compounds added to the ChEMBL database were used as test compounds, even if their targets were not represented by the approaches' knowledge base.

Our initial hypothesis was that the ML approach would begin to outperform the similarity-based approach for test compounds that were more distant to the underlying knowledge base. However, we found that this was not the case, and the similarity-based approach generally outperformed this ML approach under all three testing scenarios while having a larger coverage of the target space (4,239 targets for the similarity-based approach vs. 1,798 targets for the ML approach). Under the standard testing scenario, the similarity-based approach ranked a correct target among the top-5 targets for 88% of the queries while the ML approach ranked a correct target among the top-5 targets for 85% of the queries. In fact, for high-similarity queries, the similarity-based approach ranked a target in the top position for 95% of the queries, while the ML approach ranked the target in the top position for 90% of the queries. Under the standard time-split scenario, the overall top-5 success rates dropped by 25% for both approaches, due to an increase in low-similarity queries compared to high-similarity queries. When looking at disaggregated performances however, the performances for the standard testing scenario and the standard time-split testing scenario were comparable. Under the close-to-real-world testing scenario, the similarity-based approach had a top-5 success rate of 59%, which was 52% for the ML approach. This is due to the larger target coverage of the similarity-based approach.

---

## **P2: Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope**

Neann Mathai and Johannes Kirchmair

*International Journal of Molecular Sciences*, **21(10)**, 3585 (2020).

<https://doi.org/10.3390/ijms21103585>

### **Contributions:**

N. Mathai and J. Kirchmair conceptualized the research. N. Mathai conducted the research and analysis. N. Mathai wrote the manuscript, with contributions from J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted from: Mathai N.; J. Kirchmair, Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope *Int. J. Mol. Sci.* **2020**, *21*, 3585.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).



Article

# Similarity-Based Methods and Machine Learning Approaches for Target Prediction in Early Drug Discovery: Performance and Scope

Neann Mathai <sup>1</sup> and Johannes Kirchmair <sup>1,2,\*</sup>

<sup>1</sup> Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway; neann.mathai@uib.no

<sup>2</sup> Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

\* Correspondence: johannes.kirchmair@univie.ac.at

Received: 27 April 2020; Accepted: 16 May 2020; Published: 19 May 2020



**Abstract:** Computational methods for predicting the macromolecular targets of drugs and drug-like compounds have evolved as a key technology in drug discovery. However, the established validation protocols leave several key questions regarding the performance and scope of methods unaddressed. For example, prediction success rates are commonly reported as averages over all compounds of a test set and do not consider the structural relationship between the individual test compounds and the training instances. In order to obtain a better understanding of the value of ligand-based methods for target prediction, we benchmarked a similarity-based method and a random forest based machine learning approach (both employing 2D molecular fingerprints) under three testing scenarios: a standard testing scenario with external data, a standard time-split scenario, and a scenario that is designed to most closely resemble real-world conditions. In addition, we deconvoluted the results based on the distances of the individual test molecules from the training data. We found that, surprisingly, the similarity-based approach generally outperformed the machine learning approach in all testing scenarios, even in cases where queries were structurally clearly distinct from the instances in the training (or reference) data, and despite a much higher coverage of the known target space.

**Keywords:** target prediction; molecular similarity; machine learning; random forest; molecular fingerprints; drug discovery

## 1. Introduction

Computational methods for predicting the macromolecular targets of small molecules have become increasingly relevant and popular in recent years due to (i) the shift from the “one-drug-one-target” paradigm to “polypharmacology” [1–5], (ii) the increasing availability of chemical and biological data [6–8] and (iii) advances in algorithms and hardware technology. Depending on the types of utilized data, in silico methods for target prediction may be categorized as ligand-based, structure-based, or hybrid methods [9–12]. Ligand-based methods range from straightforward similarity-based approaches [13–21] and linear regressions [22] to more complex machine learning (ML) models such as random forests [23–25], support vector machines [25–27], self-organizing maps [28], neural and deep neural networks [25,29–34], and network-based models [35–38]. They typically use large amounts of chemical information and measured bioactivity data [12] and, as a result, have a larger coverage of the target space when compared to structure-based methods, which rely on 3D structures of macromolecules. The third type of methods, hybrid approaches such as proteochemometrics and network-based approaches, utilize chemical, biological and structural information for target prediction.

Despite the abundance of in silico methods and models for target prediction that have been published in recent years, our understanding of their value and scope under (close to) real-world

conditions remains limited [39]. In an ideal scenario, the performance of a model would be determined by large-scale prospective validation. However, the efforts and costs involved in running experiments on scales that can yield statistically meaningful conclusions are in general prohibitive. In consequence, to the best of our knowledge, the Similarity Ensemble Approach (SEA) method remains the only target prediction model that has undergone systematic experimental validation [40–42].

Most studies of new target prediction models are limited to retrospective validation [39]. Clearly, recent years have seen substantial progress in the implementation of more robust validation techniques of this kind, but one important aspect missed by most investigations is the relationship between the accuracy and reliability of predictions as a function of the distance between the compound of interest (query molecule) and the training data. In other words, reported validation studies often give a good idea of how well a model performs on the “average compound” originating from a defined dataset, but not so much about how trustworthy a prediction is for a particular compound of interest, which may or may not be structurally closely related to any of the instances in the training data. A further relevant point that is often not given the necessary consideration is target space coverage. Models trained and applied to targets for which a rich body of data is available will likely produce better performance statistics than models aiming to cover a wide target space. From the perspective of the end user, the most important question will be which method produces the most reliable predictions while covering the largest possible target space, and this question has been generally left unanswered.

This work aims to establish the value of two of the most common types of ligand-based methods for target prediction under conditions that closely resemble real-life applications: a straightforward similarity-based approach and a random forest-based ML approach, both employing Morgan2 fingerprints as representations of molecular structures. In particular, we investigate how the structural relationship between a query molecule and the molecules used for model training impact the reliability of predictions, and to what extent the individual approaches are able to cover the known target space.

## 2. Results and Discussion

### 2.1. Similarity-Based Method and Machine Learning Approach for Target Prediction

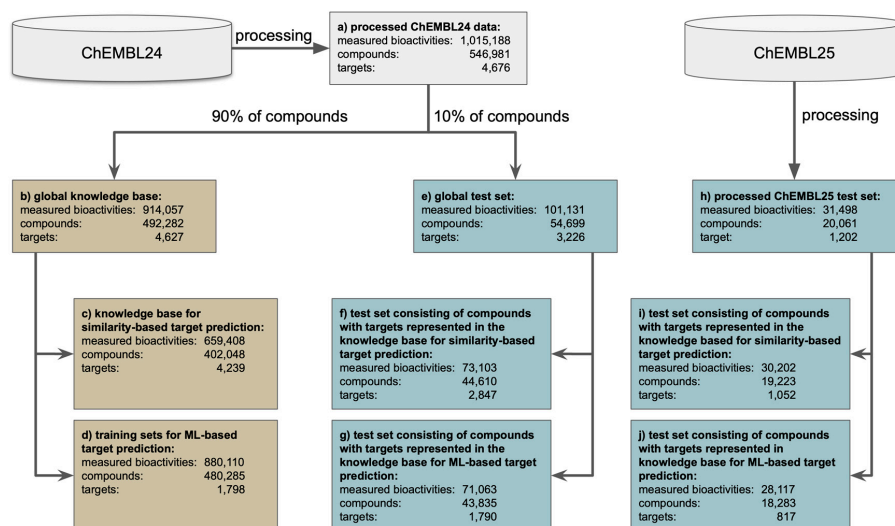
Bioactivity data for model building and validation was extracted from the ChEMBL database [43] version 24. These data were curated and processed (see Methods section for details), resulting in a “processed dataset” consisting of 1,015,188 compound-protein pairs (546,981 unique compounds and 4676 unique targets; Figure 1a). Compound-protein pairs with an activity value less than or equal to 10,000 nM were marked as “active” (732,570 bioactivities) while those with activities greater than or equal to 20,000 nM were marked as “inactive” (282,618 bioactivities). Prior to any model development, the compounds in the processed dataset (Figure 1a) were randomly assigned to a “global knowledge base” (Figure 1b) or a “global test set” (Figure 1e) at a ratio of 90:10.

The similarity-based approach uses the maximum pairwise similarities (Tanimoto coefficients (TC) derived from Morgan2 fingerprints; maxTCs) between a query molecule and the sets of ligands representing the 4239 individual proteins in the knowledge base (Figure 1c) to produce a rank-ordered list of potential targets (Figure 2). In cases where multiple proteins have the same maxTCs, the next highest TCs are considered until all proteins are ranked.

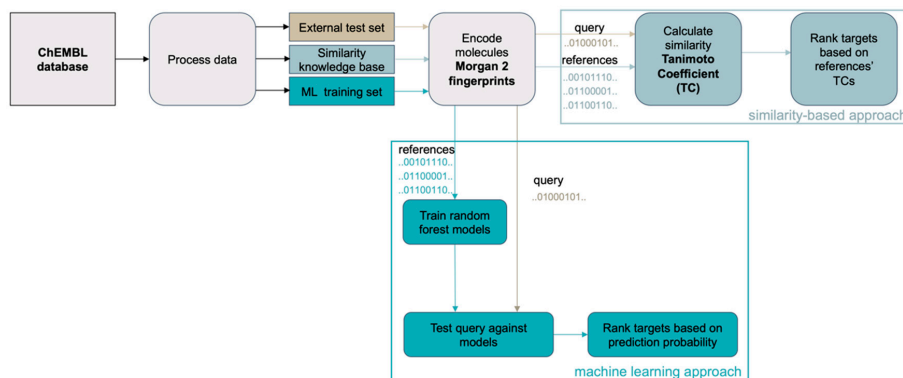
The ML approach decomposes the multi-label problem (i.e., a single query molecule may interact with many proteins) into a series of binary classification problems (i.e., a query molecule does or does not interact with a particular protein). This technique, known as binary relevance, is an intuitive and popular transformation [44] in target prediction [24,25,45]. Here, a query molecule is tested on all the target models individually and the models’ prediction probability of the active class ( $p$ -values) are then used to rank the potential targets for the query molecule (Figure 2). More specifically, random forest models were generated for each of the 1798 targets represented by a minimum of 25 ligands in the global knowledge base (Figure 1d). The individual models were trained on all active and all inactive compounds recorded for a target in the global knowledge base. Following a



widely-applied approach in target prediction for expanding chemical space coverage [9,18,33,45,46], all training sets for ML for which the number of confirmed inactive compounds did not exceed the number of confirmed active compounds by a factor of 10 (this was the case for 1793 out of the 1798 targets) were supplemented with presumed inactive compounds (i.e., randomly chosen compounds from the global knowledge base which do not have any annotation for the particular target) to give a balance of 10:1. The hyperparameters of the individual random forest classifiers were optimized during a grid search within a cross-validation framework (see Method section for details).



**Figure 1.** Creation of the knowledge base/training sets (brown) and test sets (blue) from version 24 (ChEMBL24) and version 25 (ChEMBL25) of the ChEMBL database. Measured bioactivities are counted as individual records (unique compound-protein pairs), compounds are counted as unique canonical SMILES of the preprocessed and standardized structures, and targets are counted as unique ChEMBL target IDs. The number of compounds reported in box d) does not include the presumed inactive compounds used to create the individual training sets for the generation of the ML models.

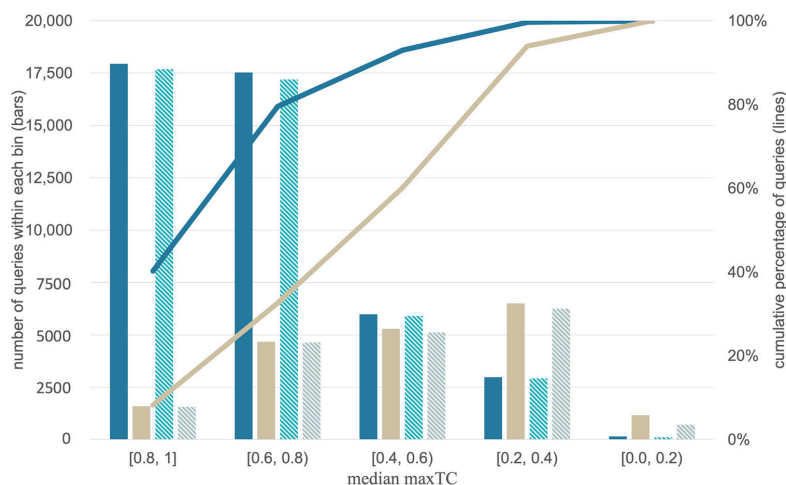


**Figure 2.** General workflow for the development and validation of the similarity-based method and the machine learning approach for target prediction.

## 2.2. Evaluation of the Scope and Performance of the Similarity-Based Method and Machine Learning Approach for Target Prediction

The scope and limitations of the individual approaches are evaluated under the following validation settings:

1. **Standard testing scenario with an external test set.** Under this scenario, the approaches are tested for their ability to predict the targets of a set of approximately 44,000 query molecules (Figure 1f,g) obtained by a single random split of the processed ChEMBL24 database prior to model development.
2. **Standard time-split validation scenario within the target space covered by the approach.** Under this scenario, the models are tested on the more than 18,000 molecules that have been newly introduced with version 25 of the ChEMBL database and have targets within the target space of the individual models (meaning that all compounds considered as queries have at least one known target that is covered by the approach's knowledge base; Figure 1i,j). This test can give a sense of how model performance will change over time [39], with the increasing alienation of chemistry from that represented by the knowledge base (as observed in Figure 3).



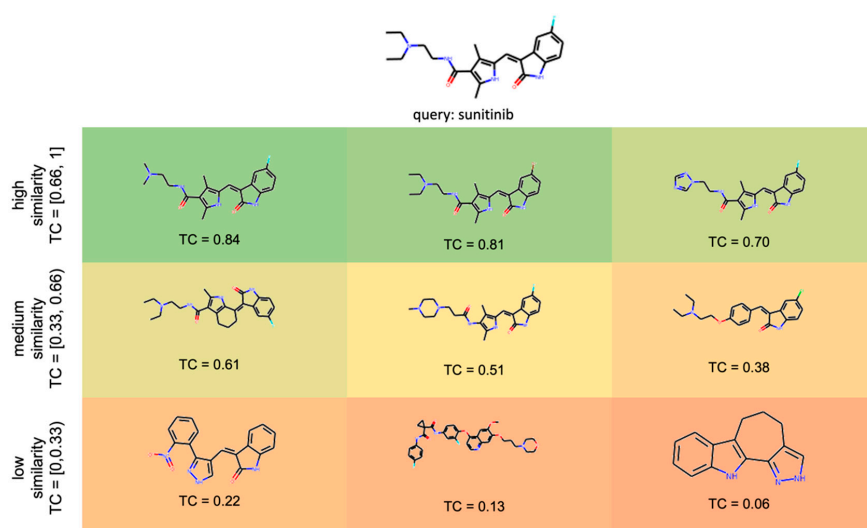
**Figure 3.** Distributions of the median maxTC values (quantifying the median molecular similarity of each query molecule and its nearest ligand of each of the query molecule's annotated targets) for the queries from the ChEMBL24 database (blue) and ChEMBL25 database (ocher and grey). The results of the similarity-approach are marked without a pattern; the results for the ML approach are shown with a pattern. The bars represent the number of queries within a median maxTC bin. The distributions show that the ChEMBL24 test set is more similar to the data of the knowledge base than the ChEMBL25 test set (which is expected as the knowledge base is a subset of the ChEBML24 database). The lines report the cumulative percentage of queries with median maxTCs greater than or equal to the values covered by a bin. For the sake of clarity, the lines are only shown for the similarity-based approach as they are almost identical with the lines for the ML approach.

3. **Close-to-real-world setting with an unbiased and comprehensive time-split dataset.** Under this scenario the methods were tested on the full set of bioactive compounds newly introduced with version 25 of the ChEMBL database (20,061 compounds; Figure 1h), regardless of whether or not any of the annotated targets is covered by the approach's knowledge base. This scenario comes closest to real-world applications of target prediction methods, as there is a good chance that the targets of the new compounds (in particular those based on new chemistry) are novel, not represented by the training data, and hence missed by the *in silico* models.

Based on our experience in working with TCs derived from Morgan2 (and some related) fingerprints, we distinguish the following classes of queries:

- “High similarity queries”: These queries share a high degree of molecular similarity with the closest ligand (of the same target) in the knowledge base (TC greater than 0.66). Chemists will identify queries of this class as structurally closely related to the nearest ligand (of the respective target) in the knowledge base.
- “Medium similarity queries”: These queries share a moderate degree of molecular similarity with the closest ligand (of the same target) in the knowledge base (TC between 0.33 and 0.66). Chemists will typically find it challenging to identify obvious similarities between a query molecule of this class and the nearest ligand (of the respective target) in the knowledge base.
- “Low similarity queries”: These queries share a low degree of molecular similarity with the closest ligand in the knowledge base (TC lower than 0.33). Chemists will unlikely identify a query molecule of this class as structurally related to any of the ligands (of the respective target) in the knowledge base.

For the sake of clarity, the discussion of the methods’ performance focuses on these three categories (Figure 4 shows examples of different TCs of a query molecule and knowledge base ligands). A more fine-grained view is provided by the graphs and supplementary tables, which use a bin width for the TC of 0.2 for the disaggregation of the results. The trends observed at both levels of granularity are consistent.



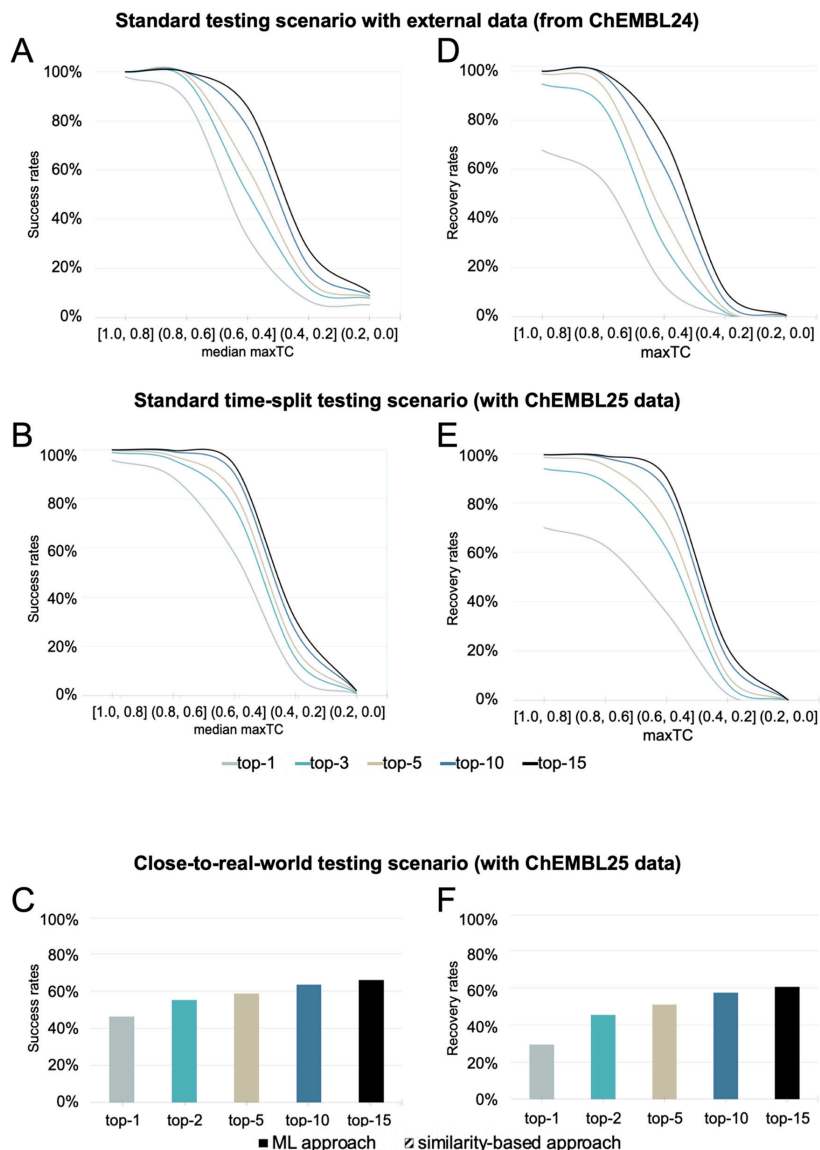
**Figure 4.** Query molecule (sunitinib) with examples of high, low, and medium similarity compounds from the ligand sets of its targets.

### 2.2.1. Evaluation of the Scope and Performance of the Similarity-Based Method

#### Performance in a Standard Testing Scenario with an External Test Set

Overall, the similarity-based approach achieved high success rates, ranking at least one known target among the top-3, top-5, and top-15 positions in 86%, 88%, and 93% of all cases (Figure 5A). The success rates were found to be strongly linked with the distance between the query molecule and the nearest ligand (for that target) in the knowledge base. For 95% of all high similarity queries (as defined in the introductory section of the Results and Discussion section), the protein ranked at the top

position was a known target. For medium similarity queries, however, the success rates were only 55%, 63%, and 82% when considering the top-3, top-5, and top-15 ranks, respectively. For low similarity queries, the success rates dropped to 10% (top-3), 12% (top-5), and 18% (top-15), respectively.



**Figure 5.** Success rates (A–C) and recovery rates (D–F) of the similarity-based approach under (A,D) the standard testing scenario with external data, (B,E) the standard time-split testing scenario, and (C,F) the close-to-real-world testing scenario. As expected, the performance under all testing scenarios drops as the queries become increasingly dissimilar to the data underlying the models. The data for these graphs are also provided in tabular format in the Supporting Information (Table S1–S4).

The similarity-based approach obtained good overall recovery rates (Figure 5D), with a strong correlation of performance and the distance between the query molecule and the compounds in the knowledge base observed here as well. The recovery rate is defined as the percentage of known bioactivities ranked among the top-*k* positions of the list of predicted targets. For high similarity queries, 92%, 97%, and 100% of the known targets were ranked among the top-3, top-5, and top-15 positions, respectively. In contrast, for medium similarity queries, the recovery rates were only 35%, 45%, and 70% when considering the top-3, top-5, and top-15 ranks, respectively. For low similarity queries, these success rates dropped to 1% (top-3), 1% (top-5), and 3% (top-15).

#### Performance in a Standard Time-Split Testing Scenario

As expected, the success and recovery rates obtained for the similarity-based approach on the set of compounds newly introduced with version 25 of the ChEMBL database (“ChEMBL 25 test set”) were generally lower than for the ChEMBL24 test set (Figure 5B). The overall success rates among the top-3, top-5, and top-15 positions were 58%, 61%, and 69% (vs. 86%, 88%, and 93% obtained for the ChEMBL24 test set, see above). However, the success rates for queries represented by structurally related ligands in the knowledge base were comparable with those obtained for the ChEMBL24 test data: for high similarity queries, a known target was ranked at the top position in 93% of all cases (vs. 95% obtained for the ChEMBL24 test set). This is contrasted by the performance on medium similarity queries which had success rates of 70%, 76%, and 88% for the ChEMBL25 test set when considering the top-3, top-5, and top-15 ranks (vs. 55%, 63%, and 82% obtained for the ChEMBL24 test set). The success rates of the low similarity queries from the ChEMBL 25 test set dropped to 5%, 7%, and 13% for the top-3, top-5, and top-15 ranks (compared to 10%, 12%, and 18% for the ChEMBL 24 test set).

In accordance with the trends observed for the success rates, the recovery rates were also lower for queries from the new data in the ChEMBL25 database. Only 47%, 53%, and 63% of the known interactions were recovered among the top-3, top-5, and top-15 targets, as opposed to a recovery rate of 72%, 79%, and 87% obtained on the ChEMBL24 test set. Again, interactions with queries, which are more structurally related to the knowledge base, had higher recovery rates than those that were more distant (Figure 5E).

#### Performance in a Close-to-Real-World Testing Scenario

In the close-to-real-world testing scenario, an additional 1296 interactions (of 838 query molecules) not represented by the knowledge base were considered in the performance assessment. The 1296 interactions correspond to 4% of all interactions newly introduced with version 25 of the ChEMBL database. In consequence, the overall success rates for the top-3, top-5, and top-15 predictions decreased to 55%, 59%, and 66%, which represents a drop by 2 to 3 percentage points compared to the standard time-split scenario (Figure 5C). Likewise, the recovery rates for the top-3, top-5, and top-15 predictions dropped to 45%, 51%, and 61%, which is a decrease by 2 to 3 percentage points compared to the standard time-split scenario (Figure 5F).

#### Scope of the Similarity-Based Approach

The similarity-based approach has the widest scope of the approaches as any target with at least one known annotated ligand is represented in the knowledge base. The similarity-based approach covers a total of 4239 targets.

#### 2.2.2. Evaluation of the Scope and Performance of the Machine-Learning Approach

In the following subsections, the performance of the ML is discussed and directly compared to the performance of the similarity-based approach. For the sake of direct comparability, in the following discussion all statements on the performance of the similarity-based approach refer to its application to the reduced target space of the ML approach (1798 proteins) rather than the full target space covered by

the similarity-based approach (4239 proteins). Note that, importantly, the target ranking performance of the similarity-based approach on the full target scope is almost identical to that on the reduced target scope, meaning that there is no noticeable drop in performance in the top-k success and recovery rates of the similarity-based approach (using identical absolute values for k) when applied to a target space of 4239 proteins instead of 1798, which is remarkable.

#### Performance in a Standard Testing Scenario with an External Test Set

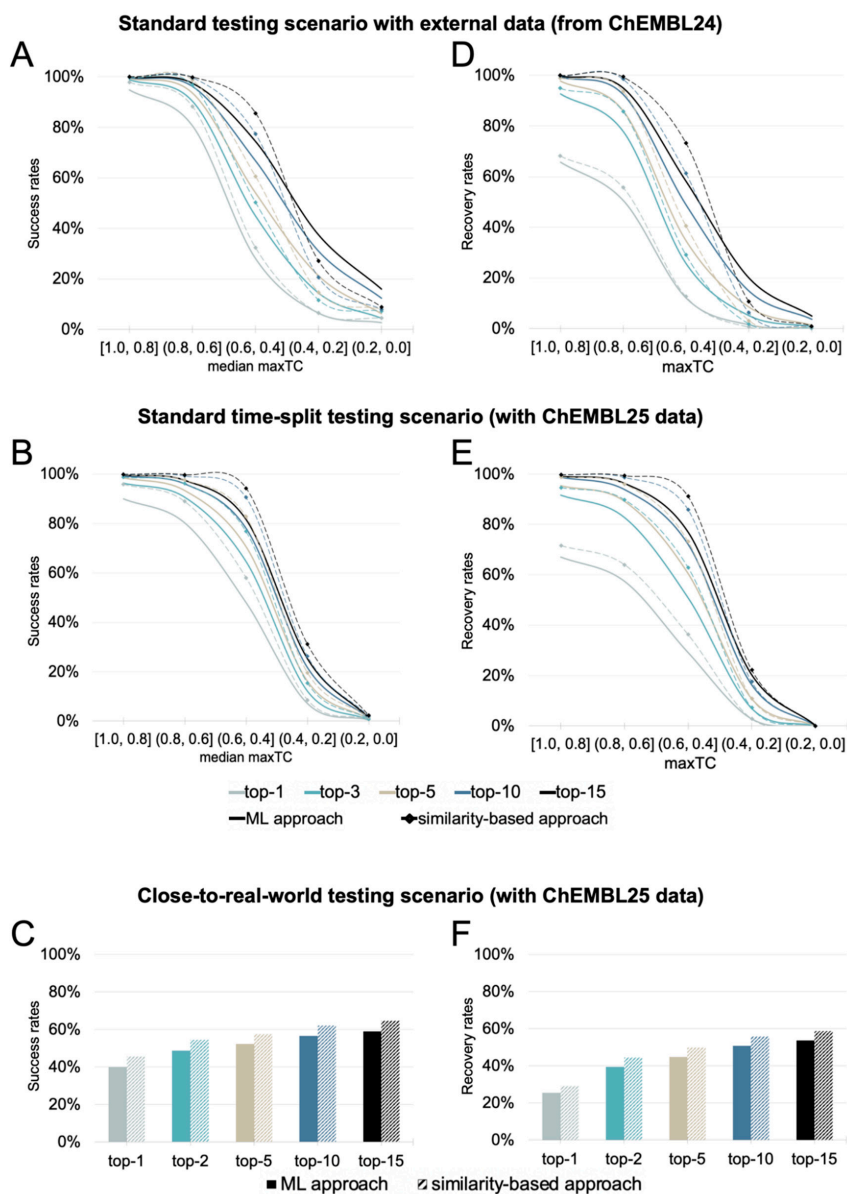
The ML approach achieved overall success rates of 82%, 86%, and 91% for the top-3, top-5, and top-15 positions, respectively (Figure 6A—solid lines), which is 1 to 3 percentage points lower than the success rates obtained by the similarity-based approach. The success rates were found to be strongly linked with the distance between the query molecule and the ligands in the knowledge base. For 90% of all high similarity queries, the known target was assigned the top rank (as compared to 95% for the similarity-based approach). For the median similarity queries, the success rates were only 49%, 57%, and 75% when considering the top-3, top-5, and top-15 ranks, respectively (which is 6 to 7 percentage points lower than the success rates of similarity-based approach). The success rates of the low similarity queries decreased to 10% (top-3), 15% (top-5), and 28% (top-15). Whereas the top-3 success rate was identical with that of the similarity-based approach, the top-5 and top-15 rates of the ML approach were 3 and 10 percentage points higher, respectively. This suggests that the ML approach may be able to predict the targets of low similarity queries better.

The recovery rates for the ML approach for high similarity queries were 88%, 94%, and 98% for the top-3, top-5, and top-15 positions, respectively (2 to 4 percentage points lower than the recovery rates of the similarity approach) (Figure 6D). For medium similarity queries, the recovery rates were only 30%, 39%, and 60% when considering the top-3, top-5, and top-15 ranks, respectively (5, 6 and 11 percentage points lower than the recovery rates of the similarity-based approach). The recovery rates for low similarity queries dropped to 3% (top-3), 5% (top-5), and 12% (top-15), which is still 2, 3, and 9 percentage points better than the values for the similarity-based approach.

#### Performance in a Standard Time-Split Testing Scenario

In line with the results obtained for the similarity-based approach, the success and recovery rates obtained for the standard time-split scenario were lower than for the standard testing scenario (Figure 6B). The overall top-3, top-5, and top-15 success rates were 53%, 57%, and 65%, respectively (vs. 82%, 86%, and 91% obtained for the ChEMBL24 test set; see above), which corresponds to 4, 3, and 1 percentage points below the time-split success rates of the similarity-based approach. For high similarity queries, a known target was ranked at the top position in 86% of all cases (which is 7 percentage points lower than the results obtained with the similarity-based approach). For medium similarity queries, the success rates were 59%, 65%, and 76% when considering the top-3, top-5, and top-15 ranks, respectively (vs. 49%, 57%, and 75% obtained for the ChEMBL24 test set). This corresponds to a drop by 11 to 13 percentage points over the similarity-based approach. For low similarity queries, the success rates were 4%, 7%, and 13% (nearly identical to the similarity-based approach in that scenario).

The recovery rates were also lower for queries from the new data in the ChEMBL25 database. Only 44%, 50%, and 60% of all the known interactions were covered among the top-3, top-5, and top-15 predictions (Figure 6E), respectively (vs. 69%, 75%, and 84% for the ChEMBL24 test set). The trends observed for the recovery rates under the standard time-split scenario for the similarity and ML approach were the same as the success rates described above.



**Figure 6.** Success rates (A–C) and recovery rates (D–F) of the ML approach (solid lines and bars) and the similarity-based approach (dashed lines and bars; reduced target scope, identical with that of the ML approach) under the (A,D) standard testing scenario with external data, (B,E) the standard time-split testing scenario, and (C,F) the close-to-real-world testing scenario. In general, the similarity-based approach shows a tendency to outperform the ML approach. As expected, the performance under all testing scenarios drops as queries become more dissimilar from the training set/knowledge base. The data for these graphs are also provided in tabular format in the Supporting Information (Table S5–S12).



## Performance in a Close-to-Real-World Testing Scenario

In the close-to-real-world testing scenario, an additional 3381 interactions (11% of the new interactions with version 25 of the ChEMBL database), which were not represented in the knowledge base, (from 1778 query molecules) were considered in the performance assessment. The overall success rates under this scenario were 49%, 52%, and 59% for the top-3, top-5, and top-15 predictions, respectively (Figure 6C). This represents a drop by 4 to 6 percentage points compared to the standard time-split scenario. The overall recovery rates for the top-3, top-5, and top-15 predictions dropped by 4 to 6 percentage points to 40%, 45%, and 54%, respectively (Figure 6F). In comparison to the similarity-based approach, the overall success rates under the close-to-real-world scenario were 6 percentage points lower for the top-3, top-5, and top-15 while the recovery rates were 5 percentage points lower.

## Scope of the ML Approach

The 90:10 split of the processed ChEMBL24 data and the requirement for a minimum of 25 ligands per target for model building resulted in a reduced scope of the ML approach over the similarity-based approach. While the similarity-based approach covers a total of 4239 targets, the ML approach covers only 1798 targets (42% of the similarity-based approach's target scope). As such, the ML approach did not cover 379 of the known targets of ChEMBL 24 queries, which resulted in the inability of the method to predict 2099 known interactions of 792 queries.

The accuracy of the ML approach can be marginally increased by using larger training sets (Table 1), at the cost of target coverage. Models based on training sets consisting of a minimum of 50, 75 or 100 ligands (and ten times as many confirmed and/or presumed inactive compounds selected as described in the Methods section) would further reduce the number of proteins to 1296, 1066 and 899, respectively. The improvements in ranking performance are explained primarily by the reduction of the number of proteins represented by the approach, making ranking an easier task.

**Table 1.** Overall success and recovery rates of the ML approach when using individual target models with a different minimum number of active compounds.

