

Using fingerprints and machine learning tools for the  
prediction of novel dual active compounds for  
leukotriene A4 hydrolase and soluble epoxide hydrolase

---

Dissertation

to obtain the doctoral degree of science

submitted to the department

Biochemistry, Chemistry and Pharmacy

at Goethe University

Frankfurt am Main

Lena Hefke

from Frankfurt am Main

Frankfurt am Main (2020)

(D30)



from the Department of Biochemistry, Chemistry and Pharmacy at Goethe University

accepted as a dissertation.

Dean: Prof. Dr. Clemens Glaubitz

Experts: Prof. Dr. Ewgenij Proschak

Prof. Dr. Steffan Knapp

Date of disputation: 12.11.2020



*Für meine Familie*



*“We are just an advanced breed of monkeys on a minor planet of a very average star. But we can understand the Universe. That makes us something very special.”*

Stephen Hawking (1942 - 2018)





## Declaration

I herewith declare that I have not previously participated in any doctoral examination procedure in a mathematics or natural science discipline.

Frankfurt am Main, 20.05.2020

---

Lena Hefke

## Author's Declaration

I herewith declare that I have produced my doctoral dissertation on the topic of:

*Using fingerprints and machine learning tools for the prediction of novel dual active compounds for Leukotriene A4 hydrolase and soluble epoxide hydrolase*

independently and using only the tools indicated therein. In particular, all references borrowed from external sources are clearly acknowledged and identified. I confirm that I have respected the principles of good scientific practice and have not made use of the services of any commercial agency in respect of my doctorate.

Frankfurt am Main, 20.05.2020

---

Lena Hefke



## Table of contents

List of abbreviations.....	i
1 Introduction.....	1
1.1 Structure-based drug design.....	2
1.2 Molecular representation.....	4
1.3 Machine learning.....	5
1.3.1 Support Vector Classification.....	7
1.3.2 Random Forest.....	8
1.3.3 Gradient boosting.....	9
1.4 Arachidonic acid cascade.....	10
1.4.1 Leukotriene A4 hydrolase.....	12
1.4.2 Soluble epoxide hydrolase.....	13
1.5 Multitarget drug design.....	14
1.6 Aim of the work.....	15
2 Methodology.....	17
2.1 Docking software.....	17
2.1.1 Molecular Operation Environment (MOE2018.0101).....	17
2.1.2 PLANTS.....	21
2.2 Compound preparation.....	23
2.2.1 Crystallized ligands.....	24
2.2.2 Active ChEMBL compounds.....	25
2.2.3 Inactive ChEMBL compounds.....	25
2.2.4 Combinatorial library.....	26
2.3 2D-fingerprints.....	27
2.4 3D-fingerprints.....	29
2.5 Fingerprint calculation.....	30
2.6 Machine learning prediction.....	32
2.7 General synthesis route.....	32
2.8 Fluorescence based LTA4H assay.....	34

2.9	sEH activity assay .....	36
3	Results and Discussion .....	37
3.1	Docking validation.....	37
3.1.1	MOE.....	37
3.1.2	PLANTS.....	41
3.2	Comparison of docking tools MOE and PLANTS.....	42
3.3	MOE docking procedure of compound batches 2 and 3 .....	43
3.4	MOE docking of compound batch 4 .....	44
3.5	PLANTS docking procedure of compound batches 2-4 .....	45
3.6	Machine learning optimization .....	46
3.7	Machine learning prediction from PLIF/MOE docking.....	51
3.8	Machine learning prediction from PLIF/PLANTS docking.....	57
3.9	Machine learning prediction from 2D-fingerprints.....	58
3.10	Synthesis and testing of selected compounds .....	62
3.11	Biological testing results.....	63
4	Conclusion.....	68
5	Summary .....	69
6	German Summary .....	74
7	Appendix .....	80
7.1	Python code for ML partitioning scheme optimization .....	80
7.2	Python code for ML parameter optimization.....	81
7.3	Python code for ML prediction.....	82
8	Bibliography.....	87
9	List of figures .....	97

## List of abbreviations

<b>Abbreviation</b>	<b>Meaning</b>
μL	Micro liter
μM	Micro molar
2D	Two-dimensional
3D	Three-dimensional
4-DMAP	4-Dimethylaminopyridine
5-HPETE	Hydroperoxyeicosatetraenoic acid
5-LO	5-lipoxygenase
Å	Ångström
AA	Arachidonic acid
ACO	Ant colony optimization
AdaBoost, ADA	Adaptive Boosting
ADMET	Absorption, distribution, metabolism, excretion, toxicity
Ala	Alanine
Asp	Aspartic acid
BTFFH	Fluoro-N,N,N',N'-bis(tetramethylen)formamidinium hexafluorophosphate
CCG	Chemical Computing Group
ChEMBL	Chemical database of the European Molecular Biology Laboratory
cmp.	Compound
CNS	Central nervous system
Confic.	Configuration
COPD	Chronic obstructive pulmonary disease
COX	Cyclooxygenase
cPLA2	Cytosolic phospholipase A2
CPU	Central Processing Unit
CV	Cross-validation
CYP450	Cytochrome P-450
Da	Dalton
DCM	Dichloromethane
DHET	Dihydroxyeicosatrienoic acid
DIPEA	N,N-diisopropylethylamine
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
ECFP	Extended Connectivity Fingerprint
EDC·HCl	1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide

EET	Epoxyeicosatrienoic acid
FCFP	Functional-Class Fingerprint
FP	Fingerprint
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
GOLD	Genetic Optimisation for Ligand Docking
GUI	Graphical user interface
HETE	Hydroxyeicosatetraenoic acid
hi	high
His	Histidine
HOBt·H <sub>2</sub> O	Hydroxy benzotriazole
HPLC	High-performance liquid chromatography
HTS	High-throughput screening
IC <sub>50</sub>	Half maximal inhibitory concentration
Ile	Isoleucine
Inc	Incorporation
KNIME	Konstanz Information Miner
lo	low
LTA <sub>4</sub>	Leukotriene A <sub>4</sub>
LTA <sub>4</sub> H	Leukotriene A <sub>4</sub> hydrolase
LTB <sub>4</sub>	Leukotriene B <sub>4</sub>
LTC <sub>4</sub>	Leukotriene C <sub>4</sub>
LTC <sub>4</sub> H	Leukotriene C <sub>4</sub> hydrolase
MACCS	Molecular Access System
MDB	Molecular Database
Met	Methionine
MgSO <sub>4</sub>	Magnesium sulfate
ML	Machine learning
MOE	Molecular Operation Environment
mol2	Tripos Mol2 file
MS Excel	Microsoft Excel
MW	Molecular weight
nm	Nanomolar
NMR	Nuclear magnetic resonance
Num.	Number
PDB	Protein Data Bank
PGH <sub>2</sub>	Prostaglandin endoperoxide H <sub>2</sub>
PGHS	Prostaglandin G/H synthase
PGP	Pro-Gly-Pro
Phe	Phenylalanine

PHOME	3-Phenyl-cyano- (6-methoxy-2-naphthalenyl) methylester-2-oxiran-acetic acid
PLANTS	Protein-Ligand ANT System
PLIF	Protein-Ligand Interaction Fingerprint
PLP	Piecewise linear potential
Pred.	Predict
Pro	Proline
ps	Picosecond
PyBOP	Benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate
QSAR	Quantitative Structure-Activity Relationship
RAM	Random-Access Memory
RBF	Radial basis function
RDKit	Open-Source Cheminformatics
rdon	Number of rotatable hydrogen bond donor groups
RF	Random Forest
rl	Ligand degree of freedom
RMSD	Root mean square deviation
RNA	Ribonucleic acid
rp	Receptor degree of freedom
SAR	Structure-activity relationship
SDF	Spatial Data File
sEH	Soluble epoxide hydrolase
Sim.	Similarity
SMARTS	SMILES arbitrary target specification
SMILES	Simplified molecular-input line-entry system
SVC	Support Vector Classification
SVL	Scientific Vector Language
SVM	Support Vector Machines
TCI	Tokyo Chemical Industry
THF	Tetrahydrofuran
Tris	Tris(hydroxymethyl)aminomethane
Trp	Tryptophan
Tyr	Tyrosine
VS	Virtual screening
XGBoost, XGB	Extreme Gradient Boosting
ZINC	Chemical database
Zn <sup>2+</sup>	Zinc ion

# 1 Introduction

The identification of lead compounds showing activity against a therapeutic target is the most important and crucial step in early-stage drug discovery in pharmaceutical and academic research. This is followed by the optimization of potency and pharmacological properties (e.g. pharmacokinetics, solubility and selectivity). High-throughput screening (HTS) is the conventional method of choice for hit identification in drug discovery. HTS comprises the testing of large compound libraries in an *in vitro* assay against pharmacologically relevant targets.<sup>1,2</sup> Reported hit rates of < 1%<sup>3-6</sup>, the high cost and time consuming character of this method shows the limitation of HTS. Progress in computational chemistry and computer-aided drug design offers an *in silico* alternative to conventional HTS. Those techniques both have the advantages of speed, cost efficiency and have become an essential part of drug design.<sup>7</sup> Virtual screening (VS), as one major *in silico* techniques, is the search of compound libraries with the goal of drug discovery. VS is widely used in design and optimization of new drugs. Two complementary areas of searching technique, namely ligand-based drug design and structure-based drug design, are part of VS.

Ligand-based drug design is based on known active ligands. Statistical methods and analytical tools are used to connect structural features to their corresponding biological effects. Some of the widely used methods in ligand-based drug design are 2D-QSAR, ligand-based virtual screening, pharmacophore generation/search and similarity search.<sup>1,8</sup>

Structure-based drug design uses the knowledge of the 3D-structure of the biological target to investigate the molecular interactions involved in protein-ligand binding.<sup>2</sup> A more detailed description of structure-based drug design follows in [section 1.1](#).

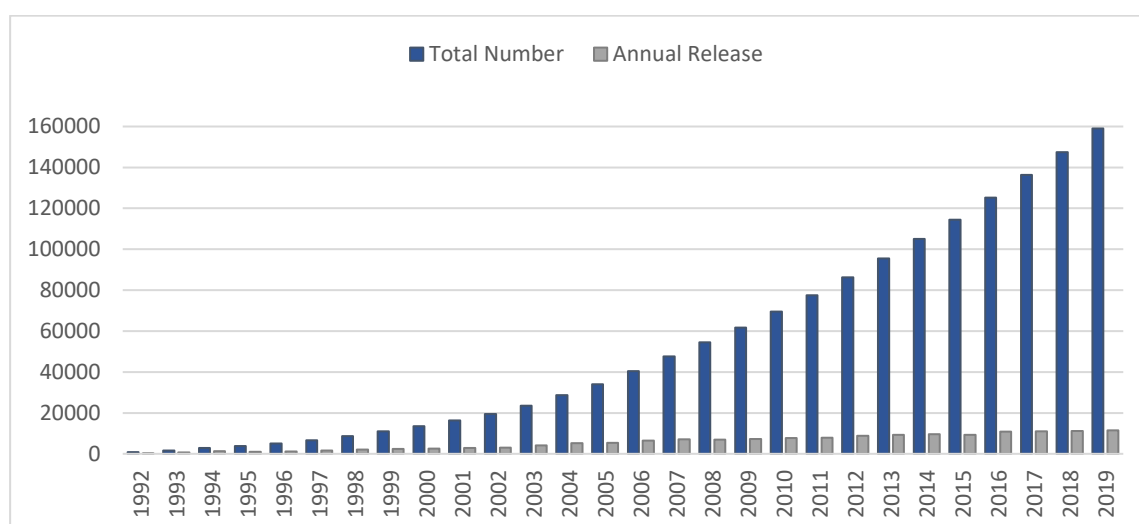
*In vitro* testing follows the hit generation of virtual screening to identify candidates for lead structure optimization.<sup>9</sup> In a first optimizing step, binding affinity is maximized through changing, adding and removing functional groups from the identified candidates. Lead compounds are optimized regarding physicochemical properties like absorption, distribution, metabolism, excretion and toxicity (ADMET). In this step bioisosteres are often used to modify physicochemical and pharmacokinetic properties. Bioisosteres are groups, which have a similar molecular shape and volume as well as approximately the



same distribution of electrons. Such groups keep or even increase the biological activity of the compounds and simultaneously improve pharmacokinetic/ physicochemical properties (e.g. solubility). The prediction of promising modifications of hit compounds, also in connection with QSAR methods, is not possible at this particular time.<sup>10,11</sup>

## 1.1 Structure-based drug design

Today, structure-based drug design plays a central role in the development of new therapeutic drugs. This progress is mainly based on the immense growth of available 3D-structures of biological targets. The key resource of 3D-structures today is the Protein Data Bank (PDB).<sup>12</sup> This open access database contains 3D-structure data for large biological molecules like proteins, DNA and RNA. In 2019, the PDB has accumulated over 150,000 available structures mainly from x-ray crystallography (~141,000 structures) but also NMR (nuclear magnetic resonance) (~12,500 structures) and electron microscopy (~4,000 structures) and it is continuously growing (**Figure 1**). This large number of structures provides the basis of structure-based drug design contributing to the design of new drugs. The structural information of protein structures and protein-ligand complexes with an increasing level of resolution can be used to understand and



**Figure 1: Histogram of overall growth of released 3D-structures in the PDB.** Since 1992, the available structures in the PDB have significantly increased. In addition, the number of structures released annually has increased to up to 10,000 new numbers every year. (Statistic taken from <https://www.rcsb.org/>, as at May 2020)

identify new targets and reaction pathways for treating diseases. The knowledge acquired in this way is used to analyze protein-ligand interactions and extract key interactions. It also enables the analysis of binding site topology, clefts and sub-pockets, electrostatic and hydrophobic properties. This information is used for the design of novel ligands containing all significant features for efficient modulation of the target of interest.<sup>13</sup>

X-ray structures, as the main source of structural information on protein targets, are merely snapshots of the protein at a certain time. Proteins are highly flexible but are frozen in one conformation during crystallization. This snapshot might not resemble the true conformation *in vivo*. The activity of a protein can be linked to its flexibility. The phenomenon of induced fit is well documented and can make the difference between a ligand being an agonist or antagonist.<sup>8,13</sup>

All structural data used for structure-based design must be carefully chosen. The determination of an X-ray structure is to a high degree a subjective interpretation of a measured electron density map by the crystallographer. This subjectivity must be handled carefully by everyone using an X-ray structure for structure-based design. Starting from an insufficient or wrong protein-ligand complex structure might jeopardize the entire drug discovery project. Not only is the resolution of an X-ray structure essential, but other crystallographic parameters (R-value<sup>14</sup>, B-value (temperature factor)<sup>15</sup>, Ramachandran violations<sup>16</sup>, structure-factors<sup>17</sup>) also contribute to the quality and suitability of a structure. The electron density map of the target of interest, especially the binding site residues and crystalized ligand atoms, should be verified manually.<sup>18,19</sup>

The first step in the drug design cycle is the hit identification. This can either be accomplished via HTS or *in silico* screening. Structure-based virtual screening involves placing compounds from a screening library in a target receptor similar to the way it happens *in vitro*. This process is called “docking”. A variety of conformations of the compound is placed into the receptor by a docking software. With this method, possible poses of the compounds are screened. After placement, the obtained poses are evaluated by “scoring functions”. Those scoring functions should find low energy poses and distinguish between the experimental determined binding mode and all other generated poses.<sup>20</sup> Scoring is used to select hits for synthesis, testing and further optimization.<sup>2,8,21,22</sup> Li et al. were able to show that reproducing the bioactive conformation is achieved in up to 60 – 80% of the cases.<sup>23,24</sup> These results are promising, but several studies show that the ultimate goal of scoring docked compounds in correspondence to their actual binding

affinity is not yet achieved.<sup>23-25</sup> Nevertheless, virtual screening is a key technique in the process of computer-aided drug design.

## 1.2 Molecular representation

Fingerprints are an efficient way of representing molecules. Fingerprints are binary strings representing chemical structures and properties. They can, for example, code the presence or absence of substructure features (e.g. MACCS fingerprint<sup>26-28</sup>) or extract chemical patterns within a specific diameter from a chemical graph (e.g. Morgan fingerprint<sup>26,29-31</sup>). Originally fingerprints were used for substructure or similarity search. In a fingerprint scheme, every bit of substructure feature is either on (=1) or off (=0). Such a binary fingerprint scheme is easy to generate, manipulate and compare. Therefore, the systematic analysis of large data sets is possible.<sup>32,33</sup> To compare two molecules based on their generated fingerprint, the Tanimoto coefficient<sup>34</sup> is the similarity measure of choice. If two molecules have  $a$  and  $b$  bits set on in their fingerprint, with  $c$  of these bits set on in both fingerprints, then the Tanimoto coefficient is defined as:

$$Tanimoto = \frac{c}{a + b - c}$$

The Tanimoto coefficient ranges between zero, no bits in common, and one, all bits are the same.<sup>33</sup> The Tanimoto coefficient was used in this work to find the most likely active pose of the docked compounds for the used data sets. In this work, 2D- and 3D-fingerprints were used to describe molecules in the data sets (see [section 2.3](#) and [section 2.4](#)).

2D-fingerprints are the most frequent type of molecular representation reported in literature.<sup>33,35</sup> They aggregate information like number of hydrogen bond donors/acceptors, number of ring systems and connectivity indices.<sup>36,37</sup> In addition to 2D-fingerprints, pharmacophoric patterns, surface properties, molecular volumes or molecular interaction fields are included in 3D-fingerprints. Those fingerprints are frequently used in 3D-QSAR studies. The complexity of 3D-fingerprints can range from single spatial patterns in a molecule to the presence or absence of many potential pharmacophore arrangements in a molecule. The fundamental assumption, that there is an underlying relationship between the molecular structure and its bioactivity, forms the basis of QSAR and the development of 2D- and 3D-fingerprint representation.<sup>37-39</sup>

### 1.3 Machine learning

Many different machine learning algorithms are described in literature, in this introduction only those used in this work are going to be described. Predictive modeling always deals with the task to develop a model based on known data to make an accurate prediction on new/unknown data. Predictions are always based on the available input data. Describing predictive modeling as a mathematical problem means to find a function  $f$  from input variables  $x$  to an output variable  $y$ .

$$y = f(x)$$

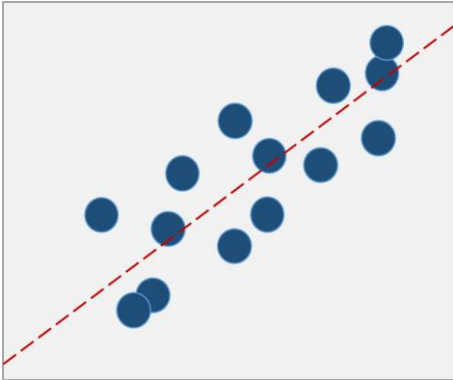
The function  $f$  is not known, therefore machine learning algorithms are trained to find the best function of mapping input variables to output variables. All functions found by machine learning algorithms are only the best possible approximation, since there is an irreducible error  $e$  that is independent of the input variables  $x$ .

$$y = f(x) + e$$

This error might be the fact, that there are not enough attributes to sufficiently characterize the best mapping function from  $x$  to  $y$  or the provided data is noisy with many outliers. The estimated mapping function always relies on the input data we can provide, and therefore, the estimations will have an error. Different machine learning algorithms make different assumptions about shape and structure of the function to learn. They differ in the way on how to best optimize a function. It is important to try different machine learning algorithms, since it is not known beforehand which approach will be best suited for the problem at hand.<sup>40</sup>

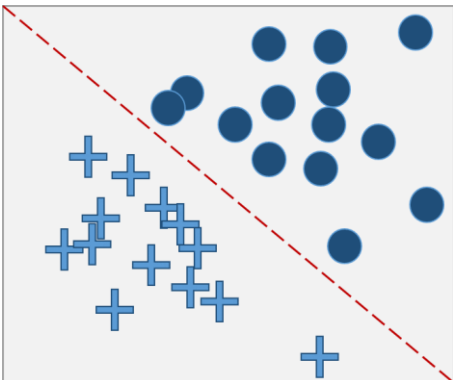
Machine learning algorithms can be classified into two groups based on the way they learn a mapping function to predict unknown data: supervised and unsupervised learning. Supervised learning means, as described above, to find a function  $f$  to map input variables  $x$  to output variables  $y$ . Labels for the input data have to be provided for supervised learning. The learning algorithm iteratively makes predictions and is corrected by altering settings in the learning algorithm. The learning stops when a mapping function is found that achieves acceptable prediction accuracy.<sup>41</sup>

Supervised learning algorithms can further be classified into regression and classification problems. In both cases, the task is to find specific relationships or structures in the available input data set to make accurate predictions about the output data. A regression problem is defined when the output variable to predict is a real or continuous value (Figure 2). Regression problems can be evaluated using root mean squared error. Different regression algorithms are used to solve various regression problems, for



**Figure 2: Simple regression problem.** Linear regression tries to fit data with the best hyper-plane to make a prediction.

example linear regression, regression trees or support vector regression. Classification problems are described if the task is to predict discrete values. Those values correspond to a finite discrete number of class labels. A classification problem is solved by drawing conclusions from the input data. A trained classification model tries to predict the class label of the output data (Figure 3). Classification problems can be evaluated using accuracy calculation.<sup>42,43</sup> Different classification algorithms can be used to solve a present problem, for example decision trees, naive Bayes, k-nearest neighbors or support vector classifier.<sup>44</sup> All classification tasks generally start with the separation of the input data



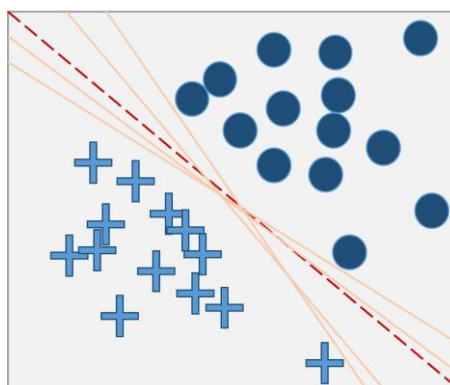
**Figure 3: Simple classification problem.** Binary classification of a data set is shown, where two classes are predicted.

into training and test sets. For the input data, the class labels are known. The training set is used to train the model, afterwards this model is used to predict the class labels for the test set. By doing so, the model accuracy can be calculated. If the accuracy reaches a satisfying level, the model can be used to predict the class labels of new/unknown data.