Minimum number of actives (number of targets represented)	Success Rates				Recovery Rates			
	25 (1798)	50 (1296)	75 (1066)	100 (899)	25 (1798)	50 (1296)	75 (1066)	100 (899)
top-1	74.23% (32,539/43,835) <sup>1</sup>	74.27% (31,897/42,946)	74.39% (31,369/42,170)	74.40% (30,754/41,336)	45.79% (32,539/71,063)	46.18% (31,897/69,066)	46.50% (31,369/67,457)	46.68% (30,754/65,880)
top-3	82.37% (36,107/43,835)	82.36% (35,370/42,946)	82.36% (34,730/42,170)	82.34% (34,035/41,336)	68.61% (48,756/71,063)	68.99% (47,649/69,066)	69.20% (46,681/67,457)	69.29% (45,651/65,880)
top-5	85.55% (37,503/43,835)	85.51% (36,725/42,946)	85.54% (36,072/42,170)	85.53% (35,353/41,336)	75.23% (53,460/71,063)	75.58% (52,197/69,066)	75.74% (51,095/67,457)	75.85% (49,970/65,880)
top-10	89.16% (39,083/43,835)	89.20% (38,308/42,946)	89.24% (37,634/42,170)	89.29% (36,909/41,336)	80.98% (57,545/71,063)	81.38% (56,205/69,066)	81.58% (55,034/67,457)	81.76% (53,863/65,880)
top-15	91.13% (39,946/43,835)	91.22% (39,175/42,946)	91.29% (38,496/42,170)	91.35% (37,759/41,336)	83.89% (59,617/71,063)	84.37% (58,270/69,066)	84.62% (57,083/67,457)	84.80% (55,866/65,880)

<sup>1</sup> The percentage indicates the success and recovery rates, while the numbers in the brackets show how many queries (success rate) or bioactivities (recovery rate) within the TC interval had a hit.

## 3. Methods

### 3.1. Data Preparation

Following a protocol closely related to that of Bosc et al. [24], the selection criteria listed below (italics indicate a ChEMBL data field and quotations indicate the value) were applied for the extraction of data from ChEMBL 24 [43]:

1. Assay covers a single protein or a protein complex (ChEMBL *confidence\_score* is 7 or 9)



2. *data\_validity\_comment* is null OR “manually validated”
3. *potential\_duplicate* is “0”
4. *standard\_type* is “Kd”, “Potency”, “AC50”, “IC50”, “Ki”, or “EC50”
5. *activity\_comment* is not “Inconclusive”, “inconclusive”, or “unspecified”
6. NOT (*standard\_relation* is null AND *activity\_comment* is not “Active” or “active”)
7. NOT (*standard\_relation* “>”, “≥”, or “>>” AND *standard\_value* less than 20,000)

This extraction procedure resulted in a dataset containing 1,482,972 bioactivity records (i.e., compound-protein pairs). Of these records, 2206 had standard\_units of “ $\mu\text{g mL}^{-1}$ ” as opposed to “nM” and therefore the standard\_value for these records were converted to nM using the topological information from *canonical\_smiles* and the Descriptors.ExactMolWt function of RDKit (RDKit: Open-source cheminformatics; version 2019.03.2.0; <http://www.rdkit.org>). The molecules were then passed through the salt and element filter described in ref. [47], and the SMILES of the remaining compounds were converted with RDKit to non-isomeric SMILES. Duplicate compound-protein pairs, resulting from multiple bioactivity values recorded in the original data or from the removal of compound stereochemistry, were consolidated by calculating the median activity value as the representative activity value for the 1,179,102 unique bioactivity records. Compound-protein pairs with activity values less than or equal to 10,000 nM were labeled “active” (732,570 bioactivities) while those with activities greater than or equal to 20,000 nM were marked as “inactive” (282,618 bioactivities). Compound-protein pairs with activity values between 10,000 nM and 20,000 nM (163,914 bioactivities) were not considered for model building or validation and were discarded. The resulting dataset (“processed dataset”) consists of 1,015,188 compound-protein pairs, comprising 546,981 unique compounds and 4676 unique targets (Figure 1a) from which the knowledge base for the similarity-based approach (Figure 1c), the training sets for the ML approach (Figure 1b), and the testing sets (Figure 1e–g) were derived. Additionally, data from the next version of the ChEMBL database (version 25) was processed as described above and new bioactivity records were used for further test sets (Figure 1h–j)

### 3.2. Development of Target Prediction Models

#### 3.2.1. Similarity-Based Approach

The pairwise similarity of each compound of the test set (query molecule) and all compounds of the knowledge base for similarity-based target prediction was quantified based on TCs derived from Morgan fingerprints with a radius of 2 and a length of 2048 bits. Morgan2 fingerprints were selected because they are closely related to the extended connectivity fingerprints [48] with a diameter of 4 bonds (ECFP4), which have been widely applied in target prediction [16,17,25,49] and virtual screening [50] and have shown to perform favorably. For example, in tests of the target prediction methods Polypharmacology Browser (PPB2) [49] and MolTarPred [16,17], the ECFP4 fingerprints obtained the best performance among a collection of different molecular fingerprints.

The proteins were assigned ranks from 1 to 4239 (the total number of proteins represented by the knowledge base) according to the maximum pairwise TC of the query molecule and any of the ligands of that protein (maxTC). In cases where multiple proteins had the same maxTC, the ranking was refined based on the distance of the query molecule to the next nearest neighbor until non-ambiguous ranks could be assigned to all proteins.

#### 3.2.2. Machine Learning Approach for Target Prediction

Random forest binary classification models were built with scikit-learn (Scikit-learn: Machine Learning in Python; version 0.20.1; <https://scikit-learn.org>) [51] for all of the 1798 targets represented in the ML knowledge base by at least 25 ligands. The training set for each model is composed of the bioactivity records from the knowledge base and supplemented with presumed inactive compounds (selected randomly following the procedure described in the section “Generation of training and test sets” of Methods) to obtain a ratio of 1:10 compounds. The number of estimators (*n* estimators) and

maximum depth of the estimators (max depth) of these models were optimized individually for each model during a grid search within a 10-fold cross-validation framework (Table 2). The best combination of parameters for each model (see the supplementary information), as measured by the average MCC score across the 10 folds, was then used to retrain the final model for each target using the complete training sets.

**Table 2.** Hyperparameters explored in the grid search of each target classification model.

Hyperparameter	Values Explored
<i>n</i> estimators: number of trees	200, 500, 1000
max depth: maximum depth of tree	25, 45, 50, 75, 100

For a given query molecule, the proteins were then ranked by prediction probability of the active class (*p*-value). Proteins assigned identical *p*-values were ranked according to the iterative process described for the similarity-based approach.

#### 4. Conclusions

This work aimed to determine the scope and limitations of two of the most commonly applied ligand-based methods for target prediction: a similarity-based approach, and a random forest-based machine learning approach, both employing Morgan2 fingerprints as molecular representations. By analyzing the performance of the approaches under three different scenarios and deconvoluting the results based on the distance of the test compounds (queries) from the training data (or knowledge base molecules), we obtained a robust and differentiated picture of the performance and reach of the approaches.

We have found that, in general, the similarity-based approach performed better than the ML approach, despite the similarity-based approach covering almost 2.5 times more proteins than the ML approach (4239 vs. 1798 proteins). Under the standard testing scenario with external data, the percentage of queries for which their target was recovered among the top-5 out of 1798 positions was 88% for the similarity-based approach and 85% for the ML approach (identical target space applied for the testing of both approaches). Under the time-split testing scenario, a drop in performance compared to the previous testing scenario was observed. Within a year's time (i.e., the difference between ChEMBL24 and ChEMBL25), the top-5 success rate performance of all the approaches dropped by an average of 25% for the new chemical and target spaces explored during that year. This reduction in performance is expected, given the evolution of the chemical and target spaces over time, and is consistent with the drop performance observed for compounds with a comparable degree of (dis-)similarity with the training data in the standard testing scenario with external data. The results indicate that the single time-split scenario, which represents a snapshot of the evolution of research in small-molecule drug discovery, may not be essential for obtaining an understanding of the robustness of a method, provided that model performance is evaluated taking into account training-to-test set distances. The third scenario, which is closest to the real-life application of the method, looks into how the method would perform, taking into account that some of the targets of new molecules may not be covered by the model. Here, minor drops in model performance were observed.

It was surprising to find that overall the similarity-based approach outperformed the ML approach, particularly for the low-similarity queries, where it was hoped that the ML models would have generalized and been able to make more reliable predictions. It is probable that the ML approach may perhaps be improved with more and diverse data and achieve better generalization with more specific training protocols for individual targets.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/1422-0067/21/10/3585/s1>, Table S1. Success rates under the standard testing scenario with external data by the similarity approach, Table S2. Recovery rates under the standard testing scenario with external data by the similarity approach, Table S3. Success rates under the time-split and close-to-real-world testing scenarios by the similarity approach, Table S4.

Recovery rates under the time-split and close-to-real-world testing scenarios with by the similarity approach, Table S5. Success rates under the standard testing scenario with external data by the similarity approach with a reduced target scope, Table S6. Recovery rates under the standard testing scenario with external data by the similarity approach with a reduced target scope, Table S7. Success rates under the time-split and close-to-real-world testing scenarios by the similarity approach with a reduced target scope, Table S8. Recovery rates under the time-split and close-to-real-world testing scenarios with by the similarity approach with a reduced target scope, Table S9. Success rates under the standard testing scenario with external data by the ML approach, Table S10. Recovery rates under the standard testing scenario with external data by the ML approach, Table S11. Success rates under the time-split and close-to-real-world testing scenarios by the ML approach, Table S12. Recovery rates under the time-split and close-to-real-world testing scenarios with by the ML approach.

**Author Contributions:** Conceptualization, N.M. and J.K.; methodology, N.M. and J.K.; software, N.M.; validation, N.M.; formal analysis, N.M. and J.K.; investigation, N.M. and J.K.; resources, J.K.; data curation, N.M.; writing—original draft preparation, N.M. and J.K.; visualization, N.M. and J.K.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Trond Mohn Foundation (BFS2017TMT01).

**Acknowledgments:** The calculations described in this work were performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

ML	Machine learning
RF	Random forest
SMILES	Simplified molecular-input line-entry system
TC	Tanimoto coefficient

## References

- Lauria, A.; Bonsignore, R.; Bartolotta, R.; Perricone, U.; Martorana, A.; Gentile, C. Drugs Polypharmacology by In Silico Methods: New Opportunities in Drug Discovery. *Curr. Pharm. Des.* **2016**, *22*, 3073–3081. [[CrossRef](#)] [[PubMed](#)]
- Lavecchia, A.; Cerchia, C. In Silico Methods to Address Polypharmacology: Current Status, Applications and Future Perspectives. *Drug Discov.* **2016**, *21*, 288–298. [[CrossRef](#)] [[PubMed](#)]
- Chaudhari, R.; Tan, Z.; Huang, B.; Zhang, S. Computational Polypharmacology: A New Paradigm for Drug Discovery. *Expert Opin. Drug Discov.* **2017**, *12*, 279–291. [[CrossRef](#)] [[PubMed](#)]
- Reddy, A.S.; Zhang, S. Polypharmacology: Drug Discovery for the Future. *Expert Rev. Clin. Pharmacol.* **2013**, *6*, 41–47. [[CrossRef](#)] [[PubMed](#)]
- Proschak, E.; Stark, H.; Merk, D. Polypharmacology by Design: A Medicinal Chemist’s Perspective on Multitargeting Compounds. *J. Med. Chem.* **2019**, *62*, 420–444. [[CrossRef](#)]
- Schneider, P.; Walters, W.P.; Plowright, A.T.; Sieroka, N.; Listgarten, J.; Goodnow, R.A., Jr.; Fisher, J.; Jansen, J.M.; Duca, J.S.; Rush, T.S.; et al. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* **2019**, *19*, 353–364. [[CrossRef](#)]
- Moffat, J.G.; Vincent, F.; Lee, J.A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.* **2017**, *16*, 531–543. [[CrossRef](#)]
- Rodrigues, T.; Bernardes, G.J.L. Machine Learning for Target Discovery in Drug Development. *Curr. Opin. Chem. Biol.* **2019**, *56*, 16–22. [[CrossRef](#)]
- Ezzat, A.; Wu, M.; Li, X.-L.; Kwoh, C.-K. Computational Prediction of Drug–Target Interactions Using Chemogenomic Approaches: An Empirical Survey. *Brief. Bioinform.* **2019**, *20*, 1337–1357. [[CrossRef](#)]
- Cortés-Ciriano, I.; Ain, Q.U.; Subramanian, V.; Lenselink, E.B.; Méndez-Lucio, O.; Ijzerman, A.P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T.E.; van Westen, G.J.P.; et al. Polypharmacology Modelling Using Proteochemometrics (PCM): Recent Methodological Developments, Applications to Target Families, and Future Prospects. *MedChemComm* **2015**, *6*, 24–50. [[CrossRef](#)]
- Reker, D.; Schneider, P.; Schneider, G.; Brown, J.B. Active Learning for Computational Chemogenomics. *Future Med. Chem.* **2017**, *9*, 381–402. [[CrossRef](#)] [[PubMed](#)]

12. Sydow, D.; Burggraaff, L.; Szengel, A.; van Vlijmen, H.W.T.; IJzerman, A.P.; van Westen, G.J.P.; Volkamer, A. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 1728–1742. [[CrossRef](#)]
13. Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. ChemMapper: A Versatile Web Server for Exploring Pharmacology and Chemical Structure Association Based on Molecular 3D Similarity Method. *Bioinformatics* **2013**, *29*, 1827–1829. [[CrossRef](#)] [[PubMed](#)]
14. Nickel, J.; Gohlke, B.-O.; Erehman, J.; Banerjee, P.; Rong, W.W.; Goede, A.; Dunkel, M.; Preissner, R. SuperPred: Update on Drug Classification and Target Prediction. *Nucleic Acids Res.* **2014**, *42*, W26–W31. [[CrossRef](#)] [[PubMed](#)]
15. Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013**, *15*, 395–406. [[CrossRef](#)] [[PubMed](#)]
16. Peón, A.; Naulaerts, S.; Ballester, P.J. Predicting the Reliability of Drug–target Interaction Predictions with Maximum Coverage of Target Space. *Sci. Rep.* **2017**, *7*, 1–11. [[CrossRef](#)] [[PubMed](#)]
17. Peón, A.; Li, H.; Ghislat, G.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P.J. MolTarPred: A Web Tool for Comprehensive Target Prediction with Reliability Estimation. *Chem. Biol. Drug Des.* **2019**, *94*, 1390–1401. [[CrossRef](#)] [[PubMed](#)]
18. Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-Based Machine Learning Methods for Predicting Drug–Target Interactions: A Brief Review. *Brief. Bioinform.* **2014**, *15*, 734–747. [[CrossRef](#)]
19. Wang, C.; Kurgan, L. Review and Comparative Assessment of Similarity-Based Methods for Prediction of Drug-Protein Interactions in the Druggable Human Proteome. *Brief. Bioinform.* **2018**, *20*, 2066–2087. [[CrossRef](#)]
20. Wang, C.; Kurgan, L. Survey of Similarity-based Prediction of Drug-Protein Interactions. *Curr. Med. Chem.* **2019**, *26*, 1. [[CrossRef](#)]
21. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for In Silico Target Fishing. *Methods* **2015**, *71*, 98–103. [[CrossRef](#)] [[PubMed](#)]
22. Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web Server for Target Prediction of Bioactive Small Molecules. *Nucleic Acids Res.* **2014**, *42*, W32–W38. [[CrossRef](#)] [[PubMed](#)]
23. Shi, H.; Liu, S.; Chen, J.; Li, X.; Ma, Q.; Yu, B. Predicting Drug-Target Interactions Using Lasso with Random Forest Based on Evolutionary Information and Chemical Structure. *Genomics* **2019**, *111*, 1839–1852. [[CrossRef](#)] [[PubMed](#)]
24. Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A.R. Large Scale Comparison of QSAR and Conformal Prediction Methods and their Applications in Drug Discovery. *J. Cheminform.* **2019**, *11*, 4. [[CrossRef](#)]
25. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J.K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441–5451. [[CrossRef](#)]
26. Ding, Y.; Tang, J.; Guo, F. Identification of Drug-Target Interactions via Multiple Information Integration. *Inf. Sci.* **2017**, *418*, 546–560. [[CrossRef](#)]
27. Keum, J.; Nam, H. SELF-BLM: Prediction of Drug-Target Interactions via Self-Training SVM. *PLoS ONE* **2017**, *12*, e0171839. [[CrossRef](#)]
28. Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of De Novo-Designed Chemical Entities Through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4067–4072. [[CrossRef](#)]
29. Gawehn, E.; Hiss, J.A.; Schneider, G. Deep Learning in Drug Discovery. *Mol. Inform.* **2016**, *35*, 3–14. [[CrossRef](#)]
30. Zhang, H.; Liao, L.; Saravanan, K.M.; Yin, P.; Wei, Y. DeepBindRG: A Deep Learning Based Method for Estimating Effective Protein-Ligand Affinity. *PeerJ* **2019**, *7*, e7362. [[CrossRef](#)]
31. Monteiro, N.R.C.; Ribeiro, B.; Arrais, J.P. Deep Neural Network Architecture for Drug-Target Interaction Prediction. In *Artificial Neural Networks and Machine Learning—ICANN 2019: Workshop and Special Sessions. Lecture Notes in Computer Science, vol 11731*; Tetko, I.V., Kůrková, V., Karpov, P., Theis, F., Eds.; Springer: Cham, Germany, 2019; Volume 11731, pp. 804–809. ISBN 9783030304928. [[CrossRef](#)]

32. Lee, K.; Kim, D. In-Silico Molecular Binding Prediction for Human Drug Targets Using Deep Neural Multi-Task Learning. *Genes* **2019**, *10*, 906. [[CrossRef](#)] [[PubMed](#)]
33. Chu, Y.-Y.; Zhang, Y.-F.; Wang, W.; Wang, X.-G.; Shan, X.-Q.; Xiong, Y.; Wei, D.-Q. DTI-CDF: A CDF Model Towards the Prediction of DTIs Based on Hybrid Features. *bioRxiv* **2019**, 657973. [[CrossRef](#)]
34. Lee, H.; Kim, W. Comparison of Target Features for Predicting Drug-Target Interactions by Deep Neural Network Based on Large-Scale Drug-Induced Transcriptome Data. *Pharmaceutics* **2019**, *11*, 377. [[CrossRef](#)] [[PubMed](#)]
35. Boezio, B.; Audouze, K.; Ducrot, P.; Taboureau, O. Network-Based Approaches in Pharmacology. *Mol. Inform.* **2017**, *36*. [[CrossRef](#)]
36. Lo, Y.-C.; Senese, S.; Damoiseaux, R.; Torres, J.Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chem. Biol.* **2016**, *11*, 2244–2253. [[CrossRef](#)]
37. Carrella, D.; Napolitano, F.; Rispoli, R.; Miglietta, M.; Carissimo, A.; Cuttillo, L.; Sirci, F.; Gregoretti, F.; Di Bernardo, D. Mantra 2.0: An Online Collaborative Resource for Drug Mode of Action and Repurposing by Network Analysis. *Bioinformatics* **2014**, *30*, 1787–1788. [[CrossRef](#)]
38. Fu, G.; Ding, Y.; Seal, A.; Chen, B.; Sun, Y.; Bolton, E. Predicting Drug Target Interactions Using Meta-Path-Based Semantic Network Analysis. *BMC Bioinform.* **2016**, *17*, 160. [[CrossRef](#)]
39. Mathai, N.; Chen, Y.; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Brief. Bioinform.* **2019**. [[CrossRef](#)]
40. Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206. [[CrossRef](#)]
41. Keiser, M.J.; Setola, V.; Irwin, J.J.; Laggner, C.; Abbas, A.I.; Hufeisen, S.J.; Jensen, N.H.; Kuijter, M.B.; Matos, R.C.; Tran, T.B.; et al. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181. [[CrossRef](#)]
42. Lounkine, E.; Keiser, M.J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J.L.; Lavan, P.; Weber, E.; Doak, A.K.; Côté, S.; et al. Large-scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012**, *486*, 361–367. [[CrossRef](#)] [[PubMed](#)]
43. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [[CrossRef](#)] [[PubMed](#)]
44. Zhang, M.-L.; Li, Y.-K.; Liu, X.-Y.; Geng, X. Binary Relevance for Multi-Label Learning: An Overview. *Front. Comput. Sci.* **2018**, *12*, 191–202. [[CrossRef](#)]
45. Cockroft, N.T.; Cheng, X.; Fuchs, J.R. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products. *J. Chem. Inf. Model.* **2019**, *59*, 4906–4920. [[CrossRef](#)]
46. Hao, M.; Bryant, S.H.; Wang, Y. Open-Source Chemogenomic Data-Driven Algorithms for Predicting Drug-Target Interactions. *Brief. Bioinform.* **2019**, *20*, 1465–1474. [[CrossRef](#)]
47. Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šicho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters. *ChemMedChem* **2018**, *13*, 564–571. [[CrossRef](#)]
48. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [[CrossRef](#)]
49. Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 10–17. [[CrossRef](#)]
50. Riniker, S.; Landrum, G.A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5*, 26. [[CrossRef](#)]
51. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.



## Chapter 6

# Using target prediction to curate compound sets for screening libraries

Screening libraries of compounds with biochemical, cell-based and/or virtual screens is routine in the early stages of a drug discovery project [90, 104]. Compound libraries used in screening campaigns are typically designed to be either focused libraries or general libraries. Focused libraries are designed to contain compounds that are likely to interact with a particular target or target family. General purpose libraries are designed to be chemically and/or biologically diverse so as to increase the chances of finding bioactive compounds for a wide range of targets of interest. In either case, it is important that libraries contain high-quality compounds which can be further optimized into a drug, agrochemical, active cosmetic ingredient etc. This means that the compounds do not exhibit “bad actor” behavior in screening assays. Small to medium-sized general-purpose libraries are of particular interest, especially for academic drug discovery projects. This is because in these settings, targets which are usually not as well established, are explored under resource constraints.

In this chapter, we present a study where we applied computational target prediction followed by an evolutionary algorithm to optimize small to medium sized compound libraries for general purpose screens. A similarity-based target prediction method, based by our earlier work presented in Chapter 5, was used to predict the macromolecular targets of a large collection of purchasable compounds. Over 1.3 million compounds with a predicted target, from the ZINC20 database, were used as the PCC from which a genetic algorithm was used to select optimal subsets to form the optimized libraries. The PCC

was made up of compounds which were labeled as “anodyne” on the ZINC web-service as they have been tested on an extensive collection of reactivity filters [119]. They are therefore assumed to be unreactive in the context of screening and suitable as a starting point for drug discovery projects. Optimized compound libraries (BonMOLière) of 1,000, 5,000, 10,000 and 15,000 compounds were generated and have been made available to the community. The smaller the size of the library the greater the improvements, as measured by the fitness function, were from the optimization process. Compared to baseline libraries, which were not optimized, the optimization resulted in improvements which ranged from +60% (for the 15,000 compound library) to +184% (for the 1,000 compound library).

### **P3: BonMOLière: Small-sized libraries of readily purchasable compounds, optimized to produce genuine hits in biological screens across the protein space**

Neann Mathai, Conrad Stork and Johannes Kirchmair

*International Journal of Molecular Sciences*, **22(15)**, 7773 (2021).

<https://doi.org/10.3390/ijms22157773>

#### **Contributions:**

N. Mathai and J. Kirchmair conceptualized the research. N. Mathai conducted the research and analysis with contributions from C. Stork. N. Mathai wrote the manuscript, with contributions from C. Stork and J. Kirchmair. J. Kirchmair supervised the work.

The following article was reprinted from: Mathai N., Stork C., J. Kirchmair, BonMOLière: Small-sized libraries of readily purchasable compounds, optimized to produce genuine hits in biological screens across the protein space *Int. J. Mol. Sci.* **2021**, *22(15)*, 7773.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).



Article

# BonMOLière: Small-Sized Libraries of Readily Purchasable Compounds, Optimized to Produce Genuine Hits in Biological Screens across the Protein Space

Neann Mathai <sup>1</sup>, Conrad Stork <sup>2</sup> and Johannes Kirchmair <sup>1,3,\*</sup>

<sup>1</sup> Computational Biology Unit (CBU) and Department of Chemistry, University of Bergen, N-5020 Bergen, Norway; neann.mathai@uib.no

<sup>2</sup> Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany; stork@zbh.uni-hamburg.de

<sup>3</sup> Division of Pharmaceutical Chemistry, Department of Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria

\* Correspondence: johannes.kirchmair@univie.ac.at

**Abstract:** Experimental screening of large sets of compounds against macromolecular targets is a key strategy to identify novel bioactivities. However, large-scale screening requires substantial experimental resources and is time-consuming and challenging. Therefore, small to medium-sized compound libraries with a high chance of producing genuine hits on an arbitrary protein of interest would be of great value to fields related to early drug discovery, in particular biochemical and cell research. Here, we present a computational approach that incorporates drug-likeness, predicted bioactivities, biological space coverage, and target novelty, to generate optimized compound libraries with maximized chances of producing genuine hits for a wide range of proteins. The computational approach evaluates drug-likeness with a set of established rules, predicts bioactivities with a validated, similarity-based approach, and optimizes the composition of small sets of compounds towards maximum target coverage and novelty. We found that, in comparison to the random selection of compounds for a library, our approach generates substantially improved compound sets. Quantified as the “fitness” of compound libraries, the calculated improvements ranged from +60% (for a library of 15,000 compounds) to +184% (for a library of 1000 compounds). The best of the optimized compound libraries prepared in this work are available for download as a dataset bundle (“BonMOLière”).

**Keywords:** optimized compound library; biological screening; purchasable compounds; evolutionary optimization; genetic algorithms; tool compounds; novel targets



**Citation:** Mathai, N.; Stork, C.; Kirchmair, J. BonMOLière: Small-Sized Libraries of Readily Purchasable Compounds, Optimized to Produce Genuine Hits in Biological Screens across the Protein Space. *Int. J. Mol. Sci.* **2021**, *22*, 7773. <https://doi.org/10.3390/ijms22157773>

Academic Editor: Isabelle Callebaut

Received: 24 June 2021

Accepted: 15 July 2021

Published: 21 July 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

A key strategy to identify bioactive compounds for biomacromolecules of interest is to screen large collections of compounds with biochemical or cell-based assays [1]. The success of such screening campaigns depends on many factors, above all, the quality and composition of the compound library: the much-cited “needle in the haystack” can only possibly be found if it actually is in the haystack. For this reason, the design of compound libraries for screening has been, and continues to be, an active field of research [2–4].

There are multiple approaches to compiling compound libraries. Focused design aims to compile a set of compounds that have an increased likelihood of being active on a particular target of interest [4–7]. In contrast, general compound libraries may be optimized for maximum chemical and/or biological diversity in order to increase the chances of identification of bioactive compounds for an arbitrary target [4–8]. In any case, besides chemical and biological diversity, the potential of compounds to be further optimized into functional molecules with desired properties (drugs, agrochemicals, or active cosmetic ingredients, in particular) should be taken into account [9]. In particular,



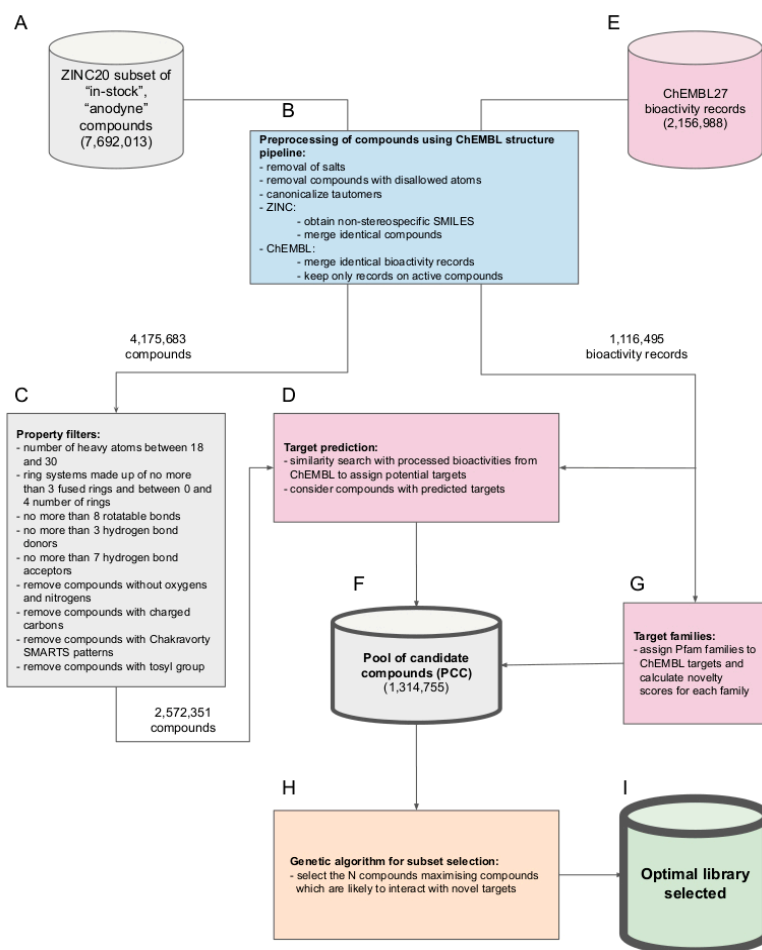
this concerns the chemistry and toxicological properties of compounds. A high-quality compound library should also not include compounds that are prone to cause false readouts in biological assays (“bad actors”) [10,11].

With the ever-growing capacities in experimental screening, and in particular with the renewed interest in phenotypic screens, there have been increased efforts in compiling diverse compound libraries [8,9,12–14]. Lahue et al. [9] detail the recent undertaking at Merck & Co. to design two libraries for phenotypic screening, consisting of 50 thousand and 250 thousand compounds, respectively. They used a combination of in-house structural alerts and PAINS patterns [15] to remove compounds that contain undesired moieties. Chemical properties of compounds, such as the molecular weight and the quantitative estimate of drug-likeness score [16] (QED; a composite score to quantify chemical beauty), were used to filter and reduce the size of the candidate pool. The candidate pool was then organized into clusters, from which compounds were randomly selected to make compound sets using a genetic algorithm. The compound sets were improved by maximizing a fitness function that captured the 3D shape diversity, bioactivity diversity, and the median QED score of a set of compounds. The optimized sets were added to the final compound library after additional quality checks and opinions garnered from in-house chemists.

Schuffenhauer et al. [17] described the process used to design the screening deck at Novartis to optimize the selection of diverse subsets for screening. Structures were first passed through quality checks, frequent hitter SMARTS patterns (based on a subset of the PAINS patterns and in-house patterns to flag nuisance compounds), molecular weight, and nitrogen and oxygen contents. The compounds were flagged based on these checks and then ranked based on aqueous solubility, cell permeability, and the number of assigned flags from the checks. Additionally, compounds were grouped, non-exclusively, into classes that described their target interaction, biological and chemical descriptors. Compounds were selected in iterative rounds in a greedy fashion, starting with the highest ranking compound which occupied the highest number of classes. This selection, based on property ranking and the class memberships of the compounds, resulted in a 2D grid with 1.5 million compounds. The *x*-axis of the grid represents the property rank and the *y*-axis represents the round in which the compound was selected. The 2D grid allows the researchers at Novartis to select their choice of how many compounds to screen from this grid, balancing the properties and diversity of the screening subset.

A small to medium-sized screening deck (1000 to 15,000 compounds) that has a high chance of producing genuine hits during screening on an arbitrary target of interest can be incredibly valuable to biochemical and cell research. This is particularly true for the research of proteins for which there is little existing knowledge to use for the compilation of focused screening sets, and for academic research which is often focused on innovative targets under tight resource constraints.

In this work, we report on the development and application of a computational method for the automated compilation of small to medium-sized compound libraries that have a high chance of producing genuine hits during experimental screens on arbitrary protein targets. We show the capacity of the new computational approach by generating a set of optimized compound libraries (“BonMOLière”) of different sizes from a subset of the ZINC20 database [18,19] (an aggregate of more than 300 commercially available compound catalogs from over 150 companies; Figure 1A). The approach utilizes (i) elaborate protocols for the preprocessing and preparation of chemical and biological data (Figure 1B), (ii) established rule sets to promote drug-likeness (Figure 1C), (iii) a validated, similarity-based approach to predict the likely targets of compounds (Figure 1D,E), and (iv) a genetic optimization algorithm (Figure 1H) to maximize coverage of the protein space (Figure 1G) when selecting subsets of compounds for a compound library, taking target novelty and target diversity into account.



**Figure 1.** Overview of the workflow followed to generate optimized compound libraries: (A) source of compounds for the generation of optimized screening libraries, (B) preprocessing of compounds, (C) removal of compounds with undesired properties, (D) target prediction, (E) source of bioactivity data for target prediction, (F) ZINC20 compounds with predicted targets, (G) assignment of Pfam families, (H) genetic algorithm for optimal subset selection, (I) optimal library selected.

## 2. Results and Discussion

All 7,692,013 compounds included in the ZINC20 subset used in this study (Figure 1A) fulfill the following key criteria (among other criteria, outlined in the Materials and Methods section):

1. The compounds are already made and readily obtainable from the manufacturer (i.e., they are part of the "in-stock" subset of the ZINC20 database).
2. The compounds are presumed benign in the context of screening with biological assays (i.e., they are also part of the "anodyne" subset of the ZINC20 database). More specifically, all of these compounds have passed an extensive collection of reactivity filters and PAINS patterns compiled and utilized by the developers of the ZINC database, meaning that they are unlikely reactive or causing pan-assay interference [20].

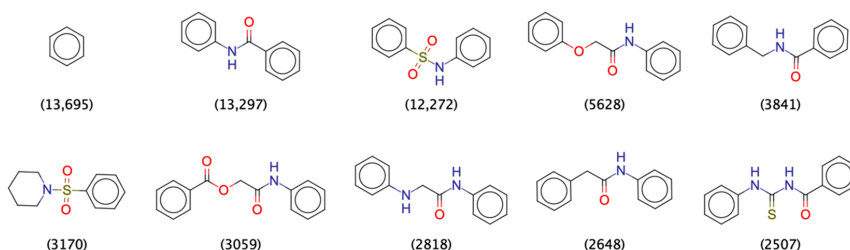
A multi-step process including the preparation of chemical and biological data (Figure 1B), the filtering for physicochemical properties to promote drug-likeness (Figure 1C), and the prediction of likely targets with a 2D similarity-based approach (Figure 1D), resulted in a pool of (1,314,755) candidate compounds (PCC). On this PCC, a genetic algorithm is applied to generate the final set of optimized screening libraries.

### 2.1. Characterization of the Pool of Candidate Compounds

In order to understand the relevance and properties of the PCC, and to enable a comparison of compound libraries prior and after optimization, we conducted a thorough characterization of the PCC.

#### 2.1.1. Physicochemical and Structural Characterization of the Pool of Candidate Compounds

The PCC consists of 1,314,755 compounds which are built on 379,690 unique Murcko scaffolds. Ninety-six percent of the scaffolds (362,707 scaffolds) represent fewer than ten compounds of the PCC. However, the remaining 16,983 scaffolds (4% of all scaffolds) have a wide distribution in terms of occurrence, ranging from 10 to 13,695 compounds per scaffold. The 10 most popular scaffolds (Figure 2) account for 60,428 compounds, with benzene being the most popular one (representing 13,695 or 1% of the PCC). Clustering of the PCC (with the Taylor-Butina algorithm and a Tanimoto similarity threshold of 0.4; see Materials and Methods) produced 28,826 clusters, 6801 of which are singletons.



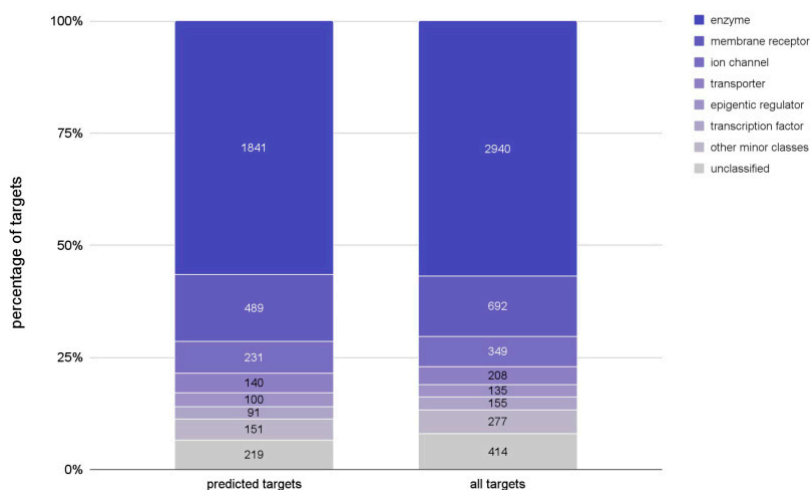
**Figure 2.** Top ten most popular Murcko scaffolds among the pool of candidate compounds. The numbers in the parentheses indicate how many compounds (out of 1,314,755) in the PCC have the scaffold.

The upper and lower boundaries of several relevant physicochemical properties of the compounds forming the PCC are set by the property filters applied previously (Figure 1C). In the case of the molecular weight, the property filters imposed an upper limit of 900 Da. The majority of the compounds forming the PCC have a molecular weight between 250 and 500 Da (Figure 4A), with the median at 342 Da. The median number of heavy atoms is 24 (Figure 4B). The majority of the compounds have 6 to 8 rotatable bonds (Figure 4C) and their median number of rings is 3 (Figure 4D). Half of all compounds have 1 hydrogen bond donor (Figure 4E) while the number of hydrogen bond acceptors per compound is more spread out (median at 4 hydrogen bond acceptors; Figure 4F). The distribution of the logP values of the compounds shows a peak near the upper filter boundary (Figure 4G), at approximately 4, with the median located at 2.90. Finally, while not utilized as a physicochemical property filter, Figure 4H shows the distribution of the QED score [16]. The QED score is a quantification of the drug-likeness of a compound, with 0 being most unfavorable and 1 being most favorable. The compounds of the PCC have a QED score distribution which is skewed towards being favorable, with a median score of 0.78. This shows that the PCC is composed of compounds with a high level of drug-likeness.

#### 2.1.2. Biological Characterization of the Pool of Candidate Compounds

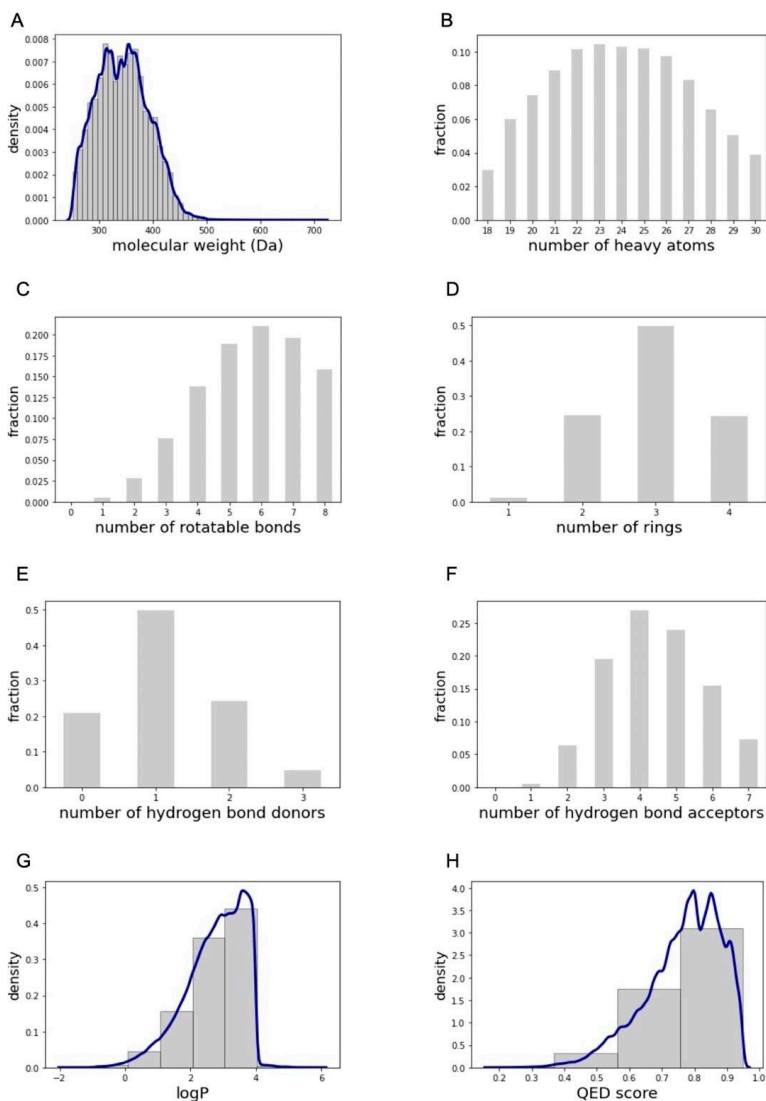
For the compounds forming the PCC, activities on a total of 3262 distinct protein targets were predicted with a target prediction model based on 2D molecular similarity. This model has been published and validated previously [21] and is built on a curated

subset of the ChEMBL27 database [22,23] (the “ChEMBL27 reference set”) that covers a total of 5170 proteins (see Materials and Methods). As shown in Figure 3, the types of targets (enzymes, membrane receptors, etc.) predicted for the PCC are a good reflection of the proteins represented in the ChEMBL27 reference set.



**Figure 3.** Types of proteins among the 3362 targets predicted for the ZINC20 compounds and the 5170 targets in the ChEMBL27 reference set. The size of the bars reflects the percentage of a target type represented while the labels are the counts of the targets for each type.

To characterize the target diversity of compound libraries, we retrieved the Pfam family classifications [24,25] of the ChEMBL proteins by scanning their sequences against the Pfam database of 18,259 families. The compounds of the PCC were predicted to be active on 3262 unique targets that represent 880 Pfam families. These predicted targets are diverse and cover over 70% of the 1214 Pfam families that represent the 5170 proteins in the ChEMBL27 reference set. Of the proteins represented in the ChEMBL27 reference set, 334 belong to more than one Pfam family and two proteins are assigned to the dummy Pfam family to group all targets for which a Pfam family could not be assigned.



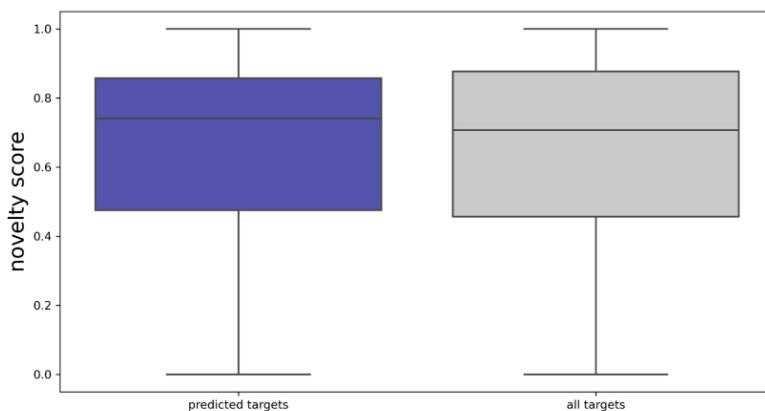
**Figure 4.** Distributions of physicochemical properties observed for the PCC: (A) molecular weight, (B) number of heavy atoms, (C) number of rotatable bonds, (D) number of rings, (E) number of hydrogen bond donors (F) number of hydrogen bond acceptors, (G) logP (note that for a very few compounds the logP value is greater than 4; this is because these logP values are calculated with RDKit (version 2020.09.1.0) [26] and may differ, to some extent, from the calculated logP values provided in the ZINC20 database), and (H) QED score.

A novelty score for each Pfam family was calculated based on how many ChEMBL bioactivities were recorded before and after the year 2010 (Equation (1)). The intention of this score is to promote, during compound library optimization, the representation of protein targets that reflect the more recent research directions (as the protein space and

the chemical space of interest evolve over time). The novelty scores were assigned to the proteins via their Pfam classification:

$$\text{Pfam novelty score} = \frac{\text{No. (bioactivities recorded in or after 2010)}}{\text{No. (bioactivities recorded in or after 2010)} + \text{No. (bioactivities recorded before 2010)}} \quad (1)$$

The distribution of the novelty scores (Figure 5) of the predicted targets of the PCC closely mirrors the distribution of the novelty scores of the full protein space of the ChEMBL27 reference set, with median novelty scores of 0.74 and 0.71, respectively. This indicates that the predicted targets are generally representative of the proteins found in the ChEMBL27 reference set, with a slight bias towards newer targets.



**Figure 5.** Distribution of novelty scores of the targets predicted for the PCC and of all targets found in the ChEMBL27 reference set.

## 2.2. Characterization of the Optimized Compound Libraries

Starting from the PCC, optimized compound libraries composed of 1000, 5000, 10,000, and 15,000 compounds were generated with the genetic algorithm described in the Materials and Methods section. As this is a subset selection problem, it is an optimization problem as an optimal subset needs to be selected which may be achieved through numerous combinations of compounds. A genetic algorithm is well suited for this as it reaches an optimal selection by selecting compounds to maximize the values of the fitness function. The algorithm optimizes the fitness of a compound set (i.e., a subset of the PCC) according to a fitness function that accounts not only for the novelty of the proteins predicted as targets of the compounds in a set but also for the number of proteins and number of times (i.e., the count) specific Pfam families are predicted for the compounds of a library (Equation (2)). This is because the more times a Pfam family is assigned to a library, via predicted targets for the compounds within a library, the higher the likelihood that proteins within this family will be a true hit when screened with this library.

$$\text{fitness score} = \frac{\left( \sum \text{Pfam family Pfam novelty score} \left( \frac{1 - 0.99^{\text{count}}}{1 - 0.99} \right) \right)}{\text{number of compounds in the library}} \quad (2)$$

Therefore, when comparing two libraries of the same size, a higher fitness score signifies a better library, enriched with:

1. more bioactive compounds as a whole
2. compounds active on proteins representing more Pfam families (maximizing target diversity)

3. compounds active on newer targets (the novelty score is higher for newer targets)

### 2.2.1. Baseline Compound Libraries

To understand the benefits of this approach (i.e., utilizing a target prediction model and a genetic algorithm to optimize the selection of compounds for the libraries), baseline compound libraries (of sizes 1000, 5000, 10,000, and 15,000) were generated. These baseline compound libraries were generated by randomly selecting sets of compounds from the 2,572,351 ZINC20 compounds which passed through the property filters (Figure 1C) irrespective of whether they are a part of the PCC or not. The selection was repeated multiple (10% of the compound library size) times and the properties of the fittest of these compound libraries are reported in Table 1.

**Table 1.** Properties of the baseline compound libraries generated from compounds before predicted targets and the genetic algorithm are used to optimize the libraries.

Library Size	Fitness Score	Number of Murcko Scaffolds	Number of Targets	Number of Pfam Families	Number of Bioactivities	Median Novelty Score
1000	0.55	963	272	127	925	0.76
5000	0.39	4516	719	275	4288	0.75
10,000	0.30	8707	910	325	8571	0.74
15,000	0.25	12,779	1051	378	12,561	0.74

### 2.2.2. Optimized Compound Libraries

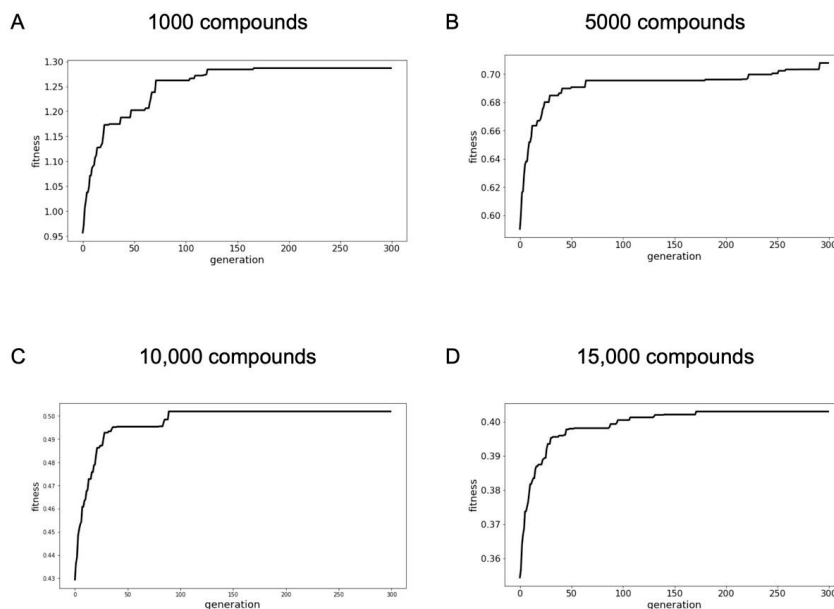
The genetic algorithm was run for 300 generations, with a population consisting of individual compound sets. The number of individuals for a population was set to 10% of the size of the library being optimized. The population evolved to reflect the fittest individuals from the previous generation (see Materials and Methods).

Within the 300 generations of evolution, the fitness increased (Table 2) and converged (Figure 6) for compound libraries of all sizes, with the biggest effects observed for the smallest compound library. That is, the fitness of the 1000-compound library improved by 34%, whereas the fitness of the 15,000-compound library increased by only 14% through the course of evolution. This shows that there are greater gains in optimization using this genetic algorithm for the smaller libraries than for larger libraries. It must be noted that, unlike the percentage improvement, the fitness score (Equation (2)) is a function of the number of compounds in a library and can therefore only be compared when they describe libraries of the same size and not libraries with different numbers of compounds.

**Table 2.** Change in properties of the fittest individual (compound library) from the first generation to the fittest individual from after 300 generations generated with population sizes that were 10% of the library size <sup>1</sup>.

Library Size	% Change in Fitness Score	Δ Number of Murcko Scaffolds	Δ Number of Compound Clusters Represented	Δ Number of Targets	Δ Number of Pfam Families	Δ Number of Bioactivities	% Change in the Median Novelty Score
1000	+34.48% (0.96→1.29)	−15 (925→910)	0 (1000 in both)	+59 (364→423)	+102 (180→282)	+395 (1679→2074)	+5.22% (0.72→0.76)
5000	+19.90% (0.59→0.71)	+17 (4211→4228)	0 (5000 in both)	+7 (984→991)	+102 (339→441)	+591 (8272→ 8863)	+0% (both at 0.75)
10,000	+16.89% (0.43→0.50)	−83 (7853→7770)	0 (10,000 in both)	+40 (1137→1177)	+79 (407→486)	+805 (16,035→16,840)	+6.11% (0.71→0.75)
15,000	+13.73% (0.35→0.40)	−189 (11,300→11,111)	0 (15,000 in both)	−81 (1362→1281)	−3 (515→512)	+769 (24,370→25,139)	−0.61% (0.75→0.74)

<sup>1</sup> The “→” symbol indicates a change in the value of a characteristic between the fittest individual at the start and the end of the evolution.



**Figure 6.** Development of the fittest population over 300 generations of a library of (A) 1000 compounds, (B) 5000 compounds, (C) 10,000 compounds, and (D) 15,000 compounds.

By the similarity principle, a more diverse compound library should reflect a more diverse set of targets on which compounds of that library are active. The similarity principle holds true for the 5000-compound library: looking at the 5000-compound library (Table 2), we see that over the 300 generations the number of Murcko scaffolds represented increased by 17. The evolution of the 1000, 10,000-, and 15,000-compound libraries, on the other hand, saw a decrease by 15, 83, and 189 Murcko scaffolds respectively. This was while the number of unique predicted targets increased by 59 (to 423) for the 1000-compound library, and by 7 (to 991) for the 10,000-compound library, indicating that compounds with more promiscuous scaffolds have been selected through the course of evolution. The 15,000-compound library did see a reduction in the number of unique targets over the course of the evolution (by 81 targets, to 1281).

There is no change in the number of molecular clusters represented at the start and the end of the evolution for all the compound libraries. This is because the PCC forms over 24 thousand clusters, therefore the limiting factor in the number of clusters represented is the size of the library as each compound within a library is from a different cluster. Notably, every compound within a library is always from a different cluster, speaking to the diversity of the libraries.

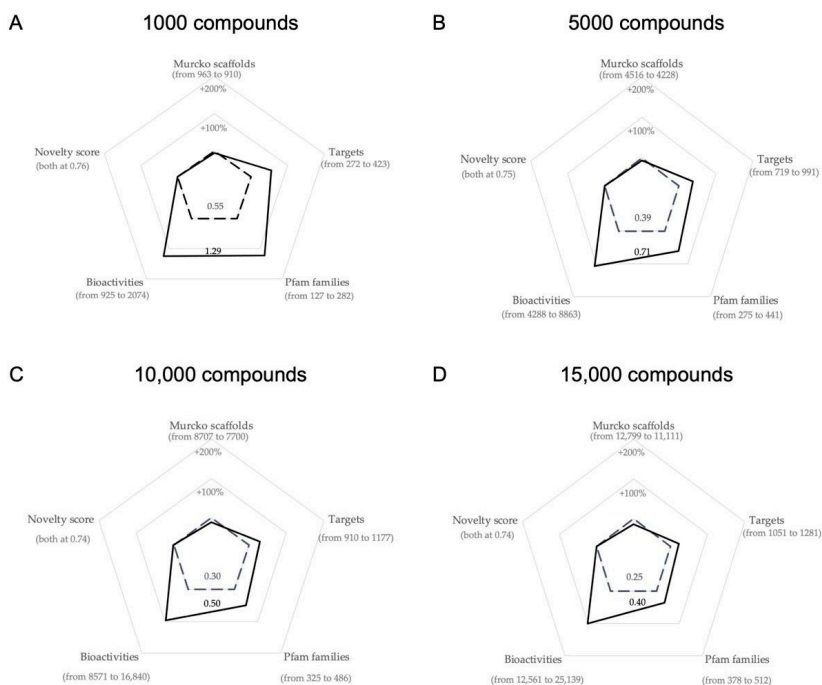
Importantly, compounds selected during the evolution are in fact predicted to be bioactive towards newer targets more often. There are 59 more targets and 102 more Pfam families represented in the 1000-compound library at the end of 300 generations. The 5000 and 10,000-compound libraries also saw, as a result of evolution, a gain by 7 and 40 predicted targets, respectively, and 102 and 79 more Pfam families, respectively. For the 1000 and 10,000-compound libraries, the evolution resulted in a +5% (from 0.72 to 0.76) and +6% (from 0.71 to 0.75) respective change in the median novelty score of the targets predicted for the most fit individuals at the start and the end of the evolution. For the 5000-compound library, the median novelty score of the predicted targets was 0.75 for the fittest individual at both the beginning and the end of the evolution. An anomaly to this trend is observed for the 15,000-compound library where there is a reduction in the number



of unique targets (by 81 targets to 1281) and consequently a very slight reduction in the median novelty score (from 0.75 to 0.74). There is, however, an improvement in the fitness of the 15,000-compound library, and this gain is acquired from the increase in predicted bioactivities (by 769 bioactivities to 25,139).

The optimization led to an increase in predicted bioactivities for all the compound libraries. That is, between the start and end of the evolution, the 1000-compound library has 395 (+24%) more predicted bioactivities, the 5000-compound library has 591 (+7%) more bioactivities, the 10,000-compound library has 805 (+5%) more bioactivities, and the 15,000-compound library has 769 (+3%) more activities. All these changes, coupled with the increase in the fitness score, show that the resulting compound libraries have got more predicted activity on novel targets.

Comparing these optimized compound libraries (Table 2) with the baseline compound libraries (Table 1), we see that the optimization shows remarkable improvements of the fitness across all libraries: +134% (0.55 vs. 1.29) for the 1000-compound library (Figure 7A), +82% (0.39 vs. 0.71) for the 5000-compound library (Figure 7B), +67% (0.30 vs. 0.50) for the 10,000-compound library (Figure 7C), and +60% (0.25 vs. 0.40) for the 15,000-compound library (Figure 7D). These improvements are driven by a steep increase in the number of predicted targets of the optimized compound libraries. When compared to the baseline compound libraries, between an additional 151 targets (+55% for the 1000-compound library) to 230 targets (+20% for the 15,000-compound library) are observed. Similarly, the number of bioactivities is also higher in the optimized libraries, between an additional 1149 bioactivities (+124% for the 1000-compound library) and 12,578 bioactivities (+100% for the 15,000-compound library), than the baseline libraries.



**Figure 7.** Radar charts visualizing the changes in the properties of the fittest library (solid black lines) of (A) 1000-compound library, (B) 5000-compound library, (C) 10,000-compound library, and (D) 15,000-compound library compared with the baseline populations (dashed black lines in each of the diagrams). The fitness values of the individual libraries are noted adjacent to the lines indicating the properties of the respective library.

### 2.2.3. Further Optimization of the Smaller-Sized Compound Libraries

As the smaller-sized compound sets were observed to benefit most from optimization by the genetic algorithm, we explored the possibility to further improve the sets of 1000 and 5000 compounds by re-running the genetic algorithm, this time with larger population sizes.

We first focus on the 1000-compound library. For this set, with a population size of 100 (Table 2), the optimization driven by the genetic algorithm yielded an increase in fitness by 34%. The larger population size of 500 (Table 3) led, over the course of its evolution, to an improvement in fitness by 46%, and with a population size of 1000, the improvement was 58% (Table 3).

Looking deeper at the 1000-compound libraries generated with population sizes of 100 (10% of the library size; Table 2) and 500 (Table 3), we see that the increase in final fitness values (1.29 vs. 1.42) correlates with an increase in the number of covered Murcko scaffolds (910 vs. 918), the number of targets (423 vs. 508), the number of covered Pfam families (282 vs. 303), and the number of predicted bioactivities (2074 vs. 2198). The increase in fitness is because the fitness function is designed to increase the score with repetitions in the predicted targets as the compound set is more likely to interact with a target when the target is predicted multiple times within the set.

The 1000-compound library generated with a population size of 1000 produced a 1000-compound library with the highest fitness score (1.56). The resulting 1000-compound library also covers more (930) scaffolds compared to 910 and 918 for the 1000-compound libraries generated with population sizes of 100 and 500, respectively. Clearly, during the course of this evolution, there is a slight increase in the number of scaffolds (from 928 to 930). This change occurred alongside an increase in the number of targets (from 433 to 710 targets from the start and end of the evolution) and Pfam families (from 198 to 323 Pfam families from the start and end of the evolution). As a result, despite only a slight increase in the number of represented scaffolds, the selected library is more populated with compounds that are predicted to interact with newer targets. This is also observed by the increase in the number of predicted bioactivities (from 1735 to 2743) and targets with higher novelty scores.

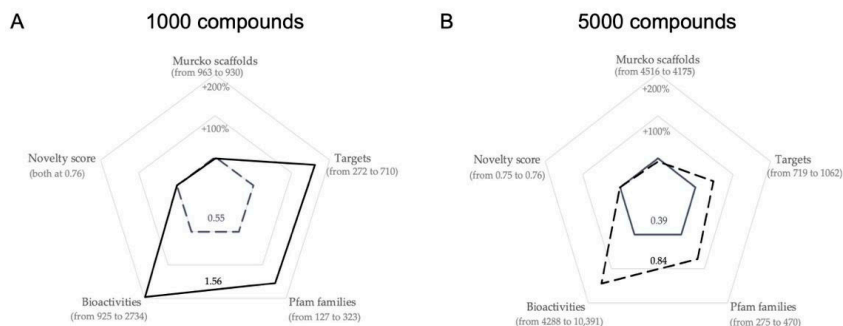
Considering the 5000-compound library, increasing the population size of the genetic algorithm from 500 individuals per generation (10% of the library size—Table 2) to 1000 and 5000 (Table 3) resulted in increased fitness of the best-scoring individuals (fitness of 0.71, 0.73 and 0.84, respectively). The improvement of the fittest individuals from the start and to the end of the evolution was also greater as the population size increased: a 20% improvement with a population size of 500, a 23% improvement when the population size was increased to 1000, and a 41% improvement for a population size of 5000. The fittest 5000-compound library generated from the population of 1000 individuals has an inverse relationship compared to the fittest library generated with the smaller population (500 individuals) on different fronts: a smaller number of scaffolds (4228 vs. 4193), fewer targets covered (991 vs. 917), and fewer Pfam families covered (441 vs. 412). The number of bioactivities, however, is higher (8863 vs. 8991), as is the median novelty score (0.75 vs. 0.76). This resulted in the improvement of the fitness score (0.71 vs. 0.73). Increasing the population size to 5000 resulted in a further increase in fitness for the 5000-compound library. Comparing the 5000-compound libraries generated with the populations of 1000 and 5000 individuals (with a fitness of 0.73 and 0.84, respectively), a different set of changes in the properties is observed with the increase in fitness. We still see that fewer and more promiscuous scaffolds are selected (4193 vs. 4175 Murcko scaffolds). However, this is coupled with more unique targets predicted (1062 vs. 917), more Pfam families covered (4210 vs. 470), and with more predicted bioactivities (8991 vs. 10391) for the compound libraries. The different modulations of the properties to achieve higher fitness is a result of the multiple parameters which must be optimized.

**Table 3.** Change in properties of the fittest individual (compound library) from the first generation to the fittest individual from after 300 generations for the 1000-compound and 5000-compound libraries using different population sizes for the genetic algorithm <sup>1</sup>.

Library Size	Population Size	% Change in Fitness Score	Δ Number of Murcko Scaffolds	Δ Number of Compound Clusters Represented	Δ Number of Targets	Δ Number of Pfam Families	Δ Number of Bioactivities	% Change in the Median Novelty Score
1000	500	+46.14% (0.97→1.42)	+6 (912→918)	0 (1000 in both)	+102 (406→508)	+101 (202→303)	+541 (1657→2198)	+2.26% (0.74→0.76)
1000	1000	+58.16% (0.99→1.56)	+2 (928→930)	0 (1000 in both)	+267 (443→710)	+125 (198→323)	+1008 (1735→2743)	+0% (both at 0.76)
5000	1000	+23.12% (0.59→0.73)	-59 (4252→4193)	0 (5000 in both)	+58 (859→917)	+15 (406→421)	+959 (8032→8991)	+1.33% (0.74→0.75)
5000	5000	+40.76% (0.60→0.84)	-17 (4192→4175)	0 (5000 in both)	+230 (832→1062)	+61 (409→470)	+2353 (8038→10,391)	+0.92% (0.75→0.76)

<sup>1</sup> The "→" symbol indicates a change in the value of a characteristic between the fittest individual at the start and the end of the evolution.

The 1000-compound library, which was optimized further with a population size of 1000, shows an improvement in fitness of 184% (0.55 vs. 1.56) over the baseline 1000-compound library (Figure 8A). This library also has an additional 438 (+161%) predicted targets, 194 (+154%) Pfam families and 1818 (+193%) predicted bioactivities compared to the baseline compound library. Likewise, the 5000-compound library, which was optimized further with a population size of 5000, has an improvement in fitness of 115% (0.39 vs. 0.84) over the baseline 5000-compound library (Figure 8B) and an additional 343 (+48%) predicted targets, 195 (+71%) Pfam families and 6103 (+142%) predicted bioactivities compared to the baseline compound library. These improvements show a clear benefit in optimizing compound libraries using this approach.



**Figure 8.** Radar charts comparing the change of properties between the baseline compound libraries and the further optimized compound libraries. The baseline compound libraries are depicted with dashed black lines for both the 1000-compound library generated with a population of 1000 (**A**), black continuous line and the 5000-compound library generated with a population size of 5000 (**B**), black continuous line.

### 3. Materials and Methods

#### 3.1. Data Sets

##### 3.1.1. ZINC20 Database

Via the ZINC20 web service [18,19], the 7,692,013 compounds annotated as “in-stock” [20] AND annotated as “anodyne” [20] AND assigned a charge state of  $-1$ ,  $0$  or  $+1$  AND assigned a calculated logP value between  $0$  and  $4$  were retrieved as SMILES strings from the ZINC20 database web service (Figure 1A) to be used as the source of compounds from which the optimized compound libraries would be generated.

##### 3.1.2. ChEMBL27 Database

From the ChEMBL27 database [22,23], the 2,156,988 bioactivity data records (i.e., compound-target pairs; Figure 1E) matching the following selection criteria (which are closely related to those used in previous works [21,27]) were retrieved:

1. Assay covers a single protein or a protein complex (ChEMBL confidence\_score is 6, 7, 8, or 9)
2. data\_validity\_comment is null OR “manually validated”
3. potential\_duplicate is “0”
4. standard\_type is “Kd”, “Potency”, “AC50”, “IC50”, “Ki”, or “EC50”
5. activity\_comment is not “Inconclusive”, “inconclusive”, or “unspecified”
6. NOT (standard\_relation is null AND activity\_comment is not “Active” or “active”)

Among the bioactivity records retrieved from the ChEMBL27 database, 4399 records had standard\_units of “ug.mL<sup>-1</sup>” as opposed to “nM” and therefore the standard\_value for these records was converted to nM using the canonical\_smiles and RDKit’s Descriptors.ExactMolWt function.

### 3.2. Chemical Structure Processing and Data Consolidation

The chemical structures from both the ZINC20 and ChEMBL27 data (Figure 1B) were standardized using the ChEMBL Structure Pipeline [28] to remove salt components (note that the compounds from the ZINC20 database do not include salt components) and solvent components and to neutralize any charges. Only compounds with molecular weight between 250 and 900 Da and composed of C, H, O, N, P, S, F, Cl, Br, and I atoms were retained. The SMILES string of the canonical tautomer of the compounds, as obtained from RDKit's TautomerEnumerator.canonicalize method, was recorded and used for further processing. In the case of the ZINC20 data, non-stereospecific SMILES were obtained and used to identify unique compounds based on their constitution (as information on the stereochemistry of the purchasable compounds is often incomplete or inaccurate).

Duplicate compounds in the processed ZINC20 data, resulting from the standardization process, were merged, resulting in 4,175,683 unique compounds. These ZINC20 compounds were further filtered for desirable molecular properties (see Molecular property filters for the ZINC20 data set).

In the case of ChEMBL27 data, when the standardization of the compounds resulted in duplicate compound-target pairs, the bioactivity records were merged and the median activity value of the merged records was set as the activity value for the compound-target pair. The 1,116,495 bioactivity records with activity values of less than or equal to 10,000 nM were labeled as "active" and were used as the reference data covering 661,839 compounds and 5170 targets for the similarity-based target prediction.

### 3.3. Filtering of the ZINC20 Subset by Molecular Properties

Following chemical structure processing, a series of molecular property filters (Figure 1C) were then used to remove any compounds with physicochemical properties that are unfavorable in the context of biochemical and cell research [11,17,29,30]. More specifically, any compounds matching any of the following criteria were removed:

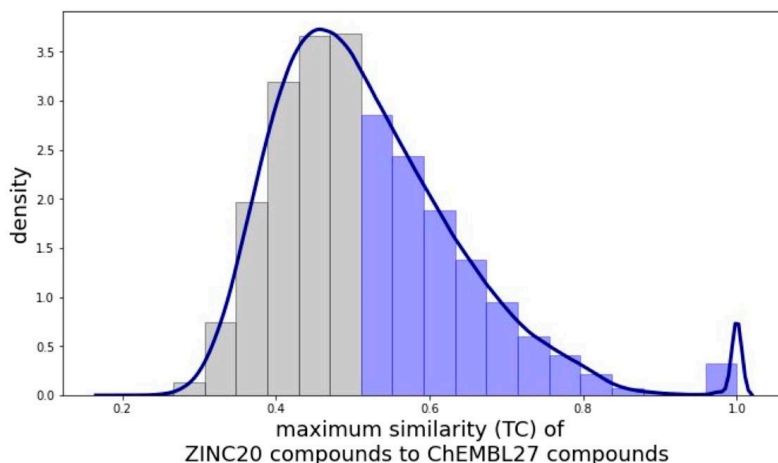
1. Less than 18 or more than 30 heavy atoms (calculated using RDKit's Lipinski.HeavyAtomCount method)
2. Less than one or more than four rings (calculated using RDKit's CalcNumRings method)
3. Ring systems with more than three fused rings (calculated using RDKit's GetRingInfo and AtomRings methods to get the ring systems and number of rings per system present a molecule)
4. More than eight rotatable bonds (rdMolDescriptors.CalcNumRotatableBonds)
5. More than three hydrogen bond donors (Lipinski.NumHDonors)
6. More than seven hydrogen bond acceptors (Lipinski.NumHAceptors)
7. Charged carbon atoms (identified using RDKit atom properties)
8. Not at least one oxygen or nitrogen atom (identified using RDKit atom properties)
9. Substructures listed in the "remove" and "extreme caution" categories of the SMARTS patterns compiled by Chakravorty et al. [31]. These SMARTS patterns were compiled from a meta-analysis of existing structural filters to identify nuisance compounds and correctly identified 57% of noisy GSK compounds in the study's validation [31]
10. Contain tosyl group (compounds which match the "S(=O)(=O)O" SMARTS pattern).

This filtering resulted in a final set of 2,572,351 ZINC20 compounds which were next subjected to target prediction.

### 3.4. Target Prediction

The targets of compounds were predicted based on their 2D molecular similarity (Figure 1D) of the query compounds (i.e., all the 2,572,351 processed compounds from ZINC20) to any of the compounds in the ChEMBL27 reference set (i.e., all the 661,839 processed compounds from the processed ChEMBL27 database with their 1,116,495 bioactivity records covering 5170 targets). More specifically, this search was performed using Morgan2 fingerprints and the search.knearest\_tanimoto\_search\_arena method implemented in chemfp [32]. Compound pairs with a Tanimoto coefficient of 0.5 or greater (Figure 9)

were retained (as we found previously that for these compound pairs the probability of predictions to be correct is 60% or higher; see Figure 3 in Ref. [21]).