In unsupervised learning on the other hand, only the input variable  $x$  is present but no corresponding output variable  $y$  or labels. <sup>42,43</sup> The goal in this class of machine learning algorithms is to discover the fundamental structure of a given data set. The most common task in unsupervised learning is clustering and reduction in dimensionality. The algorithms are left on their own devices (unsupervised) to identify and present the structure of the given data. In most unsupervised learning methods there is no way of comparing model performance, since no labels are provided in the input variable. <sup>41</sup>

### 1.3.1 Support Vector Classification

Support Vector Classification (SVC) <sup>45</sup> belongs to the machine learning methods of Support Vector Machines (SVM). Besides SVC, Support Vector Regression is also possible. <sup>45,46</sup> Like all supervised classification methods, SVC tries to find the best possible model based on the input data. In a two class or binary classification problem, the goal of an SVC model is to separate the two input classes by a mapping function. The model also has to be generalizable to work on new/unknown data. Kernel functions are used to transform the input data into high-dimensional space. SVCs handle non-linear data and successfully finds relations to successful predict the class labels of interest. <sup>47</sup> There are four basic kernels which can be used to train a SVC model: linear, polynomial, radial basis function (RBF) and sigmoid. Choosing the appropriate kernel function is necessary to produce a model, which is able to separate the input data into the desired classes. The generalizable

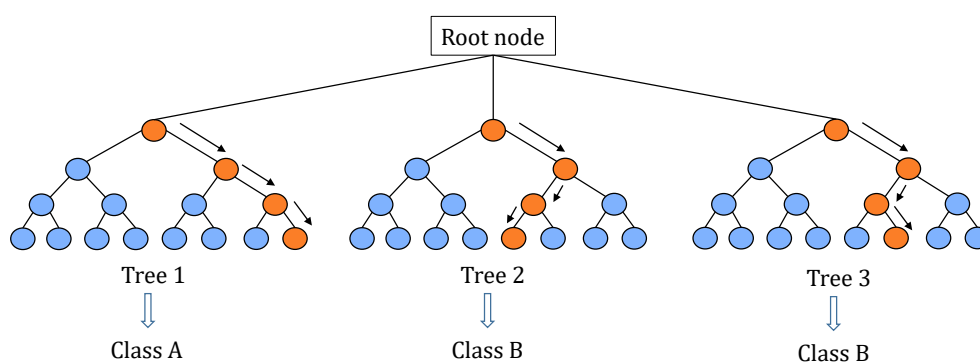


**Figure 4: Optimal separation hyperplane.** Shown are many possible linear classifier (yellow lines) to separate the binary data set, but only one that maximizes the margin (dashed red line).

model has to maximize the margin between the data points. This optimal separating hyperplane maximizes the distance between the linear classifier and the nearest data point of each class label (**Figure 4**).

### 1.3.2 Random Forest

Random Forest<sup>49</sup> is an ensemble method based on decision trees using a bagging technique. Decision trees are used to split the features of the input data into root, internal and leaf nodes, until all data is sorted into one of the desired classes. At each internal node in a decision tree, an impurity measure is used to split the data further down the tree.<sup>50</sup> The feature that best splits the input data is set as the root node.<sup>51</sup> If a large number of decision trees is generated, and a majority vote of those trees gives a final class label, it is



**Figure 5: Schematic representation of a Random Forest model.** Random Forest build of three decision trees. The majority vote of all trees leads to the final class. (Figure after Gray et al.)<sup>48</sup>

called a random forest (**Figure 5**). A random forest model is trained on data of  $n$  molecules with corresponding descriptors and class labels. The training algorithm obeys the following procedure:

1. A large number of random subsets of the training data is selected (with replacement: molecules from the training data can be selected more than once). Such a selection is called bootstrapping.
2. For each random subset, a decision tree is grown. At each internal node the best split is chosen among a randomly selected subset of descriptors. The trees are grown until no further splits are possible.
3. Repeat step 1 and 2 until an adequate number of trees is grown. The trees are generated in parallel and grown to their maximum extent, which is called bagging. For each of the  $n$

molecules in the data set a class label is predicted using the majority vote of all grown trees.<sup>52</sup>

### 1.3.3 Gradient boosting

The foundation of gradient boosting is additive modeling. The idea of additive modeling is to add many simple terms together to receive a more complicated expression. In gradient boosting this concept is used to gradually adapt an approximated function to obtain a satisfying function  $F_M(x)$ . This is achieved by adding up sub-functions to the initial function  $f_0(x)$ . The target function  $F_M(x)$  is built up in the following way:

$$\hat{y} = F_M(x) = f_0(x) + \Delta_1(x) + \dots + \Delta_M(x) = f_0(x) + \sum_{m=1}^M \Delta_m(x)$$

$F_M(x)$  accumulates the sum of  $\Delta_m$  sub-functions from  $m=1$  to  $M$  starting from an initial function  $f_0(x)$ . Those sub-functions are called weak models or weak learners in boosting terminology. In tree boosting, weak models are constructed and added in a gradual fashion, each one built to improve the overall model performance and reducing the error of the function. By building  $f_m(x)$  the previous function is not altered. The hyper-parameter  $M$  (number of stages) is defined since arbitrary growth of  $M$  leads to the risk of overfitting. Overfitting refers to a model, which models the training data too accurately. Details and noise in the training data is learned to an extent that predicting new data is negatively influenced. Boosting itself makes no determination on how to choose the weak learners or the form of the weak learners. However, for example, if all weak learners are linear models, the resulting model is also a linear model. Another hyper-parameter is  $\eta$ , the learning rate. The learning rate speeds up or slows down the overall approach of  $\hat{y}$  to  $y$ , which reduces the risk of overfitting.<sup>53-55</sup>

$$\hat{y} = F_M(x) = F_{M-1}(x) + \eta * f_m(x)$$

XGBoost (Extreme Gradient Boosting)<sup>54-56</sup> and AdaBoost (Adaptive Boosting)<sup>54,55,57</sup> are machine learning algorithms based on boosting.

AdaBoost generates decision models trying to classify the input data. In the next model the weights of misclassified data are exaggerated to give those data points a better chance of being classified correctly in the following models. The building of new models and adapting the weights of misclassified data is repeated several times to generate an ensemble of decision models. This ensemble should best classify the input data with as

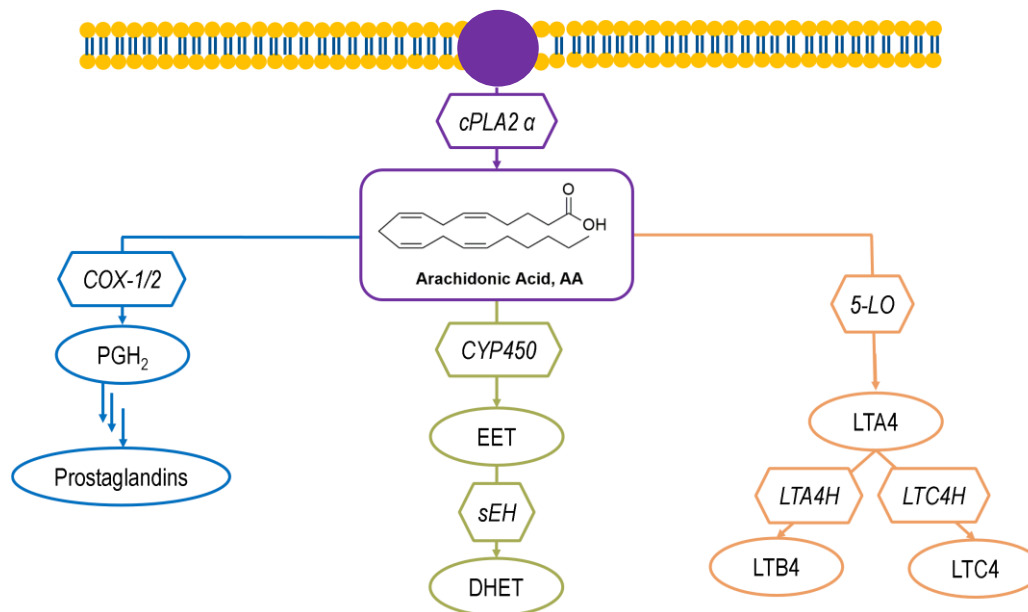


few misclassified data points as possible. The goal of this machine learning algorithm is to classify every data point perfectly. Thereby AdaBoost is vulnerable to noisy data and outliers.<sup>54,55,57</sup>

XGBoost is a gradient boosting ensemble machine learning algorithm, which was designed with attention to computational speed, model performance and enables parallel computing. XGBoost is very flexible, efficient and scalable, and can be used for regression, classification and ranking. The algorithm is based on the gradient boosting framework by J. Friedman et al.<sup>56,54</sup>. In XGBoost it starts with one naïve model, for which the errors are calculated for each data point in the input data. After that a model is built to predict those errors. This last model is then added to the ensemble of models. This cycle is repeated several times leading to the final prediction class. The model can be tuned with a few parameters. First, the number of estimators (`n_estimators`) specifies how many model cycles are generated, typical values range from 100-1,000. Second, the maximum depth of a tree (`max_depth`) can be tuned. Higher depth will allow the model to learn relations very precisely, which can lead to overfitting. The learning rate determines the impact of each tree on the final prediction. Lower values for learning rate are preferred since it makes the model robust and therefore makes the model generalize well.

## 1.4 Arachidonic acid cascade

A complex network of cellular factors controls inflammation. One main player in that network is the oxidative metabolism of arachidonic acid, the so-called arachidonic acid cascade. Arachidonic acid (AA) is a polyunsaturated fatty acid, which in its inactive form is located in membrane phospholipids.  $\text{Ca}^{2+}$  influx activates cytosolic phospholipase A2 (cPLA2), which releases AA from the phospholipids in the membrane. Free AA is then available for oxidative metabolism by three main enzymatic pathways.<sup>58-60</sup> Those pathways are the cyclooxygenase pathway, lipoxygenase pathway and cytochrome P-450 pathway (**Figure 6**).



**Figure 6: Arachidonic acid cascade.** Arachidonic acid is released from the membrane and metabolized via three enzymatic pathways, cyclooxygenase pathway, lipoxygenase pathway and cytochrome P-450 pathway.

In the cyclooxygenase pathway, cyclooxygenase (COX) (or prostaglandin G/H synthase, PGHS) catalyzes the transformation from AA into the reactive intermediate prostaglandin H<sub>2</sub> (PGH<sub>2</sub>). PGH<sub>2</sub> is converted into the biologically active prostaglandins, prostacyclin and thromboxane.<sup>58,61</sup> Prostaglandins play an important role in inflammation and thrombocyte aggregation. Well-established COX inhibitors, like the non-steroidal analgesics Aspirin and Ibuprofen, inhibit the biosynthesis of prostaglandins in the treatment of inflammation.<sup>62</sup>

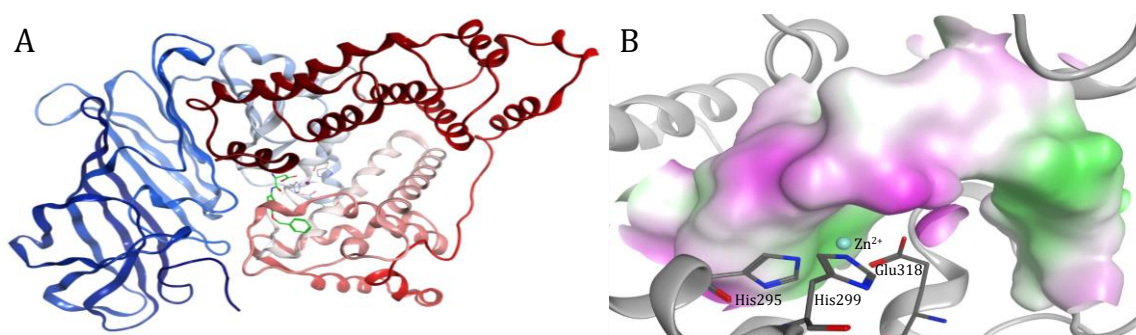
In the lipoxygenase pathway, 5-lipoxygenase (5-LO) transforms AA into the hydroperoxide 5-HPETE (hydroperoxyeicosatetraenoic acid). In the next step, 5-HPETE is transformed into the instable intermediate leukotriene A<sub>4</sub> (LTA<sub>4</sub>) by the 5-LO. LTA<sub>4</sub> is further metabolized either by leukotriene A<sub>4</sub> hydrolase (LTA<sub>4</sub>H) to LTB<sub>4</sub> or by the integral membrane protein leukotriene C<sub>4</sub> synthase to LTC<sub>4</sub>. Many acute and chronic inflammation diseases are connected with leukotrienes. These include asthma, rheumatoid arthritis, dermatitis and atherosclerosis.<sup>63</sup>

In the cytochrome P-450 (CYP450) pathway, CYP450 transforms AA into 20-hydroxyeicosatetraenoic acid (20-HETE) and epoxyeicosatrienoic acids (EETs). EETs have been determined to have anti-inflammatory effects. Transformation of EETs by

soluble epoxide hydrolase (sEH) into corresponding dihydroxyeicosatrienoic acids (DHETs) diminishes the positive anti-inflammatory effect of EETs.<sup>62</sup>

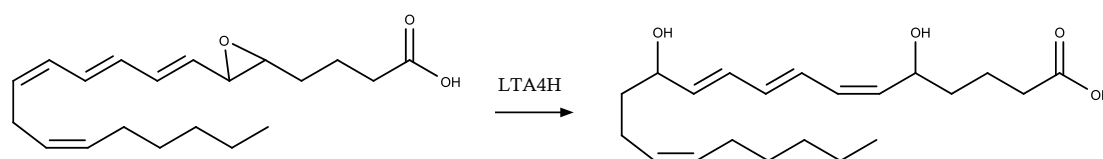
#### 1.4.1 Leukotriene A4 hydrolase

Leukotriene A4 hydrolase (LTA4H) (**Figure 7 A**) is a soluble zinc metalloenzyme, that catalyzes the hydration of LTA<sub>4</sub> into LTB<sub>4</sub> (**Figure 8**).<sup>61</sup> LTA4H has been found virtually in all cells, organs and tissues.<sup>64</sup> According to PDB code 3CHP, the zinc ion is bound by the



**Figure 7: Crystal structure of LTA4H, PDB code 3CHP.** (A) Crystal structure of LTA4H with bound ligand colored in green. Shown are the catalytic zinc ion and the amino acids His295, His299 and Glu318. (B) Binding pocket showing the location of the Zn<sup>2+</sup> ion (cyan sphere) as well as hydrophilic surface areas (purple), neutral areas (white) and hydrophobic areas (green).

amino acids His295, His299 and Glu318 (**Figure 7 B**). Haeggström et al.<sup>61</sup> were able to show that the zinc ion is essential for the LTA4H functionality. Besides the described hydrolase function, LTA4H also shows a peptidase function which is located left to zinc ion in **Figure 7 B**. This was first proposed by Haeggström et al. due to the sequence



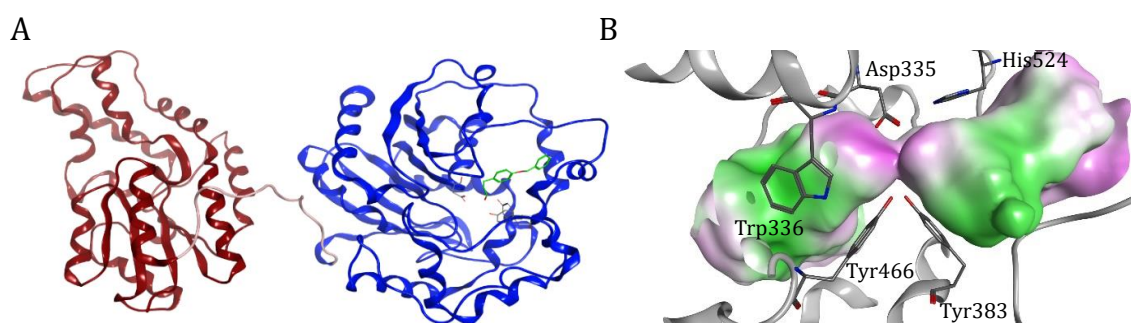
**Figure 8: Transformation from LTA<sub>4</sub> into LTB<sub>4</sub>.** LTA4H transforms LTA<sub>4</sub> into LTB<sub>4</sub>.

similarity to certain aminopeptidases.<sup>61</sup> Snelgrove et al. identified the tripeptide and neutrophil chemoattractant Pro-Gly-Pro (PGP) as the physiological substrate. PGP, which is a biomarker for COPD (chronic obstructive pulmonary disease), is hydrolyzed by extracellular LTA4H. This degradation promotes the resolution of inflammation in the lung.<sup>65</sup> Therefore, inhibition of the peptidase functionality of LTA4H has a negative effect on inflammation. On the other hand, inhibition of the hydrolase function has anti-

inflammatory effects. In the treatment of inflammation, the inhibition of LTA4H (preferably only the hydrolase functionality) is a promising strategy.

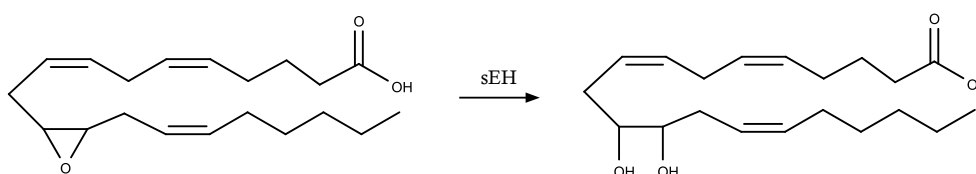
#### 1.4.2 Soluble epoxide hydrolase

The human soluble epoxide hydrolase (sEH) is a bifunctional homodimeric enzyme with a hydrolase and phosphatase function. Each monomer consists of two structural domains linked by a proline-rich segment (**Figure 9 A**). sEH can be found in the cytosol and



**Figure 9: Crystal structure of sEH, PDB code 4Y2T.** (A) Crystal structure of sEH with bound ligand colored in green. The phosphatase function is colored in red, the proline-rich linker is colored in light red and the hydrolase function is colored in blue. (B) Binding pocket showing the hydrophilic surface areas (purple), neutral areas (white) and hydrophobic areas (green).

peroxisomes and preferably hydrates aliphatic epoxides and fatty acid epoxides.<sup>66,67</sup> The catalytic activity is promoted by two tyrosine and an aspartic acid residue (Tyr383, Tyr466 and Asp335) (**Figure 9 B**). Epoxyeicosatrienoic acids (EETs) are catalytically hydrated into dihydroxyeicosatrienoic acids (DHETs) by sEH (**Figure 10**). The inhibition of sEH and therefore the increased levels of EETs have a positive biological effect and can be used in the treatment of diabetes, pain and inflammation.<sup>66</sup>



**Figure 10: Hydration of EETs into DHETs by the sEH.** Shown is the catalytic hydration of an exemplary EET into its corresponding DHET.

## 1.5 Multitarget drug design

Therapeutic drugs not only interact with their target of interest but show interactions to a variety of proteins leading to a multitarget activity profile. Those multitarget activities are mostly unintentional and can lead to negative side effects and risks. Although the unintentional interactions can be used as a chance to contribute to the overall efficacy of a drug if acknowledged within the drug design process. Complementary signaling pathways or enzymatic cascades can be modulated and increase drug effectiveness.<sup>68-70</sup>

Modern drug discovery deliberately designs small molecules with a multitarget activity profile. A beneficial multitarget activity profile is derived from additive or synergistic efficacy by modulating complementary signaling pathways or enzymatic cascades. Diseases of multifactorial nature are predestined to be targeted in a multitarget drug design. Multifactorial disorders are associated with the effect of multiple enzymes in combination with lifestyle and environmental factors. Metabolic syndrome, psychiatric or degenerative CNS disorders, infectious diseases and cancer are targeted with a multitarget approach in drug design projects. These multifunctional diseases require a treatment with a cocktail of multiple drugs due to their multifunctional nature.<sup>71-76</sup>

The classical way of therapy is to administer several target specific drugs to achieve the desired effect. Co-administration of two drugs raise safety concerns and challenges in optimizing dosage. It cannot be assumed, that two safe drugs also have a good safety profile in co-administration. Intensive safety studies, dosage ranging investigation and drug-drug interaction analysis have to be performed, which rises cost and complexity of developing combination therapy drugs.<sup>72</sup> Well-designed multitarget drugs show a range of advantages, which makes multitarget drugs a worthy field of study. Replacing several drugs with a multitarget drug reduces treatment complexity, drug side effects, pharmacokinetic complexity and drug-drug interactions. Therapeutic efficacy can be increased due to synergetic effects leading to lower drug dosages. In addition, the economic advantage of less clinical trials for a multitarget drug compared to multiple specific drugs speaks for the promising field of multitarget drug design.<sup>71-76</sup>

In this work we will demonstrate how to predict multitarget drugs. The method is applied to the drug targets LTA4H and sEH. Those enzymes are, as described above, located in the arachidonic acid cascade ([section 1.4](#)). A single-target drug can have a complex effect on

a disease network like the AA cascade (**Figure 6**). Studies have shown that the inhibition of a single pathway may lead to shunting of AA metabolites into another untargeted pathways within the AA cascade. This shunting diminishes the beneficial effects of inhibiting one pathway in the cascade.<sup>71-75,77</sup> In addition, a complex communication between different metabolizing pathways (crosstalk) could be shown.<sup>72</sup> Single-target drugs may not only lack efficacy due to this crosstalk but also show safety concerns due to unexpected target-drug interactions within the AA cascade.<sup>77</sup> Overcoming the shunting problem with a combination therapy of two drugs (combining two drugs in one single tablet) raises safety concerns (safety profile of drugs may be altered in combination therapy).<sup>78,79</sup> The design of multitarget drugs can help to solve these challenges and aims for the following improvements:

- Enhancing drug efficacy
- Improving drug safety
- Increasing patient compliance
- Eliminating drug-drug interactions

Using multitarget drugs in the AA cascade can overcome the shunting problem and improve the treatment of inflammation.<sup>71-75,77</sup>

## 1.6 Aim of the work

In this work, we describe a new method for *in silico* design of dual target compounds by combining molecular fingerprints with state-of-the-art machine learning algorithms.

Representing molecular structures using a fingerprint scheme has been done for many years (e.g. Morgan fingerprint from 1965<sup>26,29-31</sup>). Classical 2D-fingerprints are mainly used for fast similarity search in large databases. There are different types of 2D-fingerprints, which are used to describe molecular structures ([section 2.3](#)). The basic bit string scheme is the most simplified fingerprint scheme. A bit string is an array mapping the presence of a feature domain to the values 0 and 1. These values can be interpreted as on/off, valid/invalid, absent/present etc. Since there are only two possible values, they can be stored in one bit. 2D-fingerprints were developed further to include the available 3D-information of protein-ligand complexes. Those 3D-fingerprints represent the 3D-

features of molecular structures. The fundamental assumption, that there is an underlying relationship between the molecular structure and its bioactivity is the basic concept of 2D- and 3D-fingerprint similarity search.

Predictive modeling generates data-based models and extrapolates onto new and unknown data. A variety of well-established machine learning algorithms is used in this work to investigate their applicability in the process of drug design and development.

The goal is not simply to predict novel compounds, which inhibit one protein, but to predict compounds, which inhibit two different proteins at the same time. The proteins targeted in this work (LTA4H and sEH) are located in the arachidonic acid (AA) cascade and are associated with several inflammatory diseases (for example asthma, rheumatoid arthritis, dermatitis and atherosclerosis).<sup>63</sup> The AA cascade displays a strong cross-talk between the metabolic pathways. Inhibiting only one pathway may lead to AA degradation in another pathway (shunting). This shunting diminishes the beneficial effects of the administered drug. Targeting two different pathways with a single drug can overcome this phenomenon and lowers the risks of unexpected side effects or drug-drug interactions resulting from administering two different drugs.

First part of this work contains the compilation of data sets containing known-active compounds, inactive compounds and newly designed compounds from a combinatorial library. In the second part, the compiled compound data sets are represented using different fingerprints (2D and 3D). In the third part, those fingerprints are used with a variety of machine learning algorithms to predict novel active compounds against two different targets (dual active compounds). Prospective evaluation of the new method is established by synthesis and determination of the biological activity of cherry-picked compound candidates.