**Figure 9.** Distribution of the maximum similarities (quantified as Tanimoto coefficient based on Morgan fingerprints with radius 2 and length of 2048 bits) of the compounds derived from the ZINC20 data set to the compounds of the ChEMBL27 reference set for target prediction (derived from the ChEMBL27 database). The line is the kernel density estimate while the bars are the normalized histogram of the pairwise similarities. The distribution shows a large number of dissimilar pairs and a long tail as similarity increases. This observation is consistent with existing knowledge that two random compounds are more likely to be dissimilar than similar [33,34]. Of all the 2,572,351 compounds on which a similarity search was carried out, nearly half the compounds (1,257,596) had a maximum similarity of less than 0.5 to the ChEMBL27 reference set (grey bars). This means that for these compounds no likely targets could be identified by the computational approach. For the purpose of this study, these compounds were hence regarded as “dark chemical matter” [35], and since the aim of this study is to generate compound libraries with the best coverage of the target space, these compounds were discarded. The remaining 1,314,755 compounds (blue bars) were assigned the ChEMBL27 compounds’ targets as predicted targets. These 1,314,755 unique ZINC20 compounds had a coverage of 3362 predicted targets and were retained as the pool of candidate compounds (PCC) from which the final, optimized compound libraries will be generated with the genetic algorithm. The PCC had a median Tanimoto coefficient of 0.59 to the ChEMBL27 reference set and 32,032 compounds (2% of the PCC) had the same Morgan fingerprints as compounds in the ChEMBL27 reference set resulting in the peak at Tanimoto coefficient of 1.

### 3.5. Additional Descriptions of the Compounds from the ZINC20 Database

In addition to the physicochemical properties that were used to filter the compounds from the ZINC20 database, Murcko scaffolds, QED score, logP, and compound clusters were used as additional descriptions of the compound sets. The number of Murcko scaffolds was calculated using the `PandasTools.AddMurckoToFrame` function of RDKit and counting the unique SMILES strings of the Murcko scaffolds. The QED scores were calculated using the `rdkit.Chem.QED.default` function of RDKit. The logP was calculated using RDKit’s `Descriptors.MolLogP` function. To cluster compounds, Dalke Scientific’s implementation [36] of the Taylor-Butina [37,38] clustering algorithm was utilized with Morgan 2 fingerprints and a Tanimoto threshold of 0.4.

### 3.6. Calculation of the Novelty of a Target

To map the diversity of targets, the targets were assigned to a Pfam family according to their protein sequence. Pfam is a large database of protein families that are represented by

hidden Markov models which describe these families with the goal of increasing coverage with as few models as possible [24,25]. Sequences of the proteins in the ChEMBL27 reference set were retrieved from the ChEMBL27 database. The sequences were then searched against the library of Pfam hidden Markov models (Pfam-A.hmm; version 33.1) using the "Pfam\_scan.pl" script (version 1.6) with default parameters and the "-clan\_overlap" option to get a family classification (Figure 1G). All proteins for which no automatic assignment to a Pfam family could be obtained were assigned to a single "dummy" family. This is to account for the fact that, while these proteins were not assigned to a family by Pfam, and are thus different from the ones assigned to Pfam families, it is unclear how similar or different the unassigned proteins are to each other. Therefore, a conservative approach is to assign them to the same family and assume that they are similar. Bioactivity records, through their targets, were then labeled with a Pfam family.

To calculate the novelty score for the Pfam families, the dates when the bioactivity records were recorded needed to be retrieved. Wherever possible, these dates were retrieved from the ChEMBL27 database by linking an activity (using the activity\_id) to its data source (the src.src\_id and the docs.docs\_id fields) and retrieving the year of publication (the docs.year field). For 343,389 of the bioactivity records, there was no date recorded in the ChEMBL27 database, and an attempt was made to find the relevant data when the primary data source (as recorded in the src\_id field in ChEMBL database) was the PubChem Bioassay database [39] using the Assay ID (AID) which is also recorded in the ChEMBL database. Of the 1,360,528 bioactivity records (records before merging identical SMILES-target pairs) used for target prediction, 62,734 (i.e., less than 5%) did not have a date assigned as there was no date information recorded in the ChEMBL database for these records and these records had primary sources other than PubChem and dates could not be retrieved.

A novelty score (Equation (1)) for each Pfam family was then calculated. In the case of 1 of 1214 of the Pfam families where a score cannot be calculated (i.e., because the denominator could not be calculated due to a lack of dates for the bioactivity records), the average novelty score (0.71) of the scored families was assigned as the novelty score. The compounds from the PCC, along with their predicted targets, the target's Pfam families, and target novelty scores, were then passed onto a genetic algorithm (Figure 1H) for subset selection, which was used to select an optimal subset of compounds for the compound libraries (Figure 1I).

### 3.7. Genetic Algorithm for Library Generation from the Pool of Candidate Compounds

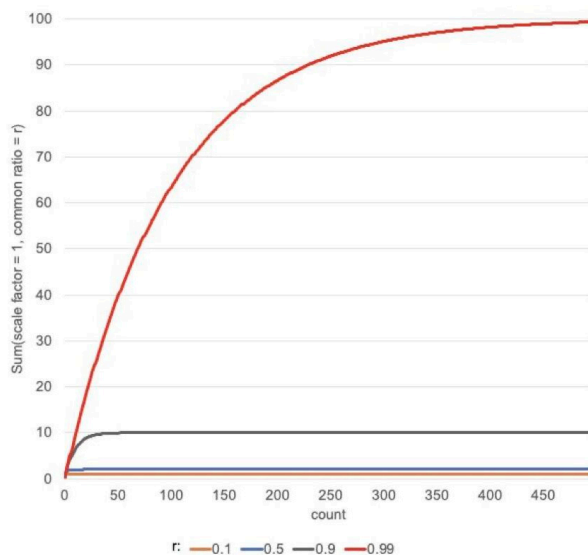
A genetic algorithm was implemented to optimize the selection of compounds for the compound libraries from the PCC. In this implementation, an individual is defined as a set of  $N$  (where  $N = 1000, 5000, 10,000,$  and  $15,000$ ) compounds. A population, composed of  $M$  individuals, then evolves over generations to produce an optimal population containing the optimal (most fit) individual which is the selected library.

#### 3.7.1. Calculation of the Fitness Function

The fitness of an individual, that is a set of compounds, is determined by the diversity and the novelty of the targets with which the compounds are predicted to interact. The fitness score (Equation (2)) of an individual was calculated to capture the properties we aimed to maximize in the set selection: we want to optimize for a set of compounds that are predicted to interact with a diverse set of targets. To capture this, the score includes a summation of the novelty of each individual Pfam family represented in the set.

When a Pfam family is represented multiple times in the predicted targets, the probability of a true interaction between the set of compounds and that family increases. Therefore, to capture the value of repeat predictions while still prioritizing diversity, the fitness function takes the form of a geometric progression (Equation (2)). This allows the score to increase with repeat family representation, however, the effect of the same family represented reduces with additional repeats. The scale factor of the geometric progression

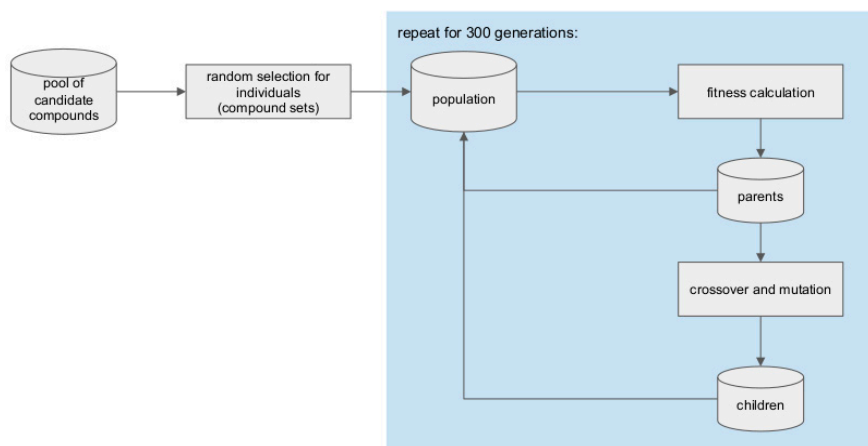
was set to the Pfam novelty score while the common ratio ( $r$ ) was set to 0.99, as at this value of  $r$  a slow plateau in the sum is observed (Figure 10). The fitness score (Equation (2)) for an individual compound set is, therefore, the sum of the geometric sums of each of the Pfam families (where the count is the number of compounds predicted to interact with the family and the scale factor is the family's novelty score), divided by the number of compounds in the individual.



**Figure 10.** Sum of a geometric progression ( $S = \frac{\text{scale factor}(1-r^{\text{count}})}{(1-r)}$ ) with a scale factor of 1 and varying values of the common ratio ( $r$ ) versus the count. When used to calculate the fitness score, a sum of a geometric progression is calculated for each Pfam family (where the novelty score is set as the scale factor, and the number of times the Pfam family is predicted is set as the count) and summed to get the fitness score (Equation (2)).

### 3.7.2. Library Optimization Procedure

The optimization of a compound library (Figure 11) begins by generating  $M$  (the population size) individuals, where each individual (i.e., compound set) is composed of  $N$  (the size of the compound library that is being generated) compounds randomly selected from the PCC. The fitness of each of the individuals in the population is calculated.



**Figure 11.** Schematic of the genetic algorithm which was implemented to select an optimal subset of compounds for the compound library.



One third of the population size is composed of selected parents from the current generation. The fittest individuals of the current generation are selected as the parents. The remaining individuals (i.e., two-thirds of the population size) are children produced by mating the parents. Each pair of parents produces four children by passing on half of their compounds (as determined by a single point crossover in the middle of each parent) to each child. When a child inherits the same compound from both parents, one of the occurrences of the compound is mutated to a randomly selected compound ensuring that the new compound which is selected does not already appear in the child. Children are also mutated, to add variation, by randomly replacing 10% of the compounds with new compounds from the PCC. The parents and children are then pooled together to form the population of size M for the next generation whose fitness is evaluated. This process is repeated over 300 generations and the fittest individual at the end of the evolution is chosen as the optimal individual for a compound library of size N compounds. The parameters of N and M are detailed in Table 4.

**Table 4.** The population size parameters used in the genetic algorithms to optimize a library of size N (N = 1000, 5000, 10,000, and 15,000).

Library Size/Size of the Individual (N)	Population Size (M)
1000	100, 500, 1000
5000	500, 1000, 5000
10,000	1000
15,000	1500

#### 4. Conclusions

In this study, we present a multi-step, computational approach for the design of small to medium-sized compound libraries that have a maximized likelihood of producing genuine hits in biological assays for an arbitrary target of interest. The approach takes multiple types of properties into account: drug-likeness, predicted bioactivities, biological space coverage, and target novelty. The hits identified by screening these compound libraries could serve as valuable tool compounds in biochemical and cell research, and some of them may also prove to be valid starting points for the development of drugs.

We have found that for all sizes of the compound libraries we generated (i.e., 1000, 5000, 10,000, and 15,000 compounds) the genetic algorithm improved the quality of the compound sets, with the individual libraries' fitnesses improving up to 58%. The genetic algorithm was initially run with populations of 10% of the size of the library. As the smaller libraries (consisting of 1000 or 5000 compounds) benefitted the most from the optimization, further evolutions with larger population sizes were run, increasing the fitness of these libraries even more. In all cases, the objective fitness values, generated from the fitness function, increased through the course of evolution.

Multiple properties of the libraries were analyzed: number of Murcko scaffolds, the number of Taylor-Butina clusters, number of predicted targets, number of Pfam families of the predicted targets, number of predicted bioactivities, and the novelty scores of the predicted targets. These properties were modulated differently during the course of the optimizations to produce fitter libraries. In some cases, more diverse compound sets (as measured by a change in the number of Murcko scaffolds) were selected for the libraries which resulted in activity on a more diverse set of predicted targets. In other cases, compounds with more promiscuous scaffolds were selected which increased the number of targets they were predicted to be bioactive on. The modulation of these multiple objectives to produce better libraries highlights the appropriateness of our fitness function.

The benefits of utilizing target prediction and a genetic algorithm to optimize compound libraries are best seen when comparing fitness values of the optimized compound libraries with that of the baseline libraries (which do not account for predicted targets and have not been evolved). The largest improvement in fitness (+184%) was observed for the 1000-compound library generated with a population size of 1000, while the small-

est improvement in fitness (+60%) was observed for the 15,000-compound library generated with a population size of 1500. The best of the optimized compound libraries prepared in this work are available for download as a dataset bundle (“BonMOLière”) from <https://doi.org/10.5281/zenodo.5114733> (accessed on 19 July 2021).

**Author Contributions:** Conceptualization, N.M., C.S. and J.K.; methodology, N.M., C.S. and J.K.; software, N.M.; validation, N.M. and C.S.; formal analysis, N.M. and J.K.; investigation, N.M. and J.K.; resources, J.K.; data curation, N.M.; writing—original draft preparation, N.M., C.S. and J.K.; visualization, N.M. and J.K.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** N.M. and J.K. are supported by the Trond Mohn Foundation (BFS2017TMT01). J.K. and C.S. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number KI 2085/1-1.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The best of the compound libraries presented in this work, including SMILES strings, ZINC IDs, and the ChEMBL IDs of the predicted targets), are available for download as a dataset bundle (“BonMOLière”) from <https://doi.org/10.5281/zenodo.5114733> (accessed on 19 July 2021) [40].

**Acknowledgments:** The authors wish to thank Andrew Dalke from Andrew Dalke Scientific Software for a license for chemfp, Steffen Hirte from the University of Vienna for valuable discussions on the fitness function, and Sophie Fischer from the University of Bergen for valuable discussions and assistance with proofreading. A portion of the calculations described in this work were performed on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway. Open Access Funding by the University of Vienna.

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

## References

1. Macarron, R.; Banks, M.N.; Bojanic, D.; Burns, D.J.; Cirovic, D.A.; Garyantes, T.; Green, D.V.S.; Hertzberg, R.P.; Janzen, W.P.; Paslay, J.W.; et al. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10*, 188–195. [CrossRef]
2. Drewry, D.H.; Macarron, R. Enhancements of Screening Collections to Address Areas of Unmet Medical Need: An Industry Perspective. *Curr. Opin. Chem. Biol.* **2010**, *14*, 289–298. [CrossRef] [PubMed]
3. Baell, J.B. Broad Coverage of Commercially Available Lead-like Screening Space with Fewer than 350,000 Compounds. *J. Chem. Inf. Model.* **2013**, *53*, 39–55. [CrossRef]
4. Paricharak, S.; Méndez-Lucio, O.; Chavan Ravindranath, A.; Bender, A.; IJzerman, A.P.; van Westen, G.J.P. Data-Driven Approaches Used for Compound Library Design, Hit Triage and Bioactivity Modeling in High-Throughput Screening. *Brief Bioinform.* **2018**, *19*, 277–285. [CrossRef] [PubMed]
5. Wassermann, A.M.; Camargo, L.M.; Auld, D.S. Composition and Applications of Focus Libraries to Phenotypic Assays. *Front. Pharmacol.* **2014**, *5*, 164. [CrossRef]
6. Petrone, P.M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J.W.; Jenkins, J.L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399–1409. [CrossRef]
7. Janzen, W.P. Screening Technologies for Small Molecule Discovery: The State of the Art. *Chem. Biol.* **2014**, *21*, 1162–1170. [CrossRef]
8. Bakken, G.A.; Bell, A.S.; Boehm, M.; Everett, J.R.; Gonzales, R.; Hepworth, D.; Klug-McLeod, J.L.; Lanfear, J.; Loesel, J.; Mathias, J.; et al. Shaping a Screening File for Maximal Lead Discovery Efficiency and Effectiveness: Elimination of Molecular Redundancy. *J. Chem. Inf. Model.* **2012**, *52*, 2937–2949. [CrossRef] [PubMed]
9. Lahue, B.R.; Glick, M.; Tudor, M.; Johnson, S.A.; Diratsouian, J.; Wildey, M.J.; Burton, M.; Mazzola, R.; Wassermann, A.M. Diversity & Tractability Revisited in Collaborative Small Molecule Phenotypic Screening Library Design. *Bioorg. Med. Chem.* **2020**, *28*, 115192. [CrossRef]
10. Stork, C.; Kirchmair, J. PAIN(S) Relievers for Medicinal Chemists: How Computational Methods Can Assist in Hit Evaluation. *Future Med. Chem.* **2018**, *10*, 1533–1535. [CrossRef]
11. Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I.H.; Frearson, J.; Wyatt, P.G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435–444. [CrossRef]

12. Spear, K.L.; Brown, S.P. The Evolution of Library Design: Crafting Smart Compound Collections for Phenotypic Screens. *Drug Discov. Today Technol.* **2017**, *61*–67. [CrossRef]
13. Haasen, D.; Schopfer, U.; Antczak, C.; Guy, C.; Fuchs, F.; Selzer, P. How Phenotypic Screening Influenced Drug Discovery: Lessons from Five Years of Practice. *Assay Drug Dev. Technol.* **2017**, *15*, 239–246. [CrossRef] [PubMed]
14. Huggins, D.J.; Venkitaraman, A.R.; Spring, D.R. Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Biol.* **2011**, *6*, 208–217. [CrossRef]
15. Baell, J.B.; Holloway, G.A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740. [CrossRef]
16. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef] [PubMed]
17. Schuffenhauer, A.; Schneider, N.; Hintermann, S.; Auld, D.; Blank, J.; Cotesta, S.; Engeloch, C.; Fechner, N.; Gaul, C.; Giovannoni, J.; et al. Evolution of Novartis' Small Molecule Screening Deck Design. *J. Med. Chem.* **2020**, *63*, 14425–14447. [CrossRef] [PubMed]
18. Irwin, J.J.; Tang, K.G.; Young, J.; Dandarchuluun, C.; Wong, B.R.; Khurelbaatar, M.; Moroz, Y.S.; Mayfield, J.; Sayle, R.A. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073. [CrossRef]
19. ZINC20. Available online: <http://zinc20.docking.org> (accessed on 26 May 2021).
20. Sterling, T.; Irwin, J.J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef]
21. Mathai, N.; Kirchmair, J. Similarity-Based Methods and Machine Learning Approaches for Target Prediction in Early Drug Discovery: Performance and Scope. *Int. J. Mol. Sci.* **2020**, *21*, 3585. [CrossRef]
22. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef] [PubMed]
23. Gaulton, A. ChEMBL\_27 SARS-CoV-2 Release. Available online: <http://chembl.blogspot.com/2020/05/chembl27-sars-cov-2-release.html> (accessed on 12 March 2021).
24. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2020**, *49*, D412–D419. [CrossRef] [PubMed]
25. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam Protein Families Database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [CrossRef] [PubMed]
26. RDKit: Open-Source Cheminformatics. Available online: <http://www.rdkit.org--version2020.09.1.0> (accessed on 8 July 2021).
27. Bosc, N.; Atkinson, F.; Felix, E.; Hersey, A.; Leach, A.R. Large Scale Comparison of QSAR and Conformal Prediction Methods and Their Applications in Drug Discovery. *J. Cheminform.* **2019**, *11*, 4. [CrossRef]
28. Patrícia Bento, A.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L.J.; De Veij, M.; Leach, A.R. An Open Source Chemical Structure Curation Pipeline Using RDKit. *J. Cheminform.* **2020**, *12*, 1–16. [CrossRef]
29. Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902. [CrossRef]
30. Pearce, B.C.; Sofia, M.J.; Good, A.C.; Drexler, D.M.; Stock, D.A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068. [CrossRef]
31. Chakravorty, S.J.; Chan, J.; Greenwood, M.N.; Popa-Burke, I.; Remlinger, K.S.; Pickett, S.D.; Green, D.V.S.; Fillmore, M.C.; Dean, T.W.; Luengo, J.I.; et al. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection. *SLAS Discov.* **2018**, *23*, 532–545. [CrossRef]
32. Dalke, A. The Chemfp Project. *J. Cheminformatics* **2019**, *11*, 76. [CrossRef]
33. Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput. Biol.* **2013**, *9*, e1003302. [CrossRef]
34. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204. [CrossRef]
35. Wassermann, A.M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F.J.; Studer, C.; Peltier, J.M.; Grippo, M.L.; Prindle, V.; Tao, J.; et al. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11*, 958–966. [CrossRef] [PubMed]
36. Chemfp Taylor Butina Implementation. Available online: [http://dalkescientific.com/writings/taylor\\_butina.py](http://dalkescientific.com/writings/taylor_butina.py) (accessed on 26 March 2021).
37. Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67. [CrossRef]
38. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750. [CrossRef]
39. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef] [PubMed]
40. Mathai, N.; Stork, C.; Kirchmair, J. BonMOLière: Small-Sized Libraries of Readily Purchasable Compounds, Optimized to Produce Genuine Hits in Biological Screens across the Protein Space; CERN: Genève, Switzerland, 2021. [CrossRef]

# Chapter 7

## Concluding discussions and future prospects

In-silico tools that can aid the identification of the macromolecular targets of small organic compounds are important to early-stage drug discovery and drug re-purposing efforts. As such, the development of target prediction methods is an area of active research. In this thesis, we analyzed the various strategies employed to evaluate target prediction methods and proposed validation strategies to overcome the limitations of only measuring and relying on generalized performance measures (**P1**). We developed and thoroughly evaluated two ligand-based target prediction methods: a similarity-based approach and a binary relevance random forest based (ML) approach (**P2**). These methods utilize the maximum amount of data available to cover as much of the target space as is possible. Finally, we applied target prediction and a genetic algorithm to curate a selection of compounds for small to medium-sized compound libraries for experimental screening (**P3**). We enriched (with respect to randomly selecting compounds for screening) these libraries with biologically diverse, bioactive compounds so as to maximize the likelihood of finding a hit for a wide range of targets. The findings presented in this thesis contribute to an area of active research to develop and apply in silico methods to predict the bioactivity of small organic compounds on biomacromolecular targets. Elucidating the interactions between compounds and targets is key to understanding the mode of action of drugs, drug repurposing, predicting side effects of possible drug candidates etc. making the process of drug development more efficient and data driven.

## 7.1 Evaluating target prediction methods

Target prediction methods have been evaluated in a variety of ways with various degrees of rigor. We analyzed existing strategies and developed new guidelines on validation strategies for target prediction methods. In an ideal scenario, a newly developed target prediction method would be evaluated using large-scale prospective testing. That is, a large volume of predicted bioactivities would be prospectively validated through experiments. However, as large-scale prospective validation is prohibitively expensive, developers of target prediction methods rely on retrospective validation. With retrospective validation, existing data on the bioactivity of compounds on targets are used as test data to evaluate the performance of a method.

There are multiple ways to utilize data to train and test a model. Models should always have performance reported based on truly external test data. That is, these data have not been used in model training in any way. To maximize the use of data and get a better sense of the generalized performance of a model, we argue that nested cross-validation should be used whenever possible. Partitioning the data in this way makes maximum use of the available data while still allowing for external data for model evaluation. With nested cross-validation, an inner-loop with partitioned data is used to evaluate the best hyperparameters for a model. Meanwhile, the outer-loop contains data which did not influence the hyperparameter selection, and is therefore external to model training, and used to measure the performance of the model. These data partitioning schemes must take computational costs into account. When a target prediction method is made up of a single model, for example, nested cross-validation repeated multiple times would give a good sense of the method's performance. This may not be feasible when a target prediction method is composed of multiple models. In all cases however, a large volume of truly external data should be used to evaluate the model's performance.

The performance of target prediction methods is often quantified using classic ML metrics. These metrics are derived from a confusion matrix of all the compounds in a test set for a model. While averaged performance metrics are indeed valuable and do measure the overall performance, they leave the user unable to decipher what the performance of the model may be with respect to their specific compound of interest. In this work, we argued that this leaves a crucial question about how the model will perform on individual compounds unanswered. That is, the performance with respect to the applicability domain of the model is not well understood. We developed guidelines to validate target prediction approaches. We advocated for performance measures to be disaggregated by how similar the test compounds are to the training compounds. This disaggregation allows for a more nuanced understanding of the performance and provides and an

easy-to-understand measure of the applicability domain. Disaggregating performance by similarity does however have to be balanced with the number of samples available per bin. Selecting a bin size which is too small may result with too few test data points in bins making the results meaningless, while a large bin size reduces the nuance sought. It is therefore important to ensure that the test data is diverse, and is reflective of the real world scenario, ensuring that a wide spectrum of compounds are tested.

## 7.2 Development and validation of target prediction methods

A key component of a generalized target prediction method, which is not focused on a particular target family for example, is having a broad target coverage. Ligand-based target prediction methods offer a broader target coverage compared to structure-based methods, as they do not rely on 3D structure of targets. Nevertheless, a limiting factor in the coverage of a ligand-based target prediction method is the amount of bioactivity data available for different targets. We developed and thoroughly validated the performance of two target prediction approaches: a similarity-based approach and a random forest binary relevance based (ML) approach. Both approaches were designed to maximize the coverage of the target space, that is, be able to make predictions for as many targets as possible.

Notably, for the first time, the performances of the approaches were tested under three test scenarios: a standard testing scenario, a standard time-split testing scenario and a close-to-real-world testing scenario. The performances were also measured with two metrics: success rates and recovery rates. The metrics and scenarios are described in detail in Section 3.5. As the success and recovery rates, for both the similarity-based and ML approach showed identical trends, this discussion focuses on the success rates to convey the differences the the performance of the two approaches. The performances were desegregated by the similarity of the test queries to their targets' ligand sets. High-similarity queries had a median TC between 0.66 and 1 to the most similar ligand(s) of the query's targets, medium-similarity queries had a median TC between 0.33 and 0.66, and low-similarity queries had a median TC between 0 and 0.33.

The knowledge base for the target prediction approaches, extracted from the ChEMBL database (version 24) and pre-processed, consisted of 492,282 compounds, 4,617 targets and 914,057 measured bioactivities. The bioactivities were labeled as "active" when the activity value of the compound-target pair was 10,000 nM or less, and as "inactive" when the activity value was 20,000 nM or more. A target needed just one active bioactivity

recorded in the knowledge base to be represented by the similarity-based approach, leading to a coverage of 4239 proteins. The ML approach was built with individual random forest models (one model per target) each of which classified a test compound as “active” or “inactive”. The random forest models were therefore built using both active and inactive bioactivity records. A model was built for all targets with at least 25 active bioactivity records, resulting in a target coverage of 1,798 proteins.

The similarity-based approach performed just as well as this ML approach. Under the standard testing scenario, the similarity-based approach identified a known target for 86%, 88% and 93% of all queries within the top-3, top-5 and top-15 ranked targets. In fact, the approach ranked a known target for 95% of all high-similarity queries at the top position. Medium-similarity queries had success rates between 55% for the queries among the top-3 to 82% of the queries in among the top-15 ranked targets. There was a further drop in success rates for low-similarity queries, with targets ranked for 10% of the queries among the top-3 to 18% among the top-15. Under the time-split scenario, the approach had overall success rates from 58%, among the top-3 targets, to 69%, among the top-15 targets, which is approximately 25 to 30 percentage points lower than the standard testing scenario. The difference in the overall success rates is due to the fact that 50% of ChEMBL 24 test data had a median TC greater than 0.8, which dropped down to 0.4 for the ChEMBL 25 test data used in the time-split scenario. That is, the ChEMBL25 test data is more different to the knowledge base than the ChEMBL24 test data. The success rates under the time-split scenario are comparable with those obtained under the standard testing scenarios for queries with the same molecular similarity. The time-split scenario may, therefore, not be essential if performances are desegregated by molecular similarity. Under the close-to-real-world testing scenario, 4% of the interactions introduced with version 25 of the ChEMBL database were not represented by the knowledge base. As a result, the success rate performance dropped by 2 to 3 percentage points over the standard time-split scenario.

Under the standard testing scenario, the ML approach ranked a target for 82%, 86% and 91% of compounds among the top-3, top-5 and top-15 compounds respectively. These are three percentage points lower than the similarity-based approach. With the ML approach, only 90% of the high-similarity compounds had a target ranked in the top position vs. 95% for the similarity-based approach. The performance of the ML approach on medium-similarity queries was up to 7 percentage points lower than the similarity-based approach (49% and 75% for the top-3 and top-15 ranks respectively). The low-similarity queries had success rates between 10% (top-3) to 28% (top-15), the latter of which is marginally better than the similarity-based approach (18% of the low-similarity queries had a target in the top-15). Just like the similarity-based approach, the ML approach had a 25 to 30 point percentage drop in performance (53% for the

top-3 to 65% for the top-5) under the time-split scenario compared to the standard testing scenario. The improved performance of the ML on low-similarity queries for the top-15 targets was not observed with the time-split validation scenario and is therefore considered to be an artifact of the data and not a generalized trend. Under the close-to-real-world testing scenario, 11% of the interactions were not represented in the knowledge base of the ML approach (as opposed to 4% for the similarity-based approach). This resulted in a further 4 to 6 percentage point drop in performance compared to the standard testing scenario.

The similarity-based approach shows better performance than this binary relevance ML approach, with a target coverage that is twice as large. This is likely due to the fact that the binary relevance decomposition of a multi-label problem does not take label dependencies into account [122] and the relationship between the targets is not well accounted for. The similarities across different targets' ligand sets, which was used to rank targets for the similarity based method, provides a better way to rank targets compared to the prediction probabilities of belonging to the active class from the random forest models, for the ML approach. The active class probabilities may be more comparable once they are calibrated using scaling techniques, such as Venn-ABERS predictors [123]. However these techniques are more data hungry as they require additional data to be set aside for calibration sets, further reducing the target scope of the approach. As more data becomes available, this ML approach may be improved by scaling the probabilities produced by individual models as well as improving the performance of the individual models themselves. An increase in quality bioactivity data would also allow for the development of multi-task prediction models with a wide target coverage (e.g. a recent multi-task target prediction classifier was built to predict bioactivity on 616 ChEMBL targets [124-126]). It will be crucial to benchmark these multitask models and compare their performance and scope to a similarity-based approach. These benchmarking efforts will push the development of state-of-the-art target prediction methods.

## 7.3 Applying target prediction to curate screening libraries

Having demonstrated the value of the similarity-based target prediction approach, both in terms of the target scope and prediction power, we applied the similarity-based approach to curate compound sets for screening libraries. Screening libraries of compounds, via bench based screens (cell-based or biochemical screens) and/or virtual screening campaigns are routine in early drug discovery projects. General screening libraries are de-



signed to be chemically and/or biologically diverse to increase the chances of finding a hit among the set of compounds for a any of a wide range of targets. The success of screening libraries varies depending on the target of interest and the composition of the library. Small to medium-sized general purpose libraries, which contain high quality compounds, have a higher-than-average chance of producing a hit on an arbitrary compound of interest, are particularly valuable to resource constrained environments such as small academic labs. Hits from these libraries could be used as starting tool compounds in the drug development process. We applied a similarity based target prediction approach followed by a genetic algorithm to optimize small to medium-sized general purpose libraries.

The process of generating the screening libraries began by gathering a pool of candidate compounds (PCC). Compounds deemed to be unreactive are retrieved from the ZINC20 [117] web service and were passed through a series of physicochemical filters to bias the pool towards compounds with drug-like properties. The targets for the retrieved compounds, which passed through the filters, were predicted using a similarity based approach and a processed knowledge base from the ChEMBL database (version 27). This resulted in the PCC of over 1.3 million candidate compounds from which the small to medium-sized (1,000, 5,000, 10,000, and 15,000 compounds) libraries were generated. As the library sizes were much smaller than the number of candidate compounds, the number of possible combinations of compounds which could form the library was too large to enumerate to find the optimum combination. As this is a subset selection problem, it is an optimization problem. As such, a genetic algorithm was utilized, to select an optimal combination of compounds for the libraries. In this implementation, a gene was a compound, an individual was a set of compounds, and a population was a collection of compound sets. A population was evolved for 300 generations to generate fitter individuals and at the end of the evolution the fittest individual (i.e. compound set) was selected as the compound library.

Defining what “optimal” is and how to measure the fitness of a set of compounds for the library is key to curating a library. We aimed to compile a library which contained compounds which have a high likelihood of interacting with a diverse range of targets with a bias towards “newer” targets of interest. To capture the diversity of the targets, the Pfam families of the targets were retrieved. Pfam is a database of evolutionary related proteins (families) which are represented by hidden Markov models [127, 128]. The diversity of the predicted targets for a set of compounds was measured by the number of different Pfam families linked to compounds by their predicted biological activity. To quantify the “newness” of a target, a novelty score was calculated for each Pfam family by looking at the dates (as recorded in ChEMBL or PubChem when dates were not available in ChEMBL) of the bioactivities of the targets that are labeled with a Pfam

family. Therefore, by getting all the unique Pfam families represented by a compound library, and summing the novelty scores of the Pfam families represented, a fitness score of an individual could be calculated. A higher fitness score would therefore signify a library which has better balanced the following:

1. a compound library which is enriched with bioactive compounds as a whole
2. a compound library which is enriched with compounds active on proteins representing more Pfam families (diversity)
3. a compound library which is enriched with compounds active on newer targets (as the novelty score is higher for newer targets)

As a whole, the library is more likely to be active on a member of a Pfam family if that family has been predicted multiple times within the library through different compounds. Therefore the fitness function was not a simple sum of the novelty scores of the unique Pfam families predicted. Rather, it was formulated as a sum of the geometric progressions where the scale factor is the novelty score of the Pfam family and the count is the number of times the Pfam family has been predicted for a set of compounds. This allowed for the benefit of the repeat predictions to be accounted for while still maximizing the diversity and trying to bias towards newer targets/Pfam families. We used the sum of the geometric progressions, as opposed to a simple sum of all the Pfam novelty scores including the repeats, as the weight of the repeated Pfams to decreased as the repeat count increased so as to still prioritize diversity.

This approach, using target prediction to compile the PCC followed by a genetic algorithm to optimize the library, lead to improvements for all library sizes when compared to randomly selecting compounds that only passed through the physicochemical filters (known as baseline libraries). The smaller the library size, the greater the improvement in fitness scores when compared to the baseline libraries. The largest improvement in fitness was observed for the 1,000-compound library (+184%) while the smallest improvement was for the 15,000-compound library (+60%).