## 2 Methodology

### 2.1 Docking software

There are many different docking software on the market, aiming at prediction of the ligand conformation in the binding site, which correspond to the experimentally determined binding mode. This pose is often the global minimum energy structure. The optimization problem of finding the global minimum energy structure is defined by an objective function and the search space. The objective function  $f$  is called “scoring function”. Scoring functions estimate the binding energy between the ligand and the receptor. The search space is defined by the degrees of freedom of the ligand ( $rl$ ) and the receptor ( $rp$ ). In most approaches solving the optimization problem, the receptor structure is kept rigid. Therefore, to find the optimal solution, the optimal orientation and position of the ligand in respect to the receptor must be found by changing the ligands translation, rotation, and torsion angles of single bonds. It comprises three translational, three rotational and  $rl$  torsional degrees of freedom. The dimension of the optimization problem equals  $n = 6 + rl$ . Small changes of amino-acid sidechains in the binding site are considered by adding one torsional degree of freedom for each rotational bond in those sidechains ( $rp$ ). This extension rises the dimension of the optimization problem to  $n = 6 + rl + rp$ . If this sidechain flexibility is not specified in the docking configuration, optimization of hydrogen atom position involved in hydrogen bonding is included in the computation. This results in  $rp = rdon$  torsional degrees of freedom, where  $rdon$  is the number of rotatable hydrogen bond donor groups, like hydroxide and ammonium ions in the binding site.<sup>80</sup> The additional degrees of freedom increases the number of computations that need to be made and therefore increases computation time.

In the following, the two docking software (MOE<sup>81</sup> and PLANTS<sup>82</sup>) used in this work are described in their theory. In [section 3.1](#) the docking validation can be found and in [sections 3.3-3.5](#) the docking procedure of the different compound batches.

#### 2.1.1 Molecular Operation Environment (MOE2018.0101)

MOE's Dock application predicts poses of small molecules in complex with their respective protein target (receptor). For each ligand, several poses are generated and scored. The ligands are placed in a small region of the receptor, called the active or binding



site. The binding site has to be defined either by a co-crystallized ligand in the complex structure or by dummy atoms placed by the site finder tool in MOE.

The docking algorithm is divided into different steps:

- Conformational Analysis,
- Placement,
- Initial Scoring,
- Refinement,
- Pharmacophore Constraints, and
- Final Scoring.

The theory of all docking algorithm steps is described in the following.

**Conformational Analysis:**

Ligand conformations can be supplied via a conformation database. If ligand conformations are not available, conformations from a single 3D-conformer can be generated by bond rotation (A set of rules for dihedral angles is based on atom types and position, with a strain energy value for each bond; bond length are kept constant).

**Placement:**

A number of poses (1,000 poses, default) is generated from the conformational input using one of the available placement methods (Triangle Matcher, Alpha PMI, Alpha Triangle, Pharmacophore, Proxy Triangle<sup>83</sup>). In this work, only the Triangle Matcher was used.

*Triangle Matcher* (default): Poses are generated by aligning ligand triplets of atoms on triplets of alpha spheres in a systematic way. The concept of alpha spheres was first introduced by Liang and Edelsbrunner.<sup>84</sup> An alpha sphere is a sphere, that contacts four atoms on its borderline and contains no atom. The four atoms are equidistant from the center of the alpha sphere. Alpha sphere radii mirror the local concavity specified by the four atoms. In a protein, different alpha sphere radii represent various parts of the protein: small spheres are located inside the protein, large spheres outside of the protein and medium spheres represent clefts and cavities.<sup>85</sup>

Initial Scoring:

All poses generated by the placement method are scored, using one of the available scoring functions (London dG<sup>86</sup>, ASE<sup>87</sup>, Affinity dG, Alpha HB, Electron Density (neglected in this work), GBVI/WSA dG). Scoring functions can be classified into three different classes: force field based, knowledge based and empirical.

Force field based scoring functions are based on classical molecular mechanics. Contacts between ligand and receptor are calculated in a pairwise manner using a Lennard-Jones potential and a Coulomb term:

$$\Delta G_{bind} = \sum_{i=1}^{ligand} \sum_{j=1}^{protein} \left( \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right)$$

The parameters  $A$  and  $B$  are defined for each individual pair of different atom type combinations,  $R$  is defined as the atomic center distance,  $q$  is defined as the partial charge on each individual atom and  $\epsilon$  is the dielectric constant.<sup>88</sup>

Knowledge based scoring functions are based on statistical observation on intermolecular ligand-receptor contacts to derive a potential, that describes the observed contact distribution. The underlying assumption is, that intermolecular contacts between certain atom types, which occur more often are energetically favorable and may contribute positively to binding affinity. The derived potentials express statistical preferences, which are collected from the knowledge base of protein-ligand complexes.<sup>89</sup>

Empirical scoring functions are not derived from any physics based energetic formulation, like the other two types of scoring functions. The underlying idea is to define a function consisting of terms related to the physical processes described with  $\Delta G_{bind}$  and estimate the functions parameters based on known affinities of protein-ligand complexes. The functional form of empirical scoring functions are generally the same, but the data used for parameterization and optimization of parameters differ.<sup>88</sup>

Which scoring function in the initial scoring leads to the best poses can be found in the docking validation [section 3.1.1](#).

*London dG*: As an empirical scoring function, London dG estimates binding free energy of ligands using a set of energy and assessment terms that are generated from datasets of measured ligands:

$$\Delta G = c + E_{flex} + \sum_{h-bonds} c_{HB} f_{HB} + \sum_{m-lig} c_M f_M + \sum_{atoms\ i} \Delta D_i$$

where  $c$  accounts the average gain or loss of rotational and translational entropy;  $E_{flex}$  is the energy lost due to ligand flexibility;  $c_{HB}$  is the energy of an ideal hydrogen bond;  $f_{HB}$  measures geometric imperfections of hydrogen bonds;  $c_M$  is the energy of an ideal metal ligation;  $f_M$  measures geometric imperfections of metal ligations and  $D_i$  describes the desolvation energy of atom  $i$ .

*ASE*: As an empirical scoring function, ASE accumulates the Gaussian function of all atom-receptor pairs and atom-alpha sphere pairs of each complex.

*Affinity dG*: As an empirical scoring function, Affinity dG uses a linear function to calculate the enthalpy contribution to the binding free energy (G):

$$\Delta G = C_{hb} f_{hb} + C_{ion} f_{ion} + C_{mlig} f_{mlig} + C_{hh} f_{hh} + C_{hp} f_{hp} + C_{aa} f_{aa}$$

where the  $f$  terms fractionally count atomic contacts of specific types and the  $C$  terms are coefficients that weight the term contributions to the affinity estimate;  $hb$  describes the interaction between hydrogen bond donor-acceptor pairs;  $ion$  accounts ionic interactions;  $mlig$  accounts metal ligation;  $hh$  describes hydrophobic interactions;  $hp$  describes interactions between hydrophobic and polar atoms;  $aa$  accounts interactions between any two atoms not described by any other term.

*Alpha HB*: As an empirical scoring function, Alpha HB is a linear combination of two terms. The first term measures the geometric fit of the ligand to the binding site. The second term measures hydrogen bond effects. A sum over all ligand atoms results in the final score.

*GBVI/WSA dG*: As a force field scoring function, GBVI/WSA dG estimates the free binding energy with weighted terms for the Coulomb energy, solvation energy and van-der-Waals contributions. GBVI/WSA dG was trained using the MMFF94x<sup>90</sup> and AMBER99<sup>91,92</sup> force field on the 99 protein-ligand complexes of the SIE training set.<sup>93</sup> The functional form of GBVI/WSA dG is a sum of terms:

$$\Delta G \approx c + \alpha \left[ \frac{2}{3} (\Delta E_{Coul} + \Delta E_{sol}) + \Delta E_{vdw} + \beta \Delta SA_{weighted} \right]$$

where  $c$  accounts the average gain or loss of rotational and translational entropy;  $\alpha$  and  $\beta$  are constants which were determined during training;  $E_{coul}$  is the coulombic electrostatic term;  $E_{sol}$  is the solvation electrostatic term;  $E_{vdw}$  is the van der Waals contribution to binding;  $SA_{weighted}$  is the surface area, weighted by exposure.

Refinement:

The Induced Fit method<sup>94</sup> or the Rigid Receptor method can further refine the poses generated by the placement method. Induced Fit adds additional computation steps to recompute conformations of flexible protein sidechains in the refinement.

By default, backbone atoms are held fixed during refinement. With rigid receptor, sidechains are also constrained. Solvation effects are calculated using the electrostatic energy term ( $E_{ele}$ , consists of a Coulomb term, a distance dependent dielectric term and a reaction field term) and a dielectric constant of 4.<sup>95</sup> The final energy is evaluated using the Generalized Born solvation model (GB/VI).<sup>96</sup>

Final Scoring:

One of the available scoring functions (London dG, ASE, Affinity dG, Alpha HB, Electron Density (neglected in this work), GBVI/WSA dG) is used to score all remaining poses. All poses are ranked according to their final score and a defined number of poses is selected for output.

All steps of the MOE docking algorithm were carried out in a KINME workflow to validate MOE docking. All possible combinations of scoring functions (Electron Density was neglected in this work) in the initial and final scoring step were analyzed. The description of the MOE docking validation can be found in [section 3.1.1](#).

### 2.1.2 PLANTS

The Protein-Ligand ANT System (PLANTS)<sup>97</sup> was developed as the first Ant colony optimization (ACO) algorithm for predicting the pose of a ligand in its receptor. The ACO is inspired by the behavior of real ants finding a shortest path between their nest and a food source. In nature, ants use pheromone trails for communicating the path between the nest and the food source. In PLANTS, an artificial ant colony is used to find a minimum energy conformation of a ligand in its binding site. These ants mimic the behavior of real ants and mark the low energy ligand conformation with a pheromone trail.

In PLANTS, the ligand flexibility is treated with  $6 + r$  degrees of freedom, which was described in [section 2.1](#). Protein flexibility is only partially considered by allowing the optimization of hydrogen atom positions that could be involved in hydrogen bonding. The search space in PLANTS is defined by the binding site size and the ligand's translational degree of freedom. Each degree of freedom is associated with a pheromone vector. Each pheromone vector of rotational and torsional degrees of freedom has 360 entries, resulting from an interval of  $1^\circ$ . The number of entries of the pheromone vector of translational degree of freedom depends on the size of the binding site. PLANTS is based on the MAX-MIN ant system, where only the best ants add to the pheromone trails and the maximum and minimum values of the pheromone are explicitly limited.

After searching the complete search space, all generated conformations (solutions) are post-processed. The set of all solutions are first sorted by increasing scoring function values. A specific number of ligand conformations (in this work 5 conformations) is selected, such that the minimal root mean square deviation (RMSD) between any of the selected conformations is larger than  $2 \text{ \AA}$ . These solutions can be rescored with more advanced and computationally more demanding scoring functions to increase the chance of finding the best ligand conformation. The best ligand conformation is defined as the conformation most similar to the experimentally determined binding mode.<sup>80,97</sup>

Three empirical scoring functions are implemented in PLANTS: a modified piecewise linear potential (PLP)<sup>98</sup> version (as well as a modified version PLP95) and a scoring function combining parts of already published scoring functions (CHEMPLP)<sup>99,100</sup>. The scoring function PLP ( $f_{PLP}$ ) uses distance-based potential and is based on scoring functions described in literature.<sup>98,101</sup>

$$f_{PLP} = f_{plp} + f_{clash-lig} + f_{tors-lig} + 0.3 * f_{score-prot} - 20.0$$

$f_{plp}$  describes steric interactions between the ligand and the receptor. Metal ions in the binding site are considered as well as the occlusion of polar atoms by nonpolar ones by distance-based potentials.  $f_{clash-lig}$  describes a simple clash term, which avoids ligand atoms to come too close to each other.  $f_{tors-lig}$  describes a torsional potential.<sup>100</sup>  $f_{score-prot}$  describes intramolecular protein-interactions, the same distance-based potential as in PLP, with an additional intra-side-chain clash term.<sup>97</sup>

The scoring function CHEMPLP has the following functional form:

$$f_{CHEMPLP} = f_{plp} + f_{chem-hb} + f_{tors-lig} + f_{clash-lig} + 0.3 * f_{score-prot} - 20.0$$

$f_{plp}$  describes the same steric interactions as characterized above, although with different parameter settings.  $f_{chem-hb}$  describes hydrogen bonding and metal-acceptor interaction between ligand and protein as implemented in GOLDS CHEMSCORE.<sup>102</sup>  $f_{tors-lig}$  and  $f_{score-prot}$  are the same as characterized above. In both scoring functions a penalty term (-20.0) is added if the ligands reference point falls outside the predefined binding site.<sup>80</sup> Excluding these ligands prevents the algorithm from finding random solutions that don't lead to active compounds, which don't interact with the binding site.

## 2.2 Compound preparation

Collecting a large set of compounds is required for constructing and training a machine learning model. For a predictive model, the more input data one can provide the better. Since machine learning models will be used for classification, data belonging to both classes (active and inactive) had to be collected. The first batch of active compounds was collected from PDB<sup>12</sup> (crystalized ligands). The second batch of active compounds were downloaded from ChEMBL database<sup>103</sup> (LTA4H ChEMBL ID: CHEMBL4618, SEH: ChEMBL ID: CHEMBL2409). In addition, a randomly picked third batch of inactive compounds came from the ChEMBL23 data set. The first, second and third batch of compounds form the basis for training machine learning models (section 3.6). Compounds, for which a prediction of activity will be done with the trained models, were generated using a combinatorial library (referred to as batch number four). In the following section all relevant sets are described, **Figure 11** summarizes the compilation of all four compound sets.

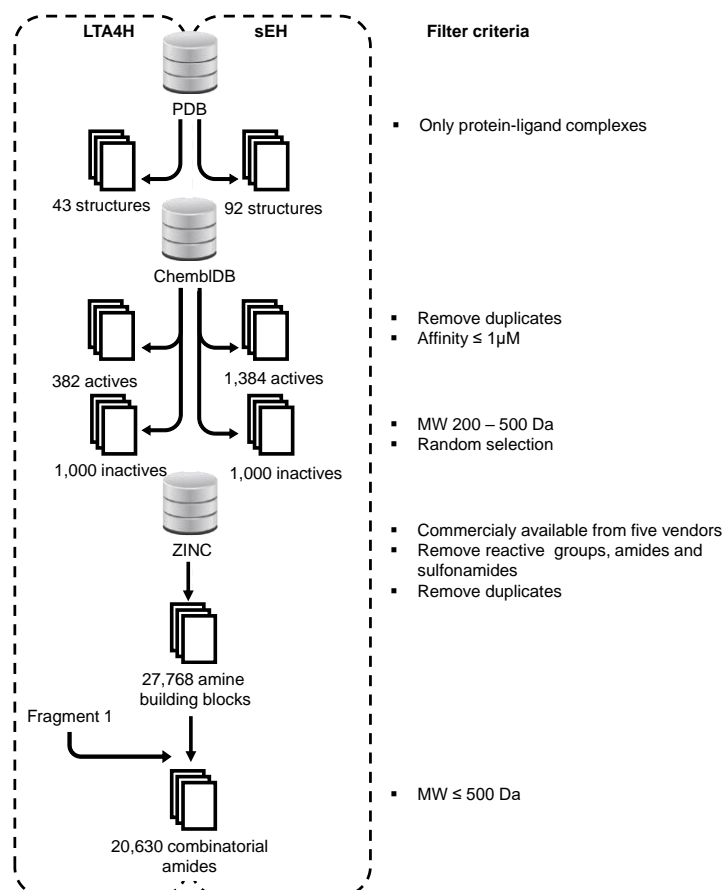


Figure 11: Compilation of data sets.

### 2.2.1 Crystallized ligands

All available complex structures for LTA4H and sEH were collected from PDB (Figure 12). The Protonate 3D node (provided by Chemical Computing Group Inc.)<sup>81</sup> was used to add hydrogen atoms to the complex structures. Complex structures for LTA4H and sEH were superposed in the MOE GUI (section 3.1.1). The ligands were separated from their receptor using the Complex Splitter node (provided by Chemical Computing Group Inc.)<sup>81</sup> in KNIME.<sup>104</sup> Hydrogen atoms were added to the ligands using the Wash node (provided by Chemical Computing Group Inc.; settings: remove lone pairs, deprotonate strong acids, protonate strong bases, add hydrogens)<sup>81</sup>. The 3D-conformation was kept constant since



Figure 12: KNIME workflow for crystallized ligand preparation.

it is presumed to be the active conformation. The final data set of crystalized ligands contains 43 LTA4H crystalized ligands and 94 sEH crystalized ligands.

### 2.2.2 Active ChEMBL compounds

The second batch of active compounds were collected from the ChEMBL database. ChEMBL database provides an open access collection of large-scale bioactivities of small molecules.<sup>103</sup> All compounds tested on our two targets were downloaded in SMILES format<sup>105</sup> with corresponding binding affinity. For LTA4H 1,022 compounds (status 2017-

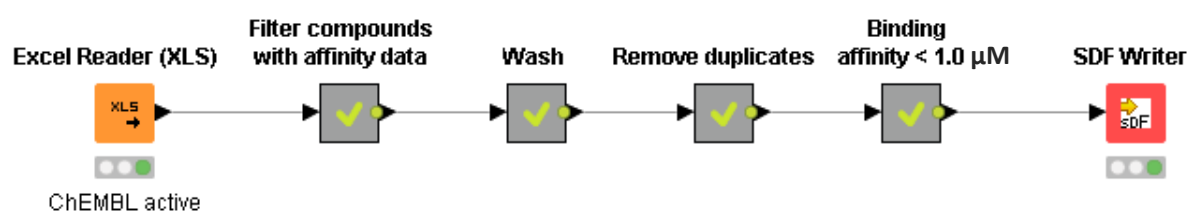


Figure 13: KNIME workflow for active ChEMBL compound preparation.

08-10) and 2,453 sEH compounds (status 2017-08-15) were downloaded from source (Figure 13). SMILES format was translated into sdf format. Hydrogens were added using the Wash node (provided by Chemical Computing Group Inc.; settings: remove lone pairs, deprotonate strong acids, remove minor components, protonate strong bases, add hydrogens)<sup>81</sup>. Duplicates of compounds were filtered, leaving 546 LTA4H compounds and 2,012 sEH compounds. As true active only compounds with a binding affinity smaller than 1.0  $\mu\text{M}$  were selected (LTA4H: 382 compounds, sEH: 1,384 compounds). To generate potential active poses of those compounds, molecular docking was conducted (sections 3.3 and 3.5).

### 2.2.3 Inactive ChEMBL compounds

The ChEMBL23 data set built the foundation of inactive compounds, which are required for the classification data set. The entire ChEMBL23 data set, containing 1,727,112 molecules (status 2017-05-30), was downloaded into KNIME (Figure 14). A molecule property filter was applied to restrict the molecular weight between 200 g/mol and 500 g/mol, comparable to the active compounds. Hydrogens were added using the Wash node (provided by Chemical Computing Group Inc.)<sup>81</sup>. From the filtered 1,371,762 molecules, 1,000 were randomly selected using the Row Sampling node (provided by KNIME AG).<sup>104</sup> 3D-conformations were generated using the docking procedure described



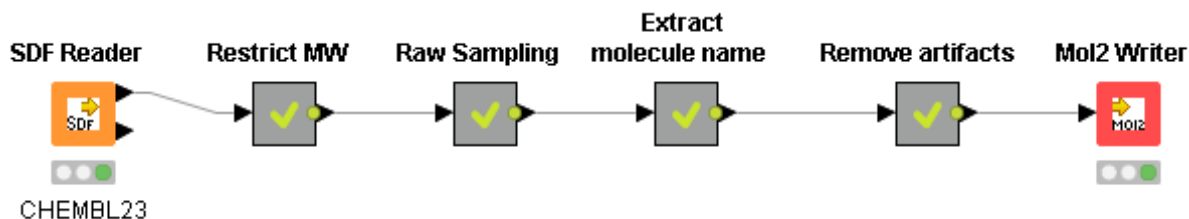


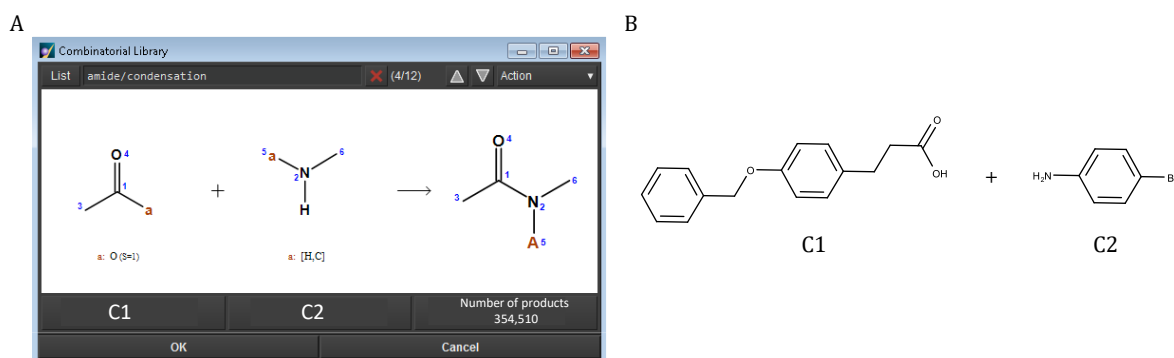
Figure 14: KNIME workflow for inactive ChEMBL compound preparation.

in section 3.3. 4,768 poses were generated for LTA4H, for sEH, 4,792 poses were generated.

#### 2.2.4 Combinatorial library

In general, a combinatorial library is the result of combinatorial chemistry, where a chemical reaction generates many related compounds.<sup>106</sup> These libraries can be generated via chemical synthesis or as in this case virtually by computer software.

We designed a combinatorial library specially to identify new dual active compounds for LTA4H and sEH. The key fragment used as the constant reaction partner was 3-[4-(benzyloxy)phenyl]propionic acid **C1**. The corresponding alcohol (3-[4-(benzyloxy)phenyl]propanol) was initially identified by Amano et al.<sup>107</sup> as a fragment which binds to sEH and exhibits moderate potency and ligand efficacy. Using this fragment, the combinatorial library was focused on generating possible lead compounds inhibiting the targets of interest LTA4H and sEH. As reaction partners, amine building blocks were downloaded from ZINC database.<sup>108</sup> From six vendors (Acros, Alfa-Aesar, Apollo Scientific, Fluorochem, Sigma Aldrich, TCI) purchasable amine building blocks were downloaded (status 2017-10-13). 354,510 fragments in SMILE format were collected from the libraries and loaded into KNIME. After that, several filtering steps were performed. First, compounds with potential reactive groups were excluded (metals, phosphor-, N/O/S-N/O/S single bonds, thiols, acyl halide, Michael Acceptors, azides and esters)<sup>81</sup> using the *MOE Descriptors* node (provided by Chemical Computing Group Inc.)<sup>81</sup>, reducing the number of building blocks to 242,305. Amides and sulfonamides were filtered out, as well as duplicate structures reducing the number to 27,768 building blocks. The library was generated using the Combinatorial Library application in the MOE GUI. The reaction amide/condensation was selected. For compound **1 (C1)** the key fragment 3-[4-(Benzyloxy)phenyl]propionic acid was selected. For reaction partner compound **2 (C2)** the amine building blocks were set (**Figure 15**). The resulting



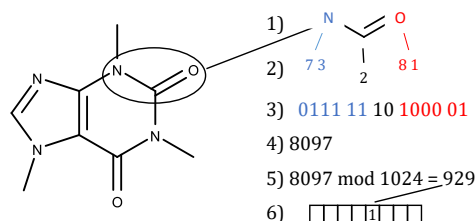
**Figure 15: Condensation reaction between C1 and C2.** (A) MOE GUI for building the combinatorial library. (B) C1, key fragment 3-[4-(benzyloxy)phenyl]propionic acid. C2, exemplary amine building block.

combinatorial library contained 25,153 compounds. A molecular weight filter was applied to limit the molecular weight to smaller than 500 g/mol. Tertiary amides, which were generated through the condensation reaction, were excluded. The final combinatorial library contains 20,630 related but unique compounds.

### 2.3 2D-fingerprints

2D-fingerprints were originally developed for similarity search in large chemical libraries. The 2D-fingerprint is an abstraction of the ligand whereby information is lost, however it allows it to make compounds comparable.<sup>109</sup> Information coded into 2D-fingerprints are extracted from atom and bond types, and graph distances derived from chemical graphs. This information is stored as bits in a Bit-String that serves as a fingerprint scheme. Each bit in the fingerprint corresponds to a chemical property, for example a substructure feature. If the feature is present in the chemical structure, the corresponding bit is set on (=1). Based on fingerprint similarity (Tanimoto coefficient<sup>110</sup>) to a biologically active compound, those fingerprints can be used to identify new active compounds or more potent compounds to a target of interest.<sup>32,111,112</sup> In this work, four different 2D-fingerprints (AtomPair<sup>113</sup>, FeatMorgan<sup>29</sup>, Morgan<sup>114</sup> and MACCS<sup>28</sup>) have been used to describe the compounds in the assembled data sets found in [section 2.2](#).

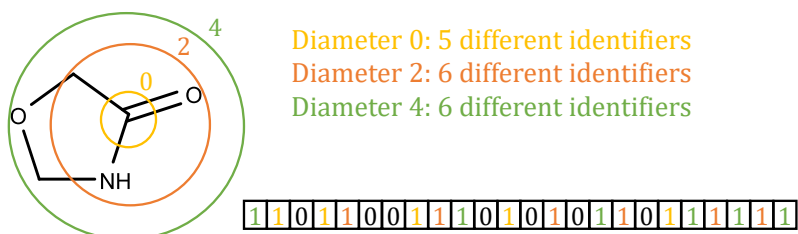
The AtomPair fingerprint (RDKit) describes two atoms with atom descriptions and the distance between the two atoms. The atom description includes its chemical atom type, the number of non-hydrogen atoms and the number of bonding  $\pi$  electrons (**Figure 16**).<sup>26,113,115</sup> The following steps are conducted for every pair of heavy atoms



**Figure 16: General construction of the AtomPair fingerprint.** (Figure after Jelínek et al.)<sup>115</sup>

when creating an AtomPair fingerprint: 1) Extraction of given pair of heavy atoms; 2) encoding of descriptors, atom type, number of bonds for both atoms and their topological distance; 3) encoded descriptors are converted into a bit string; 4) bit string is concatenated into one number; 5) The number is hashed into the index space; 6) bit in the corresponding fingerprint scheme is set on (=1).

The Morgan fingerprint (RDKit) is an ECFP-like circular fingerprint (Extended-connectivity fingerprints). Structures are represented by assigning numbers to heavy atoms combining several connectivity features (element type, number of heavy atoms, number of hydrogens, charge etc.). Those substructure features are translated into a

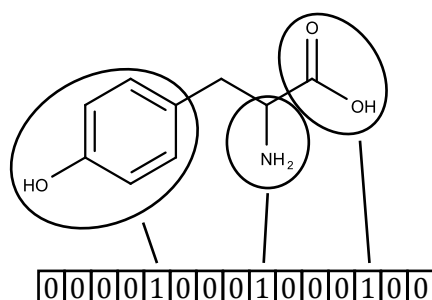


**Figure 17: General construction of the Morgan fingerprint.** (Figure after Rogers et al.)<sup>29</sup>

fingerprint scheme (**Figure 17**).<sup>26,29-31</sup> First, every heavy atom is assigned a hashed integer identifier register various atom properties. Second, iteration captures the circular neighborhood around each heavy atom and is encoded into a hashed integer number. The diameter specifies the maximum diameter of the circular neighborhood considered for every heavy atom of the molecular structure.

The FeatMorgan fingerprint (RDKit) is a FCFP-like (functional-class fingerprint) circular fingerprint based on the Morgan algorithm. The FCFP is an abstraction of the ECFP fingerprint, where atom identifiers are a set of pharmacophoric identifiers (hydrogen-bond acceptor and donor, negatively and positively ionizable, aromatic, halogen).<sup>26,29</sup>

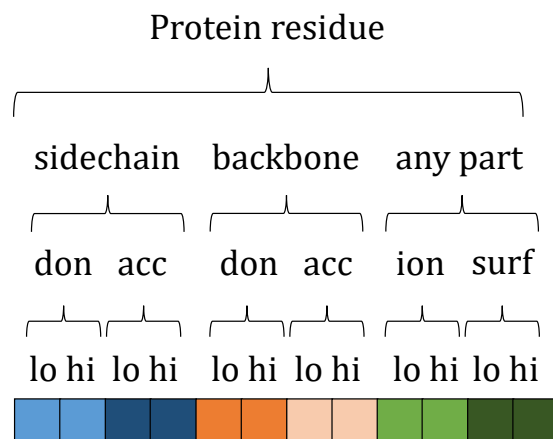
The MACCS fingerprint (Molecular ACCESS System) (RDKit) is translated into SMARTS pattern, corresponding to 166 MACCS keys describing possible substructures. The SMARTS patterns are used to describe a molecular structure in a fingerprint scheme (Figure 18).<sup>26-28</sup>



**Figure 18: General construction of the MACCS fingerprint.** A hypothetical 15-bit fingerprint representing a MACCS fingerprint. Three bits are set on (=1) because the substructures they represent are present in the hypothetical molecular structure. (Figure after Cereto-Massagué et al.<sup>28</sup>)

## 2.4 3D-fingerprints

3D-fingerprints are derived from 2D-fingerprints and include three-dimensional information (pharmacophoric patterns, surface properties, molecular volumes or molecular interaction fields) with the purpose to better represent the actual 3D-conformation. In this work, the 3D-fingerprint Protein-Ligand Interaction Fingerprint (PLIF)<sup>116</sup> was used to represent the compounds in the assembled data sets found in section 2.2. The PLIF from MOE2018.0101 summarizes the interaction between ligands and receptors using a fingerprint scheme. Ten potential contacts (sidechain hydrogen bonds (donor or acceptor), backbone hydrogen bonds (donor or acceptor), solvent hydrogen bonds (donor or acceptor), ionic attraction, metal ligation and arene contacts) and surface contacts) are integrated in the current PLIF version (Figure 19). The algorithm analyses the compound poses in the receptor structure and records all occurring interactions. Each amino acid residue involved in any of the collected interactions is classified into categories by following a general scheme, which is shown in Figure 19. If a fingerprint for a specific pose has less than 50 bits set all interactions of this pose are included in the fingerprint. If the aforementioned fingerprint exceeds 50 bits



**Figure 19: General scheme of the Protein-Ligand Interaction Fingerprint.** The scheme shows sample interactions taken from the MOE2018 manual.

interactions will only make it into the fingerprint, if at least two compounds in the docking show that interaction. This restriction is applied to emphasize bits belonging to a common binding mode. For each interaction type a low and high bit can be set. The minimum thresholds for low and high bits are by default 0.5 kcal/mol and 1.5 kcal/mol, respectively. If the occurring interaction outreaches the low bit threshold, the bit is set (1). If the occurring interaction also outreaches the high bit threshold, the bit is also set (1). This results in possible bit patterns of 00 (neither low or high threshold is reached), 10 (only low threshold is reached), or 11 (both low and high thresholds are reached), respectively.<sup>81,117</sup>

## 2.5 Fingerprint calculation

The five fingerprints described in [sections 2.3 and 2.4](#) were used to represent the molecular structure of the compounds described in [section 2.2](#). Calculation of the fingerprints was carried out in a KNIME workflow (version KNIME 3.6.1<sup>104</sup>).

The four compound data sets include the crystalized ligands (43 LTA4H ligands, 94 sEH ligands), active ChEMBL compounds (382 LTA4H compounds, 1,384 sEH compounds), inactive ChEMBL compounds (1,000 compounds) and the combinatorial library (20,631 compounds). All ligands and compounds were combined into one data set for LTA4H (22,056 compounds) and one data set for sEH (23,107 compounds) as preparation for the fingerprint calculation.

The *RDKit Fingerprint* node was used to calculate the AtomPair, Morgan, FeatMorgan and MACCS fingerprint. The following settings were set for the different types of fingerprints. For the AtomPair fingerprint, the fixed-length bit string was set to 1,024, which is the standard bit string length. The maximum path length (distance between atoms) for an atom pair was set to default 30. For the Morgan and FeatMorgan fingerprint, the bit string length was set to 1,024, with a radius of 2. For the MACCS fingerprint, the RDKit implementation of the 166 public MACCS keys were used. Afterwards, the bit string fingerprints were expanded into individual columns with one column for each bit in the bit string (*Fingerprints Expander node* provided by Erlwood Cheminformatics). This step was necessary for the following predictions of new active compounds using machine learning models ([sections 3.7-3.9](#)).

The *PLIF Scores* node (provided by Chemical Computing Group Inc.)<sup>81</sup> was used to calculate the PLIF fingerprint. Default settings were used for calculation. The minimum thresholds for low and high bits are the following:

- Sidechain H-bond donor/acceptor: low: 0.5 kcal/mol, high: 1.5 kcal/mol
- Backbone H-bond donor/acceptor: low: 0.5 kcal/mol, high: 1.5 kcal/mol
- Solvent H-bond donor/acceptor: low: 0.5 kcal/mol, high: 1.5 kcal/mol
- Ionic attraction: low: 0.5 kcal/mol, high: 1.5 kcal/mol
- Surface contacts: low: 20 Å<sup>2</sup>, high: 50 Å<sup>2</sup>
- Metal ligation: low: 0.5 kcal/mol, high: 3.5 kcal/mol
- Arene contacts: low: 0.5 kcal/mol, high: 1.5 kcal/mol

The bit length is set to a maximum of 250 bits, if this number is exceeded following bits will be discarded.

The PLIF fingerprint requires 3D-conformations of all compounds. Except for co-crystallized ligands, docking had to be performed to generate docking poses of all compounds. The docking procedure is described in [sections 3.3-3.5](#). Resulting from this procedure five poses for each compound were generated. PLIF similarity between the crystallized ligands and the compound poses was used to select one pose per compound. The pose with the highest similarity to any of the crystallized ligands was chosen. Noise in the following machine learning step was reduced, since potentially incorrect docking poses were eliminated.

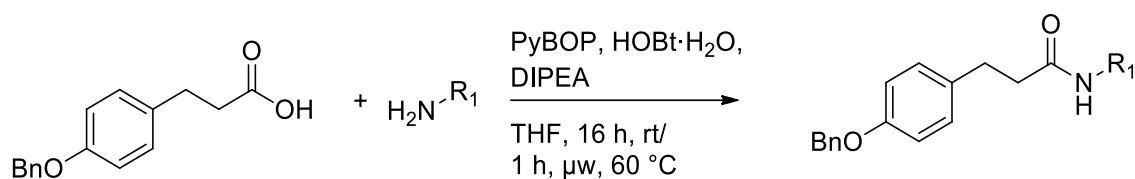
## 2.6 Machine learning prediction

For each target of interest (LTA4H and sEH), fingerprint (PLIF, AtomPair, FeatMorgan, Morgan and MACCS) and machine learning algorithm (SVC, RF, XGB and ADA) a prediction model was built. Optimized parameters and partitioning schemes (splitting data set into training and test set) were used for model building. Parameters were optimized to achieve best possible prediction accuracies. A description of the optimization process can be found in [section 3.6](#). The optimized models were used to predict the class labels (active or inactive) of the combinatorial library compounds containing the five calculated fingerprints (2D and 3D). All optimizations and predictions were made in Jupyter notebook<sup>118</sup> using the scikit-learn python package.<sup>119</sup> Jupyter notebook is an open source web application to create code, equations and visualizations. Scikit-learn is an open source machine learning library for the python programming language. The library contains applications for supervised and unsupervised learning, model selection and evaluation, visualization and many more in the context of machine learning. The code used in this work for all optimization steps and the final predictions can be found in the appendix [section 7.1 - 7.3](#).

## 2.7 General synthesis route

The results of the machine learning predictions are data sets of compounds, which are predicted to be active on the targets of interest using different fingerprints and machine learning algorithms. For each molecular representation type (2D and 3D) one fingerprint with one machine learning method was selected. From those compounds predicted to be active ([section 3.7 and 3.9](#)), a selection of compounds was cherry picked for synthesis to validate the method developed in this work. Feasibility of synthesis, estimated solubility and uniqueness of the compounds (amongst the selected compounds and compared to known inhibitors) were used as guidelines for cherry picking. The synthesis of the compounds was conducted by Kerstin Hiesinger. The reagent 3-(4-(benzyloxy)phenyl)propionic acid was synthesized by Felix Zhu. Three different procedures form the base of the synthesis of the 14 selected compounds ([sections 3.7 and 3.9](#)).

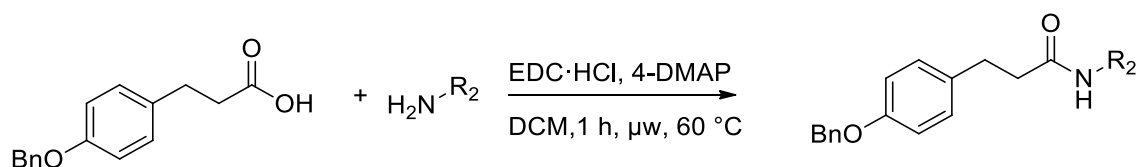
#### Procedure A:



**Scheme 1: General synthesis route of procedure A.**

1.1 eq 3-(4-(benzyloxy)phenyl)propionic acid, 1.1 eq PyBOP (benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate) as the coupling reagent, 0.5-1.1 eq HOBT·H<sub>2</sub>O (hydroxy benzotriazole) to improve efficiency of the synthesis and 1.0 eq corresponding amine were dissolved in absolute THF (Tetrahydrofuran). Further, 1.5-3.0 eq DIPEA (N,N-diisopropylethylamine) was added and the mixture stirred either for 16 hours at room temperature or 1 hour at 60 °C under microwave irradiation (**Scheme 1**). After solvent removal, the residue was dissolved in ethyl acetate and was washed three times with demineralized water and one time with brine. The organic phase was dried over MgSO<sub>4</sub> (magnesium sulfate) and filtered. Under reduced pressure the solvent was removed, and the obtained oil was purified via column chromatography. A solid was obtained as the final product.

#### Procedure B:

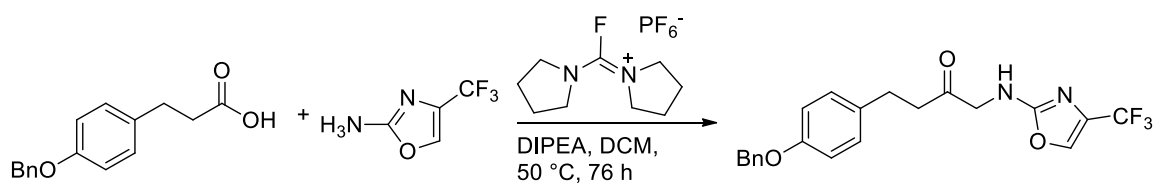


**Scheme 2: General synthesis route of procedure B.**

1.0 eq 3-(4-(benzyloxy)phenyl)propionic acid, 1.0-1.3 eq amine derivative, 1.2 eq EDC·HCl (1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide) as the coupling reagent and a catalytic amount of 4-DMAP (4-Dimethylaminopyridine) were dissolved in absolute DCM (dichloromethane). The mixture was heated to 60 °C for 1 hour under microwave irradiation (**Scheme 2**). Under reduced pressure the solvent was removed, and the residue was purified via column chromatography. The obtained solid was purified further with preparative HPLC (high-performance liquid chromatography) to gain purities over 95%.



Procedure C:



**Scheme 3: General synthesis route of procedure C.**

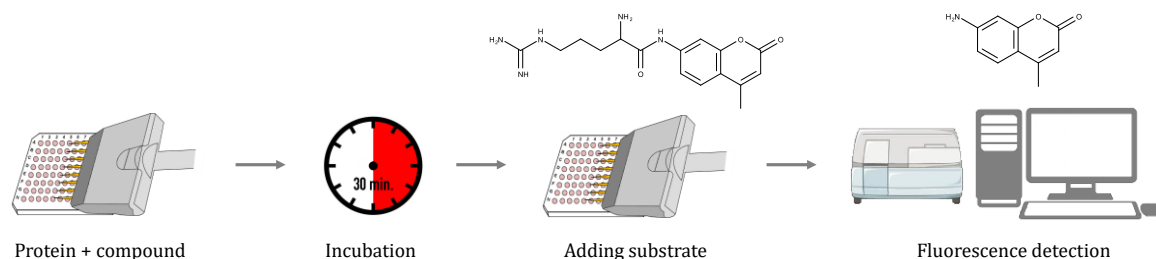
Under argon atmosphere 1.5 eq 3-(4-(benzyloxy)phenyl)propionic acid, 1.5 eq BTFFH (Fluoro-N,N,N',N'-bis(tetramethylen)formamidinium hexafluorophosphate) as the coupling reagent and 4.5 eq DIPEA (N,N-diisopropylethylamine) were dissolved in 3 mL absolute DCM (dichloromethane). The mixture was heated to 50 °C for 4 hours under microwave irradiation. Additionally, 1.0 eq 4-trifluoromethyl-oxazol-2-ylamine, dissolved in 3 mL absolute DCM, was added and heated to 50 °C for 72 hours under microwave irradiation (**Scheme 3**). The reaction mixture was diluted with ethyl acetate and washed three times with demineralized water. The precipitate, which was generated during the washing steps, was filtered. The organic phase was dried over MgSO<sub>4</sub> (magnesium sulfate), filtered and the solvent was evaporated. The residue was purified with column chromatography and further with preparative HPLC.

The results of the synthesis can be found in [section 3.10](#).

## 2.8 Fluorescence based LTA4H assay

The inhibitory activity of the synthesized compounds was tested for method validation on the target receptor LTA4H. All biochemical testing was carried out by Kerstin Hiesinger and Lilia Weizel.

Besides the identified tripeptide PGP from Snelgrove et al.<sup>65</sup>, Orning et al.<sup>120</sup> could show that LTA4H preferential hydrolyzes tripeptides with an arginine residue on the N-terminal end. This knowledge was used to design a fluorescence-based assay to determine IC<sub>50</sub> values for the synthesized compounds. The activity assay was performed according to the protocol published by Wittmann et al.<sup>121</sup> with minor modifications (**Figure 20**). The assay is based on the hydrolyzation of the non-fluorescent substrate L-Arginine-7-amido-



**Figure 20: Schematic representation of the fluorescence based LTA4H assay.** Protein and compound are incubated for 30 minutes, afterwards the substrate is added. Fluorescence is detected, inhibitory activities are determined.

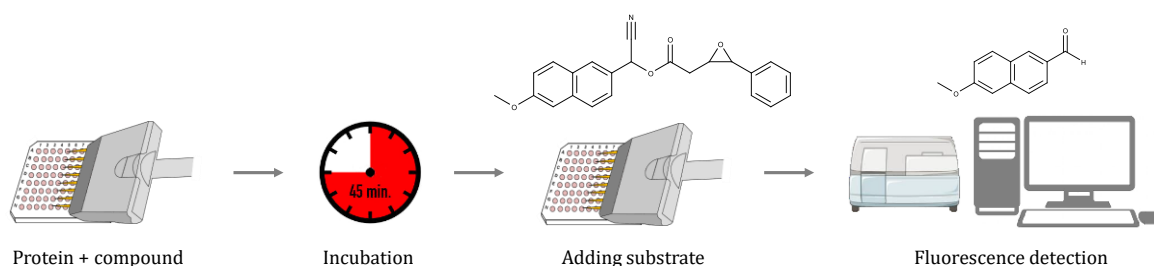
4-methylcoumarin is hydrolyzed by LTA4H into the fluorescent substrate 7-amido-4-methylcoumarin. The assay is conducted in a 96-well plate where the protein as well as the compound that is to be examined are pre-incubated. An increase of fluorescence with an extinction of 360 nm and an emission of 465 nm is detected after admitting the substrate. If the compound inhibits the protein, the fluorogenic substrate hydrolysis is blocked. Therefore, the increase in fluorescence should be lower. The hydrolysis of the substrate is detected over a period of 30 minutes. For reference a buffer control (without protein and without inhibitor) and a protein control (without inhibitor) was placed on each plate. To determine inhibitory activities of the compounds, the assay was conducted using at least five different compound concentrations. The optimal protein concentration of 0.3  $\mu\text{M}$  (50 mM Tris (tris(hydroxymethyl)aminomethane; component of buffer solution), 500 mM NaCl, pH 8 with 0.01% Triton-X 100 (detergent to prevent compound aggregation)) and 200  $\mu\text{M}$  of the substrate was used. A blank control (1% pure DMSO) as well as a positive control (1% pure DMSO with protein) was used. All measurements were performed in three independent experiments and in triplicates. Percent inhibition was calculated by referencing the slope in the linear phase of the reactions to the slopes of buffer and protein controls in MS Excel. For further fitting GraphPad Prism 7 was used (sigmoidal dose response curve fit, variable slope with 4 parameters).

Results of the activity assay can be found in [section 3.11](#).

## 2.9 sEH activity assay

The inhibitory activity of the synthesized compounds was tested for method validation on the target receptor sEH. All biochemical testing was carried out by Kerstin Hiesinger and Lilia Weizel.

Hydrolase activity of sEH was experimentally determined using an adaption of the fluorescence-based assay described by Lukin et al.<sup>122</sup> and Hahn et al.<sup>123</sup> (**Figure 21**). It is based on the hydrolysis of the non-fluorescent substrate PHOME (3-phenyl-cyano- (6-methoxy-2-naphthalenyl) methylester-2-oxiran-acetic acid) by sEH into the fluorescent substance 6-methoxy-2-naphthaldehyde. The substrate hydrolysis is monitored via fluorescence change at an extinction of 330 nm and an emission of 465 nm using a Tecan



**Figure 21: Schematic representation of the fluorescence based sEH assay.** Protein and compound are incubated for 45 minutes, afterwards the substrate PHOME is added. Fluorescence is detected, inhibitory activities are determined.

Infinite F200 Pro plate reader. A dilution series of the compounds to be tested in DMSO was generated. For each concentration 1  $\mu\text{L}$  of the compound and a mixture of recombinant human full length sEH (3 nM) in Bis-Tris buffer (pH 7) with 0.1 mg/ml BSA and 0.01% Triton-X 100 was added into the wells. The plates were incubated for 45 minutes at room temperature. The reaction was started by adding 10  $\mu\text{L}$  PHOME solution (50  $\mu\text{L}$ ) and monitored for 45 minutes. A blank control (1% pure DMSO) as well as a positive control (1% pure DMSO with protein) was used. All measurements were performed in three independent experiments and in triplicates. The termination of inhibitory activities was described above in the LTA4H activity assay ([section 2.8](#)).