When looking at the change in fitness through the course of an evolution, again, the smaller the library size, the greater the improvements. When the genetic algorithm was run with population sizes of 10% of the size of the library, the fitness improved by +34%, +20%, +17% and +14% for the 1,000, 5,000, 10,000 and 15,000-compound library respectively. Additional evolutionary runs were carried out for the 1,000 and 5,000-compound libraries, increasing the population size to further improve the fitness of the libraries. The 1,000-compound library had an improvement of +58% in the fitness score, during the course of the evolution, when generated with a population of 1,000

individuals. Similarly the 5,000-compound library had an improvement of +41% in the fitness score when generated with a population of 5,000 individuals.

Beyond the objective fitness function values of the libraries, the change in the number of Murcko scaffolds, the number of Taylor-Butina clusters [129, 130], number of predicted targets, number of Pfam families of the predicted targets, number of predicted bioactivities, and the novelty scores of the predicted targets were analyzed. These five properties were modulated differently during the course of the different evolutionary runs to produce fitter libraries. In some cases, more diverse compound sets (as measured by a change in the number of Murcko scaffolds) were selected for the libraries which resulted in activity on a more diverse set of predicted targets. In other cases, compounds with more promiscuous scaffolds were selected which increased the number of targets they are predicted to be active on. In all cases, every compound within a library was from a different Taylor-Butina cluster (clustered with a TC threshold of 0.4) which speaks to chemical space diversity of the libraries. The modulation of these multiple objectives to produce better libraries highlights the appropriateness of our fitness function and the use of the genetic algorithm for this optimization task. The optimized compound libraries, BonMOLière, of the different sizes have been made available to the community.

## 7.4 Concluding remarks

In this thesis we present the development, validation and application of large-scale target prediction methods. Understanding the interactions between compounds and their biomolecular targets gives insights into the mode of action of compounds. This allows for a deeper understanding on what targets a compound modulates to achieve an effect, including the side effects it may elicit. Target prediction is particularly useful to drug repositioning efforts, where the compounds which have gone through the thorough and costly process of safety trials are approved to treat alternative ailments. Predicting the biomolecular targets of small molecules is a key tool in early drug discovery, the focus of this thesis, as well other chemical industries such as cosmetics and agrochemicals.

We developed and thoroughly validated two target prediction approaches: a similarity-based approach and a ML approach. These approaches were designed to fully maximize the use of the available data and to develop target prediction methods with a large target coverage. We showed that the similarity-based approach performed just as well as the binary-relevance random forest based (ML) approach, while having a target coverage which was more than twice the size. The development of methods in the exciting domain of target prediction will continue to be fueled as the amount of bioactivity data for model

building increases. The possibility of developing a large-scale target prediction method which uses well calibrated ML models (such as via conformal prediction models, scaling methods for binary relevance approaches or a multitask model with a wide target scope) becomes more of a reality with more data. In our opinion it is worthwhile to further develop machine learning models for target prediction which generalize better than a similarity-based approach to increase the performance for the prediction of the targets of low-similarity queries. The development of well calibrated ML models would allow for the development of hybrid target prediction approaches, which capitalize on accuracy of similarity-based approaches for high-similarity queries and would use well calibrated and generalized ML approaches for low-similarity queries.

Having evaluated the performance of two target prediction approaches, we applied the similarity-based target prediction approach to curate general purpose screening libraries which have a higher-than-average chance of producing hits on an arbitrary target of interest. Readily purchasable compounds with predicted targets form the PCC from which compounds are selected for screening libraries. We aimed to produce libraries that were enriched with compounds that are likely to be bioactive on diverse targets. As this selection is a subset selection problem, a genetic algorithm is used to optimize the selection. Applying target prediction followed by a genetic algorithm for compound selection lead to improvements ranging from +60% (for a library of 15,000 compounds) to +184% (for a library of 1,000 compounds) in the the fitness of the libraries. The optimized libraries have been made available to the community. It would be valuable to run large-scale prospective validations using these enriched libraries which would test the performance of the target prediction and in-turn demonstrate the value of utilizing target prediction to enrich screening libraries. As a result of wanting to be agnostic to chemical vendors, the BonMOLière data set was optimized with compounds that may be sourced from any of the 150 companies featured on the ZINC20 database. Repeating this study from popular vendors, such as MolPort [131] or Enamine [132], would provide the community with a compound library which could be more easily sourced. Of course, over time vendor stocks change and any compound library will, over time, become outdated. An easily accessible tool which allows a user to upload a user defined super set of compounds on which target prediction followed by the genetic algorithm subset selection could be conducted, returning an optimized library to a user, would be a valuable tool to the community. Such a tool would be especially useful for small academic labs for whom running large screens is prohibitive.

Over the course of this thesis, we developed two target prediction approaches, which maximized the use of available data to ensure a large coverage of the target space. These approaches were thoroughly validated and showed that the similarity-based approach performs just as well as the binary relevance random forest based ML approach, while

covering a much larger target space. There is therefore scope to build better ML models for target prediction while still ensuring that a wide target scope is maintained. We have also demonstrated the value of applying target prediction on a large-scale to curate screening libraries, and provided these enriched libraries to the community. We hope that the insights from this work will be used to push the field of target prediction, and therefore data driven drug design, further.

# Bibliography

- [1] Seyhan, A. A. Lost in translation: the valley of death across preclinical and clinical divide—identification of problems and overcoming obstacles. *Translational Medicine Communications* **2019**, *4*, 1–19.
- [2] Ban, T. A. The role of serendipity in drug discovery. *Dialogues in clinical neuroscience* **2006**, *8*, 335.
- [3] Hansch, C. Drug research or the luck of the draw. *Journal of Chemical Education* **1974**, *51*, 360.
- [4] Curtis, M. Drug discovery challenges now the low hanging fruit has been harvested. <https://www.bps.ac.uk/publishing/blog/september-2018/drug-discovery-challenges-now-the-low-hanging-frui>, (accessed on 2021-04-06).
- [5] Williams, M. Productivity shortfalls in drug discovery: contributions from the pre-clinical sciences? *Journal of Pharmacology and Experimental Therapeutics* **2011**, *336*, 3–8.
- [6] Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research* **2015**, *38*, 1686–1701.
- [7] Wouters, O. J.; McKee, M.; Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Journal of the American Medical Association* **2020**, *323*, 844–853.
- [8] DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics* **2016**, *47*, 20–33.
- [9] DiMasi, J. A. Research and development costs of new drugs. *Journal of the American Medical Association* **2020**, *324*, 517–517.

- [10] U.S.F.D.A, The drug development process. <https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process>, (accessed on 2021-04-06).
- [11] Bajorath, J. Computer-aided drug discovery. *F1000Research* **2015**, *4*.
- [12] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational methods in drug discovery. *Pharmacological Reviews* **2014**, *66*, 334–395.
- [13] Leelananda, S. P.; Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **2016**, *12*, 2694–2718.
- [14] Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug discovery* **2017**, *16*, 531–543.
- [15] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.
- [16] Chaudhari, R.; Tan, Z.; Huang, B.; Zhang, S. Computational polypharmacology: a new paradigm for drug discovery. *Expert Opinion on Drug Discovery* **2017**, *12*, 279–291.
- [17] Zhou, H.; Gao, M.; Skolnick, J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific Reports* **2015**, *5*, 11090.
- [18] Cereto-Massagué, A.; Ojeda, M. J. M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in silico target fishing. *Methods* **2015**, *71*, 98–103.
- [19] Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A web server for target prediction of bioactive small molecules. *Nucleic Acids Research* **2014**, *42*, W32–W38.
- [20] Peón, A.; Naulaerts, S.; Ballester, P. J. Predicting the Reliability of Drug-target Interaction Predictions with Maximum Coverage of Target Space. *Scientific Reports* **2017**, *7*, 3820.
- [21] Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in silico target prediction to multi-target drug design: Current databases, methods and applications. *Journal of Proteomics* **2011**, *74*, 2554–2574.
- [22] Park, K. A review of computational drug repurposing. 2019; <https://synapse.koreamed.org/DOIx.php?id=10.12793/tcp.2019.27.2.59>.

- [23] Vilar, S.; Hripcsak, G. The role of drug profiles as similarity metrics: Applications to repurposing, adverse effects detection and drug-drug interactions. *Briefings in Bioinformatics* **2017**, *18*, 670–681.
- [24] Sam, E.; Athri, P. Web-based drug repurposing tools: a survey. *Briefings in Bioinformatics* **2017**, 1–18.
- [25] Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **2018**, *557*, S55–S55.
- [26] Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews* **2019**, *119*, 10520–10594.
- [27] Rodrigues, T.; Bernardes, G. J. Machine learning for target discovery in drug development. *Current opinion in chemical biology* **2020**, *56*, 16–22.
- [28] Wang, C.; Kurgan, L. Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Briefings in Bioinformatics* **2019**, *20*, 2066–2087.
- [29] Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. 2018.
- [30] Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today* **2015**, *20*, 318–331.
- [31] Huang, H.; Zhang, G.; Zhou, Y.; Lin, C.; Chen, S.; Lin, Y.; Mai, S.; Huang, Z. Reverse screening methods to search for the protein targets of chemopreventive compounds. 2018; [www.frontiersin.org](http://www.frontiersin.org).
- [32] Lotfi Shahreza, M.; Ghadiri, N.; Mousavi, S. R.; Varshosaz, J.; Green, J. R. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics* **2018**, *19*, 878–892.
- [33] Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J., et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* **2015**, *6*, 24–50.
- [34] Ezzat, A.; Wu, M.; Li, X.-L.; Kwok, C.-K. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics* **2019**, *20*, 1337–1357.



- [35] Kaushik, A. C.; Mehmood, A.; Dai, X.; Wei, D.-Q. A comparative chemogenic analysis for predicting Drug-Target Pair via Machine Learning Approaches. *Scientific Reports* **2020**, *10*, 1–11.
- [36] Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. ChemMapper: A versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. *Bioinformatics* **2013**, *29*, 1827–1829.
- [37] Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *The AAPS Journal* **2013**, *15*, 395–406.
- [38] Nickel, J.; Gohlke, B. O.; Erehman, J.; Banerjee, P.; Rong, W. W.; Goede, A.; Dunkel, M.; Preissner, R. SuperPred: Update on drug classification and target prediction. *Nucleic Acids Research* **2014**, *42*, 26–31.
- [39] Peón, A.; Li, H.; Ghislat, G.; Leung, K. S.; Wong, M. H.; Lu, G.; Ballester, P. J. MolTarPred: A web tool for comprehensive target prediction with reliability estimation. *Chemical Biology and Drug Design* **2019**, cbdd.13516.
- [40] Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: A brief review. *Briefings in Bioinformatics* **2013**, *15*, 734–747.
- [41] Wang, C.; Kurgan, L. Survey of similarity-based prediction of drug-protein interactions. *Current Medicinal Chemistry* **2020**, *27*, 5856–5886.
- [42] Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of Cheminformatics* **2019**, *11*, 4.
- [43] Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science* **2018**, *9*, 5441–5451.
- [44] Ding, Y.; Tang, J.; Guo, F. Identification of drug-target interactions via multiple information integration. *Information Sciences* **2017**, *418–419*, 546–560.
- [45] Keum, J.; Nam, H. SELF-BLM: Prediction of drug-Target interactions via self-Training SVM. *PLoS ONE* **2017**, *12*, e0171839.
- [46] Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Molecular Informatics* **2016**, *35*, 3–14.

- [47] Zhang, H.; Liao, L.; Saravanan, K. M.; Yin, P.; Wei, Y. DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. *PeerJ* **2019**, *7*, e7362.
- [48] Monteiro, N. R.; Ribeiro, B.; Arrais, J. P. Deep neural network architecture for drug–target interaction prediction. International Conference on Artificial Neural Networks. 2019; pp 804–809.
- [49] Lee, K.; Kim, D. In-Silico molecular binding prediction for human drug targets using deep neural multi-task learning. *Genes* **2019**, *10*, 906.
- [50] Lee, H.; Kim, W. Comparison of target features for predicting drug–target interactions by deep neural network based on large-scale drug-induced Transcriptome data. *Pharmaceutics* **2019**, *11*, 377.
- [51] Lo, Y. C.; Senese, S.; Li, C. M.; Hu, Q.; Huang, Y.; Damoiseaux, R.; Torres, J. Z. Large-Scale Chemical Similarity Networks for Target Profiling of Compounds Identified in Cell-Based Chemical Screens. *PLoS Computational Biology* **2015**, *11*, e1004153.
- [52] Lo, Y.-C.; Torres, J. Z. Chemical Similarity Networks for Drug Discovery. *Special Topics in Drug Discovery* **2016**,
- [53] Lo, Y. C.; Senese, S.; Damoiseaux, R.; Torres, J. Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chemical Biology* **2016**, *11*, 2244–2253.
- [54] Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. Facing the challenges of structure-based target prediction by inverse virtual screening. *Journal of Chemical Information and Modeling* **2014**, *54*, 1676–1686.
- [55] Schomburg, K. T.; Rarey, M. What is the potential of structure-based target prediction methods? *Future Medicinal Chemistry* **2014**, *6*, 1987–1989.
- [56] Pantsar, T.; Poso, A. Binding affinity via docking: fact and fiction. *Molecules* **2018**, *23*, 1899.
- [57] Hwang, H.; Dey, F.; Petrey, D.; Honig, B. Structure-based prediction of ligand–protein interactions on a genome-wide scale. *Proceedings of the National Academy of Sciences* **2017**, *114*, 13685–13690.
- [58] Lavecchia, A.; Cerchia, C. In silico methods to address polypharmacology: Current status, applications and future perspectives. *Drug Discovery Today* **2016**, *21*, 288–298.

- [59] Ezzat, A.; Wu, M.; Li, X.-L.; Kwok, C.-K. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics* **2018**,
- [60] Reker, D.; Schneider, P.; Schneider, G.; Brown, J. Active learning for computational chemogenomics. *Future Medicinal Chemistry* **2017**, *9*, 381–402.
- [61] Yang, S.-Q.; Ye, Q.; Ding, J.-J.; Lu, A.-P.; Chen, X.; Hou, T.-J.; Cao, D.-S. Current advances in ligand-based target prediction. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, e1504.
- [62] González-Medina, M.; Naveja, J. J.; Sánchez-Cruz, N.; Medina-Franco, J. L. Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Advances* **2017**, *7*, 54153–54163.
- [63] EMBL-EBI, ChEMBL. <https://www.ebi.ac.uk/chembl/>, (accessed on 2021-04-28).
- [64] Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M., et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **2019**, *47*, D930–D940.
- [65] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E., et al. The ChEMBL database in 2017. *Nucleic Acids Research* **2017**, *45*, D945–D954.
- [66] Katsila, T.; Spyroulias, G. A.; Patrinos, G. P.; Matsoukas, M. T. Computational approaches in target identification and drug discovery. *Computational and Structural Biotechnology Journal* **2016**, *14*, 177–184.
- [67] Chen, X.; Yan, C. C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics* **2016**, *17*, 696–712.
- [68] Vanhaelen, Q.; Mamoshina, P.; Aliper, A. M.; Artemov, A.; Lezhnina, K.; Ozerov, I.; Labat, I.; Zhavoronkov, A. Design of efficient computational workflows for in silico drug repurposing. *Drug Discovery Today* **2017**, *22*, 210–222.
- [69] Brown, A. S.; Patel, C. J. A review of validation strategies for computational drug repositioning. *Briefings in Bioinformatics* **2018**, *19*, 174–177.
- [70] Liu, H.; Sun, J.; Guan, J.; Zheng, J.; Zhou, S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **2015**, *31*, i221–i229.

- [71] Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target prediction utilising negative bioactivity data covering large chemical space. *Journal of Cheminformatics* **2015**, *7*, 1–16.
- [72] Mahmud, S. H.; Chen, W.; Meng, H.; Jahan, H.; Liu, Y.; Hasan, S. M. Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Analytical Biochemistry* **2020**, *589*, 113507.
- [73] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- [74] Jastrzebski, S.; Leśniak, D.; Czarnecki, W. M. Learning to smile (s). *arXiv preprint arXiv:1602.06289* **2016**,
- [75] Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A., et al. QSAR without borders. *Chemical Society Reviews* **2020**, *49*, 3525–3564.
- [76] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.
- [77] Willett, P. The calculation of molecular structural similarity: principles and practice. *Molecular Informatics* **2014**, *33*, 403–413.
- [78] Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology* **2007**, *25*, 197–206.
- [79] Keiser, M. J. et al. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.
- [80] Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361–367.
- [81] Gfeller, D.; Michielin, O.; Zoete, V. Shaping the interaction landscape of bioactive molecules. *Bioinformatics* **2013**, *29*, 3073–3079.
- [82] Park, Y.; Marcotte, E. M. Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods* **2012**, *9*, 1134–1136.

- [83] Robinson, M. C.; Glen, R. C., et al. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *Journal of Computer-Aided Molecular Design* **2020**, 1–14.
- [84] Liu, X.; Vogt, I.; Haque, T.; Campillos, M. HitPick: A web server for hit identification and target prediction of chemical screenings. *Bioinformatics* **2013**, *29*, 1910–1912.
- [85] Hamad, S.; Adornetto, G.; Naveja, J. J.; Chavan Ravindranath, A.; Raffler, J.; Campillos, M. HitPickV2: a web server to predict targets of chemical compounds. *Bioinformatics* **2019**, *35*, 1239–1240.
- [86] Awale, M.; Reymond, J.-L. L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *Journal of Chemical Information and Modeling* **2019**, *59*, 10–17.
- [87] Cockroft, N. T.; Cheng, X.; Fuchs, J. R. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products. *Journal of Chemical Information and Modeling* **2019**, acs.jcim.9b00489.
- [88] Huwe, C. M. Synthetic library design. *Drug Discovery Today* **2006**, *11*, 763–767.
- [89] Langer, T.; Hoffmann, R.; Bryant, S.; Lesur, B. Hit finding: towards ‘smarter’ approaches. *Current Opinion in Pharmacology* **2009**, *9*, 589–593.
- [90] Paricharak, S.; Méndez-Lucio, O.; Chavan Ravindranath, A.; Bender, A.; IJzerman, A. P.; van Westen, G. J. Data-driven approaches used for compound library design, hit triage and bioactivity modeling in high-throughput screening. *Briefings in Bioinformatics* **2018**, *19*, 277–285.
- [91] Wassermann, A. M.; Camargo, L. M.; Auld, D. S. Composition and applications of focus libraries to phenotypic assays. *Frontiers in Pharmacology* **2014**, *5*, 164.
- [92] Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chemical Biology* **2012**, *7*, 1399–1409.
- [93] Janzen, W. P. Screening technologies for small molecule discovery: the state of the art. *Chemistry & Biology* **2014**, *21*, 1162–1170.
- [94] Bakken, G. A.; Bell, A. S.; Boehm, M.; Everett, J. R.; Gonzales, R.; Hepworth, D.; Klug-McLeod, J. L.; Lanfear, J.; Loesel, J.; Mathias, J., et al. Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *Journal of Chemical Information and Modeling* **2012**, *52*, 2937–2949.

- [95] Spear, K. L.; Brown, S. P. The evolution of library design: crafting smart compound collections for phenotypic screens. *Drug Discovery Today: Technologies* **2017**, *23*, 61–67.
- [96] Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Current Opinion in Chemical Biology* **2010**, *14*, 347–361.
- [97] Stork, C.; Kirchmair, J. PAIN (S) relievers for medicinal chemists: how computational methods can assist in hit evaluation. 2018.
- [98] Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **2008**, *3*, 435.
- [99] Schneider, P.; Schneider, G. Privileged Structures Revisited. *Angewandte Chemie International Edition* **2017**, *56*, 7971–7974.
- [100] Lahue, B. R.; Glick, M.; Tudor, M.; Johnson, S. A.; Diratsouian, J.; Wildey, M. J.; Burton, M.; Mazzola, R.; Wassermann, A. M. Diversity & tractability revisited in collaborative small molecule phenotypic screening library design. *Bioorganic & Medicinal Chemistry* **2020**, *28*, 115192.
- [101] Follmann, M.; Briem, H.; Steinmeyer, A.; Hillisch, A.; Schmitt, M. H.; Haning, H.; Meier, H. An approach towards enhancement of a screening library: The Next Generation Library Initiative (NGLI) at Bayer—against all odds? *Drug Discovery Today* **2019**, *24*, 668–672.
- [102] Lisurek, M.; Rupp, B.; Wichard, J.; Neuenschwander, M.; von Kries, J. P.; Frank, R.; Rademann, J.; Kühne, R. Design of chemical libraries with potentially bioactive molecules applying a maximum common substructure concept. *Molecular Diversity* **2010**, *14*, 401–408.
- [103] Schuffenhauer, A. et al. Evolution of Novartis' Small Molecule Screening Deck Design. *Journal of Medicinal Chemistry* **2020**, acs.jmedchem.0c01332.
- [104] Baell, J. B. Broad coverage of commercially available lead-like screening space with fewer than 350,000 compounds. *Journal of Chemical Information and Modeling* **2013**, *53*, 39–55.
- [105] Gillet, V. J. Diversity selection algorithms. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 580–589.
- [106] Huggins, D. J.; Venkitaraman, A. R.; Spring, D. R. Rational methods for the selection of diverse screening compounds. *ACS Chemical Biology* **2011**, *6*, 208–217.

- [107] Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of iterative screening with stepwise compound selection based on Novartis in-house HTS data. *ACS Chemical Biology* **2016**, *11*, 1255–1264.
- [108] Gaulton, A. ChEMBL 24 Released! 2018; <http://chembl.blogspot.com/2018/05/chembl-24-released.html>, (accessed on 2021-04-28).
- [109] Gaulton, A. ChEMBL 25 and new web interface released. 2019; <http://chembl.blogspot.com/2019/03/chembl-25-and-new-web-interface-released.html>, (accessed on 2021-04-28).
- [110] Gaulton, A. ChEMBL 27 SARS-CoV-2 release. 2020; <http://chembl.blogspot.com/2020/05/chembl27-sars-cov-2-release.html>, (accessed on 2021-04-28).
- [111] Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem chemical structure standardization. *Journal of Cheminformatics* **2018**, *10*, 1–40.
- [112] Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org>, (accessed on 2021-02-16).
- [113] Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* **2020**, *12*, 1–15.
- [114] Dalke, A. The chemfp project. *Journal of Cheminformatics* **2019**, *11*, 1–21.
- [115] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
- [116] Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling* **2020**,
- [117] ZINC20, ZINC20. 2020; <http://zinc20.docking.org>, (accessed on 2021-02-16).
- [118] Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry* **2010**, *53*, 2719–2740.
- [119] Sterling, T.; Irwin, J. J. ZINC 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

- [120] Daylight Chemical Information Systems, I. SMARTS™—A Language for Describing Molecular Patterns. **2007**,
- [121] Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I., et al. Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS DISCOVERY: Advancing Life Sciences R&D* **2018**, *23*, 532–545.
- [122] Alvares-Cherman, E.; Metz, J.; Monard, M. C. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications* **2012**, *39*, 1647–1655.
- [123] Mervin, L. H.; Afzal, A. M.; Engkvist, O.; Bender, A. Comparison of Scaling Methods to Obtain Calibrated Probabilities of Activity for Protein–Ligand Predictions. *Journal of Chemical Information and Modeling* **2020**, *60*, 4546–4559.
- [124] Flíx, E. Multi-task neural network on ChEMBL with PyTorch 1.0 and RDKit. 2019; <http://chembl.blogspot.com/2019/05/multi-task-neural-network-on-chembl.html>, (accessed on 2021-05-13).
- [125] Flíx, E. Target predictions in the browser with RDKit MinimalLib (JS) and ONNX.js. 2021; <http://chembl.blogspot.com/2021/03/target-predictions-in-browser-with.html>, (accessed on 2021-05-13).
- [126] Flíx, E. Multitask Target prediction with RDKit MinimalLib (JS) and ONNX.js. 2021; [https://eloyfelix.github.io/rdkitjs\\_onnx\\_multitask/](https://eloyfelix.github.io/rdkitjs_onnx_multitask/), (accessed on 2021-05-13).
- [127] El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A., et al. The Pfam protein families database in 2019. *Nucleic acids research* **2019**, *47*, D427–D432.
- [128] Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L.; Tosatto, S. C.; Paladin, L.; Raj, S.; Richardson, L. J., et al. Pfam: The protein families database in 2021. *Nucleic Acids Research* **2021**, *49*, D412–D419.
- [129] Taylor, R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 59–67.
- [130] Butina, D. Unsupervised data base clustering based on daylight’s fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 747–750.



[131] Inc., M. MolPort. [www.molport.com](http://www.molport.com), (accessed on 2021-07-28).

[132] Ltd, E. Enamine. [www.enamine.net](http://www.enamine.net), (accessed on 2021-07-28).

# Acronyms

**2D** two-dimensional.

**3D** three-dimensional.

**ADME** absorption, distribution, metabolism and elimination/excretion.

**AI** artificial intelligence.

**ATC** Anatomical Therapeutic Codes.

**CADD** computer-aided drug discovery.

**ECFP** Extended Connectivity Fingerprint.

**FN** false negative.

**FP** false positive.

**MCC** Matthews correlation coefficient.

**MCP** Mondrian conformal predictor.

**ML** machine learning.

**MQN** Molecular Quantum Numbers.

**NMR** nuclear magnetic resonance.

**PAINS** pan assay interference compounds.

**PCC** pool of candidate compounds.

**QSAR** quantitative structure-activity relationship.

**SMARTS** SMILES ARbitrary Target Specification.

**SMILES** Simplified Molecular-Input Line-Entry System.

**TC** Tanimoto coefficient.

**TN** true negative.

**TP** true positive.

**Xfp** Shape and Pharmacophore Fingerprint.

# Supporting information for P2

This appendix contains the supporting information for the publication: Mathai, N. and Kirchmair, J. (2020) Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope, *International Journal of Molecular Sciences*, 21(10), 3585. doi: 10.3390/ijms21103585.



Supporting information for

# Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope

Neann Mathai<sup>1</sup> and Johannes Kirchmair<sup>1,2,\*</sup>

<sup>1</sup> Department of Chemistry and Computational Biology Unit (CBU), University of Bergen, N-5020 Bergen, Norway

<sup>2</sup> Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

\* Correspondence: johannes.kirchmair@univie.ac.at

The data of success and recovery rates presented in the graphs of the paper are reported below. The percentage indicates the success and recovery rates, while the numbers in the brackets show how many queries (success rate) or bioactivities (recovery rate) within the TC interval had a hit.

## Similarity-based approach

Standard testing scenario with external data

**Table S1.** Success rates under the standard testing scenario with external data by the similarity approach

median maxTC	[0.8, 1]	[0.6, 0.8)	[0.4, 0.6)	[0.2, 0.4)	[0.0, 0.2)	overall
top-1	97.85% (17560/ 17946)	88.29% (15486/ 17539)	32.94% (1973/ 5989)	6.98% (208/ 2982)	5.19% (8/ 154)	78.98% (35235/ 44610)
top-3	99.74% (17899/ 17946)	96.61% (16945/ 17539)	50.63% (3032/ 5989)	12.17% (363/ 2982)	7.79% (12/ 154)	85.75% (38251/ 44610)
top-5	99.96% (17939/ 17946)	98.66% (17304/ 17539)	60.96% (3651/ 5989)	15.19% (453/ 2982)	8.44% (13/ 154)	88.23% (39360/ 44610)
top-10	99.98% (17943/ 17946)	99.70% (17486/ 17539)	77.63% (4649/ 5989)	21.09% (629/ 2982)	9.09% (14/ 154)	91.28% (40721/ 44610)
top-15	99.98% (17943/ 17946)	99.88% (17518/ 17539)	85.72% (5134/ 5989)	27.80% (829/ 2982)	10.39% (16/ 154)	92.89% (41440/ 44610)

**Table S2.** Recovery rates under the standard testing scenario with external data by the similarity approach

maxTC	[0.8, 1]	[0.6, 0.8)	[0.4, 0.6)	[0.2, 0.4)	[0.0, 0.2)	overall
top-1	67.82% (18368/ 27084)	55.34% (15480/ 27972)	12.65% (1341/ 10603)	0.66% (45/ 6767)	0.15% (1/ 677)	48.20% (35235/ 73103)
top-3	94.73% (25658/ 27084)	85.51% (23919/ 27972)	29.05% (3080/ 10603)	1.86% (126/ 6767)	0.15% (1/ 677)	72.20% (52784/ 73103)
top-5	98.91% (26788/ 27084)	93.59% (26179/ 27972)	40.48% (4292/ 10603)	3.16% (214/ 6767)	0.30% (2/ 677)	78.62% (57475/ 73103)
top-10	99.87% (27049/ 27084)	98.39% (27523/ 27972)	60.95% (6463/ 10603)	6.34% (429/ 6767)	0.44% (3/ 677)	84.08% (61467/ 73103)
top-15	99.96% (27074/ 27084)	99.32% (27782/ 27972)	72.76% (7715/ 10603)	10.64% (720/ 6767)	0.59% (4/ 677)	86.58% (63295/ 73103)

Standard time-split and close-to-real world testing scenarios

**Table S3.** Success rates under the time-split and close-to-real-world testing scenarios by the similarity approach

median maxTC	time-split scenario					overall	close-to- real-world scenario
	[0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
top-1	95.59% (1518/ 1588)	88.61% (4147/ 4680)	57.81% (3055/ 5285)	8.68% (565/ 6510)	0.69% (8/ 1160)	48.34% (9293/ 19223)	46.32% (9293/ 20061)
top-3	98.74% (1568/ 1588)	95.62% (4475/ 4680)	76.31% (4033/ 5285)	15.44% (1005/ 6510)	1.03% (12/ 1160)	57.71% (11093/ 19223)	55.30% (11093/ 20061)
top-5	99.75% (1584/ 1588)	97.37% (4557/ 4680)	82.65% (4368/ 5285)	19.43% (1265/ 6510)	1.38% (16/ 1160)	61.33% (11790/ 19223)	58.77% (11790/ 20061)
top-10	99.87% (1586/ 1588)	99.08% (4637/ 4680)	90.26% (4770/ 5285)	26.31% (1713/ 6510)	1.81% (21/ 1160)	66.21% (12727/ 19223)	63.44% (12727/ 20061)
top-15	99.87% (1586/ 1588)	99.62% (4662/ 4680)	93.72% (4953/ 5285)	31.08% (2023/ 6510)	1.98% (23/ 1160)	68.91% (13247/ 19223)	66.03% (13247/ 20061)

**Table S4.** Recovery rates under the time-split and close-to-real-world testing scenarios with by the similarity approach

maxTC	time-split scenario					overall	close-to-real-world scenario
	[0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
top-1	70.10% (1611/ 2298)	62.63% (4426/ 7067)	35.74% (2983/ 8346)	2.61% (273/ 10462)	0.00% (0/ 2029)	30.77% (9293/ 30202)	29.50% (9293/ 31498)
top-3	94.04% (2161/ 2298)	88.67% (6266/ 7067)	61.71% (5150/ 8346)	6.88% (720/ 10462)	0.00% (0/ 2029)	47.34% (14297/ 30202)	45.39% (14297/ 31498)
top-5	98.69% (2268/ 2298)	95.39% (6741/ 7067)	72.12% (6019/ 8346)	10.31% (1079/ 10462)	0.00% (0/ 2029)	53.33% (16107/ 30202)	51.14% (16107/ 31498)
top-10	99.70% (2291/ 2298)	98.40% (6954/ 7067)	85.15% (7107/ 8346)	16.93% (1771/ 10462)	0.00% (0/ 2029)	60.01% (18123/ 30202)	57.54% (18123/ 31498)
top-15	99.74% (2292/ 2298)	99.19% (7010/ 7067)	90.47% (7551/ 8346)	21.62% (2262/ 10462)	0.00% (0/ 2029)	63.29% (19115/ 30202)	60.69% (19115/ 31498)



## Similarity-based approach - reduced scope

Standard testing scenario with external data

**Table S5.** Success rates under the standard testing scenario with external data by the similarity approach with a reduced target scope

<b>median maxTC</b>	<b>[0.8, 1]</b>	<b>[0.6, 0.8]</b>	<b>[0.4, 0.6]</b>	<b>[0.2, 0.4]</b>	<b>[0.0, 0.2)</b>	<b>overall</b>
top-1	97.81% (17307/ 17695)	88.21% (15177/ 17205)	32.28% (1904/ 5899)	6.43% (188/ 2923)	4.42% (5/ 113)	78.89% (34581/ 43835)
top-3	99.76% (17652/ 17695)	96.65% (16629/ 17205)	50.18% (2960/ 5899)	11.56% (338/ 2923)	7.08% (8/ 113)	85.75% (37587/ 43835)
top-5	99.96% (17688/ 17695)	98.70% (16981/ 17205)	60.54% (3571/ 5899)	14.64% (428/ 2923)	7.96% (9/ 113)	88.23% (38677/ 43835)
top-10	99.98% (17692/ 17695)	99.72% (17157/ 17205)	77.37% (4564/ 5899)	20.56% (601/ 2923)	7.96% (9/ 113)	91.30% (40023/ 43835)
top-15	99.98% (17692/ 17695)	99.88% (17185/ 17205)	85.61% (5050/ 5899)	27.10% (792/ 2923)	8.85% (10/ 113)	92.91% (40729/ 43835)

**Table S6.** Recovery rates under the standard testing scenario with external data by the similarity approach with a reduced target scope

maxTC	[0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]	overall
top-1	68.08% (18074/ 26550)	55.61% (15174/ 27287)	12.59% (1292/ 10265)	0.61% (40/ 6528)	0.23% (1/ 433)	48.66% (34581/ 71063)
top-3	94.92% (25201/ 26550)	85.72% (23391/ 27287)	29.00% (2977/ 10265)	1.81% (118/ 6528)	0.23% (1/ 433)	72.74% (51688/ 71063)
top-5	98.96% (26275/ 26550)	93.73% (25576/ 27287)	40.52% (4159/ 10265)	3.12% (204/ 6528)	0.46% (2/ 433)	79.11% (56216/ 71063)
top-10	99.87% (26516/ 26550)	98.48% (26872/ 27287)	61.27% (6289/ 10265)	6.31% (412/ 6528)	0.69% (3/ 433)	84.56% (60092/ 71063)
top-15	99.96% (26540/ 26550)	99.37% (27114/ 27287)	73.27% (7521/ 10265)	10.65% (695/ 6528)	0.92% (4/ 433)	87.07% (61874/ 71063)