Results of the activity assay can be found in [section 3.11](#).

## 3 Results and Discussion

### 3.1 Docking validation

Two docking software (MOE and PLANTS) were validated in the following section. The use of various docking software to generate docking poses is used to analyze the influence of docking poses on the way to predict novel dual active compounds. The general function for the docking software can be found in [section 2.1](#). The generation of docking poses was necessary since the 3D-fingerprint called Protein-Ligand Interaction Fingerprint (PLIF; provided in the drug discovery software package MOE 2018.0101<sup>81</sup>) requires poses of the compounds in the targets of interest. This fingerprint was used as an example of 3D-fingerprints. The results of 3D-fingerprints as a molecular representation were compared to the simpler 2D-fingerprint molecular representation.

For validation, a re-docking was conducted, where the crystallized ligands are docked into their corresponding receptors. Since the binding modes of the ligands are known, the accuracy of the docking software was evaluated based on the RMSD between the crystallized ligand and the top five ranked docking poses. A pose with an RMSD smaller than 1.0 Å was defined as accurate in this work. This is a stricter criterion for the validation of docking poses than the general standard of an RMSD of less than 2.0 Å.<sup>124,125</sup>

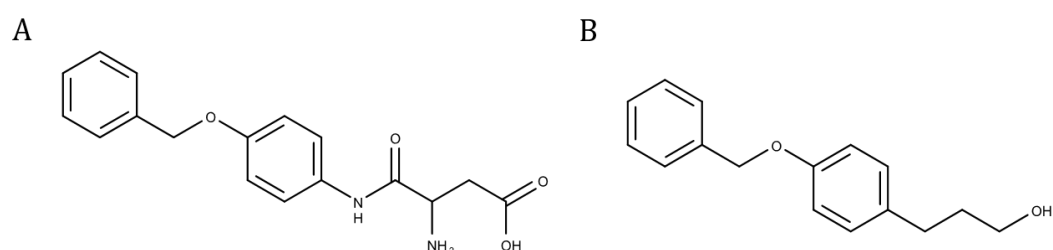
#### 3.1.1 MOE

MOE<sup>81</sup> docking validation was implemented in a KNIME workflow. The *PDB Downloader* node (provided by Vernalis)<sup>126</sup> was used to download 45 LTA4H and 94 sEH PDB structures from the Protein Data Bank<sup>12</sup> ([section 2.2.1](#)). Superposition of crystal structures was conducted with the MOE GUI. The structures were loaded into MOE as an MDB file.

First, sequence and structural alignment using default setting was conducted. In a second step, the current alignment was used to superpose all residues and ligands from the 43 LTA4H crystal structures (apo-structures with PDB code 3B7S and 3B7T were removed) and 94 sEH crystal structures. In MOE 2018 the default sequence alignment settings are the following:

Blosum62 scoring matrix was used to score the alignments between the protein sequences. Penalty parameters for “gap start” and “extend” (10 and 2 respectively) are taken from low-level group-to-group Needleman-Wunsch calculations. The alignment is built up using the tree-based method, which is an all-against-all strategy to build the initial alignment.<sup>127</sup>

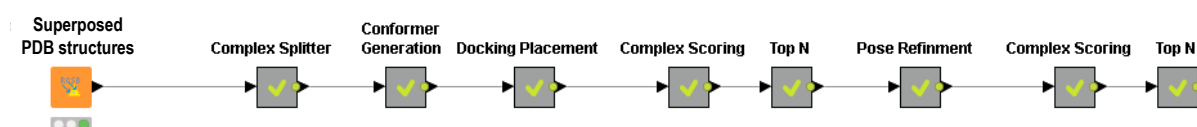
The structural superposition used only alpha carbon atoms for the calculation of the superposition. After alignment and superposition, the average RMSD between all LTA4H structures was 0.271 Å. The structures had an overall sequence identity of 96% and only differ in the first 23 places. The average RMSD between all sEH structures was 0.368 Å. PDB code 3CHP<sup>128</sup> was chosen as the structure for LTA4H docking. The structure has a resolution of 2.1 Å and is co-crystallized with the ligand 4BO (Figure 22 A). For sEH, PDB code 4Y2T<sup>107</sup> was chosen as the structure for docking. The structure has a resolution of 2.4 Å and is co-crystallized with the ligand 49Q (Figure 22 B).



**Figure 22: Co-crystallized ligands of LTA4H and sEH.** (A) Ligand 4BO of crystal structure 3CHP (LTA4H). (B) Ligand 49Q of crystal structure 4Y2T (sEH).

Figure 23 shows the KNIME workflow of the following docking validation steps. The superposed structures were reloaded into KNIME in *mol2* format. The binding site was defined with dummy atoms using the position of all crystalized ligands.

Using the *Complex Splitter* node (provided by Chemical Computing Group Inc.)<sup>81</sup>, receptor and ligands were split from the complex structure.



**Figure 23: KNIME workflow describing the MOE docking validation.**

Before docking, a maximum of 100 conformations of every ligand was generated with the *Conformations* node (provided by Chemical Computing Group Inc.)<sup>81</sup> using the LowModeMD method. The LowModeMD search method generates conformations using a short  $\sim 0.5$  ps molecular dynamics run.<sup>129,130</sup> All settings were left at default with an energy window of 7.0 kcal/mol (conformations will be discarded if their potential energy is greater than  $E_{min} + s$ , where  $E_{min}$  is the lowest energy among the generated conformations and  $s$  is the specified energy window) and an RMSD limit of 0.25 Å (conformations with a RMSD less than the specified RMSD limit are considered duplicates and will be discarded). 722 conformations were generated for LTA4H ligands and 1,391 conformations for sEH ligands. For LTA4H an average of 16.8 conformers per compound and 14.8 conformers per compound for sEH were obtained.

The generated conformers were docked into their corresponding receptor structure using the *Docking placement* node (provided by Chemical Computing Group Inc.)<sup>81</sup>. For each receptor 20 poses were generated using the *Triangle Matcher* placement method (section 2.1.1).

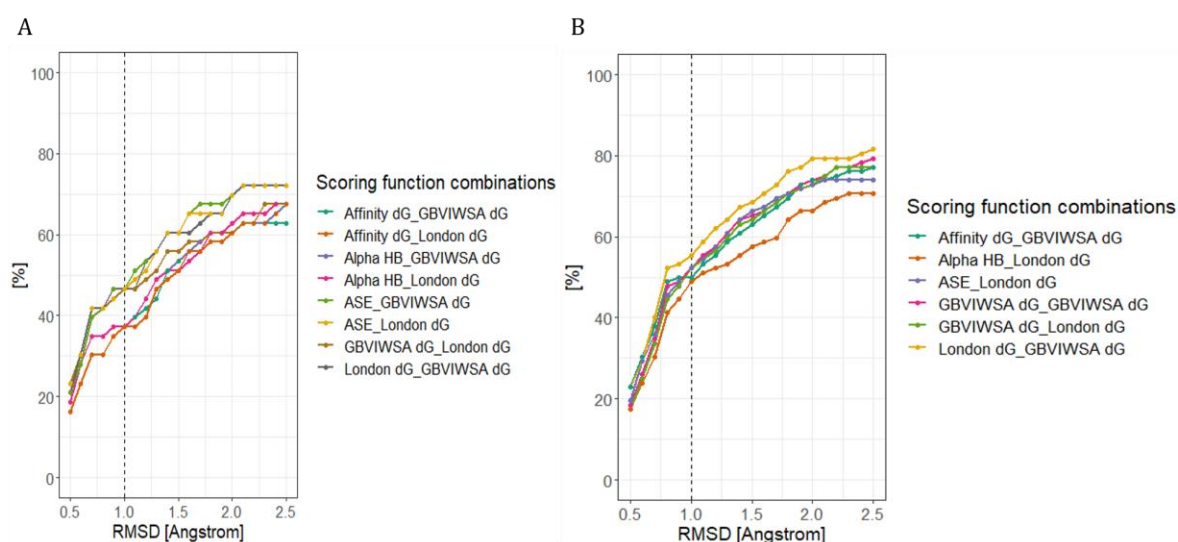
In the next step all generated poses were scored using the *Complex Scoring* node (provided by Chemical Computing Group Inc.)<sup>81</sup>. All five MOE scoring functions (London dG, ASE, Affinity dG, Alpha HB, GBVI/WSA dG) were used for scoring. The 10 best scoring poses (highest ranked) for each receptor and scoring function were filtered (*Top N* node, provided by Chemical Computing Group Inc.)<sup>81</sup> for further processing.

The generated poses may still be strained and hydrogen bonds not optimally oriented. Therefore the remaining poses were refined using the *Pose Refinement* node (provided by Chemical Computing Group Inc.)<sup>81</sup>. Pose refinement uses a force field minimization of the poses generated in the docking placement. All settings were set as default using the Amber 10:EHT force field<sup>95</sup> and fixed receptor atoms (see section 2.1.1 for more details).

The refined poses needed to be re-scored (since a pose minimization was conducted in the previous step) using the *Complex scoring* node. All five scoring functions were used for scoring. The number of poses was further reduced to the five best scoring poses using the *Top N* node.

Applying this workflow consisting of conformer generation, docking placement, complex scoring, Top N filter, pose refinement, complex scoring and Top N filter (**Figure 23**), all possible combinations of scoring functions for placement and refinement provided in

MOE could be validated. The scoring functions used in the placement and refinement step have great impact on the resulting selection of re-docked poses. For this reason, all possible combinations of scoring functions were analyzed by calculating the RMSD between crystalized ligand and the top five re-docked ligand poses. Of the top five re-docked poses, the one with the smallest RMSD to its crystalized ligand was used in the final assessment. Those results are shown in **Figure 24** for a selection of scoring function combinations.



**Figure 24: MOE docking validation for LTA4H and sEH.** Percentage of docking poses for (A) LTA4H and (B) sEH within different RMSDs of the crystalized ligand position. Results for a selection of scoring function combinations is shown.

For each protein-ligand complex (43 for LTA4H and 94 for sEH) 25 poses, one for each scoring function combination, were collected. As the final validation step, the number of poses within a RMSD of 1.0 Å and 2.0 Å to the crystallized ligands were analyzed for each scoring function combination to identify the optimal combination for both target receptors (LTA4H and sEH). In both cases the default settings with London dG for placement and GBVI/WSA dG for refinement resulted in poses with a minimum RMSD in the re-docking. In case of LTA4H, 46.5% of the re-docked ligands have an RMSD  $\leq$  1.0 Å (**Figure 24 A**). In the re-docking of sEH, 55.4% of the ligands have an RMSD  $\leq$  1.0 Å (**Figure 24 B**). Within an RMSD  $\leq$  2.0 Å, 69.8% (LTA4H) to 79.4% (sEH) of the bioactive ligand conformations can be reproduced in the re-docking. Results for all scoring function combinations tested can be found in the appendix Tables A1 and A2.



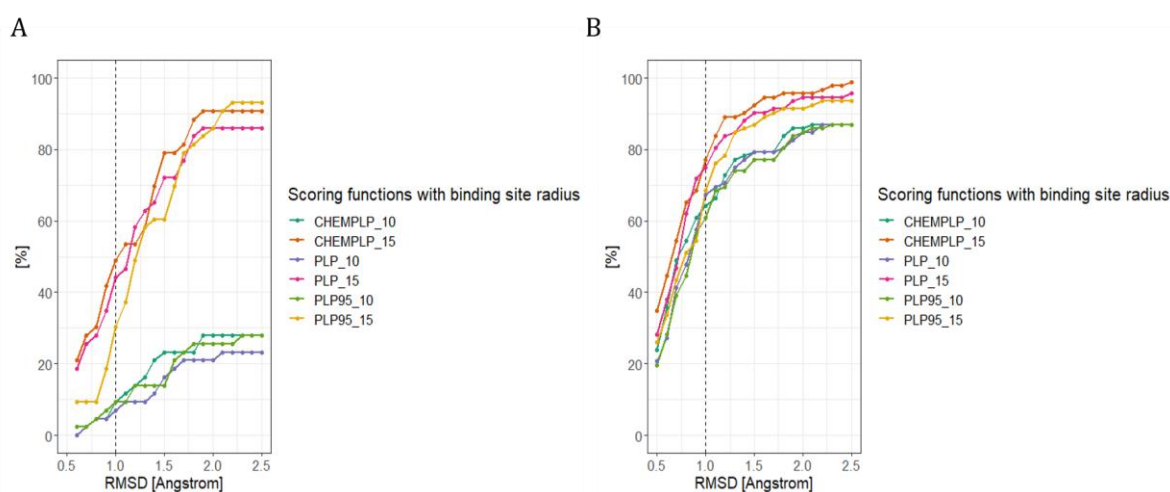
### 3.1.2 PLANTS

The PLANTS docking algorithm was validated as a second docking tool. PLANTS docking was included in a KNIME Workflow. The 43 LTA4H crystal structures and 93 sEH crystal structures were used for docking validation. Each crystalized ligand was re-docked into the corresponding receptor. For each crystal structure, an individual configuration file had to be created. In these configuration files the input protein and ligand file, the binding site center and radius, number of output poses and the scoring function had to be defined.

For LTA4H the zinc ion in the binding site was chosen as the center of the binding site for each individual receptor structure. For sEH a manually selected dummy atom in the approximated center of the binding site was chosen.

Different binding site radii were set (10, 15, 20 and 25Å) and all three available scoring functions were tested (CHEMPLP, PLP, PLP95). For each crystalized ligand, five docking poses were generated. This led to 12 different docking set ups which had to be defined in the configuration file.

A KNIME workflow was used for the configuration file generation. The RMSD between crystalized ligand and the five re-docked ligand poses was calculated for each combination of scoring function and binding site radius. Following the MOE docking validation, of the top five re-docked poses, the one with the smallest RMSD to its crystalized ligand was used in the final assessment. **Figure 25 A** shows a selection of results for the target LTA4H. The scoring function CHEMPLP with a binding site radius of



**Figure 25: PLANTS docking validation for LTA4H and sEH.** Percentage of docking poses for (A) LTA4H and (B) sEH within different RMSDs of the crystalized ligand position. Given are the three different scoring functions with the binding site radius after the underline.



15 Å resulted in a docking pose with a minimal RMSD. 48.8% of the re-docked ligands can be found within an RMSD of 1.0 Å. Within an RMSD of 2.0 Å, 90.7% of the re-docked ligands can be found.

The same combination of scoring function and binding site radius (CHEMPLP with a radius of 15Å) led to the best results for sEH (**Figure 25 B**). Results for all tested scoring function and binding site radii can be found in appendix Tables A3 and A4.

### 3.2 Comparison of docking tools MOE and PLANTS

The accuracies of the docking tools MOE and PLANTS were validated by re-docking the crystalized ligands in the binding sites of LTA4H and sEH. The accuracies were determined by how often the re-docked structures were found in a 1.0 Å and 2.0 Å radius of the crystalized ligands. **Table 1** shows the results of the docking accuracies for MOE and PLANTS docking.

Comparing both targets, it can be said that docking into the LTA4H binding site is more difficult than docking in the sEH binding site. When using MOE docking, ~9% more ligands are found within an RMSD of 1.0 Å (compared to the crystalized ligands) for sEH compared to LTA4H (**Table 1**, columns 1 and 3). In PLANTS docking the difference is even greater with ~30% (**Table 1**, columns 2 and 4).

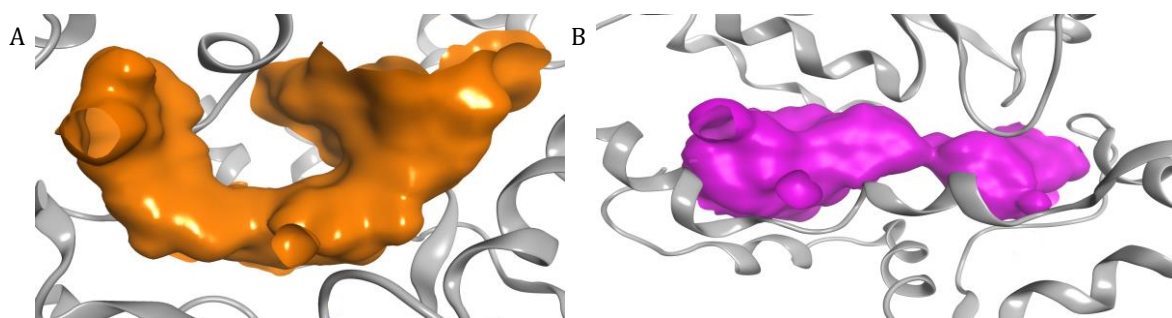
**Table 1: Comparison of docking accuracy between MOE and PLANTS.** The docking accuracy (RMSD between crystalized ligands and re-docked ligands) of MOE and PLANTS are shown. The percentage of re-docked ligands with an RMSD of smaller 1.0 and 2.0 Å are shown

LTA4H		sEH	
MOE	PLANTS	MOE	PLANTS
<i>RMSD ≤ 1.0 Å</i>		<i>RMSD ≤ 1.0 Å</i>	
46.5%	48.8%	55.4%	77.2%
<i>RMSD ≤ 2.0 Å</i>		<i>RMSD ≤ 2.0 Å</i>	
69.8%	90.7%	79.4%	95.7%

MOE and PLANTS perform nearly with the same docking accuracy on LTA4H and an RMSD smaller 1.0 Å (**Table 1**, columns 1 and 2). Looking at an RMSD of smaller 2.0 Å PLANTS docking leads to 90% correct docking poses, where on the other hand MOE docking only

reached 70% correct docking poses (**Table 1**, columns 1 and 2). The differences between MOE and PLANTS are even more pronounced when looking at the sEH docking accuracies. PLANTS produced correct poses in 77% of the cases with an RMDS smaller 1.0 Å, MOE reaches 55% (**Table 1**, columns 3 and 4).

One explanation for the large docking accuracy differences between LTA4H and sEH lies in analyzing the different binding sites. Both binding sites have approximately the same volume (LTA4H: 4,298 Å<sup>3</sup>, sEH: 4,258 Å<sup>3</sup>). The main difference is the shape of the binding sites. The LTA4H binding site has an L-shape form which seems to be more difficult to dock the compounds into (**Figure 26 A**). The sEH binding site is nearly linear, which makes it easier to dock the compounds into the binding site (**Figure 26 B**).



**Figure 26: Binding site shape of LTA4H and sEH.** (A) L-shaped binding site of LTA4H (PDB: 3CHP). (B) Linear shaped binding site of sEH (PDB: 4Y2T).

The validation of MOE and PLANTS docking tools lead to the statement, that PLANTS is the better docking tool taking docking accuracy (**Table 1**) and calculation time ([sections 3.3-3.5](#)) into account. Docking the combinatorial library of 20,630 compounds is two times faster in PLANTS (14 days) than in MOE (30 days).

### 3.3 MOE docking procedure of compound batches 2 and 3

All compounds were docked using the batch docking function in MOE. A batch file generated in MOE is an SVL (Scientific Vector Language) source code containing the docking configurations. The configurations were defined in the MOE GUI docking tool. The general docking settings were set as follows: The receptor was defined using the prepared receptor structure for 3CHP (LTA4H) and 4Y2T (sEH). The binding site was defined by

the volume occupied by all crystalized ligands (after superposing all available complex structures, see [section 2.2.1](#)). This space was filled with dummy atoms using the *Site Finder* function in the MOE GUI ([section 2.1.1](#)). Ligand structures were loaded as an MDB file. The method for placement and refinement, as well as the corresponding scoring functions were set according to the results of the docking validation ([section 3.1.1](#)):

- Triangle Matcher as placement method with the scoring function London dG, generation of 10 poses.
- Rigid receptor as refinement method with the scoring function GBVI/WSA dG, generation of 5 poses.

Following specified docking time corresponds to a working machine with an i7 core, 3.30GHz CPU and 64 GB RAM.

Docking of the ChEMBL active compounds resulted in 1,910 poses for 382 LTA4H compounds with an approximated calculation time of four days. Docking of 1,384 sEH active ChEMBL compounds resulted in 6,920 docking poses. Calculation time for sEH was approximately eight days.

Docking of 1,000 inactive ChEMBL compounds in the LTA4H receptor resulted in 4,768 poses. For 21 compounds, less than five poses were obtained after pose refinement. For sEH, 4,792 poses were generated, after pose refinement two compounds obtained less than five poses. Calculation time was approximately seven days for each receptor.

### 3.4 MOE docking of compound batch 4

After designing the focused combinatorial library containing 20,630 compounds, poses were generated in a docking procedure. Since the combinatorial library contained a common (phenoxyethyl)benzene key fragment, a different docking approach was applied (compared to the docking of batches 2 and 3). Both ligands in the used receptor structures (3CHP, 4Y2T) contained the same fragment, see **Figure 15 B (C1)**. Therefore, a template-based docking procedure was conducted.

In a template-based docking one part of the compounds to dock is fixed according to the position and orientation of the overlapping structural part found in the crystal structure.

Template based docking saves computation time because the complexity of the optimization problem is dramatically reduced.

In our case, the key fragment C1 was selected as the template. The position and orientation of the amine building blocks C2 were optimized during docking. An MOE batch file was generated for each of the receptors. According to the result of the MOE docking performance study settings were set as follows:

The substructure placement method was combined with the scoring function London dG (maximum of 30 poses) and rigid receptor refinement method with the scoring function GBVI/WSA dG (maximum of 5 poses). As a result, 103,107 poses were generated for LTA4H and 103,100 poses for sEH. Calculation time was approximately four weeks for each receptor.

### 3.5 PLANTS docking procedure of compound batches 2-4

For PLANTS docking a configuration file (config.txt) containing all settings for docking was generated for both targets of interest (LTA4H and sEH) and data set batches 2-4 (ChEMBL active and inactive compounds, as well as compounds from the combinatorial library). According to the results of PLANTS docking validation ([section 3.1.2](#)) the scoring function CHEMPLP with a binding site radius of 15Å was chosen. The configuration file contains input location for the receptor and the compounds to be docked, an output file location and binding site center (LTA4H: zinc ion in the binding pocket; sEH: manually selected dummy atom). The docking settings defined in the configuration (binding site radius, number of output poses and the scoring function) were identical in each configuration file. All PLANTS dockings were executed using the following command in the command line:

```
PLANTS1.1_mingwm --mode screen config.txt.
```

Following specified docking time corresponds to a working machine with an i7 core, 3.30GHz CPU and 64 GB RAM.

The prepared active ChEMBL compounds ([section 2.2.2](#)) for LTA4H (382 compounds) and sEH (1,384 compounds) were docked into their respective receptors using individual

configuration files. Calculation time was approximately 5 hours for LTA4H and 17 hours for sEH.

The 1,000 prepared inactive ChEMBL compounds (section 2.2.3) were docked into their respective receptor using individual configuration files. For each compound, 5 poses were generated by the PLANTS docking algorithm, leading to 5,000 compound poses for LTA4H and sEH. Calculation time was approximately 12 hours for each receptor.

The generated combinatorial library containing 20,630 related and unique compounds (section 2.2.4) was docked in each receptor of interest (LTA4H and sEH). For each compound, 5 poses were generated by the PLANTS docking algorithm, leading to 103,155 compound poses for LTA4H and sEH. Calculation time was approximately 2 weeks for each receptor.

### 3.6 Machine learning optimization

Each of the four machine learning algorithms (SVC, Random Forest, XGBoost, AdaBoost) were optimized for each of the five fingerprints (PLIF, AtomPair, Morgan, FeatMorgan, MACCS) and the two targets of interest (LTA4H and sEH). This resulted in 48 model optimizations (Figure 27). All models were built using the scikit-learn python package in a Jupyter notebook.<sup>118,119</sup>

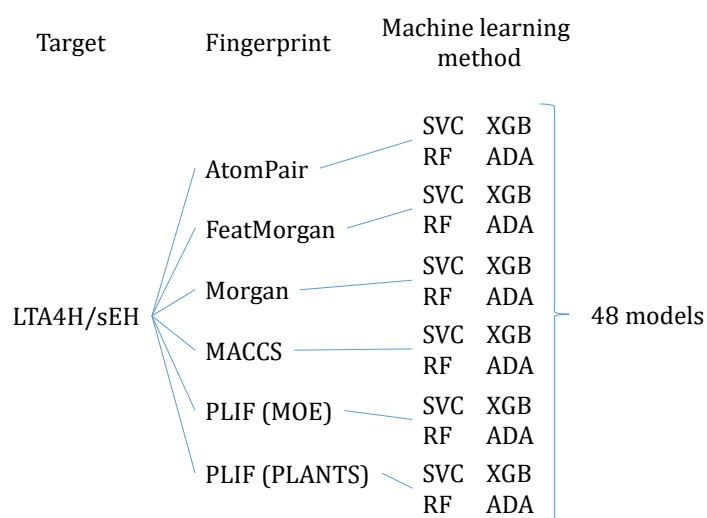


Figure 27: Scheme of the models built and used for prediction.

The models were evaluated by 10-fold cross validation and accuracy was used as the primary measure of model performance. Model optimization was conducted on the data set containing the crystalized ligands, active and inactive ChEMBL compounds (in total 1,425 LTA4H compounds and 2,476 sEH compounds).

The data set was split into training- and test sets. This partitioning scheme was evaluated in a grid search between 75% and 95% of the training set size. Since classification for all compounds is available in this data set the different machine learning algorithms could be optimized by maximizing the accuracy of the model. Accuracy was calculated using default

**Table 2: Optimal partitioning scheme for LTA4H and sEH.** Partitioning between 75% and 95% training set size was optimized to achieve maximal accuracy. The range of accuracy for the four tested machine learning algorithms (SVC, Random Forest, XGBoost, AdaBoost) is shown.

<i>LTA4H</i>	Training set size [%]	Test set size [%]	Accuracy	Mean accuracies
<b>PLIF   MOE</b>	90	10	0.76 – 0.84	0.77
<b>PLIF   PLANTS</b>	80	20	0.77 – 0.80	0.78
<b>AtomPair</b>	75	25	0.96 – 0.98	0.97
<b>Morgan</b>	85	15	0.98 – 0.99	0.98
<b>FeatMorgan</b>	75	25	0.95 – 0.98	0.97
<b>MACCS</b>	85	15	0.97 – 0.99	0.98
<i>sEH</i>				
<b>PLIF   MOE</b>	90	10	0.62 – 0.65	0.63
<b>PLIF   PLANTS</b>	80	20	0.62 – 0.67	0.64
<b>AtomPair</b>	85	15	0.90 – 0.97	0.94
<b>Morgan</b>	75	25	0.93 – 0.96	0.94
<b>FeatMorgan</b>	80	20	0.89 – 0.94	0.92
<b>MACCS</b>	80	20	0.92 – 0.95	0.93

settings for the four machine learning algorithms. The results of the partitioning scheme optimizations are shown in **Table 2**. For each of the four machine learning algorithms individual parameters were optimized in a grid search to achieve maximal accuracy. The parameter for SVC were left default. For Random Forest classification the number of estimators (number of trees) was optimized.

For XGBoost classification the parameters,

- max depth (maximum tree depth, increasing this value will make the model more complex and more likely to overfit),
- learning rate (learning rate regulates the contribution of each tree),
- number of estimators (number of boosting stages) and
- alpha (L1 regularization term on weights, increasing this value will make the model more conservative)

were optimized. For AdaBoost classification, the number of estimators (maximum number of estimators at which boosting is terminated) was optimized. Parameters optimized for the machine learning algorithms by the grid search can be found in **Table 3**.

**Table 3: Parameters optimized by grid search.** Parameters and range of tested values with step size are specified.

Machine learning algorithm	Parameter	Tested values
XGBoost	Max depth	10 – 200   10
	Learning rate	0.01, 0.001
	Estimators	100 – 1,000   100
	Alpha	0.0, 0.005
Random Forest	Estimators	10 – 1,000   10
AdaBoost	Estimators	10 – 200   10

Optimal model parameters were chosen empirically based on maximal accuracy and computational cost. Optimized parameters and accuracy results can be found in **Table 4**.

**Table 4: Optimized parameters for the used machine learning algorithms.** RF: number of estimators; XGB: max depth, learning rate, number of estimators, alpha; ADA: number of estimators. SVC default parameters not shown. Accuracy shows the mean value from 10-fold cross validation and the standard deviation.

<b>LTA4H</b>			<b>sEH</b>		
<b>Fingerprint</b>	<b>Optimized parameters</b>	<b>Accuracy</b>	<b>Fingerprint</b>	<b>Optimized parameters</b>	<b>Accuracy</b>
<b><i>PLIF/MOE</i></b>			<b><i>PLIF/MOE</i></b>		
SVC	Default	0.75 ± 0.02	SVC	Default	0.610 ± 0.004
RF	130	0.77 ± 0.03	RF	130	0.60 ± 0.03
XGB	10, 0.01, 600, 0.005	0.78 ± 0.04	XGB	20, 0.01, 800, 0.0	0.60 ± 0.03
ADA	190	0.77 ± 0.03	ADA	50	0.63 ± 0.01
<b><i>PLIF/PLANTS</i></b>			<b><i>PLIF/PLANTS</i></b>		
SVC	Default	0.68 ± 0.02	SVC	Default	0.59 ± 0.01
RF	30	0.76 ± 0.03	RF	190	0.59 ± 0.03
XGB	40, 0.01, 500, 0.005	0.75 ± 0.03	XGB	30, 0.01, 700, 0.005	0.58 ± 0.03
ADA	80	0.74 ± 0.03	ADA	70	0.63 ± 0.03
<b><i>AtomPair</i></b>			<b><i>AtomPair</i></b>		
SVC	Default	0.95 ± 0.02	SVC	Default	0.93 ± 0.01
RF	30	0.96 ± 0.02	RF	490	0.94 ± 0.01



XGB	10, 0.01, 700, 0.005	0.97 ± 0.02	XGB	10, 0.01, 900, 0.005	0.95 ± 0.01
ADA	180	0.97 ± 0.02	ADA	180	0.93 ± 0.02
<b><i>FeatMorgan</i></b>			<b><i>FeatMorgan</i></b>		
SVC	Default	0.95 ± 0.01	SVC	Default	0.90 ± 0.02
RF	240	0.98 ± 0.01	RF	600	0.95 ± 0.01
XGB	10, 0.01, 800, 0.005	0.98 ± 0.01	XGB	20, 0.01, 800, 0.005	0.95 ± 0.01
ADA	90	0.95 ± 0.02	ADA	200	0.93 ± 0.01
<b><i>Morgan</i></b>			<b><i>Morgan</i></b>		
SVC	Default	0.96 ± 0.02	SVC	Default	0.92 ± 0.02
RF	180	0.98 ± 0.01	RF	520	0.95 ± 0.01
XGB	10, 0.01, 1,000, 0.005	0.98 ± 0.01	XGB	10, 0.01, 1,000, 0.005	0.95 ± 0.01
ADA	110	0.98 ± 0.02	ADA	60	0.91 ± 0.02
<b><i>MACCS</i></b>			<b><i>MACCS</i></b>		
SVC	Default	0.96 ± 0.02	SVC	Default	0.93 ± 0.02
RF	60	0.97 ± 0.02	RF	220	0.94 ± 0.02
XGB	10, 0.01, 300, 0.005	0.96 ± 0.02	XGB	10, 0.01, 1,000, 0.005	0.95 ± 0.02
ADA	30	0.94 ± 0.02	ADA	200	0.91 ± 0.02

Overall, regardless of used fingerprint and machine learning algorithm, the prediction accuracy for LTA4H is slightly better than for sEH. Comparing the accuracy of the PLIF fingerprint with the 2D-fingerprints, an accuracy difference of approximately 0.2 (for LTA4H) and 0.3 (for sEH) was obtained (**Table 4**).

These results show that using the 2D-fingerprints to distinguish between active and inactive compounds results in 20% to 30% better results compared to the 3D-fingerprint. For the PLIF fingerprint, a maximal accuracy of 0.78 with the machine learning algorithm XGBoost for LTA4H is achieved. For sEH, the machine learning algorithm AdaBoost gives the maximal accuracy of 0.63. In general, accuracies using the different machine learning algorithms hardly differ. They range from a minimal difference in accuracy of 0.02 (LTA4H|AtomPair, LTA4H|Morgan, sEH|AtomPair) to a maximum difference of 0.08 for LTA4H|PLIF PLANTS.

Comparing the accuracy results between the targets LTA4H and sEH using the 3D-fingerprint and the two different docking software (PLIF|MOE and PLIF|PLANTS). Accuracies are on average almost 20% better for LTA4H compared to sEH. The average accuracy difference for LTA4H between MOE and PLANTS with 4.56% and for sEH with 2.05% is comparatively smaller.

The results show that there is a significant difference in 2D- and 3D-fingerprints accuracies but no significant difference using various docking software and various machine learning algorithms.

### 3.7 Machine learning prediction from PLIF/MOE docking

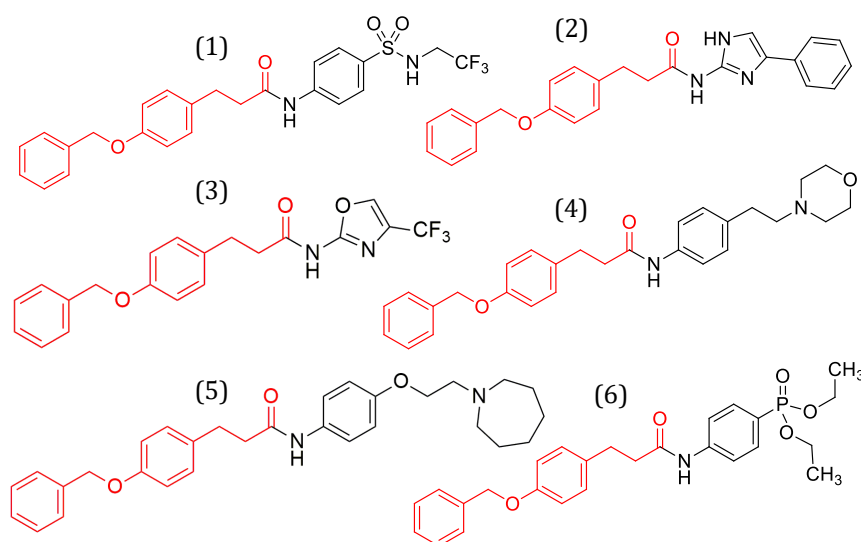
The optimized models ([section 3.6](#)) were used to predict the class labels (active or inactive) of the combinatorial library compounds containing the calculated PLIF fingerprint. The results for the prediction of novel dual active compounds using PLIF fingerprint and MOE docking poses can be found in **Table 5**.

**Table 5: Number of compounds predicted to be active using PLIF fingerprint and MOE docking.** For the target of interest (LTA4H and sEH) each fingerprint and each machine learning algorithm, the number of compounds to be predicted active, number of compounds to be predicted active with a fingerprint similarity of  $\geq 0.5$  against the crystalized ligands and the number of compounds predicted to be dual active are given.

LTA4H				sEH				
FP	Pred. active	Pred. active FP sim.	Pred. active conf.	FP	Pred. active	Pred. active FP sim.	Pred. active conf	Pred. dual activity
<i>PLIF</i>				<i>PLIF</i>				
SVC	73	42	-	SVC	20,601	18,360	-	39
RF	659	462	163	RF	17,134	15,482	14,353	115
XGB	1,475	812	-	XGB	17,378	15,476	-	601
ADA	1,821	1,041	-	ADA	19,206	17,273	-	874

For sEH, the number of predicted active compounds includes almost the entire combinatorial data set (20,630). The search was focused by restricting the fingerprint similarity against the crystalized ligands to a Tanimoto similarity minimum of 0.5. This led to a reduction of the number of compounds of 29.9% for LTA4H and 9.6% for sEH (Table 5, column 3 and 7). Since the reduction of compounds via Tanimoto similarity is smaller for sEH compared to LTA4H the following statement can be made: compounds predicted to be active on sEH are very similar to the crystalized sEH ligands but compounds predicted to be active on LTA4H are less similar to the crystalized LTA4H ligands. The confidence of the prediction for the Random Forest model was restricted to be a minimum of 0.7. Results can be found in Table 5, column 4 and 8 for Random Forest. This restriction led to a compound reduction for LTA4H of 45.4% and 6.6% for sEH. These reduction results show that the Random Forest prediction for LTA4H generates many assumed to be active compounds with a low prediction confidence. The predictions made for sEH resulted in fewer assumed to be active compounds with a low prediction confidence since the compound reduction is less severe. To extract novel dual active compounds, the sets of compounds predicted to be active for LTA4H and sEH were compared. Compounds predicted to be active on both targets were collected. Those counts can be found in Table 5 in the rightmost column. The result of the PLIF fingerprint using Random Forest for dual active prediction resulted in 115 compounds. From those 115 compounds a manual selection (cherry picking) of six compounds was made for

chemical synthesis and biological testing (**Figure 28**). Feasibility of synthesis, estimated solubility and uniqueness of the compounds (amongst the selected compounds and compared to known inhibitors) were used as guidelines for cherry picking. The synthesis ([section 3.10](#)) and following testing ([section 3.11](#)) of the selected compounds was used to validate the new method presented in this work.



**Figure 28: Selection of six compounds for chemical synthesis using PLIF fingerprint and Random Forest.** The red part of the structures is the key fragment C1. The coupled black part derives from the amine building blocks. Those building blocks correspond to the following ZINC compounds: (1) ZINC03888838, (2) ZINC05908671, (3) ZINC04198753, (4) ZINC19623909, (5) ZINC04938424, (6) ZINC01708144

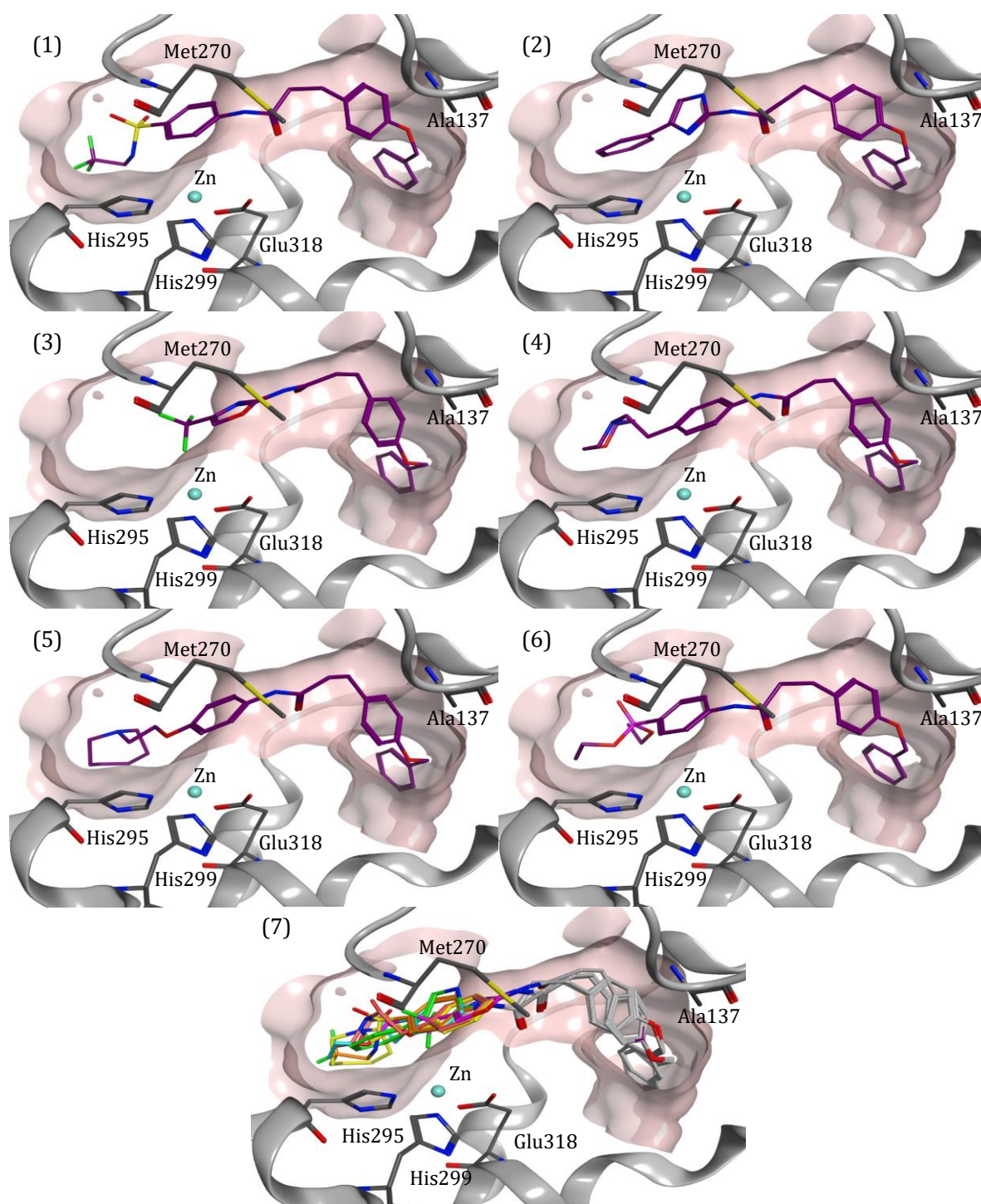
Molecular properties are used as a first assessment to evaluate the lead likeness of the predicted and synthesized compounds. The six synthesized compounds are structurally heterogeneous but show some shared structural features. All compounds have a minimum of three aromatic rings, whereas two aromatic rings originate from the key fragment C1. All compounds share an aromatic ring directly attached to the amide (**Figure 28**). The molecular weight ranges from 390 g/mol to 490 g/mol (a limit of 500 g/mol was set in the design of the combinatorial library, see [section 2.2.4](#)). Compounds 4 and 5 contain an aliphatic ring system. All compounds incorporate at least five heteroatoms (O, N, S, F or P) (**Figure 28**). **Table 6** shows the properties of the Lipinski's rule of five, which serves as a rule of thumb to evaluate the compounds pharmacokinetic (molecular weight < 500 g/mol, cLogP < 5.0, H-bond donors < 5, H-bond

acceptors < 10). Further, the number of rotatable bonds (< 10) is listed, which was defined by Veber et al.<sup>131</sup> as an additional criterion for oral bioavailability of compounds. The rightmost column in **Table 6** shows the count of rule violation. Those calculated properties do not predict if a compound is pharmacologically active, but if the compound has possible favorable pharmacokinetic properties. Compound **5** has two violations coming from cLogP and the number of rotational bonds. Compounds **2** and **4** have no violations concerning the number of rotatable bonds. Veber et al.<sup>131</sup> could show, that a low number of rotational bonds has a positive impact on bioavailability.

**Table 6: Molecule properties of six selected compounds for synthesis using PLIF fingerprint and Random Forest.** Compounds 1 to 6; molecular weight in Da; calculated LogP; Number of hydrogen bond donors; number of hydrogen bond acceptors; number of rotatable bonds.

<b>Cmp.</b>	<b>MW [Da]</b>	<b>cLogP</b>	<b>H-bond donors</b>	<b>H-bond acceptors</b>	<b>Rotatable bond</b>	<b>Rule violation</b>
<b>1</b>	492.5	4.87	2	5	14	1
<b>2</b>	397.5	5.72	2	3	8	1
<b>3</b>	390.4	4.17	1	3	11	1
<b>4</b>	444.6	4.11	1	4	10	0
<b>5</b>	472.6	5.59	1	3	11	2
<b>6</b>	467.5	4.66	1	5	14	1

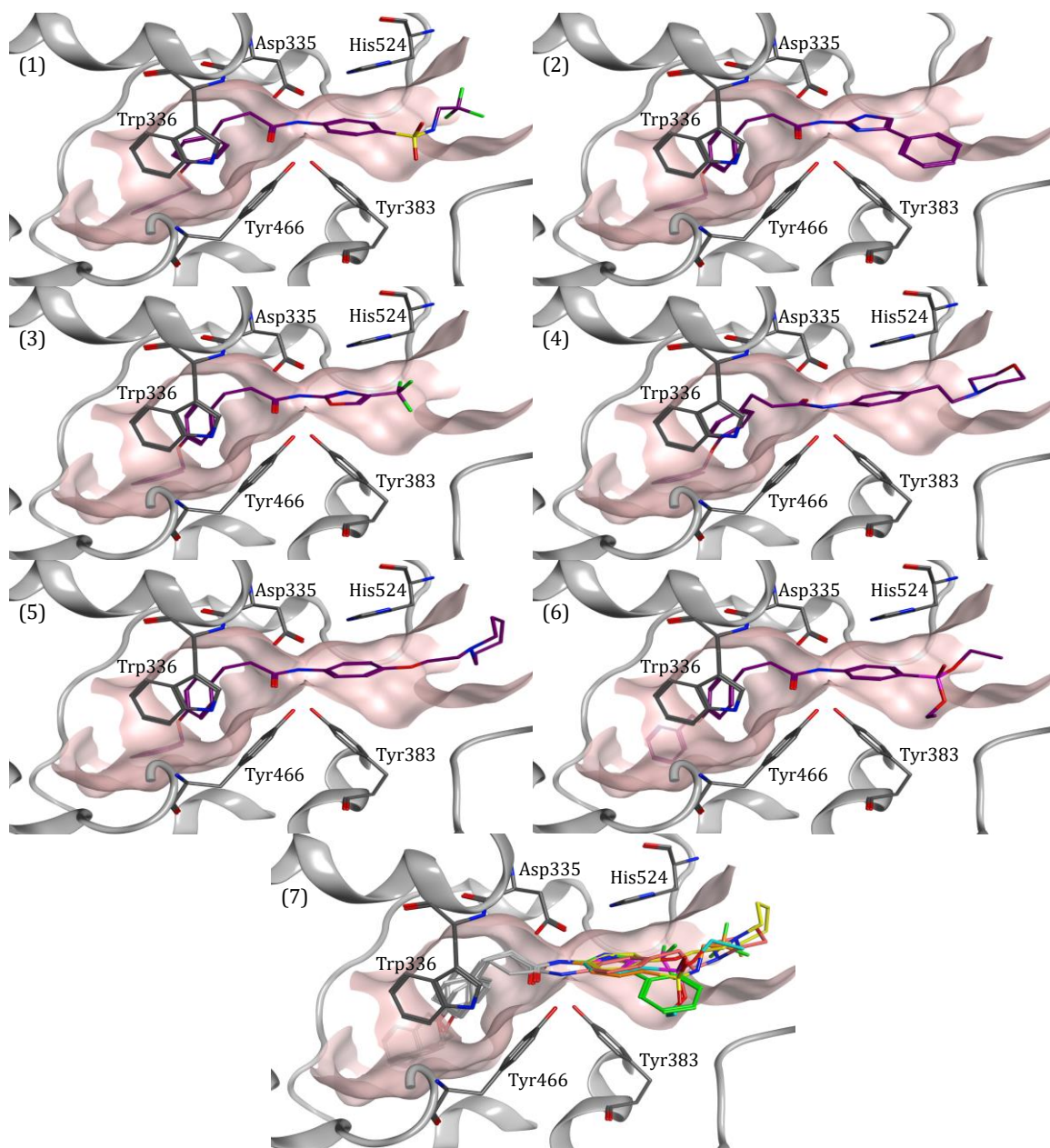
In **Figure 29** the pose of the six synthesized compounds docked with MOE in the LTA4H binding site are shown. Compounds **1,2, 4, 5** and **6** form hydrogen bonds to Met270 over the nitrogen atom of the amid bond. Further, all three aromatic rings are forming pi interactions with receptor amino acids. Only compound **3** forms an interaction to the zinc ion, although nearly 50% of the crystalized ligands form a strong interaction to the catalytic zinc ion. All docking poses pictured in this work have to be considered with care. In the process of docking many restrictions and simplifications are made (e.g. rigid receptor, scoring functions; [section 2.1](#)). Therefore, the presented docking poses are merely a suggestion on how the compounds may bind in the binding site. Slight differences in orientation and location of the compounds can lead to different interactions between compound and receptor. To give more information about the actual binding mode and formed interactions, crystallization of those compounds would have to be performed, which is out of scope of this work. Currently, only inhibitory activity can give more insights in the SAR.