Standard time-split and close-to-real world testing scenarios

**Table S7.** Success rates under the time-split and close-to-real-world testing scenarios by the similarity approach with a reduced target scope

median maxTC	time-split scenario					overall	close-to- real-world scenario
	[0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
top-1	95.84% (1496/ 1561)	88.91% (4122/ 4636)	58.04% (2971/ 5119)	8.51% (532/ 6251)	0.70% (5/ 716)	49.92% (9126/ 18283)	45.49% (9126/ 20061)
top-3	98.78% (1542/ 1561)	96.23% (4461/ 4636)	76.75% (3929/ 5119)	15.37% (961/ 6251)	0.84% (6/ 716)	59.61% (10899/ 18283)	54.33% (10899/ 20061)
top-5	99.74% (1557/ 1561)	97.58% (4524/ 4636)	82.77% (4237/ 5119)	19.53% (1221/ 6251)	1.54% (11/ 716)	63.17% (11550/ 18283)	57.57% (11550/ 20061)
top-10	99.87% (1559/ 1561)	99.29% (4603/ 4636)	90.56% (4636/ 5119)	26.25% (1641/ 6251)	1.96% (14/ 716)	68.11% (12453/ 18283)	62.08% (12453/ 20061)
top-15	99.87% (1559/ 1561)	99.72% (4623/ 4636)	94.30% (4827/ 5119)	31.04% (1940/ 6251)	2.23% (16/ 716)	70.91% (12965/ 18283)	64.63% (12965/ 20061)

**Table S8.** Recovery rates under the time-split and close-to-real-world testing scenarios with by the similarity approach with a reduced target scope

maxTC	time-split scenario					overall	close-to-real-world scenario
	[0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
top-1	71.53% (1583/ 2213)	63.97% (4368/ 6828)	36.22% (2904/ 8018)	2.75% (271/ 9860)	0.00% (0/ 1198)	32.46% (9126/ 28117)	28.97% (9126/ 31498)
top-3	94.62% (2094/ 2213)	89.81% (6132/ 6828)	62.91% (5044/ 8018)	7.22% (712/ 9860)	0.00% (0/ 1198)	49.73% (13982/ 28117)	44.39% (13982/ 31498)
top-5	98.78% (2186/ 2213)	95.88% (6547/ 6828)	73.15% (5865/ 8018)	10.87% (1072/ 9860)	0.00% (0/ 1198)	55.73% (15670/ 28117)	49.75% (15670/ 31498)
top-10	99.77% (2208/ 2213)	98.68% (6738/ 6828)	85.88% (6886/ 8018)	17.58% (1733/ 9860)	0.00% (0/ 1198)	62.47% (17565/ 28117)	55.77% (17565/ 31498)
top-15	99.77% (2208/ 2213)	99.36% (6784/ 6828)	91.16% (7309/ 8018)	22.28% (2197/ 9860)	0.00% (0/ 1198)	65.79% (18498/ 28117)	58.73% (18498/ 31498)

## ML approach

Standard testing scenario with external data

**Table S9.** Success rates under the standard testing scenario with external data by the ML approach

<b>median maxTC</b>	<b>[0.8, 1]</b>	<b>[0.6, 0.8]</b>	<b>[0.4, 0.6]</b>	<b>[0.2, 0.4]</b>	<b>[0.0, 0.2]</b>	<b>overall</b>
top-1	94.93% (16797/ 17695)	80.73% (13889/ 17205)	28.16% (1661/ 5899)	6.47% (189/ 2923)	2.65% (3/ 113)	74.23% (32539/ 43835)
top-3	98.82% (17487/ 17695)	90.32% (15540/ 17205)	44.96% (2652/ 5899)	14.47% (423/ 2923)	4.42% (5/ 113)	82.37% (36107/ 43835)
top-5	99.38% (17586/ 17695)	93.57% (16098/ 17205)	54.08% (3190/ 5899)	21.28% (622/ 2923)	6.19% (7/ 113)	85.55% (37503/ 43835)
top-10	99.73% (17648/ 17695)	96.30% (16569/ 17205)	66.86% (3944/ 5899)	31.06% (908/ 2923)	12.39% (14/ 113)	89.16% (39083/ 43835)
top-15	99.84% (17667/ 17695)	97.49% (16773/ 17205)	74.47% (4393/ 5899)	37.46% (1095/ 2923)	15.93% (18/ 113)	91.13% (39946/ 43835)

**Table S10.** Recovery rates under the standard testing scenario with external data by the ML approach

<b>maxTC</b>	<b>[0.8, 1]</b>	<b>[0.6, 0.8]</b>	<b>[0.4, 0.6]</b>	<b>[0.2, 0.4]</b>	<b>[0.0, 0.2]</b>	<b>overall</b>
top-1	65.62% (17423/ 26550)	50.49% (13777/ 27287)	12.06% (1238/ 10265)	1.53% (100/ 6528)	0.23% (1/ 433)	45.79% (32539/ 71063)
top-3	92.63% (24594/ 26550)	77.73% (21209/ 27287)	25.54% (2622/ 10265)	5.02% (328/ 6528)	0.69% (3/ 433)	68.61% (48756/ 71063)
top-5	97.65% (25927/ 26550)	85.82% (23418/ 27287)	34.60% (3552/ 10265)	8.53% (557/ 6528)	1.39% (6/ 433)	75.23% (53460/ 71063)
top-10	99.30% (26364/ 26550)	92.33% (25193/ 27287)	48.76% (5005/ 10265)	14.81% (967/ 6528)	3.70% (16/ 433)	80.98% (57545/ 71063)
top-15	99.66% (26461/ 26550)	94.86% (25885/ 27287)	57.99% (5953/ 10265)	19.87% (1297/ 6528)	4.85% (21/ 433)	83.89% (59617/ 71063)

Standard time-split and close-to-real world testing scenarios

**Table S11.** Success rates under the time-split and close-to-real-world testing scenarios by the ML approach

	time-split scenario					overall	close-to-real-world scenario
	median [0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
maxTC							
top-1	90.01% (1405/ 1561)	80.20% (3718/ 4636)	47.86% (2450/ 5119)	7.04% (440/ 6251)	0.42% (3/ 716)	43.84% (8016/ 18283)	39.96% (8016/ 20061)
top-3	96.28% (1503/ 1561)	90.55% (4198/ 4636)	64.43% (3298/ 5119)	12.01% (751/ 6251)	0.42% (3/ 716)	53.34% (9753/ 18283)	48.62% (9753/ 20061)
top-5	98.40% (1536/ 1561)	93.21% (4321/ 4636)	70.38% (3603/ 5119)	15.95% (997/ 6251)	0.70% (5/ 716)	57.22% (10462/ 18283)	52.15% (10462/ 20061)
top-10	99.30% (1550/ 1561)	96.05% (4453/ 4636)	77.93% (3989/ 5119)	21.52% (1345/ 6251)	1.12% (8/ 716)	62.05% (11345/ 18283)	56.55% (11345/ 20061)
top-15	99.62% (1555/ 1561)	97.43% (4517/ 4636)	81.46% (4170/ 5119)	25.05% (1566/ 6251)	1.26% (9/ 716)	64.63% (11817/ 18283)	58.91% (11817/ 20061)

**Table S12.** Recovery rates under the time-split and close-to-real-world testing scenarios with by the ML approach

	time-split scenario					overall	close-to-real-world scenario
	maxTC [0.8, 1]	[0.6, 0.8]	[0.4, 0.6]	[0.2, 0.4]	[0.0, 0.2]		
top-1	67.06% (1484/ 2213)	57.45% (3923/ 6828)	29.13% (2336/ 8018)	2.77% (273/ 9860)	0.00% (0/ 1198)	28.51% (8016/ 28117)	25.45% (8016/ 31498)
top-3	91.55% (2026/ 2213)	82.70% (5647/ 6828)	50.86% (4078/ 8018)	7.03% (693/ 9860)	0.00% (0/ 1198)	44.26% (12444/ 28117)	39.51% (12444/ 31498)
top-5	95.39% (2111/ 2213)	89.22% (6092/ 6828)	60.50% (4851/ 8018)	10.75% (1060/ 9860)	0.00% (0/ 1198)	50.20% (14114/ 28117)	44.81% (14114/ 31498)
top-10	98.73% (2185/ 2213)	93.85% (6408/ 6828)	72.11% (5782/ 8018)	16.33% (1610/ 9860)	0.08% (1/ 1198)	56.86% (15986/ 28117)	50.75% (15986/ 31498)
top-15	99.14% (2194/ 2213)	96.28% (6574/ 6828)	76.93% (6168/ 8018)	20.23% (1995/ 9860)	0.17% (2/ 1198)	60.22% (16933/ 28117)	53.76% (16933/ 31498)



Results of the internal grid search for each target model

target_id	mcc_score	n_estimators	max_depth
0	CHEMBL1770047	0.980064077	200
1	CHEMBL2814	0.980064077	200
2	CHEMBL5499	1	200
3	CHEMBL5495	0.962666667	200
4	CHEMBL2873	0.634807394	200
5	CHEMBL2898	0.984852814	500
6	CHEMBL3021	0.849401601	200
7	CHEMBL3751	0.929465678	200
8	CHEMBL2543	0.634422835	1000
9	CHEMBL3047	1	200
10	CHEMBL4632	1	200
11	CHEMBL4241	0.82716209	200
12	CHEMBL4134	0.388461209	200
13	CHEMBL4133	0.620683769	200
14	CHEMBL3545	0.864042596	200
15	CHEMBL3938	0.194230605	200
16	CHEMBL3886	0.263568129	200
17	CHEMBL4693	0.929401601	1000
18	CHEMBL2390812	0.929465678	200
19	CHEMBL3708265	0.865085311	1000
20	CHEMBL2104	0.690384868	1000
21	CHEMBL2169716	0.846108912	200
22	CHEMBL5847	0.723696283	200
23	CHEMBL5866	0.24978616	500
24	CHEMBL2109233	0.949401601	500
25	CHEMBL4752	0.980123362	200
26	CHEMBL3123	0.31641682	500
27	CHEMBL4677	0.83037867	200
28	CHEMBL3150	1	200
29	CHEMBL4564	0.9403108	200
30	CHEMBL3399914	0.689840801	500
31	CHEMBL4331	0.391928323	200
32	CHEMBL3490	0.6334926	1000
33	CHEMBL2029197	0.454036332	200
34	CHEMBL3578	0.845254218	1000
35	CHEMBL3632452	0.693374456	200
36	CHEMBL1923	1	200
37	CHEMBL3032	0.330170823	1000
38	CHEMBL3027	0.949512228	200
39	CHEMBL3799	0.86994639	200
40	CHEMBL2560	0.960246723	200
41	CHEMBL2897	0.375071192	200
42	CHEMBL5686	0.880185228	200
43	CHEMBL2801	0.235758226	200

target_id	mcc_score	n_estimators	max_depth
44	CHEMBL5724	0.650219475	200
45	CHEMBL15905	0.62281739	500
46	CHEMBL2146350	0.934433006	200
47	CHEMBL15648	0.883945234	200
48	CHEMBL12349	0.250047461	200
49	CHEMBL1250368	0.960246723	200
50	CHEMBL4411	0.980178373	200
51	CHEMBL3413	0.91525271	200
52	CHEMBL3817	0.762192723	500
53	CHEMBL15970	0.222834406	1000
54	CHEMBL1293271	0.606372487	200
55	CHEMBL13491	0.641769776	200
56	CHEMBL13911	0.940535118	200
57	CHEMBL4090	0.464971006	200
58	CHEMBL13322	0.960356745	200
59	CHEMBL13564	0.845146233	500
60	CHEMBL13987	0.167125804	200
61	CHEMBL16079	1	200
62	CHEMBL13638328	1	500
63	CHEMBL1075294	0.808766551	200
64	CHEMBL13638329	1	200
65	CHEMBL2146301	0.347304177	500
66	CHEMBL1086648	0.929793253	200
67	CHEMBL15174	0.849320439	200
68	CHEMBL2935	0.969436507	1000
69	CHEMBL12113	0.780830743	200
70	CHEMBL3038484	0.842485005	200
71	CHEMBL5432	0.299808975	200
72	CHEMBL15867	1	200
73	CHEMBL1293191	0.476845288	200
74	CHEMBL13125	0.16093596	200
75	CHEMBL2189127	0.859586505	200
76	CHEMBL3290	0.167125804	200
77	CHEMBL2182	0.431143928	200
78	CHEMBL2558	0.366562659	200
79	CHEMBL13688	0.920918223	500
80	CHEMBL13663	0.92833668	200
81	CHEMBL13297641	0.849182657	1000
82	CHEMBL13646	0.949659206	200
83	CHEMBL12967	0.760613743	200
84	CHEMBL1805	0.88571728	200
85	CHEMBL1982	1	200
86	CHEMBL13253	0.639345787	200
87	CHEMBL2362976	0.341494615	200
88	CHEMBL3414411	0.949710389	200

target_id	mcc_score	n_estimators	max_depth
89 CHEMBL2111474	0.816209221	500	25
90 CHEMBL3308969	0.980229556	200	25
91 CHEMBL1743122	-0.004962917	200	25
92 CHEMBL2069156	0.890399056	500	25
93 CHEMBL1845	0.83103654	500	25
94 CHEMBL2794	0.866368158	200	25
95 CHEMBL3384	0.28571728	200	25
96 CHEMBL1870	0.969480833	200	25
97 CHEMBL4613	0.910169501	200	25
98 CHEMBL4788	0.898994395	500	25
99 CHEMBL5705	0.76160689	1000	25
100 CHEMBL5925	0.855801293	200	25
101 CHEMBL4619	0.980229556	200	25
102 CHEMBL1075282	0.955777335	200	25
103 CHEMBL4577	0.167332005	200	25
104 CHEMBL4522	0.167332005	500	25
105 CHEMBL5989	0.890399056	200	25
106 CHEMBL4531	0.863416731	200	75
107 CHEMBL3927	0.640406619	200	25
108 CHEMBL5310	0.499477857	500	25
109 CHEMBL1075029	1	200	25
110 CHEMBL6115	0.925790674	200	25
111 CHEMBL3769292	0.980229556	500	50
112 CHEMBL4607	0.875441063	200	25
113 CHEMBL3928	0.960459111	200	25
114 CHEMBL4731	0.855178803	500	45
115 CHEMBL242731	0.965373592	200	25
116 CHEMBL3038491	0.852382527	200	25
117 CHEMBL4720	0.904050282	200	25
118 CHEMBL1628461	0.955288883	200	25
119 CHEMBL5698	0.223366231	1000	25
120 CHEMBL3712868	1	200	25
121 CHEMBL5657	0.901386486	200	45
122 CHEMBL5362	0.896120775	200	25
123 CHEMBL2570	0.980277297	200	25
124 CHEMBL4924	0.055841558	200	50
125 CHEMBL5068	0.960554594	200	25
126 CHEMBL3259478	0.945650889	200	25
127 CHEMBL2111462	0.960554594	500	75
128 CHEMBL1965712	0.955288883	200	25
129 CHEMBL3830	0.167524673	200	25
130 CHEMBL3883323	0.955841558	200	25
131 CHEMBL5941	0.16174796	200	25
132 CHEMBL3534	0.886796098	200	25
133 CHEMBL3503	0.736726056	200	45

target_id	mcc_score	n_estimators	max_depth
134 CHEMBL1820	0.878391458	200	75
135 CHEMBL5579	0.227946787	200	100
136 CHEMBL4554	0.915173067	500	45
137 CHEMBL1075134	0.980277297	1000	45
138 CHEMBL4597	0.111683115	200	25
139 CHEMBL2724	0.890631367	500	25
140 CHEMBL12146298	0.409123231	500	45
141 CHEMBL5914	0.082568366	200	75
142 CHEMBL5931	0.8622576	200	25
143 CHEMBL2331070	1	200	25
144 CHEMBL2321628	1	200	25
145 CHEMBL2592	0.813413063	500	75
146 CHEMBL1075132	0.867161362	500	25
147 CHEMBL5873	1	200	100
148 CHEMBL2966	0.801332295	200	45
149 CHEMBL4586	0.786396395	500	25
150 CHEMBL3154	0.940965799	500	50
151 CHEMBL5165	1	200	25
152 CHEMBL3638333	1	200	25
153 CHEMBL1743126	0	200	25
154 CHEMBL4117	0.980321933	200	25
155 CHEMBL3569	0.940965799	1000	25
156 CHEMBL3097982	0.798197583	500	25
157 CHEMBL4000	1	200	25
158 CHEMBL4108	0.940965799	200	25
159 CHEMBL4125	0.906754804	200	45
160 CHEMBL2111381	0.929261954	200	25
161 CHEMBL2095942	1	200	25
162 CHEMBL6029	0.887057417	200	45
163 CHEMBL3834	0.111916275	200	25
164 CHEMBL2024	0.848336746	500	45
165 CHEMBL3754	0.799150196	200	50
166 CHEMBL3430892	0.926285544	500	25
167 CHEMBL4245	0.960727513	500	25
168 CHEMBL5536	1	500	25
169 CHEMBL5538	0.92876537	200	25
170 CHEMBL5869	0.830850232	200	100
171 CHEMBL5810	0.223832549	1000	45
172 CHEMBL5766	0.450058157	500	45
173 CHEMBL4664	0.980363756	200	25
174 CHEMBL2285348	0.926285544	200	25
175 CHEMBL4701	0.757863879	200	25
176 CHEMBL5749	0.304196306	200	25
177 CHEMBL4972	0.882288042	200	25
178 CHEMBL5072	0.210202294	500	45

target_id	mcc_score	n_estimators	max_depth
179 CHEMBL5699	0.304196306	200	50
180 CHEMBL5155	1	200	25
181 CHEMBL5427	0.210202294	200	25
182 CHEMBL4335	1	200	25
183 CHEMBL1075021	0.980363756	200	25
184 CHEMBL6197	1	200	25
185 CHEMBL1075111	0.960727513	200	25
186 CHEMBL1075022	0.980363756	200	25
187 CHEMBL1649055	0.873654119	200	25
188 CHEMBL1615386	0.985194275	200	75
189 CHEMBL5135	0.725444354	200	25
190 CHEMBL2004	0.980403025	200	25
191 CHEMBL4197	0.851003947	200	25
192 CHEMBL1287610	1	200	25
193 CHEMBL5476	0.429058476	200	25
194 CHEMBL5865	0.965641661	200	25
195 CHEMBL1075167	0.29624578	200	25
196 CHEMBL5440	0.885948441	200	25
197 CHEMBL3783	0.885087106	500	25
198 CHEMBL2775	0.921694515	200	25
199 CHEMBL1293240	0.429058476	200	25
200 CHEMBL5554	0.46871884	500	50
201 CHEMBL3736	1	1000	25
202 CHEMBL5518	0.404447839	200	25
203 CHEMBL1795127	0.62749467	200	50
204 CHEMBL2653	1	200	25
205 CHEMBL4592	0.821830916	200	25
206 CHEMBL3559639	0.936414229	200	25
207 CHEMBL3638322	1	200	25
208 CHEMBL5010	0.950880296	500	25
209 CHEMBL3345	0.961306804	200	25
210 CHEMBL1741210	0.342541059	500	25
211 CHEMBL2111387	0.898004919	200	25
212 CHEMBL5381	0.950963598	1000	25
213 CHEMBL4342	0.980403025	500	25
214 CHEMBL5511	0.408286579	500	45
215 CHEMBL4576	0.437405682	200	25
216 CHEMBL3406	0.960879933	200	45
217 CHEMBL5742	0.796331351	500	25
218 CHEMBL5360	0.833254403	500	25
219 CHEMBL4852	0.29290542	1000	25
220 CHEMBL3913	0.892420382	200	25
221 CHEMBL3243909	0.876806773	500	100
222 CHEMBL3112376	0.980439967	200	25
223 CHEMBL2951	0.960356006	200	25

target_id	mcc_score	n_estimators	max_depth
224 CHEMBL4941	0.980439967	200	25
225 CHEMBL5627	0.365155551	500	25
226 CHEMBL1293260	0.670588898	500	45
227 CHEMBL4931	0.441924918	200	25
228 CHEMBL5311	0.729243034	200	25
229 CHEMBL2073709	0.985319469	200	25
230 CHEMBL2074	0.883853253	200	25
231 CHEMBL4400	0.956061191	200	45
232 CHEMBL5106	0.934420092	200	100
233 CHEMBL4098	0.838011444	200	25
234 CHEMBL13137290	0.802195837	1000	25
235 CHEMBL13194	0.627318701	200	75
236 CHEMBL5038	0.926780741	200	25
237 CHEMBL1697668	0.68131593	200	100
238 CHEMBL5784	0.368905048	200	50
239 CHEMBL1255137	0.856340893	200	25
240 CHEMBL2052039	1	200	25
241 CHEMBL3831325	1	200	25
242 CHEMBL3707464	0.946305959	200	25
243 CHEMBL1075108	0.804363183	500	25
244 CHEMBL3638365	1	200	25
245 CHEMBL3751648	1	200	25
246 CHEMBL5756	0.936543969	200	25
247 CHEMBL5620	0.932443917	200	100
248 CHEMBL4487	0.65076756	200	50
249 CHEMBL6148	0.652737619	200	75
250 CHEMBL4840	0.960949563	200	25
251 CHEMBL2548	0.985356396	200	25
252 CHEMBL3972	0.960949563	500	25
253 CHEMBL4897	0.52520757	500	45
254 CHEMBL2368	0.980474782	500	25
255 CHEMBL3137268	0.965898905	200	25
256 CHEMBL2284	0.369694066	200	25
257 CHEMBL1667684	0.985391256	200	25
258 CHEMBL5070	0.961667808	200	25
259 CHEMBL1075097	0.980507649	200	25
260 CHEMBL2111366	0.93179781	200	45
261 CHEMBL2052038	1	200	25
262 CHEMBL2860	0.781893513	200	25
263 CHEMBL4708	0.55125706	200	25
264 CHEMBL4367	0.352477731	1000	100
265 CHEMBL4589	1	200	25
266 CHEMBL4226	0.56442989	200	25
267 CHEMBL1795148	1	200	25
268 CHEMBL3383	0.836214633	200	45

target_id	mcc_score	n_estimators	max_depth
269	CHEMBL1075028	0.865979963	200
270	CHEMBL13475	0.505767717	200
271	CHEMBL1938219	0.985356396	200
272	CHEMBL3835	0.314646222	200
273	CHEMBL5330	0.26443032	200
274	CHEMBL4676	0.913803578	500
275	CHEMBL5862	0.929673568	500
276	CHEMBL3879857	0.967438203	200
277	CHEMBL3621035	0.885472314	500
278	CHEMBL303	0.944259754	200
279	CHEMBL1971	0.881523243	200
280	CHEMBL4526	0.232810768	1000
281	CHEMBL2498	0.926471162	200
282	CHEMBL5351	0.854451055	200
283	CHEMBL4305	0.818941968	200
284	CHEMBL1795138	0.98542422	200
285	CHEMBL1293257	0.718764383	1000
286	CHEMBL1795167	0.873889324	200
287	CHEMBL2079848	1	1000
288	CHEMBL2366471	0.98542422	200
289	CHEMBL4420	0.951387166	200
290	CHEMBL1795135	0.55032489	1000
291	CHEMBL4088	0.757207608	500
292	CHEMBL4087	0.936811385	1000
293	CHEMBL3430879	0.951387166	200
294	CHEMBL1921664	1	200
295	CHEMBL1932895	0.965962946	200
296	CHEMBL5332	0.692567965	200
297	CHEMBL3562166	1	200
298	CHEMBL2150838	1	200
299	CHEMBL4614	0.921887359	200
300	CHEMBL3593	0.812318332	200
301	CHEMBL3344	0.74449155	1000
302	CHEMBL3415	0.947362382	500
303	CHEMBL3666	0.900751077	200
304	CHEMBL4349	0.973681191	500
305	CHEMBL5274	0.325609423	500
306	CHEMBL1075165	0.783209083	200
307	CHEMBL3866	0.985454366	200
308	CHEMBL4532	0.960523052	1000
309	CHEMBL1293230	0.20050054	200
310	CHEMBL3797017	0.970910872	500
311	CHEMBL1293317	0.512141514	200
312	CHEMBL3744	0.777798894	200
313	CHEMBL1075261	1	500

target_id	mcc_score	n_estimators	max_depth
314	CHEMBL16040	0.905971438	200
315	CHEMBL13236	0.698492038	500
316	CHEMBL3724	0.939890327	500
317	CHEMBL13348	0.646412775	200
318	CHEMBL3430878	0.956455124	200
319	CHEMBL2111345	0.6653128905	200
320	CHEMBL3232700	1	200
321	CHEMBL250	0.845679784	200
322	CHEMBL4163	1	200
323	CHEMBL5114	0.922536236	200
324	CHEMBL2977	0.954405285	200
325	CHEMBL1944499	0.878304542	1000
326	CHEMBL4521	0.861708883	200
327	CHEMBL5408	0.616045775	200
328	CHEMBL2111464	0.71878159	1000
329	CHEMBL3636	0.954405285	200
330	CHEMBL3513	0.729049661	500
331	CHEMBL16030	0.985485041	200
332	CHEMBL2623	0.816668556	500
333	CHEMBL5378	0.798182386	1000
334	CHEMBL3638324	1	200
335	CHEMBL3855	0.867147122	200
336	CHEMBL3436	1	200
337	CHEMBL3326	0.913050826	1000
338	CHEMBL1795117	0.962090299	200
339	CHEMBL3638323	0.968964466	200
340	CHEMBL2150837	0.940252994	1000
341	CHEMBL3656	0.988288571	200
342	CHEMBL5630	0.853194779	500
343	CHEMBL5831	0.985513157	200
344	CHEMBL2021745	1	200
345	CHEMBL2417350	0.985513157	200
346	CHEMBL3122	0.937162048	200
347	CHEMBL2140	0.913050826	200
348	CHEMBL2878	0.926860203	200
349	CHEMBL4843	0.988288571	200
350	CHEMBL3038472	0.985485041	200
351	CHEMBL1795119	0.956539471	200
352	CHEMBL1903	0.88077892	200
353	CHEMBL4940	0.708537744	500
354	CHEMBL5215	1	200
355	CHEMBL2304401	0.944535069	200
356	CHEMBL312387	1	200
357	CHEMBL1743183	1	200
358	CHEMBL4954	0.241121411	500

target_id	mcc_score	n_estimators	max_depth
359 CHEMBL5836	0.421827625	200	25
360 CHEMBL4948	0.384582771	200	25
361 CHEMBL1293307	0.676947564	1000	25
362 CHEMBL1667681	0.985539892	200	25
363 CHEMBL2439944	0.938979636	200	25
364 CHEMBL1907591	0.898779246	200	25
365 CHEMBL3638364	1	200	25
366 CHEMBL1909044	0.918794786	500	100
367 CHEMBL4017	0.956619677	200	75
368 CHEMBL5774	0.891294786	200	25
369 CHEMBL5786	1	200	25
370 CHEMBL4518	0.985539892	200	25
371 CHEMBL4983	0.92633332	200	25
372 CHEMBL2146315	0.369430474	1000	50
373 CHEMBL5358	0.642641744	200	75
374 CHEMBL1940	0.729656948	200	45
375 CHEMBL2163181	0.842983521	200	25
376 CHEMBL4844	0.945121951	200	25
377 CHEMBL3107	0.959475914	200	50
378 CHEMBL4729	0.092394433	200	25
379 CHEMBL1649054	0.988345211	200	100
380 CHEMBL1169598	0.670117966	200	45
381 CHEMBL1628482	0.405516025	500	50
382 CHEMBL2366503	1	200	25
383 CHEMBL3351204	1	200	25
384 CHEMBL5608	0.549235594	200	75
385 CHEMBL4304	0.828102389	500	45
386 CHEMBL2146300	0.440643718	500	100
387 CHEMBL3091268	0.780113511	500	75
388 CHEMBL3541	0.841097522	200	75
389 CHEMBL4211	0.919977489	200	45
390 CHEMBL5983	0.91915158	500	100
391 CHEMBL2717	0.766819045	200	25
392 CHEMBL6172	0.842910655	200	50
393 CHEMBL2424504	0.704404862	500	75
394 CHEMBL3317335	0.988396569	200	25
395 CHEMBL3492	0.866089037	500	25
396 CHEMBL2428	0.669500951	200	25
397 CHEMBL4001	0.974000934	200	25
398 CHEMBL2010635	0.971225529	200	25
399 CHEMBL4440	0.988396569	200	25
400 CHEMBL2553	0.353421889	500	50
401 CHEMBL3738	1	200	25
402 CHEMBL4114	0.841282789	200	45
403 CHEMBL2111374	0.948066495	200	50

target_id	mcc_score	n_estimators	max_depth
404 CHEMBL2366408	0.852618903	500	25
405 CHEMBL3176	0.915829137	200	25
406 CHEMBL5189	0.866450029	1000	25
407 CHEMBL2721	0.84745483	200	75
408 CHEMBL4391	0.874412698	200	25
409 CHEMBL2135	0.97861594	200	25
410 CHEMBL1290	0.805381197	500	75
411 CHEMBL2649	0.472101597	1000	25
412 CHEMBL3456	0.928767483	500	25
413 CHEMBL5463	0.988443328	200	25
414 CHEMBL1293311	0.350517225	200	25
415 CHEMBL2446	1	200	25
416 CHEMBL1764946	0.80495978	200	45
417 CHEMBL4769	0.878657496	500	75
418 CHEMBL3357	0.466877561	1000	50
419 CHEMBL1075214	0.891473043	1000	100
420 CHEMBL1938211	0.781943472	200	45
421 CHEMBL4101	0.715441464	200	25
422 CHEMBL2366461	0.865254277	200	25
423 CHEMBL2964	0.467283897	200	50
424 CHEMBL3351190	0.935225967	200	25
425 CHEMBL2424	0.921082183	200	25
426 CHEMBL4223	0.389012988	200	25
427 CHEMBL1795168	0.712758864	200	75
428 CHEMBL4937	0.634866648	200	25
429 CHEMBL2366481	0.814856981	200	25
430 CHEMBL1886	0.957442688	200	25
431 CHEMBL5503	0.867596302	500	25
432 CHEMBL3559643	0.985656036	200	25
433 CHEMBL3313832	0.86384026	500	25
434 CHEMBL5600	0.793323931	200	100
435 CHEMBL1293312	0.044232587	500	25
436 CHEMBL2659	0.97412121	200	25
437 CHEMBL4973	0.810789766	200	25
438 CHEMBL1293244	0.688571817	1000	100
439 CHEMBL3190	0.953860695	200	45
440 CHEMBL3166	0.693062829	200	25
441 CHEMBL2605	0.846484298	200	75
442 CHEMBL3559681	0.933631603	200	45
443 CHEMBL4578	0.545536301	200	50
444 CHEMBL1741208	0.274981165	500	50
445 CHEMBL3243910	1	200	25
446 CHEMBL1806	0.584661668	200	45
447 CHEMBL4802	0.813190952	500	25
448 CHEMBL2023	0.924153491	200	75

target_id	mcc_score	n_estimators	max_depth
449 CHEMBL3392	0.815291561	1000	45
450 CHEMBL2094139	0.903222256	200	25
451 CHEMBL4014	0.524776101	1000	25
452 CHEMBL1744522	0.890465663	200	25
453 CHEMBL3399	0.78989062	200	25
454 CHEMBL3981	0.429094663	200	45
455 CHEMBL4503	0.968115477	1000	25
456 CHEMBL4909	0.949764972	200	25
457 CHEMBL1961783	0.940967597	200	25
458 CHEMBL2129	0.919212988	200	25
459 CHEMBL2046259	0.97697217	200	25
460 CHEMBL3108638	0.951868804	200	25
461 CHEMBL4631	0.849883294	200	25
462 CHEMBL2111421	0.870923115	200	25
463 CHEMBL3621	0.962687888	200	50
464 CHEMBL3638335	1	200	25
465 CHEMBL5868	0.676252468	500	50
466 CHEMBL3252	0.809434774	500	45
467 CHEMBL4187	0.749254722	200	25
468 CHEMBL5533	0.833415023	500	45
469 CHEMBL5649	0.924265994	200	25
470 CHEMBL4900	0.769541605	200	75
471 CHEMBL5035	0.900878681	500	25
472 CHEMBL402	0.905012444	500	45
473 CHEMBL1615381	0.908431785	200	25
474 CHEMBL1938210	0.818619761	500	25
475 CHEMBL2458	0.828516608	200	100
476 CHEMBL2569	0.98850612	200	25
477 CHEMBL6005	0.70586356	200	45
478 CHEMBL2634	0.533644309	500	25
479 CHEMBL2514	0.961357199	200	25
480 CHEMBL2528	0.977050667	200	50
481 CHEMBL1255165	1	200	25
482 CHEMBL5031	0.9109035	500	25
483 CHEMBL2007628	0.448605825	500	25
484 CHEMBL5107	0.951290287	200	25
485 CHEMBL4877	0.973067135	200	25
486 CHEMBL3108640	0.83337389	500	50
487 CHEMBL4818	0.954101335	200	25
488 CHEMBL317	0.823899224	200	25
489 CHEMBL4403	0.977050667	500	25
490 CHEMBL5255	0.934858033	1000	25
491 CHEMBL1615387	0.099848374	200	75
492 CHEMBL2095190	0.903079614	500	25
493 CHEMBL2111407	0.870454852	200	45

target_id	mcc_score	n_estimators	max_depth
494 CHEMBL4092	0.879125081	200	25
495 CHEMBL1914272	0.977916667	200	25
496 CHEMBL6113	0.891634992	500	50
497 CHEMBL3813	0.920451548	200	25
498 CHEMBL1211413	0.863446958	500	25
499 CHEMBL2708	0.801928568	1000	25
500 CHEMBL15500	0.802913392	1000	75
501 CHEMBL3761	0.965631323	200	25
502 CHEMBL4454	0.424091061	200	25
503 CHEMBL3883316	0.829964357	200	75
504 CHEMBL4525	0.633921061	1000	25
505 CHEMBL1075315	0.850339755	200	45
506 CHEMBL2885	0.585752523	200	25
507 CHEMBL3565	0.988561489	200	25
508 CHEMBL3638337	1	200	25
509 CHEMBL2872	0.656355566	200	25
510 CHEMBL2647	0.954245954	200	25
511 CHEMBL3548	0.980739223	200	75
512 CHEMBL3429	0.883067733	200	50
513 CHEMBL6191	0.252727125	1000	50
514 CHEMBL3313836	1	200	25
515 CHEMBL3137262	0.956054077	200	25
516 CHEMBL263	0.916601423	500	25
517 CHEMBL3430888	0.967492589	200	25
518 CHEMBL4360	0.977122977	200	25
519 CHEMBL2221341	0.917595371	500	25
520 CHEMBL2250	0.470370857	1000	100
521 CHEMBL2468	0.556355566	500	75
522 CHEMBL2309	1	200	25
523 CHEMBL4202	0.403494502	200	50
524 CHEMBL5903	0.675010304	200	25
525 CHEMBL2010636	0.9663718	200	25
526 CHEMBL2034805	0.95141144	200	25
527 CHEMBL2094126	0.866923442	500	50
528 CHEMBL15043	0.807466481	500	50
529 CHEMBL2433	0.186241732	200	25
530 CHEMBL2535	0.931180361	200	25
531 CHEMBL2552	0.88614806	200	25
532 CHEMBL4497	0.825961455	200	50
533 CHEMBL4208	0.878451494	1000	25
534 CHEMBL2052036	0.954379607	500	45
535 CHEMBL2321630	0.938282325	500	45
536 CHEMBL2010631	0.673024289	500	25
537 CHEMBL1275212	1	200	25
538 CHEMBL3863	0.675201296	200	25