**Figure 29: Binding poses (MOE docking) of synthesized compounds using PLIF and Random Forest in the receptor structure LTA4H.** The right part of the structures is the key fragment C1. The coupled left part derives from the amine building blocks. Those building blocks correspond to the following ZINC compounds: (1) ZINC03888838, (2) ZINC05908671, (3) ZINC04198753, (4) ZINC19623909, (5) ZINC04938424, (6) ZINC01708144. Image (7) show the overlay of all 6 compounds in the sEH binding site, the key fragment is colored in light gray, the building blocks are colored as follows: orange: compound (1), green: compound (2), magenta: compound (3), light red: compound (4) yellow: compound (5), azure: compound. (6).



In **Figure 30** the docking poses of the same synthesized compounds (**1-6**) are shown in the binding site of sEH. The key fragment C1 is located to the left of the binding site, according to the crystalized ligand in receptor PDB 4Y2T (**Figure 9**). The amide bond of the compounds is in the catalytic region of the amino acid residues Tyr383, Tyr466 and



**Figure 30: Binding poses (MOE docking) of synthesized compounds using PLIF and Random Forest in the receptor structure sEH.** The left part of the structures is the key fragment C1. The coupled right part derives from the amine building blocks. Those building blocks correspond to the following ZINC compounds: (1) ZINC03888838, (2) ZINC05908671, (3) ZINC04198753, (4) ZINC19623909, (5) ZINC04938424, (6) ZINC01708144. Image (7) show the overlay of all 6 compounds in the sEH binding site, the key fragment is colored in light gray, the building blocks are colored as follows: orange: compound (1), green: compound (2), magenta: compound (3), light red: compound (4) yellow: compound (5), azure: compound (6).

Asp355. The amine building blocks extend to the right. Compounds **1-3** and compounds **5** and **6** form an arene-hydrogen bond from the amino acid residue Trp336 to a CH<sub>2</sub>-group of the key fragment. Compound **4** forms arene-hydrogen interaction with Trp336 with one of the benzene rings of the key fragment C1. Compounds **2-6** form an arene-arene interaction of one aromatic group of their amine building block with the amino acid residue His524.

### 3.8 Machine learning prediction from PLIF/PLANTS docking

The results for the prediction of novel dual active compounds using PLIF fingerprint and PLANTS docking poses can be found in **Table 7**. The number of compounds predicted to be active by the different machine learning algorithms is listed in column 2 and 5 of **Table 7**. The number of compounds was restricted by fingerprint similarity (**Table 7**, column 6). A minimal Tanimoto similarity of 0.5 was set as a limit.

**Table 7: Number of compounds predicted to be active using PLIF fingerprint and PLANTS docking.** For the target of interest (LTA4H and sEH) each fingerprint and each machine learning algorithm, the number of compounds to be predicted active, number of compounds to be predicted active with a fingerprint similarity of  $\geq 0.5$  against the crystalized ligands and the number of compounds predicted to be dual active are given.

LTA4H			sEH			
Fingerprint	Predicted active	Predicted active fingerprint sim.	Fingerprint	Predicted active	Predicted active fingerprint sim.	Predicted dual activity
<i>PLIF</i>			<i>PLIF</i>			
SVC	168	90	SVC	18,038	13,890	52
RF	6,030	2,856	RF	9,609	7,210	982
XGB	5,900	3,148	XGB	8,952	6,807	977
ADA	5,633	3,729	ADA	10,019	7,708	1,304



When comparing the results of the MOE and PLANTS docking some differences can be observed.

The docking accuracy was already discussed in [section 3.2](#). The number of compounds predicted to be active is higher for LTA4H in the PLANTS docking compared to the MOE docking. For sEH, the number of compounds predicted to be active is smaller by using the PLANTS docking compared to the MOE docking. The restriction of fingerprint similarity led to a reduction of predicted active compounds for LTA4H of 55% and 76% for sEH. The number of compounds of predicted dual activity is larger compared to the MOE docking. Around 1,000 compounds are predicted to be dual active using Random Forest, XGBoost and ADABOOST (MOE in comparison: RF 115 compounds, XGB 601 compounds, ADA 874 compounds). A manual selection of compounds for synthesis is difficult. Further restrictions or grouping scaffolds would be advisable to reduce this number.

### 3.9 Machine learning prediction from 2D-fingerprints

Optimized machine learning models ([section 3.6](#)) were used to predict the class labels (active or inactive) of the compounds in the combinatorial library. In **Table 8**, the results for the 2D-fingerprints can be found. The table is built the same way as **Table 5** and **Table 7**. The fingerprint similarity was also restricted to a minimum of 0.5. Those results can be found in column 3 and 6 in **Table 8**. This restriction led to a compound reduction for LTA4H of 69% and 79% for sEH. To extract novel dual active compounds, the sets of compounds predicted to be active for LTA4H and sEH were compared. Compounds predicted to be active on both targets were collected. Those results can be found in column 7 in **Table 8**. For the Morgan and FeatMorgan fingerprints no dual active compounds were obtained. For the MACCS fingerprint in combination with the machine learning algorithm SVC and ADA, more than 2,000 dual active compounds were predicted. It is not feasible to manually inspect and cherry pick candidates for synthesis from these amounts of compounds. For the AtomPair fingerprint, a reasonable number of dual active compounds was obtained. Looking at the different machine learning algorithms:

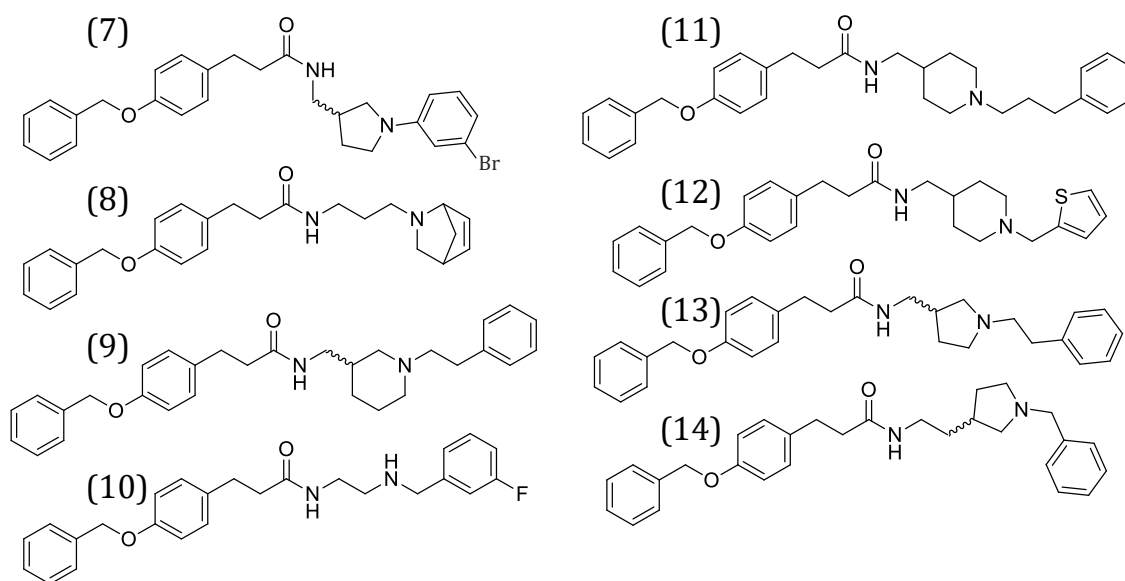
- SVC: 422 compounds
- RF: 116 compounds
- XGB: 53 compounds
- ADA: 107 compounds

**Table 8: Number of compounds predicted to be active using 2D-fingerprints.** For the target of interest (LTA4H and sEH) each fingerprint and each machine learning algorithm, the number of compounds to be predicted active, number of compounds to be predicted active with a fingerprint similarity of  $\geq 0.5$  against the crystalized ligands and the number of compounds predicted to be dual active are given.

LTA4H			sEH			
Fingerprint	Predicted active	Predicted active fingerprint sim.	Fingerprint	Predicted active	Predicted active fingerprint sim.	Predicted dual activity
<i>AtomPair</i>			<i>AtomPair</i>			
SVC	6,187	2,004	SVC	4,434	2,297	422
RF	3,046	1,157	RF	2,203	864	116
XGB	2,774	1,129	XGB	1,832	641	53
ADA	6,739	1,726	ADA	2,434	1,173	107
<i>FeatMorgan</i>			<i>FeatMorgan</i>			
SVC	20	15	SVC	4,320	836	2
RF	38	27	RF	338	90	0
XGB	301	92	XGB	640	109	0
ADA	275	71	ADA	2,861	294	0

<b>Morgan</b>			<b>Morgan</b>			
SVC	16,986	1,399	SVC	818	7	0
RF	10,890	1,208	RF	626	3	1
XGB	6,525	899	XGB	457	4	0
ADA	7,387	880	ADA	818	7	0
<b>MACCS</b>			<b>MACCS</b>			
SVC	7,344	7,344	SVC	8,469	8,456	3,389
RF	619	619	RF	2,910	2,910	68
XGB	337	337	XGB	4,151	4,151	5
ADA	5,175	5,175	ADA	9,887	9,875	2,783

From the set of 116 compounds predicted to be active with Random Forest, eight compounds were cherry picked for synthesis. This set was chosen for a better comparison to the results of the selected compounds from the 3D-fingerprint. Feasibility of synthesis, estimated solubility and uniqueness of the compounds (amongst the selected compounds and compared to known inhibitors) were used as guidelines for cherry picking. The structure of the selected eight compounds can be found in **Figure 31**. Six of the eight compounds (compound **7**, **9**, **11-14**) exhibit an *N*-substituted piperidine or pyrrolidine moiety. These saturated heterocycles are common elements in diverse series of LTA4H<sup>132</sup>



**Figure 31: Selection of eight compounds for chemical synthesis using AtomPair fingerprint and Random Forest.** The left part of the molecules is the key fragment C1. The right part of the structures derives from the amine building blocks from ZINC. Those building blocks correspond to the following ZINC compounds: (7) ZINC19481317, (8) ZINC49585019, (9) ZINC04384302, (10) ZINC32919586, (11) ZINC11889049, (12) ZINC11888991, (13) ZINC55363946, (14) ZINC02511731

and sEH<sup>133</sup> inhibitors. All compounds incorporate three aromatic rings, two aromatic rings originate from the key fragment C1, the third aromatic ring is located on the opposite end (**Figure 31**). In addition, all compounds incorporate at least four heteroatoms (O, N, S, F or Br). **Table 9** shows the molecular properties of the Lipinski's rule of five (comparable to **Table 9**). The calculated LogP values lead to most rule violations. Only compounds **8**, **10** and **12** have a cLogP value smaller than 5.0. Further, compounds **8** and **12** have no rule violations at all. Whether there is a correlation between the molecular properties analyzed in this section and the experimentally determined inhibitory activity will be discussed in [section 3.11](#).

After synthesis and testing, those results are compared with the results of the tested compounds from the PLIF fingerprint ([section 3.12](#)).

**Table 9: Molecule properties of eight selected compounds for synthesis using AtomPair fingerprint and Random Forest.** Compounds **7** to **14**; molecular weight in Da; calculated LogP; Number of hydrogen bond donors; number of hydrogen bond acceptors; number of rotatable bonds.

<b>Cmp.</b>	<b>MW [Da]</b>	<b>cLogP</b>	<b>H-bond donors</b>	<b>H-bond acceptors</b>	<b>Rotatable bond</b>	<b>Rule violation</b>
<b>7</b>	492.1	6.35	1	3	10	1
<b>8</b>	390.2	3.91	1	3	10	0
<b>9</b>	456.3	5.49	1	3	11	2
<b>10</b>	406.2	4.52	2	3	12	1
<b>11</b>	470.3	5.85	1	3	12	2
<b>12</b>	448.2	4.40	1	3	10	0
<b>13</b>	442.3	5.13	1	3	11	2
<b>14</b>	442.3	5.33	1	3	11	2

### 3.10 Synthesis and testing of selected compounds

The general synthesis procedures A, B and C, carried out by Kerstin Hiesinger, are described in [section 2.7](#). In **Table 10** synthesized compounds with synthesis procedure and yields are listed.

Not all of the 14 cherry picked compounds using the two different kinds of fingerprints could be synthesized. Compound **2** could not be isolated from the reaction mixture, compound **8** interfered with assay reagents and for compound **10** the reactant was not purchasable in a reasonable time period.

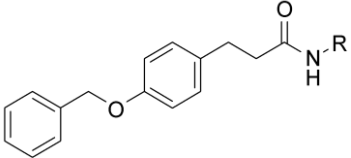
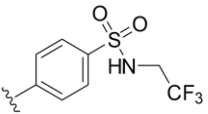
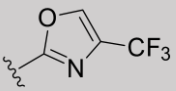
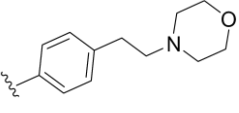
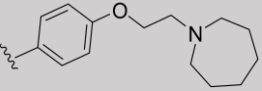
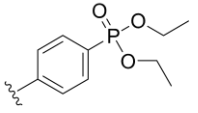
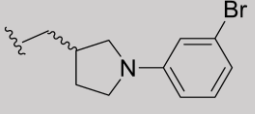
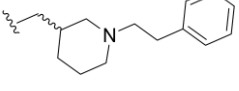
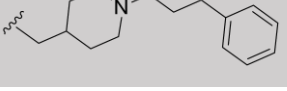
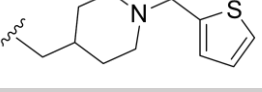
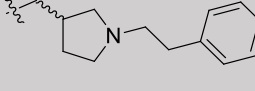
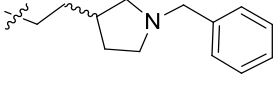
**Table 10: List of synthesized compounds.** Listed are the synthesized compounds **1** to **14** with the corresponding synthesis procedures and yield [%]. For \* different coupling reagents were tested, the coupling reagent CDI (1,1'-Carbonyldiimidazole) yielded 80% conversion but the product could not be isolated.

Compound	Procedure	Yield [%]
<b>1</b>	B	20%
<b>2</b>	*	could not be isolated
<b>3</b>	C	35%
<b>4</b>	A	71%
<b>5</b>	B	60%
<b>6</b>	A	23%
<b>7</b>	B	35%
<b>8</b>	A	interference with assay
<b>9</b>	B	75%
<b>10</b>	-	reactant not purchasable
<b>11</b>	A	76%
<b>12</b>	B	87%
<b>13</b>	A	54%
<b>14</b>	B	28%

### 3.11 Biological testing results

For the 11 synthesized compounds binding activity values were determined for the targets of interest, LTA4H and sEH. The testing procedure is described in [sections 2.8 and 2.9](#) and was performed by Kerstin Hiesinger and Lilia Weizel. Compounds **1-6** from the PLIF fingerprint and compounds **7-14** from the 2D-fingerprint calculations are listed in [Table 11](#) with their established binding affinity values for LTA4H and sEH. In the following, three predicted and found to be dual active compounds are discussed based on the generated docking poses.

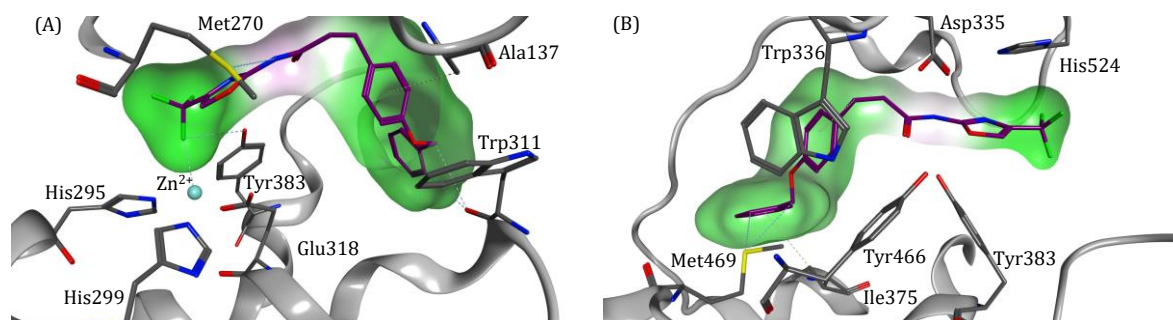
**Table 11: Inhibitory activity values of synthesized compounds.** Shown are the compound number, the compound structure and the IC<sub>50</sub> or % inhibition values on LTA4H and sEH.

Compound nr.	R	LTA4H (IC <sub>50</sub> or % inhibition)	sEH (IC <sub>50</sub> or % inhibition)
			
1		9% at 10 μM	3% at 10 μM
3		0.57 ± 0.08 μM	0.32 ± 0.01 μM
4		30% at 100 μM	54% at 100 μM
5		16% at 10 μM	1.3 ± 0.1 μM
6		4.7 ± 0.9 μM	4.2 ± 0.8 μM
7		4% at 10 μM	9.7 ± 3.5 μM
9		0.69 ± 0.08 μM	16.75 ± 1.11 μM
11		18.3 ± 0.8 μM	7% at 10 μM
12		0.75 ± 0.09 μM	0.5 ± 0.2 μM
13		0.67 ± 0.04 μM	4% at 10 μM
14		3.2 ± 0.6 μM	28.40 ± 0.97 μM

Receptor atoms and secondary structure elements are colored in grey, docked compounds are colored in dark magenta, compound surfaces are colored by lipophilicity (hydrophilic: magenta, neutral: white, hydrophobic: green) and the zinc ion in the LTA4H structures is colored in cyan. Contacts/interactions between receptor residues and the docked compounds are indicated by dashed lines (with an energy minimum of 0.2 kcal/mol). Interaction strength are calculated in MOE, they range from weak (< 0,5 kcal/mol) to medium (< 1,0 kcal/mol) and strong (> 1,0 kcal/mol) interactions. Hydrophobic contacts are calculated based on a geometric model, ionic interactions are calculated using R-Field electrostatics and hydrogen bonds are calculated by first determining donor and acceptor heavy atoms followed by analyzing the 3D-geometry between donor and acceptor heavy atoms.

Using the 3D-fingerprint PLIF for the prediction of novel dual active compounds, two selected and synthesized compounds show dual activity on the targets of interest (compound **3** and **6**). In general, the binding activity results show, that LTA4H is more tolerable concerning different variations of ring, substitution patterns and N-coupled lipophilic moiety as long it contains an ionizable tertiary amine.

Compound **3** shows very good dual activity of 0.57  $\mu\text{M}$  on LTA4H and 0.32  $\mu\text{M}$  on sEH. The docking poses of compound **3** in receptor LTA4H and sEH are shown in **Figure 32**. Interactions with an energy minimum of 0.2 kcal/mol between compound and receptor are indicated by dashed lines. In receptor LTA4H, compound **3** forms a strong interaction

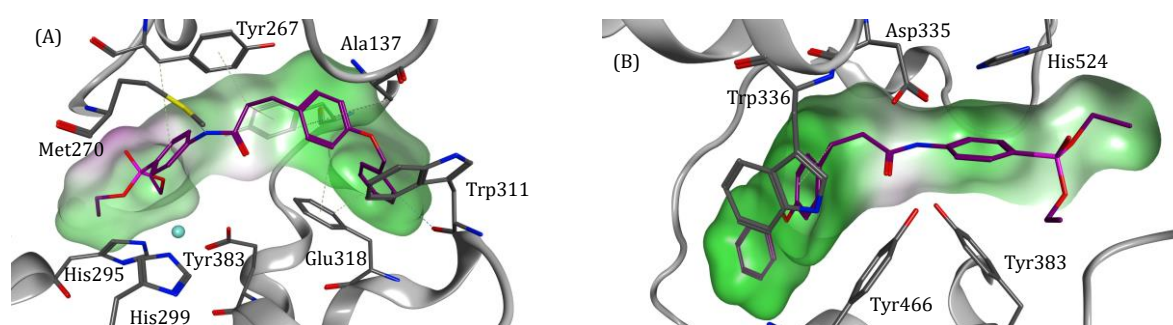


**Figure 32: Compound 3 docked in target structures LTA4H and sEH.** (A) LTA4H binding pocket showing the location of the Zn<sup>2+</sup> ion (cyan sphere), docked compound **3** colored in dark magenta as well as the lipophilicity surface area of compound **3** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line. (B) sEH binding pocket with docked compound **6** colored in dark magenta as well as the lipophilic surface area of compound **3** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line.



between the amide nitrogen atom and amino acid Met270 as well as a strong pi interaction between one of the aromatic rings of the key fragment C1 and amino acid Ala137 (**Figure 32 A**). A weak interaction between the CF<sub>3</sub> moiety and the zinc ion<sup>134</sup> and a weak interaction between residue Trp311 and the key fragment is recognized in the docking pose. Compound **3** is located deep in the hydrophobic pocket with the hydrophobic key fragment C1 (**Figure 32 A**, on the right). In receptor sEH (**Figure 32 B**), compound **3** forms one medium-strength pi interaction to amino acid Trp336. Compound **3** with its oxazole moiety shows, a previously unknown chemotype for LTA4H and sEH. This novel scaffold shows the great value and the potential this new method holds in the field of drug discovery.