target_id	mcc_score	n_estimators	max_depth
539 CHEMBL2283	0.92648194	200	25
540 CHEMBL1987	0.907939782	200	25
541 CHEMBL2857	0.874480948	500	75
542 CHEMBL1764940	0.928796834	200	25
543 CHEMBL1323	0.866363311	200	100
544 CHEMBL4773	0.988610676	200	25
545 CHEMBL2810	0.9103538	200	25
546 CHEMBL1628468	0.961703271	200	25
547 CHEMBL1725216	0.961954028	500	25
548 CHEMBL1741163	0.425243522	200	25
549 CHEMBL1770034	0.653543509	500	50
550 CHEMBL3337	0.959257649	200	25
551 CHEMBL1293302	0.398947002	500	25
552 CHEMBL1907588	0.751964753	500	45
553 CHEMBL3721308	1	200	25
554 CHEMBL3433	0.862056542	500	45
555 CHEMBL2888	0.934222437	200	75
556 CHEMBL5014	0.439228793	200	50
557 CHEMBL3775	0.625292893	1000	25
558 CHEMBL4037	0.734599645	1000	75
559 CHEMBL2319	0.864382689	200	25
560 CHEMBL1929	0.819599818	200	25
561 CHEMBL1770046	1	200	25
562 CHEMBL4527	0.494307915	1000	75
563 CHEMBL5337	0.834219793	200	50
564 CHEMBL2331047	0.947493665	200	25
565 CHEMBL5573	0.256443796	500	100
566 CHEMBL5299	0.957025307	200	25
567 CHEMBL1837	0.929046789	200	25
568 CHEMBL5844	0.990453403	200	25
569 CHEMBL3038477	0.979108884	200	25
570 CHEMBL2052031	0.825806648	200	75
571 CHEMBL1293287	0.412404731	200	75
572 CHEMBL5951	0.969017547	200	45
573 CHEMBL2689	0.79610472	200	45
574 CHEMBL3085	0.853894091	500	25
575 CHEMBL3338	0.84955158	200	25
576 CHEMBL4898	0.566982226	200	25
577 CHEMBL4036	0.735563409	500	75
578 CHEMBL5785	0.507298903	500	50
579 CHEMBL257	0.970202281	500	45
580 CHEMBL2526	0.991751661	200	25
581 CHEMBL5419	0.876862418	200	100
582 CHEMBL1795091	0.282182662	200	25
583 CHEMBL5487	1	200	25

target_id	mcc_score	n_estimators	max_depth
584 CHEMBL5938	0.812504884	200	25
585 CHEMBL6157	0.606490716	200	45
586 CHEMBL1918	0.89392392	200	45
587 CHEMBL13935	0.28266891	1000	75
588 CHEMBL2007629	0.19568607	1000	45
589 CHEMBL4556	0.927669881	200	25
590 CHEMBL12125	0.938583879	200	25
591 CHEMBL13758064	0.990482786	200	25
592 CHEMBL1293229	0.495531719	200	25
593 CHEMBL2063	0.979165022	200	25
594 CHEMBL1075269	1	500	25
595 CHEMBL1275221	0.9571289	1000	25
596 CHEMBL2480	0.903864446	200	25
597 CHEMBL5028	0.962040669	200	100
598 CHEMBL3980	0.855641418	1000	45
599 CHEMBL2094127	0.819015505	200	75
600 CHEMBL6144	0.730058332	200	25
601 CHEMBL5805	0.951297567	200	25
602 CHEMBL5185	0.982290586	200	25
603 CHEMBL1926488	0.913736582	500	45
604 CHEMBL5695	0.87661925	1000	100
605 CHEMBL401	0.840050605	200	25
606 CHEMBL1908385	0.780389603	200	25
607 CHEMBL12360	0.851996382	500	45
608 CHEMBL4730	0.942392667	1000	45
609 CHEMBL1935	0.96085582	200	100
610 CHEMBL2128	0.818872771	1000	75
611 CHEMBL3056	0.865299697	200	25
612 CHEMBL1795186	0.868875755	200	45
613 CHEMBL2164	0.721652665	200	45
614 CHEMBL5575	0.953886747	200	25
615 CHEMBL4892	0.846812623	200	25
616 CHEMBL3117	0.921607343	200	25
617 CHEMBL13106	0.882556346	500	50
618 CHEMBL3160	0.926572871	500	45
619 CHEMBL3385	0.774744845	200	25
620 CHEMBL2096680	0.817827503	200	50
621 CHEMBL329	0.9619886	200	100
622 CHEMBL5412	0.926619816	500	25
623 CHEMBL5464	0.823274872	200	25
624 CHEMBL5445	0.888663194	500	45
625 CHEMBL3637	0.9223382025	200	25
626 CHEMBL4244	1	200	25
627 CHEMBL5568	0.701195613	500	25
628 CHEMBL1795139	0.886288049	500	100

target_id	mcc_score	n_estimators	max_depth
629 CHEMBL2321614	0.876381433	200	50
630 CHEMBL3399	1	200	25
631 CHEMBL2111288	0.945927362	500	25
632 CHEMBL1169596	0.981119388	200	25
633 CHEMBL5661	0.983688586	200	25
634 CHEMBL4267	0.767361812	200	75
635 CHEMBL3961	0.793479431	200	45
636 CHEMBL5952	0.958693375	200	75
637 CHEMBL1808	0.935158071	1000	50
638 CHEMBL4601	0.733863695	200	25
639 CHEMBL3832645	1	200	45
640 CHEMBL2363017	1	200	25
641 CHEMBL4674	0.897020891	200	25
642 CHEMBL1741201	0.780840229	200	45
643 CHEMBL4317	0.908463524	200	25
644 CHEMBL5921	0.945996149	1000	45
645 CHEMBL3667	0.959488073	200	25
646 CHEMBL3114	0.923438142	1000	25
647 CHEMBL5524	0.822909918	200	75
648 CHEMBL1671613	0.954375116	200	75
649 CHEMBL1781	0.893164251	200	75
650 CHEMBL3740	0.963613568	500	45
651 CHEMBL3623	0.794433189	500	100
652 CHEMBL3784	0.698202866	500	25
653 CHEMBL1075152	0.984686176	500	50
654 CHEMBL6056	1	200	25
655 CHEMBL2186	0.793741551	500	25
656 CHEMBL1293289	0.857956552	500	100
657 CHEMBL2079849	0.914214578	200	25
658 CHEMBL330	0.949638727	1000	45
659 CHEMBL5406	0.873519912	200	25
660 CHEMBL4033	0.982491787	200	25
662 CHEMBL1873	0.844043608	200	45
663 CHEMBL5517	0.983777004	200	45
664 CHEMBL5850	0.913231002	1000	45
665 CHEMBL2095167	0.902300502	500	25
666 CHEMBL1293286	0.372340259	200	25
667 CHEMBL1075280	0.928686816	200	25
668 CHEMBL1293304	0.035598676	500	25
669 CHEMBL2095198	0.907953694	500	25
670 CHEMBL5158	0.970810885	200	25
671 CHEMBL2176859	0.973739535	200	25
672 CHEMBL291	0.89616743	200	25
673 CHEMBL2938	0.907389272	200	25

target_id	mcc_score	n_estimators	max_depth
674 CHEMBL4330	0.968536956	200	25
675 CHEMBL12361	0.838190398	500	75
676 CHEMBL3714531	0.99284074	200	25
677 CHEMBL5200	0.941095293	500	25
678 CHEMBL3023	0.900633767	200	100
679 CHEMBL2095172	0.849795893	200	25
680 CHEMBL2577	0.991908677	200	25
681 CHEMBL3430885	0.943775166	200	25
682 CHEMBL1803	0.97764058	1000	25
683 CHEMBL2095197	0.868183968	200	25
684 CHEMBL1861	0.983836662	200	25
685 CHEMBL5480	0.870295906	1000	75
686 CHEMBL2096980	0.875554053	1000	25
687 CHEMBL3478	0.950989889	500	25
688 CHEMBL3616354	0.870146338	200	75
689 CHEMBL1909484	0.482910336	500	25
690 CHEMBL3438	0.846538196	200	25
691 CHEMBL4977	0.875170319	200	100
692 CHEMBL5879	0.915817285	200	25
693 CHEMBL4150	0.746710188	200	25
694 CHEMBL5335	0.991927712	500	75
695 CHEMBL2664	0.886418889	1000	75
696 CHEMBL1785	0.911021231	200	100
697 CHEMBL4567	0.951044181	1000	50
698 CHEMBL2390810	0.889671545	500	100
699 CHEMBL3251	0.303876425	500	25
700 CHEMBL3272	0.765247064	200	50
701 CHEMBL2095202	0.363015774	200	50
702 CHEMBL2181	0.950855539	1000	25
703 CHEMBL5973	0.731721314	200	100
704 CHEMBL3883328	0.9838914	200	25
705 CHEMBL2978	0.919320181	500	75
706 CHEMBL614865	0.911436008	200	45
707 CHEMBL4237	0.791561923	500	25
708 CHEMBL3132741	0.96873563	500	50
709 CHEMBL3055	0.609856067	1000	50
710 CHEMBL1293239	0.64449664	1000	100
711 CHEMBL3263	0.93682778	200	45
712 CHEMBL2107838	0.991954328	200	25
713 CHEMBL3455	0.889934801	200	75
714 CHEMBL3198	0.822869973	500	75
715 CHEMBL1787	0.969723152	200	45
716 CHEMBL3906	0.741693978	200	25
717 CHEMBL3902	0.971384307	200	25
718 CHEMBL1293269	0.622500322	1000	50



target_id	mcc_score	n_estimators	max_depth
719 CHEMBL3060	0.838973116	500	25
720 CHEMBL2065	0.88460846	1000	45
721 CHEMBL1628481	0.869949119	200	25
722 CHEMBL3714704	1	200	25
723 CHEMBL3721	0.493328764	1000	75
724 CHEMBL5769	0.904438043	500	50
725 CHEMBL307	0.992907248	200	25
726 CHEMBL5678	0.902721869	200	25
727 CHEMBL2188	0.974040251	200	25
728 CHEMBL4225	0.661373736	500	25
729 CHEMBL3616	0.931148022	500	25
730 CHEMBL4444	0.929220225	200	50
731 CHEMBL4894	0.991962725	1000	25
732 CHEMBL3425388	0.985470432	200	25
733 CHEMBL2411	0.92732532	200	25
734 CHEMBL3753	0.88827607	500	75
735 CHEMBL2189162	0.960276396	200	25
736 CHEMBL4718	0.792076486	200	25
737 CHEMBL3250	0.951673874	200	50
738 CHEMBL4355	0.698054033	1000	45
739 CHEMBL4781	0.948860223	200	45
740 CHEMBL1795180	0.916606796	200	25
741 CHEMBL5939	0.790497386	1000	25
742 CHEMBL3898	0.984910901	200	25
743 CHEMBL2854	0.95403321	200	100
744 CHEMBL1882	0.984910901	200	25
745 CHEMBL1250375	0.950364359	1000	25
746 CHEMBL1163116	0.88125226	200	50
747 CHEMBL2095183	0	200	45
748 CHEMBL2163176	0.932876462	200	50
749 CHEMBL3638334	0.917475684	200	25
750 CHEMBL4791	0.903557126	1000	100
751 CHEMBL4896	0.958556564	200	25
752 CHEMBL4530	0.90260078	1000	75
753 CHEMBL3774295	0.874222275	200	100
754 CHEMBL2343	0.947184876	200	25
755 CHEMBL2146297	0.573298116	200	45
756 CHEMBL1287623	0.916676498	200	100
757 CHEMBL3998	0.845143208	200	25
758 CHEMBL5122	0.652032997	200	45
759 CHEMBL4617	0.859541122	1000	25
760 CHEMBL5662	0.949518238	1000	45
761 CHEMBL3921	0.917723267	1000	25
762 CHEMBL2111377	0.952863556	200	50
763 CHEMBL6069	0.919234144	200	45

target_id	mcc_score	n_estimators	max_depth
764 CHEMBL2318	0.843334306	500	50
765 CHEMBL13562165	0.985909924	200	25
766 CHEMBL5767	0.913933055	1000	25
767 CHEMBL5306	0.986637251	200	25
768 CHEMBL5852	0.933318805	200	100
769 CHEMBL3430907	0.957120457	1000	25
770 CHEMBL1681611	1	200	25
771 CHEMBL1641350	0.986652021	200	25
772 CHEMBL4800	0.973304041	500	25
773 CHEMBL4464	0.934139113	200	75
774 CHEMBL2758	0.514658252	200	25
775 CHEMBL2157	0.406096411	1000	45
776 CHEMBL5221	0.795742031	1000	25
777 CHEMBL4767	0.979614246	200	25
778 CHEMBL1169599	1	500	45
779 CHEMBL5558	0.964426568	500	45
780 CHEMBL4408	0.978886676	200	25
781 CHEMBL2490	0.965462225	1000	25
782 CHEMBL2474	0.88862476	1000	50
783 CHEMBL5319	0.625924817	500	45
784 CHEMBL5966	0.889086317	200	100
785 CHEMBL1741200	0.517283214	500	25
786 CHEMBL2749	0.887305534	500	25
787 CHEMBL3305	0.91416934	200	25
788 CHEMBL4566	0.908358496	200	25
789 CHEMBL2366512	0.927869426	200	75
790 CHEMBL4145	0.849546276	500	50
791 CHEMBL2593	0.929636156	200	45
792 CHEMBL2787	0.854865659	200	25
793 CHEMBL1741161	0.409434528	200	45
794 CHEMBL3038471	0.965229021	200	25
795 CHEMBL3349	0.980391707	200	25
796 CHEMBL2886	0.973367932	200	25
797 CHEMBL2919	0.884547605	200	50
798 CHEMBL3933	0.993718028	200	25
799 CHEMBL1944	0.933687128	500	45
800 CHEMBL1641358	0.992982973	200	25
801 CHEMBL2889	0.559530079	500	50
802 CHEMBL1287628	0.81393588	500	100
803 CHEMBL5024	0.939281836	200	25
804 CHEMBL3746	0.805273698	500	50
805 CHEMBL281	0.982339931	1000	25
806 CHEMBL5525	0.92701914	200	25
807 CHEMBL5401	0.80227298	200	50
808 CHEMBL3638350	0.992989563	200	45

target_id	mcc_score	n_estimators	max_depth
809 CHEMBL1075317	0.962791914	200	25
810 CHEMBL3712907	0.870115703	200	45
811 CHEMBL1835	0.736541924	500	50
812 CHEMBL3313	1	200	25
813 CHEMBL3259470	0.986714233	200	25
814 CHEMBL2253	1	1000	50
815 CHEMBL2069161	0.94711195	200	50
816 CHEMBL3655	0.920639936	200	45
817 CHEMBL5262	0.981035785	200	25
818 CHEMBL3402	0.603189331	500	45
819 CHEMBL2027	0.944490202	200	45
820 CHEMBL5720	0.894864768	200	25
821 CHEMBL2252	0.925560229	1000	25
822 CHEMBL2563	0.87130426	200	50
823 CHEMBL1795194	0.438550416	200	100
824 CHEMBL4338	0.974465878	200	25
825 CHEMBL2046264	0.993737494	200	25
826 CHEMBL2499	0.973188345	500	25
827 CHEMBL2095174	0.408524195	200	100
828 CHEMBL320	0.932724662	200	25
829 CHEMBL6101	0.225152206	200	25
830 CHEMBL3780	0.964991585	200	25
831 CHEMBL4086	0.954886128	200	25
832 CHEMBL3832644	0.993743687	200	25
833 CHEMBL1904	0.907307215	500	100
834 CHEMBL2083	0.810152674	200	100
835 CHEMBL5261	0.72527499	500	75
836 CHEMBL5285	0.862961937	1000	25
837 CHEMBL1822	0.957133776	500	45
838 CHEMBL4421	0.971240808	200	25
839 CHEMBL2447	0.98749948	200	25
840 CHEMBL3897	0.9158134	200	25
841 CHEMBL1907	0.882971941	200	75
842 CHEMBL4270	0.931669805	200	25
843 CHEMBL6084	0.775592521	1000	50
844 CHEMBL5451	0.910991208	200	100
845 CHEMBL3905	0.585308535	200	50
846 CHEMBL2321613	0.97475073	200	25
847 CHEMBL3868	0.932168989	500	25
848 CHEMBL1949	0.929528338	500	45
849 CHEMBL5131	0.803829333	200	25
850 CHEMBL2111363	0.945896995	1000	45
851 CHEMBL3459	0.731519329	1000	100
852 CHEMBL1795126	0.98224285	200	25
853 CHEMBL2145	0.939324543	200	25

target_id	mcc_score	n_estimators	max_depth
854 CHEMBL1907604	0.876742587	200	75
855 CHEMBL13119	0.969150658	200	25
856 CHEMBL14468	0.936912785	500	45
857 CHEMBL15631	0.8964671	200	75
858 CHEMBL13476	0.862375454	500	45
859 CHEMBL4376	0.789516739	200	50
860 CHEMBL2716	0.799582448	500	100
861 CHEMBL1667665	0.993778069	200	25
862 CHEMBL5963	0.913285516	200	50
863 CHEMBL14161	0.863900362	1000	50
864 CHEMBL2096912	0.926281506	200	25
865 CHEMBL2008	0.964808988	500	50
866 CHEMBL2391	0.898380996	500	25
867 CHEMBL3771	0.933210262	200	50
868 CHEMBL1681620	0.971708203	200	100
869 CHEMBL4465	0.934067444	200	100
870 CHEMBL4660	0.925491863	1000	25
871 CHEMBL3038506	0.884128056	200	45
872 CHEMBL5736	0.993783376	200	25
873 CHEMBL3038489	0.970333159	200	25
874 CHEMBL3368	0.952954014	200	45
875 CHEMBL3403	0.874736451	200	45
876 CHEMBL3411	0.321775229	200	45
877 CHEMBL3314	0.836882743	200	25
878 CHEMBL6120	0.965234845	200	45
879 CHEMBL2085	0.76888915	200	25
880 CHEMBL2055	0.970111578	200	25
881 CHEMBL3614	0.863705399	200	50
882 CHEMBL5293	0.959826078	200	100
883 CHEMBL5701	0.720972858	200	25
884 CHEMBL2725	0.894503043	1000	25
885 CHEMBL1907590	0.931122111	500	100
886 CHEMBL1075323	0.943580337	1000	25
887 CHEMBL2095189	0.830276971	1000	25
888 CHEMBL1782	0.926436053	200	50
889 CHEMBL3832642	0.929103087	200	25
890 CHEMBL3038499	0.964801749	200	25
891 CHEMBL5704	0.98819151	200	25
892 CHEMBL1293256	0.47867149	500	25
893 CHEMBL3238	0.712168874	1000	75
894 CHEMBL3831283	0.994392796	200	25
895 CHEMBL1255164	0.9503342	200	25
896 CHEMBL4027	0.925929215	200	45
897 CHEMBL2123	0.905047025	500	50
898 CHEMBL4878	0.944037006	200	25

target_id	mcc_score	n_estimators	max_depth
899 CHEMBL5530	0.660179	1000	25
900 CHEMBL1741203	0.641213372	200	25
901 CHEMBL1907592	0.943978889	500	100
902 CHEMBL2488	0.960346603	200	25
903 CHEMBL4029	0.933155861	200	75
904 CHEMBL4383	0.979434357	200	25
905 CHEMBL1907603	0.919892661	1000	45
906 CHEMBL4919	0.916154928	200	25
907 CHEMBL3085613	0.799152037	500	25
908 CHEMBL3308976	0.98527875	200	25
909 CHEMBL3081	0.885834526	200	50
910 CHEMBL461	0.951911407	200	25
911 CHEMBL4430	0.929485545	200	50
912 CHEMBL2016430	0.994415495	500	25
913 CHEMBL6166	0.764252337	1000	25
914 CHEMBL3891	0.89954944	1000	75
915 CHEMBL4132	0.909541165	500	25
916 CHEMBL2231	0.951521057	200	25
917 CHEMBL2096911	0.800112837	200	75
918 CHEMBL2397	0.963515903	200	45
919 CHEMBL1293249	0.466686086	200	75
920 CHEMBL3508	0.966490036	500	45
921 CHEMBL2336	0.951320807	200	50
922 CHEMBL5328	0.811467002	200	45
923 CHEMBL4696	0.522679338	500	75
924 CHEMBL4699	0.911919136	200	25
925 CHEMBL5582	0.940861202	200	25
926 CHEMBL5493	0.912203969	200	25
927 CHEMBL1919	0.984523528	200	45
929 CHEMBL3762	0.817012306	200	45
930 CHEMBL5689	0.975291683	200	25
931 CHEMBL3374	0.956812247	200	50
932 CHEMBL3100	0.964215434	200	75
933 CHEMBL2003	0.974083635	500	25
934 CHEMBL3268	0.888545284	500	50
935 CHEMBL3234	0.761446981	1000	50
936 CHEMBL2706	0.964221144	200	25
937 CHEMBL2037	0.995306523	200	25
938 CHEMBL2828	0.770294235	500	25
939 CHEMBL4816	0.889023614	200	25
940 CHEMBL4081	0.964638425	500	100
941 CHEMBL3180	0.815620377	200	45
942 CHEMBL3004	0.774258355	200	100
943 CHEMBL5545	0.960321368	200	100

target_id	mcc_score	n_estimators	max_depth
944 CHEMBL2366456	0.767582644	200	100
945 CHEMBL15062	0.902844406	1000	25
946 CHEMBL4835	0.837466147	200	75
947 CHEMBL1250413	0.380375847	1000	25
948 CHEMBL4899	0.894061097	500	50
949 CHEMBL1287620	0.306300783	500	25
950 CHEMBL3310	0.90521419	500	45
951 CHEMBL13544	0.709558891	1000	25
952 CHEMBL4062	0.976580049	200	50
953 CHEMBL1293263	0.798634708	500	25
954 CHEMBL4903	0.573960989	500	100
955 CHEMBL3788	0.750411041	200	50
956 CHEMBL4687	0.958169848	200	75
957 CHEMBL368338	1	200	25
958 CHEMBL5282	0.734230219	1000	25
959 CHEMBL4394	0.937913666	200	100
960 CHEMBL1841	0.49544723	1000	25
961 CHEMBL1075284	0.958195278	500	45
962 CHEMBL5800	0.732226683	200	50
963 CHEMBL3038474	0.981089514	200	25
964 CHEMBL3514	0.902977915	500	75
965 CHEMBL1793	0.450696572	200	75
966 CHEMBL2439	0.886664029	1000	100
967 CHEMBL5141	0.972474732	200	50
968 CHEMBL241	0.779278504	500	45
969 CHEMBL1255150	0.967508271	200	25
970 CHEMBL2096987	0.969311338	200	25
971 CHEMBL3038488	0.934951246	500	25
972 CHEMBL4921	0.944052823	200	50
973 CHEMBL2803	0.847262213	500	45
974 CHEMBL3819	0.950951987	500	25
975 CHEMBL1741221	0.288390218	1000	25
976 CHEMBL16137	0.827967955	200	50
977 CHEMBL5387	0.896150442	200	45
978 CHEMBL3360	0.94096299	500	45
979 CHEMBL12575	0.940404056	1000	25
980 CHEMBL2830	0.891609658	500	45
981 CHEMBL2821	0.777693461	500	50
982 CHEMBL4016	0.813143294	1000	75
983 CHEMBL4329	0.964269136	500	25
984 CHEMBL2756	0.834510881	200	100
985 CHEMBL3832641	0.921189187	200	25
986 CHEMBL4652	0.995969895	200	25
987 CHEMBL2169736	0.728651426	1000	25
988 CHEMBL5936	0.986539197	200	50

target_id	mcc_score	n_estimators	max_depth
989 CHEMBL4774	0.917596673	200	25
990 CHEMBL5103	0.922749471	500	25
991 CHEMBL2095164	0.915472092	200	50
992 CHEMBL2304402	0.947933383	200	25
993 CHEMBL1255126	0.840093405	200	75
994 CHEMBL3807	0.732440602	200	100
995 CHEMBL3201	0.373994805	200	45
996 CHEMBL2095226	0.901208017	1000	45
997 CHEMBL3590	0.9824072	200	25
998 CHEMBL4761	0.934327726	200	50
999 CHEMBL3361	0.924718244	500	25
1000 CHEMBL4714	0.93602488	1000	25
1001 CHEMBL2189110	0.919066324	200	50
1002 CHEMBL4828	0.944376481	1000	25
1003 CHEMBL4073	0.912861472	200	100
1004 CHEMBL3009	0.867571317	1000	45
1005 CHEMBL2850	0.75162145	200	50
1006 CHEMBL4140	0.917971056	1000	25
1007 CHEMBL4461	0.920083765	500	45
1008 CHEMBL5260	0.87477571	200	25
1009 CHEMBL3582	0.855789488	500	75
1010 CHEMBL4358	0.838539758	500	100
1011 CHEMBL3037	0.894624698	500	45
1012 CHEMBL3996	0.957052215	200	25
1013 CHEMBL2329	0.983682697	500	25
1014 CHEMBL2288	0.689915595	1000	45
1015 CHEMBL2111416	0.96535166	200	50
1016 CHEMBL278	0.978548979	500	45
1017 CHEMBL3589	0.912579875	500	45
1018 CHEMBL4188	0.939974188	500	75
1019 CHEMBL4683	0.851598664	200	50
1020 CHEMBL1075101	0.991297573	200	25
1021 CHEMBL2146303	0.717831251	500	25
1022 CHEMBL1955	0.669214008	200	50
1023 CHEMBL2265	0.811748155	500	45
1024 CHEMBL2734	0.937375811	200	25
1025 CHEMBL2096675	0.942304839	200	45
1026 CHEMBL4218	0.302384993	500	25
1027 CHEMBL4471	0.789602299	200	50
1028 CHEMBL1293243	0.564845164	200	25
1029 CHEMBL2096666	0.851163919	200	45
1030 CHEMBL3151	0.387290097	200	25
1031 CHEMBL2489	0.983745469	500	75
1032 CHEMBL3766	0.971946737	200	45
1033 CHEMBL3426	0.876737175	500	100

target_id	mcc_score	n_estimators	max_depth
1034 CHEMBL252	0.875439071	500	25
1035 CHEMBL1856	0.984142008	1000	45
1036 CHEMBL2366	0.489667899	200	45
1037 CHEMBL5485	0.894486524	200	25
1038 CHEMBL2072	0.968263805	200	25
1039 CHEMBL3437	0.911351717	200	25
1040 CHEMBL2146305	0.038225491	500	25
1041 CHEMBL2868	0.954714156	200	25
1042 CHEMBL3959	0.829510727	1000	25
1043 CHEMBL2007624	0.450525136	500	25
1044 CHEMBL1850	0.880445972	200	50
1045 CHEMBL3815	0.884805567	500	25
1046 CHEMBL5011	0.847860663	200	45
1047 CHEMBL1255149	0.944074043	200	50
1048 CHEMBL6089	0.389780854	200	50
1049 CHEMBL1293292	0.98056085	200	25
1050 CHEMBL2611	0.98177868	1000	45
1051 CHEMBL3530	0.977194231	200	25
1052 CHEMBL4959	0.992507938	200	25
1053 CHEMBL5247	0.984495958	200	25
1054 CHEMBL3419	0.926260172	500	100
1055 CHEMBL2029198	0.533854833	500	25
1056 CHEMBL5469	0.508254581	500	25
1057 CHEMBL4779	0.832595899	1000	75
1058 CHEMBL6145	0.976259065	500	75
1059 CHEMBL3969	0.826357697	200	25
1060 CHEMBL4804	0.820391998	200	75
1061 CHEMBL5719	0.93458862	200	25
1062 CHEMBL1961790	0.45100865	1000	75
1063 CHEMBL1781862	0.989045641	200	25
1064 CHEMBL3067	0.973814699	500	45
1065 CHEMBL5398	0.738594896	1000	25
1066 CHEMBL3202	0.834791391	500	50
1067 CHEMBL1997	0.97383649	200	45
1068 CHEMBL3359	0.886217063	200	25
1069 CHEMBL2617	0.94391864	200	45
1070 CHEMBL5192	0.830020789	500	45
1071 CHEMBL309	0.906900881	200	45
1072 CHEMBL4803	0.700783209	1000	25
1073 CHEMBL4596	0.898090958	500	45
1074 CHEMBL2203	0.976911204	1000	50
1075 CHEMBL1907587	0.922231597	200	100
1076 CHEMBL1902	0.901976271	200	50
1077 CHEMBL1075319	0.916477861	500	25
1078 CHEMBL1293319	0.31066099	1000	25

target_id	mcc_score	n_estimators	max_depth
1079 CHEMBL2216739	0.784856467	500	25
1080 CHEMBL3802	0.985146437	200	50
1081 CHEMBL3401	0.647883235	200	25
1082 CHEMBL5776	0.869308469	200	50
1083 CHEMBL2459	0.90477293	200	25
1084 CHEMBL3691	0.818494952	1000	75
1085 CHEMBL2146312	0.310197426	1000	25
1086 CHEMBL1615382	0.384545531	500	25
1087 CHEMBL5377	0.762819693	200	25
1088 CHEMBL2730	0.919231716	200	100
1089 CHEMBL315	0.767590666	200	50
1090 CHEMBL1864	0.953187444	1000	25
1091 CHEMBL4123	0.696087197	500	45
1092 CHEMBL1795101	0.967074464	200	25
1093 CHEMBL1907593	0.815438712	500	25
1094 CHEMBL2108	0.996295001	200	25
1095 CHEMBL4234	0.96059526	1000	25
1096 CHEMBL2096661	0.957666591	500	25
1097 CHEMBL1293266	0.342112335	200	25
1098 CHEMBL1993	0.596034961	200	25
1099 CHEMBL4079	0.892575618	200	25
1100 CHEMBL3816	0.949442178	200	50
1101 CHEMBL4482	0.892771209	1000	50
1102 CHEMBL1075051	0.925762245	200	45
1103 CHEMBL5339	0.90013657	200	45
1104 CHEMBL2146296	0.417182498	200	25
1105 CHEMBL3976	0.834571883	500	25
1106 CHEMBL1853	0.922489775	500	50
1107 CHEMBL5457	0.969880681	200	45
1108 CHEMBL2536	0.966043585	500	25
1109 CHEMBL5203	0.965633496	1000	100
1110 CHEMBL1075140	0.978791075	200	25
1111 CHEMBL2373	0.937859016	200	25
1112 CHEMBL5443	0.959202308	200	50
1113 CHEMBL5160	0.745727583	1000	25
1114 CHEMBL1907595	0.914002174	200	25
1115 CHEMBL2731	0.43822529	200	45
1116 CHEMBL4653	0.962822036	200	25
1117 CHEMBL6184	0.909730417	1000	50
1118 CHEMBL5804	0.62540281	500	50
1119 CHEMBL3687	0.576906436	1000	45
1120 CHEMBL2882	0.865839813	200	45
1121 CHEMBL3392948	0.847303203	500	25
1122 CHEMBL5331	0.829624039	500	100
1123 CHEMBL3038482	0.867904595	500	45

target_id	mcc_score	n_estimators	max_depth
1124 CHEMBL3399910	0.980945344	500	25
1125 CHEMBL13912	0.870153589	200	45
1126 CHEMBL2778	0.968296549	200	25
1127 CHEMBL1892	0.967865603	200	25
1128 CHEMBL1293267	0.603451706	500	25
1129 CHEMBL2304404	0.847914101	1000	75
1130 CHEMBL2955	0.747792808	500	45
1131 CHEMBL1293222	0.64206501	500	25
1132 CHEMBL3879842	0.987020529	200	25
1133 CHEMBL2000	0.877999235	500	25
1134 CHEMBL1952	0.87497723	1000	25
1135 CHEMBL3719	0.899074478	200	50
1136 CHEMBL1798	0.946869325	500	75
1137 CHEMBL15543	0.634992429	1000	25
1138 CHEMBL15443	0.854083496	500	75
1139 CHEMBL5617	0.447598562	200	25
1140 CHEMBL2094116	0.897693715	500	50
1141 CHEMBL2782	0.870465845	1000	45
1142 CHEMBL4780	0.865462731	500	25
1143 CHEMBL2457	0.502708943	200	25
1144 CHEMBL3038510	0.971671214	200	25
1145 CHEMBL1849	0.868424926	200	75
1146 CHEMBL1917	0.94239116	200	75
1147 CHEMBL16175	0.603761949	200	25
1148 CHEMBL2095184	0.974279949	200	75
1149 CHEMBL4336	0.962489346	200	50
1150 CHEMBL1860	0.939842121	500	50
1151 CHEMBL4600	0.930882294	200	100
1152 CHEMBL2567	1	200	25
1153 CHEMBL3247	0.969285733	200	100
1154 CHEMBL4895	0.765040667	500	100
1155 CHEMBL3025	0.836984273	500	45
1156 CHEMBL4398	0.916945411	1000	25
1157 CHEMBL4426	0.28584674	200	25
1158 CHEMBL3474	0.893410463	200	75
1159 CHEMBL2781	0.942841821	1000	50
1160 CHEMBL4869	0.688345673	500	50
1161 CHEMBL4506	0.768608395	1000	45
1162 CHEMBL3836	0.894864312	500	100
1163 CHEMBL2111367	0.938456774	500	50
1164 CHEMBL16164	0.890009798	500	50
1165 CHEMBL1741194	0.27070743	500	25
1166 CHEMBL2345	0.893613973	200	25
1167 CHEMBL1075145	0.947589724	500	100
1168 CHEMBL5263	0.967671276	1000	45

target_id	mcc_score	n_estimators	max_depth
1169 CHEMBL5314	0.883953747	200	45
1170 CHEMBL4489	0.953771163	200	100
1171 CHEMBL5570	0.985187737	500	25
1172 CHEMBL3157	0.960925863	1000	45
1173 CHEMBL4666	0.767163116	200	25
1174 CHEMBL5498	0.895616639	500	25
1175 CHEMBL1804	0.954784717	200	50
1176 CHEMBL3991	0.962440352	200	45
1177 CHEMBL5932	0.896731429	500	45
1178 CHEMBL2007	0.783635372	200	25
1179 CHEMBL2002	0.964893916	200	25
1180 CHEMBL1941	0.794187996	1000	25
1181 CHEMBL5027	0.525127502	200	50
1182 CHEMBL4068	0.94159965	1000	75
1183 CHEMBL3273	0.991371891	200	25
1184 CHEMBL5522	0.924134166	200	25
1185 CHEMBL3831281	0.994163341	200	45
1186 CHEMBL3529	0.938828246	200	25
1187 CHEMBL5067	0.926472474	200	50
1188 CHEMBL5979	0.654313072	200	25
1189 CHEMBL4789	0.807857914	200	100
1190 CHEMBL1907602	0.85376466	500	25
1191 CHEMBL5747	0.709905899	500	75
1192 CHEMBL362982	0.590229865	1000	25
1193 CHEMBL2331053	0.971410925	1000	45
1194 CHEMBL2096618	0.886780419	1000	45
1195 CHEMBL2285	0.87452206	500	25
1196 CHEMBL2902	0.924290393	200	50
1197 CHEMBL5462	0.977478141	200	45
1198 CHEMBL1293316	0.593694862	200	45
1199 CHEMBL1801	0.877332917	200	50
1200 CHEMBL3776	0.96569423	500	25
1201 CHEMBL4080	0.983298251	200	25
1202 CHEMBL3254	0.935175	200	25
1203 CHEMBL2871	0.956008503	200	45
1204 CHEMBL265	0.855587948	200	25
1205 CHEMBL3768	0.956442762	1000	100
1206 CHEMBL5501	0.376500404	200	25
1207 CHEMBL6154	0.908178044	1000	50
1208 CHEMBL3332	0.981644437	200	25
1209 CHEMBL1311	0.939750625	200	75
1210 CHEMBL5508	0.941028641	200	45
1211 CHEMBL3464	0.805905756	1000	45
1212 CHEMBL1293246	0.514064337	200	25
1213 CHEMBL2414	0.935747757	200	100

target_id	mcc_score	n_estimators	max_depth
1214 CHEMBL1900	0.905144376	200	25
1215 CHEMBL2095194	0.976480244	200	75
1216 CHEMBL12219	0.855908295	200	100
1217 CHEMBL4377	0.21815398	200	50
1218 CHEMBL1968	0.907847002	200	50
1219 CHEMBL3714130	0.961464901	200	45
1220 CHEMBL13920	0.918758736	500	25
1221 CHEMBL2095178	0.968211212	200	25
1222 CHEMBL1246299	0.640950662	1000	100
1223 CHEMBL4051	0.937145488	1000	50
1224 CHEMBL1246310	0.425023816	500	75
1225 CHEMBL4372	0.817846411	1000	500
1226 CHEMBL16152	0.514254657	500	50
1227 CHEMBL5669	0.95291971	200	45
1228 CHEMBL12996	0.921991672	1000	75
1229 CHEMBL3048	0.903104729	200	50
1230 CHEMBL326	0.919534625	200	75
1231 CHEMBL5036	1	200	25
1232 CHEMBL3540	1	200	25
1233 CHEMBL5413	0.980133546	500	25
1234 CHEMBL3797	0.210151465	200	50
1235 CHEMBL1921666	0.957731299	200	100
1236 CHEMBL3559	0.913140178	200	25
1237 CHEMBL3038511	0.965279198	200	25
1238 CHEMBL3922	0.94019161	200	45
1239 CHEMBL1293313	0.586268696	200	25
1240 CHEMBL1995	0.962874073	1000	50
1241 CHEMBL4204	0.858303939	200	45
1242 CHEMBL5491	0.938390352	200	45
1243 CHEMBL3108645	0.579097225	1000	45
1244 CHEMBL5697	0.862798129	500	50
1245 CHEMBL2470	0.928242537	500	25
1246 CHEMBL5007	0.404976803	200	25
1247 CHEMBL4698	0.988538433	200	25
1248 CHEMBL5471	0.938854063	500	45
1249 CHEMBL4681	0.88860684	500	45
1250 CHEMBL4111	0.98335121	1000	25
1251 CHEMBL2094130	0.912509699	200	45
1252 CHEMBL2016	0.940675155	500	25
1253 CHEMBL3318	0.81460617	500	75
1254 CHEMBL299	0.894319616	1000	25
1255 CHEMBL2061	0.909129513	200	25
1256 CHEMBL209	0.960389803	200	100
1257 CHEMBL298	0.969670492	1000	25
1258 CHEMBL2508	0.981555582	500	25

target_id	mcc_score	n_estimators	max_depth
1259 CHEMBL2094128	0.858908017	200	25
1260 CHEMBL2107	0.948698282	1000	45
1261 CHEMBL4191	0.907945637	1000	25
1262 CHEMBL324	0.86615457	500	25
1263 CHEMBL4793	0.928106733	1000	100
1264 CHEMBL4618	0.950032156	1000	45
1265 CHEMBL3942	0.937051038	500	100
1266 CHEMBL4824	0.991045665	200	25
1267 CHEMBL3778	0.866542454	1000	25
1268 CHEMBL1881	0.887722592	200	75
1269 CHEMBL2111389	0.913215746	500	45
1270 CHEMBL1907601	0.894525294	200	45
1271 CHEMBL4026	0.472835977	200	100
1272 CHEMBL1293275	0.626654158	1000	25
1273 CHEMBL5373	0.96256639	1000	100
1274 CHEMBL6136	0.901720415	200	45
1275 CHEMBL4641	0.936195183	200	45
1276 CHEMBL5353	0.986927932	1000	25
1277 CHEMBL3286	0.893109169	500	25
1278 CHEMBL1075257	0.52747196	200	25
1279 CHEMBL319	0.829007877	1000	25
1280 CHEMBL3535	0.98058643	500	25
1281 CHEMBL2959	0.903947053	500	50
1282 CHEMBL1293293	0.904434691	500	25
1283 CHEMBL1907594	0.862670165	200	50
1284 CHEMBL3833	0.967758464	500	25
1285 CHEMBL1075228	0.961333909	500	25
1286 CHEMBL2146316	0.768495329	500	25
1287 CHEMBL4975	0.933982974	500	100
1289 CHEMBL4478	0.938494165	500	100
1290 CHEMBL7951116	0.971751245	200	25
1291 CHEMBL1741195	0.86665415	1000	45
1292 CHEMBL3066	0.951308086	200	45
1293 CHEMBL4070	0.936685305	500	45
1294 CHEMBL3486	0.888138055	200	45
1295 CHEMBL3358	0.785360118	200	25
1296 CHEMBL6080	0.947665943	1000	100
1297 CHEMBL3018	0.954799825	500	25
1298 CHEMBL1741215	0.469599508	500	25
1299 CHEMBL1741180	0.331553587	200	25
1300 CHEMBL2635	0.498082653	500	25
1301 CHEMBL1795087	0.676449273	500	25
1302 CHEMBL1293247	0.622515327	1000	50
1303 CHEMBL3231	0.897735286	500	45

target_id	mcc_score	n_estimators	max_depth
1304 CHEMBL4128	0.883953254	200	50
1305 CHEMBL5017	0.91633207	200	100
1306 CHEMBL5414	0.908171636	1000	45
1307 CHEMBL2274	0.92445932	1000	25
1308 CHEMBL4662	0.934272632	1000	25
1309 CHEMBL6007	0.8432868	200	25
1310 CHEMBL3351	0.939511687	500	25
1311 CHEMBL2335	0.96958704	500	25
1312 CHEMBL2527	0.913557583	200	100
1313 CHEMBL3983	0.914850743	500	25
1314 CHEMBL5815	0.996107024	200	25
1315 CHEMBL4198	0.877567445	1000	100
1316 CHEMBL3864	0.633105951	200	45
1317 CHEMBL3038469	0.854362288	500	25
1318 CHEMBL4768	0.913897728	500	25
1319 CHEMBL4374	0.354889159	1000	25
1320 CHEMBL4801	0.948603489	200	25
1321 CHEMBL4393	0.924688108	200	25
1322 CHEMBL3072	0.923967265	500	25
1323 CHEMBL2041	0.784176743	500	75
1324 CHEMBL2095231	0.944355012	200	45
1325 CHEMBL3223	0.944067959	1000	45
1326 CHEMBL2496	0.963462876	500	100
1327 CHEMBL4462	0.875386216	500	25
1328 CHEMBL4224	0.773490293	1000	25
1329 CHEMBL1980	0.81242525	1000	100
1330 CHEMBL1293236	0.426048496	1000	25
1331 CHEMBL2525	0.940750616	500	45
1332 CHEMBL2736	0.917010275	1000	50
1333 CHEMBL2363	0.94057439	1000	25
1334 CHEMBL4018	0.8132558	200	45
1335 CHEMBL2789	0.9700103	200	25
1336 CHEMBL3181	0.943794696	200	50
1337 CHEMBL2590	0.886965461	500	75
1338 CHEMBL3714079	0.98229119	200	25
1339 CHEMBL1942	0.81048996	200	25
1340 CHEMBL3975	0.936874281	200	25
1341 CHEMBL5424	0.915348984	500	50
1342 CHEMBL5441	0.923479362	500	45
1343 CHEMBL2069	0.934476828	1000	50
1344 CHEMBL1741207	0.491912547	1000	25
1345 CHEMBL221	0.660847016	500	75
1346 CHEMBL2362978	0.661349396	1000	25
1347 CHEMBL5393	0.868551737	500	75
1348 CHEMBL4315	0.968432184	200	100

target_id	mcc_score	n_estimators	max_depth
1349 CHEMBL2207	0.982157928	200	25
1350 CHEMBL1836	0.933311893	200	100
1351 CHEMBL312	0.872535651	1000	25
1352 CHEMBL3138	0.922542544	1000	100
1353 CHEMBL2851	0.973200951	200	45
1354 CHEMBL3468	0.927896171	200	50
1355 CHEMBL1293237	0.422391806	200	45
1356 CHEMBL3729	0.852946413	200	25
1357 CHEMBL4685	0.798727274	500	75
1358 CHEMBL4077	0.931265392	500	50
1359 CHEMBL5077	0.902197304	200	75
1360 CHEMBL3308	0.669230775	200	25
1361 CHEMBL4414	0.936728604	200	100
1362 CHEMBL2094121	0.948911176	1000	45
1363 CHEMBL2073	0.646966647	1000	25
1364 CHEMBL1901	0.960328369	500	25
1365 CHEMBL1075232	0.993070519	200	25
1366 CHEMBL1741179	0.630649224	1000	25
1367 CHEMBL3553	0.865615195	500	100
1368 CHEMBL4599	0.940923455	200	25
1369 CHEMBL1878	0.965711764	1000	45
1370 CHEMBL1075126	0.987922934	1000	25
1371 CHEMBL1868	0.850300483	200	75
1372 CHEMBL2094122	0.964048447	500	100
1373 CHEMBL287	0.821413344	200	75
1374 CHEMBL2858	0.964000822	500	45
1375 CHEMBL4657	0.875891789	1000	25
1376 CHEMBL1899	0.882718464	1000	25
1377 CHEMBL4588	0.931694353	200	45
1378 CHEMBL4303	0.920615071	500	25
1379 CHEMBL1744525	0.955839112	500	50
1380 CHEMBL3524	0.901454133	200	45
1381 CHEMBL2568	0.947449968	200	25
1382 CHEMBL2916	0.844772989	200	25
1383 CHEMBL2622	0.856756152	500	75
1384 CHEMBL5646	0.967111254	200	75
1385 CHEMBL3045	0.963934254	200	75
1386 CHEMBL2068	0.928829309	200	75
1387 CHEMBL5365	0.457180116	500	45
1388 CHEMBL5555	0.954766753	500	100
1389 CHEMBL1907605	0.863476533	500	50
1390 CHEMBL5112	0.987121119	500	25
1391 CHEMBL3431938	0.663056409	1000	25
1392 CHEMBL3869	0.909789655	200	25
1393 CHEMBL5409	0.976283295	200	75

target_id	mcc_score	n_estimators	max_depth
1394 CHEMBL4633	0.956407843	1000	75
1395 CHEMBL1276	0.817706555	1000	25
1396 CHEMBL2094120	0.958366614	500	100
1397 CHEMBL1921	0.900421767	500	25
1398 CHEMBL4508	0.980253235	200	50
1399 CHEMBL1741193	0.353514977	500	25
1400 CHEMBL4477	0.929226639	500	100
1401 CHEMBL2028	0.971530125	1000	25
1402 CHEMBL3795	0.965452653	500	50
1403 CHEMBL5971	0.973357931	500	45
1404 CHEMBL3974	0.973309736	200	45
1405 CHEMBL4427	0.962173624	500	45
1406 CHEMBL2179	0.833726167	200	25
1407 CHEMBL1821	0.806797556	500	25
1408 CHEMBL3587	0.92653291	1000	25
1409 CHEMBL254	0.934548825	1000	50
1410 CHEMBL4777	0.835268245	500	45
1411 CHEMBL1875	0.974499044	500	45
1412 CHEMBL5313	0.287157699	200	25
1413 CHEMBL4608	0.884100731	200	75
1414 CHEMBL2820	0.95732083	200	45
1415 CHEMBL4102	0.779279671	1000	25
1416 CHEMBL4507	0.479186377	500	50
1417 CHEMBL202	0.897352594	500	25
1418 CHEMBL2413	0.945028407	200	45
1419 CHEMBL4321	0.914015837	1000	50
1420 CHEMBL3832643	0.997085356	200	45
1421 CHEMBL2993	0.883049295	500	45
1422 CHEMBL1916	0.845766292	200	25
1423 CHEMBL4481	0.878116516	1000	25
1424 CHEMBL2327	0.89457372	1000	45
1425 CHEMBL1163101	0.954856346	1000	45
1426 CHEMBL3230	0.807539655	1000	50
1427 CHEMBL2608	0.471034935	500	25
1428 CHEMBL3568	0.923232859	500	25
1429 CHEMBL3142	0.951408454	1000	50
1430 CHEMBL2111432	0.978517882	200	100
1431 CHEMBL2637	0.816194764	1000	25
1432 CHEMBL3471	0.788251106	1000	25
1433 CHEMBL3785	0.952435824	500	25
1434 CHEMBL1825	0.841308891	500	45
1435 CHEMBL2035	0.806427278	500	25
1436 CHEMBL1790	0.956212125	200	45
1437 CHEMBL5645	0.974828063	200	25
1438 CHEMBL4625	0.892749493	200	25



target_id	mcc_score	n_estimators	max_depth
1439 CHEMBL2487	0.868820649	500	100
1440 CHEMBL3116	0.959148097	1000	45
1441 CHEMBL4074	0.938315196	200	25
1442 CHEMBL301	0.854643541	500	100
1443 CHEMBL2208	0.912520799	200	75
1444 CHEMBL3943	0.873891447	500	25
1445 CHEMBL4203	0.831957814	1000	100
1446 CHEMBL3858	0.979653627	500	25
1447 CHEMBL304	0.935409808	200	25
1448 CHEMBL2949	0.671362183	1000	25
1449 CHEMBL4649	0.986654017	500	75
1450 CHEMBL1762	0.974596961	200	25
1451 CHEMBL2292	0.874056772	1000	25
1452 CHEMBL3137261	0.969361771	200	45
1453 CHEMBL248	0.901194398	200	100
1454 CHEMBL5857	0.771886424	200	25
1455 CHEMBL3356	0.673669077	500	50
1456 CHEMBL1926	0.934384187	200	45
1457 CHEMBL5147	0.930580009	500	45
1458 CHEMBL2598	0.985740191	200	25
1459 CHEMBL1293194	0.977930288	200	100
1460 CHEMBL1945	0.966281434	1000	25
1461 CHEMBL3510	0.898135597	500	50
1462 CHEMBL2581	0.899369761	200	25
1463 CHEMBL288	0.909616712	200	25
1464 CHEMBL227	0.87468423	200	50
1465 CHEMBL4644	0.932910579	200	45
1466 CHEMBL1913	0.847369599	500	100
1467 CHEMBL1981	0.91755577	200	45
1468 CHEMBL4227	0.958442867	200	25
1469 CHEMBL4429	0.970802509	500	25
1470 CHEMBL3375	0.863312138	200	25
1471 CHEMBL4893	0.903067606	500	25
1472 CHEMBL2047	0.911565496	1000	50
1473 CHEMBL3764	0.988627984	200	50
1474 CHEMBL1163125	0.851131839	200	25
1475 CHEMBL4766	0.988765595	200	25
1476 CHEMBL1966	0.921499899	500	25
1477 CHEMBL4439	0.94368849	500	50
1478 CHEMBL3769	0.931016441	200	45
1479 CHEMBL5080	0.981316969	200	100
1480 CHEMBL283	0.89184608	200	50
1481 CHEMBL1075322	0.466061052	1000	25
1482 CHEMBL5137	0.963513099	200	45
1483 CHEMBL2652	0.97184014	200	25

target_id	mcc_score	n_estimators	max_depth
1484 CHEMBL1983	0.911744097	1000	25
1485 CHEMBL5076	0.962093188	200	25
1486 CHEMBL1994	0.952377662	500	100
1487 CHEMBL1936	0.855969431	500	50
1488 CHEMBL3629	0.953707823	200	25
1489 CHEMBL2337	0.967557569	200	75
1490 CHEMBL3602	0.918059967	1000	100
1491 CHEMBL1898	0.926098597	500	100
1492 CHEMBL3501	0.966884131	200	45
1493 CHEMBL1946	0.970956727	500	25
1494 CHEMBL3710	0.9536675	200	50
1495 CHEMBL5619	0.538821676	200	25
1496 CHEMBL4143	0.298714207	1000	50
1497 CHEMBL3649	0.942342237	1000	25
1498 CHEMBL4306	0.920756257	1000	100
1499 CHEMBL2534	0.931340375	500	25
1500 CHEMBL2049	0.963695215	200	45
1501 CHEMBL1795085	0.343439919	1000	25
1502 CHEMBL3012	0.965362389	500	75
1503 CHEMBL2492	0.829624586	500	50
1504 CHEMBL2146304	0.342116011	1000	25
1505 CHEMBL2094108	0.879370637	200	75
1506 CHEMBL1811	0.971381584	1000	25
1507 CHEMBL3572	0.918967596	1000	50
1508 CHEMBL2431	0.969588136	200	45
1509 CHEMBL258	0.831708488	200	50
1510 CHEMBL1867	0.864320608	500	25
1511 CHEMBL3522	0.973264244	1000	100
1512 CHEMBL2808	0.929062066	200	45
1513 CHEMBL3798	0.991519599	500	25
1514 CHEMBL1827	0.951763905	1000	75
1515 CHEMBL1741165	0.543699575	500	25
1516 CHEMBL2326	0.895345999	200	50
1517 CHEMBL223	0.903236864	1000	25
1518 CHEMBL3199	0.930824453	1000	75
1519 CHEMBL3973	0.969667337	1000	25
1520 CHEMBL1741213	0.510562218	1000	25
1521 CHEMBL2425	0.954176138	1000	25
1522 CHEMBL232	0.917836338	200	25
1523 CHEMBL1907599	0.969169426	200	45
1524 CHEMBL1844	0.911482341	200	75
1525 CHEMBL5542	0.384499351	1000	25
1526 CHEMBL4093	0.946493003	1000	25
1527 CHEMBL1907598	0.954624145	1000	100
1528 CHEMBL2434	0.970500876	500	100

target_id	mcc_score	n_estimators	max_depth
1529 CHEMBL4581	0.925165747	1000	25
1530 CHEMBL3706	0.955937388	1000	50
1531 CHEMBL4501	0.944089751	1000	50
1532 CHEMBL1907600	0.9111651342	1000	45
1533 CHEMBL4561	0.978107812	1000	75
1534 CHEMBL285	0.948262627	500	25
1535 CHEMBL1908389	0.985157454	200	25
1536 CHEMBL1829	0.905752857	1000	25
1537 CHEMBL4072	0.880386619	200	45
1538 CHEMBL2056	0.848171519	200	75
1539 CHEMBL6009	0.969843407	200	25
1540 CHEMBL1741186	0.887551143	1000	25
1541 CHEMBL1947	0.740392768	1000	45
1542 CHEMBL1907589	0.889230479	200	25
1543 CHEMBL332	0.93188808	500	45
1544 CHEMBL2007626	0.444406751	1000	25
1545 CHEMBL3820	0.964858852	500	50
1546 CHEMBL6054	0.994524452	200	100
1547 CHEMBL5407	0.953061794	200	25
1548 CHEMBL1908	0.960571443	200	100
1549 CHEMBL331	0.961553458	200	50
1550 CHEMBL1977	0.517687676	1000	25
1551 CHEMBL5251	0.934497577	1000	45
1552 CHEMBL4179	0.936365524	1000	75
1553 CHEMBL3772	0.973272188	200	45
1554 CHEMBL213	0.936885099	200	100
1555 CHEMBL1889	0.950013716	500	50
1556 CHEMBL211	0.873004395	1000	50
1557 CHEMBL3713062	0.996639152	500	25
1558 CHEMBL3473	0.983020383	1000	25
1559 CHEMBL4979	0.955249167	500	100
1560 CHEMBL4247	0.935438725	500	45
1561 CHEMBL1075104	0.94169262	200	45
1562 CHEMBL1937	0.930251489	1000	50
1563 CHEMBL4308	0.991730065	200	100
1564 CHEMBL1985	0.981792879	200	100
1565 CHEMBL2366517	0.933591907	200	25
1566 CHEMBL5391	0.413697787	1000	25
1567 CHEMBL4261	0.662250883	200	25
1568 CHEMBL2276	0.929595036	200	75
1569 CHEMBL275	0.938666104	1000	25
1570 CHEMBL2722	0.960469111	200	75
1571 CHEMBL1855	0.990369356	500	50
1572 CHEMBL231	0.885365305	1000	100
1573 CHEMBL2093869	0.911269909	200	25

target_id	mcc_score	n_estimators	max_depth
1574 CHEMBL1795098	0.544252387	1000	25
1575 CHEMBL4552	0.983365815	500	45
1576 CHEMBL1906	0.946200836	500	100
1577 CHEMBL4158	0.689465439	1000	100
1578 CHEMBL1293234	0.533115396	500	25
1579 CHEMBL16141	0.982328344	1000	25
1580 CHEMBL2362981	0.299883153	500	25
1581 CHEMBL4908	0.98177603	200	25
1582 CHEMBL2243	0.933979749	1000	25
1583 CHEMBL2998	0.984262104	500	25
1584 CHEMBL246	0.968140479	200	45
1585 CHEMBL4142	0.966867998	200	25
1586 CHEMBL2146309	0.212481724	200	25
1587 CHEMBL3222	0.957332383	500	50
1588 CHEMBL1795086	0.555958489	1000	45
1589 CHEMBL2147	0.928000363	500	25
1590 CHEMBL4691	0.99707338	200	25
1591 CHEMBL4422	0.942351624	200	25
1592 CHEMBL3192	0.91894085	500	100
1593 CHEMBL5652	0.983238572	500	25
1594 CHEMBL5136	0.968214104	200	25
1595 CHEMBL3892	0.922163095	500	25
1596 CHEMBL4124	0.940151505	1000	25
1597 CHEMBL4441	0.976145698	1000	25
1598 CHEMBL2334	0.857325803	200	75
1599 CHEMBL4302	0.848721483	500	45
1600 CHEMBL215	0.884934301	500	25
1601 CHEMBL321	0.921331094	200	75
1602 CHEMBL2001	0.980419735	1000	75
1603 CHEMBL3837	0.897250157	200	45
1604 CHEMBL2362980	0.603597296	200	25
1605 CHEMBL4980	0.923835733	1000	45
1606 CHEMBL3145	0.907873645	500	50
1607 CHEMBL2973	0.9494102	1000	100
1608 CHEMBL1859	0.672430166	500	25
1609 CHEMBL5658	0.902219061	200	100
1610 CHEMBL229	0.91971036	500	100
1611 CHEMBL4630	0.931848427	1000	25
1612 CHEMBL2185	0.9142844	1000	50
1613 CHEMBL249	0.949056211	200	25
1614 CHEMBL322	0.89877598	1000	25
1615 CHEMBL5102	0.954126782	1000	25
1616 CHEMBL302	0.934746058	500	50
1617 CHEMBL1951	0.801063997	500	25
1618 CHEMBL2903	0.495492539	500	25

target_id	mcc_score	n_estimators	max_depth
1619 CHEMBL318	0.913383597	500	45
1620 CHEMBL6140	0.991715021	200	25
1621 CHEMBL335	0.838930106	500	25
1622 CHEMBL1862	0.913311817	1000	25
1623 CHEMBL3910	0.945647969	200	45
1624 CHEMBL269	0.957373576	1000	25
1625 CHEMBL2695	0.947377581	500	50
1626 CHEMBL2146302	0.732406665	500	45
1627 CHEMBL1991	0.953532696	200	25
1628 CHEMBL4860	0.958317428	1000	45
1629 CHEMBL3979	0.959057724	200	50
1630 CHEMBL3759	0.94117341	200	75
1631 CHEMBL3948	0.985566477	500	75
1632 CHEMBL2815	0.939388481	500	45
1633 CHEMBL216	0.881717496	500	25
1634 CHEMBL4523	0.965668473	200	25
1635 CHEMBL280	0.923006863	1000	25
1636 CHEMBL274	0.963374128	500	25
1637 CHEMBL1974	0.871563728	500	75
1638 CHEMBL1907596	0.933886487	500	25
1639 CHEMBL2392	0.380301146	1000	25
1640 CHEMBL210	0.838995338	500	100
1641 CHEMBL1293226	0.331590258	1000	45
1642 CHEMBL1914	0.842561324	200	100
1643 CHEMBL3884	0.978592694	200	45
1644 CHEMBL262	0.87845074	500	100
1645 CHEMBL333	0.922743193	200	25
1646 CHEMBL3130	0.911059454	500	45
1647 CHEMBL3571	0.952052292	500	50
1648 CHEMBL230	0.844217982	1000	50
1649 CHEMBL338	0.949231538	500	45
1650 CHEMBL1978	0.909893231	200	75
1651 CHEMBL2742	0.973865592	500	75
1652 CHEMBL2014	0.982684485	200	100
1653 CHEMBL286	0.985356841	200	50
1654 CHEMBL1800	0.988520414	500	25
1655 CHEMBL208	0.956251929	500	25
1656 CHEMBL1833	0.873153567	200	50
1657 CHEMBL1963	0.367338568	1000	50
1658 CHEMBL2366516	0.866674195	1000	50
1659 CHEMBL267	0.89567584	1000	50
1660 CHEMBL242	0.86958573	1000	100
1661 CHEMBL339	0.896970324	1000	50
1662 CHEMBL245	0.921779554	200	25
1663 CHEMBL3650	0.940227978	200	100

target_id	mcc_score	n_estimators	max_depth
1664 CHEMBL4153	0.934156283	1000	100
1665 CHEMBL219	0.918652894	200	100
1666 CHEMBL268	0.962727058	500	50
1667 CHEMBL4015	0.967052814	200	50
1668 CHEMBL1741217	0.541149087	500	45
1669 CHEMBL4616	0.973744222	500	45
1670 CHEMBL2409	0.940886294	200	100
1671 CHEMBL5514	0.374358195	200	50
1672 CHEMBL2366505	0.876690458	500	25
1673 CHEMBL3105	0.954222377	500	75
1674 CHEMBL3267	0.924247481	1000	75
1675 CHEMBL4354	0.949194445	200	25
1676 CHEMBL3229	0.930056242	1000	50
1677 CHEMBL3155	0.917547803	200	75
1678 CHEMBL1871	0.924783037	1000	25
1679 CHEMBL1824	0.953920709	200	75
1680 CHEMBL6035	0.553608422	1000	50
1681 CHEMBL1741219	0.531884938	200	45
1682 CHEMBL2564	0.912981734	1000	75
1683 CHEMBL4361	0.627861746	500	25
1684 CHEMBL4282	0.958917503	500	45
1685 CHEMBL2954	0.957209039	200	100
1686 CHEMBL1293248	0.397481235	1000	25
1687 CHEMBL4159	0.34131616	500	25
1688 CHEMBL247	0.918931105	1000	50
1689 CHEMBL4722	0.929180804	200	25
1690 CHEMBL270	0.959378961	500	100
1691 CHEMBL2835	0.96572562	200	100
1692 CHEMBL2148	0.944639761	200	50
1693 CHEMBL273	0.929613022	200	50
1694 CHEMBL204	0.877143046	500	75
1695 CHEMBL239	0.936237515	500	25
1696 CHEMBL238	0.908033017	1000	50
1697 CHEMBL1865	0.949346309	1000	25
1698 CHEMBL5763	0.880077869	200	100
1699 CHEMBL2094135	0.957588497	1000	75
1700 CHEMBL2039	0.876692103	1000	45
1701 CHEMBL4805	0.973336174	500	50
1702 CHEMBL255	0.930101705	200	45
1703 CHEMBL1741220	0.479916147	200	45
1704 CHEMBL3880	0.713421185	1000	50
1705 CHEMBL3227	0.938508445	1000	25
1706 CHEMBL3952	0.960420729	1000	75
1707 CHEMBL5023	0.943599942	500	50
1708 CHEMBL206	0.907218304	500	75

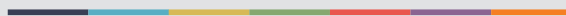
target_id	mcc_score	n_estimators	max_depth
1709 CHEMBL2034	0.970283511	1000	45
1710 CHEMBL4550	0.995250108	500	25
1711 CHEMBL1293255	0.546895178	500	45
1712 CHEMBL260	0.931550139	500	50
1713 CHEMBL4333	0.957130559	500	100
1714 CHEMBL1957	0.966149303	1000	45
1715 CHEMBL5113	0.926300414	500	45
1716 CHEMBL220	0.877581803	200	45
1717 CHEMBL2599	0.973608674	200	100
1718 CHEMBL313	0.94727531	1000	45
1719 CHEMBL259	0.96736559	200	25
1720 CHEMBL3397	0.626284682	1000	45
1721 CHEMBL5162	0.404927023	500	45
1722 CHEMBL4040	0.689301736	1000	75
1723 CHEMBL5071	0.973071885	200	50
1724 CHEMBL236	0.924269872	500	100
1725 CHEMBL4296	0.957299517	200	100
1726 CHEMBL225	0.895967492	1000	45
1727 CHEMBL4792	0.977778574	200	25
1728 CHEMBL4794	0.946648082	500	75
1729 CHEMBL222	0.935505117	500	50
1730 CHEMBL1075189	0.396703062	500	50
1731 CHEMBL289	0.650871139	200	45
1732 CHEMBL1293299	0.181187769	500	25
1733 CHEMBL3563	0.571176085	200	45
1734 CHEMBL5145	0.970099952	500	75
1735 CHEMBL3242	0.913057924	200	100
1736 CHEMBL1075138	0.37753813	1000	45
1737 CHEMBL3717	0.95706696	200	100
1738 CHEMBL1293258	0.47924441	1000	45
1739 CHEMBL3622	0.628797657	500	75
1740 CHEMBL2971	0.930749441	1000	45
1741 CHEMBL4829	0.986140331	1000	100
1742 CHEMBL235	0.937878627	1000	25
1743 CHEMBL1293224	0.627481644	1000	50
1744 CHEMBL244	0.960250072	500	100
1745 CHEMBL214	0.91917391	1000	100
1746 CHEMBL325	0.952682577	200	45
1747 CHEMBL284	0.969493357	1000	50
1748 CHEMBL2007625	0.405270057	500	50
1749 CHEMBL237	0.9222118655	200	45
1750 CHEMBL4078	0.881533809	200	100
1751 CHEMBL3594	0.904686862	200	75
1752 CHEMBL243	0.972124372	1000	50
1753 CHEMBL4235	0.947637188	1000	45

target_id	mcc_score	n_estimators	max_depth
1754 CHEMBL4409	0.977227018	200	45
1755 CHEMBL1233	0.920552253	200	45
1756 CHEMBL13371	0.943060661	200	45
1757 CHEMBL2842	0.924583284	500	100
1758 CHEMBL1264	0.94380751	500	75
1759 CHEMBL1224	0.909725054	500	50
1760 CHEMBL1256	0.920607405	200	100
1761 CHEMBL261	0.900264664	200	75
1762 CHEMBL3024	0.49946193	1000	45
1763 CHEMBL5990	0.475226982	500	45
1764 CHEMBL234	0.918570046	500	100
1765 CHEMBL1287622	0.215108752	500	25
1766 CHEMBL1226	0.924686557	500	45
1767 CHEMBL1205	0.913891802	1000	45
1768 CHEMBL1228	0.929043053	500	45
1769 CHEMBL1293227	0.504143589	500	25
1770 CHEMBL344	0.962519435	1000	100
1771 CHEMBL218	0.911731012	200	50
1772 CHEMBL1075094	0.432412217	1000	50
1773 CHEMBL4005	0.952303122	500	100
1774 CHEMBL1293254	0.278445607	500	45
1775 CHEMBL251	0.928187675	500	45
1776 CHEMBL6110	0.499054841	500	50
1777 CHEMBL4822	0.951972653	200	45
1778 CHEMBL203	0.935905909	500	50
1779 CHEMBL217	0.92006975	200	75
1780 CHEMBL5896	0.560030571	200	45
1781 CHEMBL4096	0.38814151	1000	25
1782 CHEMBL253	0.919895528	500	45
1783 CHEMBL240	0.729548701	500	45
1784 CHEMBL279	0.932638599	1000	45
1785 CHEMBL1293238	0.238154705	1000	25
1786 CHEMBL1741209	0.40891305	500	50
1787 CHEMBL340	0.710327672	1000	45
1788 CHEMBL1784	0.341937689	1000	25
1789 CHEMBL1293278	0.347656494	1000	50
1790 CHEMBL3577	0.3765550185	500	45
1791 CHEMBL6032	0.447834638	500	50
1792 CHEMBL1293277	0.575628997	500	75
1793 CHEMBL1293232	0.276454568	500	25
1794 CHEMBL1293294	0.578581347	1000	75
1795 CHEMBL5567	0.648636406	1000	75
1796 CHEMBL1293303	0.33805268	1000	45
1797 CHEMBL1293231	0.675461812	1000	50





Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



[uib.no](http://uib.no)

ISBN: 9788230847794 (print)  
9788230857410 (PDF)