Compound **6** has a moderate dual activity of 4.7 μM on LTA4H and 4.2 μM on sEH, respectively. The phosphonate ester moiety of compound **6** was already identified by Kim et al.<sup>135</sup> to be tolerated by sEH especially with alkyl groups on the phosphonate function, as existing in compound **6**. In receptor LTA4H, compound **6** forms a strong hydrogen bond to amino acid Met270 over the nitrogen atom of the amid bond. Further, all three aromatic rings form pi hydrogen bonds to different amino acids (weak: Tyr267, Glu318, Trp311; strong: Ala137). Analyzing the lipophilic surface of compound **6** in the LTA4H receptor structure (**Figure 33 A**), shows how well the compound fills the hydrophobic pocket on the right (see **Figure 7** for binding pocket surface) with the hydrophobic key fragment C1. The lipophilic part of the phosphonate moiety fits in the slightly lipophilic peptidase

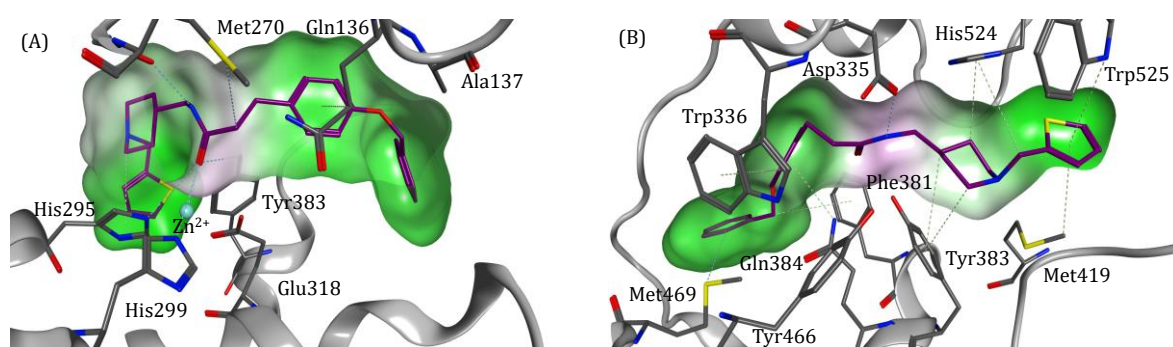


**Figure 33: Compound 6 docked in target structures LTA4H and sEH.** (A) LTA4H binding pocket showing the location of the Zn<sup>2+</sup> ion (cyan sphere), docked compound **6** colored in dark magenta as well as the lipophilicity surface area of compound **6** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line. (B) sEH binding pocket with docked compound **6** colored in dark magenta as well as the lipophilic surface area of compound **6** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line.

pocket on the left. **Figure 33 B** shows compound **6** docked in receptor structure sEH. A medium-strength pi interaction between compound **6** and amino acid residue Trp336 is recognized.

Using the 2D-AtomPair fingerprint for the prediction of novel dual active compounds, one selected and synthesized compound (compound **12**) shows dual activity on the targets of interest. Binding affinity data for sEH indicate, that sEH is more restrictive in hosting different ring and substitution patterns.

Compound **12** shows good dual activity of 0.75  $\mu\text{M}$  on LTA4H and 0.5  $\mu\text{M}$  on sEH. The docked pose of compound **12** in target LTA4H shows a strong interaction of the amide carbonyl to the zinc ion and amino acid residues His295 and Tyr383 (**Figure 34 A**). A weak pi interaction between the compound and amino acid Gln136 and a weak interaction to amino acid Met270 is recognized. The hydrophobic key fragment C1 is located in the hydrophobic pocket on the right (see **Figure 7** for binding pocket surface). **Figure 34 B** shows compound **12** in the target sEH. Both aromatic rings of the key fragment C1 form pi interactions to different amino acids (weak: Met469 and Phe381; medium: Gln384; strong: Trp336). The piperidine moiety shows a strong pi-hydrogen interaction with amino acid His524 and several weak interactions with residues Phe381, Met419 and Trp525 (**Figure 34 B**). A strong interaction between the amide nitrogen and Asp335 is recognized in the docking pose. The lipophilic surface area of compound **12** fits



**Figure 34: Compound 12 docked in target structures LTA4H and sEH.** (A) LTA4H binding pocket showing the location of the Zn<sup>2+</sup> ion (cyan sphere), docked compound **12** colored in dark magenta as well as the lipophilicity surface area of compound **12** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line. (B) sEH binding pocket with docked compound **6** colored in dark magenta as well as the lipophilic surface area of compound **12** (hydrophilic: magenta, neutral: white, hydrophobic: green); interactions between compound and receptor indicated by dashed line.

perfectly in the lipophilic surface area of the receptor (see **Figure 7** for binding pocket surface).

Comparing the results of the two different fingerprint strategies was one of the goals of this work. Both strategies (2D-fingerprints and 3D-fingerprints) result in the identification of novel dual active compounds. Compound **3** using the PLIF fingerprint and compound **12** using the AtomPair fingerprint show very good dual activity on both targets of interest, LTA4H and sEH. Binding affinities of 0.57  $\mu\text{M}$  on LTA4H and 0.32  $\mu\text{M}$  on sEH for compound **3** and 0.75  $\mu\text{M}$  on LTA4H and 0.5  $\mu\text{M}$  on sEH for compound **12** still offer room for optimization. The oxazole moiety of compound **3** is a previously unknown scaffold in LTA4H and sEH inhibition.

## 4 Conclusion

The goal of this work was to predict novel dual active compounds for the targets LTA4H and sEH. The prediction was realized using 2D- and 3D-fingerprints in combination with machine learning algorithms.

2D-fingerprints are solely based on the information of previously published active ligands. This ligand-based strategy is dependent on a large set of diverse active ligand structures. In case of this study, the data availability was adequate for LTA4H and sEH. Results are biased towards the previously identified chemotypes. A careful investigation is necessary to overcome this bias. We created a model that generalizes adequately using expert knowledge and manual investigation.

Using 2D-fingerprints has a major advantage over 3D-fingerprints regarding calculation time. The ligand-based strategy is very fast in generating results since only the ligand structure information is needed. 3D-fingerprints are based on the information contained in X-ray structures with various ligands. This structure-based strategy is less biased by chemotypes compared to 2D-fingerprints. The results of the 3D-fingerprint (PLIF) show that 3D-fingerprints open up the possibility to identify novel scaffolds (compound **3**). The structurally diverse compounds predicted in this work demonstrate this advantage in favor of 3D-fingerprints. 3D-fingerprints are limited by the requirement of 3D-poses of the ligands in the targets of interest, since docking introduces a multitude of error sources.

These error sources propagate through the following steps. The identification of docking poses, which are close to the experimentally determined binding mode is greatly flawed due to restrictions and simplifications of currently available docking software (e.g. rigid receptor, scoring functions; [section 2.1](#))<sup>25</sup> Docking is a very time consuming process, which counts among the disadvantages of using the structure-based approach.

Machine learning algorithms profit from a large amount of data. In the case of drug discovery the availability of active ligands is required to generate good and reliable results. Without numerous published active compounds, machine learning algorithms will possibly fail to predict novel active compounds.<sup>136</sup>

Target selection is a crucial step in the presented method. As described in the introduction ([section 1.4](#)), LTA4H and sEH interact with similar ligands (arachidonic acid epoxides), which leads to similar binding sites concerning the hydrophobicity patterns. It is unclear whether this new method is applicable to completely dissimilar targets. The analysis of abstraction to dissimilar targets is subject of further research.

The method developed in this work resulted in the identification of three novel dual potential lead compounds. Those compounds can serve as a starting point for further optimization regarding binding affinity, solubility and other pharmacological and physicochemical properties. Especially the identification of a novel scaffold (compound 3) as a dual active compound inhibiting LTA4H and sEH shows the potential this method holds.

## 5 Summary

The aim of this work was to establish a new way of predicting novel dual active compounds by combining classical fingerprint representation with state-of-the-art machine learning algorithms. Advantages and disadvantages of the applied 2D- and 3D-fingerprints were investigated. Further, the impact of various machine learning algorithms was analyzed. The new method developed in this work was used to predict compounds, which inhibit two different targets (LTA4H<sup>61</sup> and sEH<sup>66</sup>) involved in the same disease pattern (inflammation). The development of multitarget drugs has become more important in recent years. Many widespread diseases like metabolic syndrome, or cancer

are of a multifactorial nature, which makes them hard to be treated effectively with a single drug.<sup>71-76</sup> The new *in silico* method presented in this work can help to accelerate the design and development of multitarget drugs, saving time and efforts.

The nowadays readily available access to a large number of 3D-structures of biological targets and published activity data of millions of synthesized compounds enabled this study and was used as a starting point for this work. Four different data sets were compiled:

- crystalized ligands from the PDB<sup>12</sup> (LTA4H: 43 ligands, sEH: 94 ligands),
- active compounds from ChEMBL23<sup>103</sup> (LTA4H: 382 compounds, sEH: 1,384 compounds),
- inactive compounds from ChEMBL23<sup>103</sup> (LTA4H/sEH: 1,000 compounds),
- as well as newly designed compounds using a combinatorial library (20,630 compounds).

Those data sets were collected and processed using an automated KNIME<sup>104</sup> workflow. This automation has the advantage of allowing easy change and update of compound sources and adapted processing ways.

In a next step, the compounds from the compiled data sets were represented using a variety of well-established 2D- and 3D-fingerprints. All those fingerprints share the same underlying bit string scheme but vary in the way they describe the molecular structure. Especially the difference between 2D- and 3D-fingerprints was investigated. 2D-fingerprints are solely based on ligand information. 3D-fingerprints, on the other hand, are based on X-ray structure information of protein-ligand complexes. One major difference between 2D- and 3D-fingerprints usage is the need for a 3D-conformation (pose) of the compound in the targets of interest when using 3D-fingerprints.<sup>33,37</sup> This additional step is time-consuming and brings further uncertainties to the method. To investigate the impact of pose generation on the predictions made in the final step, two different docking software were used for the pose generation. The two docking software, MOE<sup>81</sup> and PLANTS<sup>80,97</sup>, were first validated on the crystalized ligands by conducting a re-docking of the ligands into their corresponding receptor structure. Overall, PLANTS docking had a higher docking accuracy compared to MOE. Looking at the targets of interest, PLANTS docking leads to 48.4% (LTA4H)/ 77.2% (sEH) correct docking poses, where on the other hand MOE docking reaches 46.5% (LTA4H)/ 55.4% (sEH).

Furthermore, calculation time speaks in favor of PLANTS docking, since docking the large combinatorial library is two times faster compared to MOE docking. Not only the time-consuming docking step, but also the actual fingerprint calculation is much slower for the 3D-fingerprint (PLIF)<sup>116</sup> compared to the 2D-fingerprints used in this work (AtomPair<sup>113</sup>, Morgan<sup>114</sup>, FeatMorgan<sup>29</sup> and MACCS<sup>28</sup>).

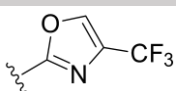
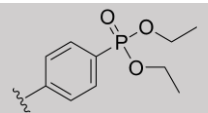
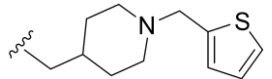
Based on the calculated fingerprints state-of-the-art machine learning algorithms were used to predict novel dual active compounds. Prior to the predictions, machine learning parameters were optimized. Parameter optimization was conducted on the data sets of crystalized ligands and active and inactive ChEMBL compounds (1,425 LTA4H compounds and 2,476 sEH compounds in total). First, the partitioning scheme was optimized (splitting the data set into training- and test sets) for each target and each fingerprint (PLIF (MOE, PLANTS), AtomPair, Morgan, FeatMorgan, MACCS) with each of the four machine learning algorithms (SVC<sup>45</sup>, RF<sup>49</sup>, XGB<sup>54-56</sup> and ADA<sup>54,55,57</sup>) resulting in 48 individual model optimizations. The models were evaluated by 10-fold cross validation and accuracy as the primary measure of model performance was maximized. Second, individual parameters of the four machine learning algorithms were optimized in a grid search to achieve maximal accuracy using the optimized partitioning scheme. Overall accuracies, regardless of fingerprint and machine learning algorithm, are slightly better for LTA4H than for sEH.

The results show that 2D-fingerprints can generally better distinguish between active and inactive compounds in 20% to 30% of the cases. Summarizing all the results show that there is a significant difference between 2D- and 3D-fingerprints but no significant difference using various machine learning algorithms. After optimization, the optimized partitioning schemes and parameters were used to predict possible active compounds for LTA4H and sEH from the combinatorial library data set. For each of the two targets, five fingerprints, two types of docking software and four machine learning algorithms optimized models were built and used for prediction, resulting in 48 different models.

The goal to predict dual active compounds was realized by comparing the set of predicted to be active compounds for LTA4H and sEH. For the 3D-fingerprint PLIF (MOE docking) the machine learning algorithm Random Forest was chosen, from which compounds for synthesis and testing were selected. Of 115 predicted to be active compounds, six compounds were cherry picked. Feasibility of synthesis, estimated solubility and

uniqueness of compounds were used as guidelines for cherry picking. The six synthesized compounds are structurally heterogeneous, but share an aromatic ring directly attached to the amide. From the six selected compounds one compound (compound **2**) could not be isolated from the reaction mixture. From the remaining five compounds, two compounds (compounds **3** and **6**) showed very good/moderate dual inhibitory activity (Table 12). Compound **3** contains an oxazole moiety, a previously unknown scaffold in LTA4H and sEH inhibition.

**Table 12: Inhibitory activity values of dual active compounds.** Shown are the compound number, the compound structure and the IC<sub>50</sub> values on LTA4H and sEH.

Compound nr.	R	LTA4H (IC <sub>50</sub> )	sEH (IC <sub>50</sub> )
Compound <b>3</b>		0.57 ± 0.08 μM	0.32 ± 0.01 μM
Compound <b>6</b>		4.7 ± 0.9 μM	4.2 ± 0.8 μM
Compound <b>12</b>		0.75 ± 0.09 μM	0.5 ± 0.2 μM

Of the 2D-fingerprints, the AtomPair fingerprint in combination with the machine learning algorithm Random Forest was chosen from which compounds were selected for synthesis and testing. The AtomPair fingerprint generated a reasonable number of predicted to be active compounds. Random Forest was chosen for better comparison with the 3D-fingerprint. 116 compounds were predicted to be dual active against LTA4H and sEH. Eight compounds were cherry picked for synthesis and testing. Two of those compounds could not be realized (compound **8** interfered with assay reagents; for compound **10** the reactant was not purchasable in a reasonable time period). All remaining compounds exhibit an N-substituted piperidine or pyrrolidine moiety, which are common elements in diverse series of LTA4H and sEH inhibitors. One of those compounds (compound **12**) showed good dual inhibitory activity (Table 12).

The three predicted novel dual active compounds were analyzed in detail using the generated docking poses in the receptor structures of LTA4H and sEH. All compounds

form strong interactions between the amide group and various amino acids. Further, compound **3** forms a strong pi interaction between one of the aromatic rings of the key fragment C1 and amino acid Ala137 (LTA4H) (**Figure 32 A**). Compound **6** forms a medium-strength pi interaction to amino acid Trp336 (sEH) (**Figure 33 B**). Compound **12** forms one strong interaction from the amide carbonyl to the zinc ion (**Figure 34 A**).

In this work it was possible to show advantages and disadvantages of using 2D- and 3D-fingerprints in combination with machine learning algorithms. Both strategies (2D: ligand-based, 3D: structure-based) lead to the prediction of novel dual active compounds with moderate to very good inhibitory activity (compounds **3,6** and **12**). 2D-fingerprints have an advantage in calculation time and solely the need of ligand structures. On the other hand, this ligand-based strategy is biased towards chemotypes and did not generate novel scaffolds. 3D-fingerprints are time consuming since a docking needs to be performed. However, 3D-fingerprints generated diverse dual active compounds with novel scaffolds and a compound with very good activity on both targets (compound **3**).

Regarding the machine learning algorithms, these show almost identical prediction accuracies but vary in amount of predicted active compounds. The Random Forest algorithm, as a very simple machine learning algorithm, generated a manually manageable number of dual active compound and led to the prediction of three novel dual active compounds. In general, machine learning algorithms profit from large data sets. Without enough data machine learning algorithms will possibly fail due to lack of diverseness in the data set.

The targets used in this work interact with similar ligands, which leads to similar binding sites concerning the hydrophobic patterns. Whether this method will work with two completely different targets is unclear and was out of scope of this work.

The method developed in this work is able to predict dual active compounds with very good inhibitory activity and novel (previously unknown) scaffolds inhibiting the targets LTA4H and sEH. This contribution to *in silico* drug design is promising and can be used for the prediction of novel dual active compounds. Those compounds can further be optimized regarding binding affinity, solubility and further pharmacological and physicochemical properties.



## 6 German Summary

Ziel dieser Arbeit ist es neuartige Verbindungen vorherzusagen, die nicht nur ein Einzelnes, sondern zugleich zwei unterschiedliche Proteine inhibieren. Die Zielproteine dieser Arbeit (Leukotrien A4 Hydrolase (LTA4H)<sup>61</sup> und lösliche Epoxid Hydrolase (sEH)<sup>66</sup>) befinden sich in der Arachidonsäure (AA) Kaskade und werden mit verschiedenen inflammatorischen Erkrankungen in Verbindung gebracht (z.B. Asthma, Rheumatoide Arthritis, Dermatitis und Atherosklerose).<sup>63</sup> Die AA Kaskade zeigt eine intensive Kommunikation zwischen den einzelnen Metabolisierungswegen. Die Inhibition von nur einem Metabolisierungsweg lässt den metabolischen Abbau von AA über die anderen beiden Metabolisierungswege zu. Dadurch werden positive Auswirkungen von verabreichten Wirkstoffen verringert. Werden jedoch zwei verschiedene Metabolisierungswege gleichzeitig von einem Wirkstoff inhibiert kann dieses Phänomen überwunden werden. Dies kann über die Gabe von mehreren Wirkstoffen oder einen Wirkstoff, der mehrere Proteine inhibiert erreicht werden (dualer Wirkstoff). Ein dualer Wirkstoff minimiert die Gefahr unvorhersehbarer Wirkstoffinteraktionen, die durch die Gabe von zwei verschiedenen Wirkstoffen hervorgerufen werden können.<sup>71-76</sup>

Mögliche Wirkstoff-Kandidaten sollen in dieser Arbeit durch eine neue *in silico* Methode vorhergesagt werden, wobei molekulare Fingerabdrücke mit Algorithmen des Maschinellen Lernens (ML) kombiniert werden.

Der heutzutage leichte Zugang zu einer großen Anzahl von biologischen 3D Strukturen (Proteine, RNA und DNA), sowie veröffentlichter Aktivitätsdaten von Millionen von synthetisierten Verbindungen, ermöglichten diese Arbeit und wurden als Ausgangspunkt verwendet. Aus öffentlich zugänglichen Datenbanken wurden folgende Datensätze zusammengestellt:

- Co-kristallisierte Liganden aus der PDB<sup>12</sup> (LTA4H: 43 Liganden, sEH: 94 Liganden),
- Aktive Verbindungen aus der ChEMBL<sup>23103</sup> (LTA4H: 382 Verbindungen, sEH: 1.384 Verbindungen),
- Inaktive Verbindungen aus der ChEMBL<sup>23103</sup> (LTA4H/sEH: 1.000 Verbindungen),

- Neu designte Verbindungen aus einer kombinatorischen Bibliothek (20.630 Verbindungen).

Über einen automatisierten KNIME<sup>104</sup> Workflow wurden diese Datensätze zusammengestellt und weiterverarbeitet. Die Automatisierung hat den Vorteil, dass Änderungen und Updates von Datenquellen einfach zu integrieren sind. Des Weiteren ist ein Anpassen des Bearbeitungsprozesses möglich.

Neu designte Verbindungen wurden über eine kombinatorische Bibliothek generiert. Eine kombinatorische Bibliothek ist das Ergebnis kombinatorischer Chemie, wobei über eine chemische Reaktion eine Vielzahl von strukturähnlichen Verbindungen erzeugt wird. Diese Bibliotheken können durch chemische Synthese, oder wie in diesem Fall virtuell, mit Hilfe einer Software generiert werden. In dieser Arbeit wurde eine kombinatorische Bibliothek speziell zur Identifizierung neuer dual aktiver Verbindungen für LTA4H und sEH aufgebaut. Ein Schlüsselfragment, (3-[4-(Benzyloxy)Phenyl]Propionsäure), wurde konstant gehalten und mit Amin-Bindungspartnern über eine Amid-Kondensation kombiniert. Die resultierende Bibliothek an strukturverwandten Amiden beinhaltet 20,630 einzigartige Verbindungen.

Im nächsten Schritt wurden die Verbindungen in den zusammengestellten Datensätzen mit Hilfe einer Auswahl an 2D- und 3D-Fingerabdrücken dargestellt. Das heißt, die molekularen Strukturen der Verbindungen wurden in ein Bit-String Schema übersetzt. Die 2D- und 3D-Fingerabdrücke unterscheiden sich in der Art und Weise wie sie die molekularen Strukturen beschreiben. In dieser Arbeit wurden speziell die Unterschiede zwischen 2D- und 3D-Fingerabdrücken untersucht. 2D-Fingerabdrücke basieren ausschließlich auf Ligandinformationen (Ligand-basierter Ansatz); 3D-Fingerabdrücke basieren auf 3D-Strukturinformationen von Protein-Ligand Komplexen (strukturbasierter Ansatz). 3D-Fingerabdrücke benötigen also eine 3D-Konformation (Pose) der Verbindungen in den Zielproteinen.<sup>33,37</sup> Die Erstellung von 3D-Konformationen wird mit Hilfe von Docking (virtuelles Platzieren der Verbindungen in den Zielproteinen) realisiert, was jedoch ein zeitaufwändiger und fehleranfälliger Schritt ist. Um den Einfluss der generierten 3D-Konformationen auf die abschließende Vorhersage neuer Verbindungen zu analysieren, wurden zwei unterschiedliche Docking-Programme (MOE<sup>81</sup> und PLANTS<sup>80,97</sup>) für die Generierung der Posen verwendet und verglichen. Diese Docking-Programme wurden zunächst validiert, indem ein sogenanntes

Re-Docking durchgeführt wurde. Dabei werden die bereits Co-kristallisierten Liganden mit Hilfe der Docking-Programme in ihre entsprechenden Proteine gedockt und die Reproduktion der experimentell bestimmten Bindemodi ausgewertet. Insgesamt hat PLANTS eine höhere Docking-Genauigkeit im Vergleich zu MOE. PLANTS Docking führte zu 48,4 % (LTA4H)/ 77,2 % (sEH) richtiger Docking-Posen (innerhalb eines RMSD von 1Å), wohingegen MOE lediglich 46,5 % (LTA4H)/ 55,4 % (sEH) erreicht. Auch die Berechnungszeit der Docking-Posen spricht für das PLANTS- Docking, da die Generierung von 3D-Konformationen der kombinatorischen Bibliothek doppelt so schnell wie MOE-Docking ist. Des Weiteren ist auch die eigentliche Berechnung des 3D-Fingerabdrucks (PLIF)<sup>116</sup> zeitaufwändiger als die Berechnung der verschiedenen 2D-Fingerabdrücke (AtomPair<sup>113</sup>, Morgan<sup>114</sup>, FeatMorgan<sup>29</sup> und MACCS<sup>28</sup>).

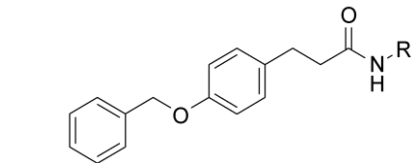
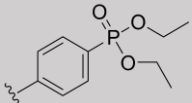
Nach der Generierung der Fingerabdrücke (KNIME Workflow) wurden Algorithmen des Maschinellen Lernens verwendet, um neue dual aktive Verbindungen vorherzusagen. Zunächst wurden die Parameter für jeden der verwendeten Algorithmen (SVC<sup>45</sup>, RF<sup>49</sup>, XGB<sup>54-56</sup> und ADA<sup>54,55,57</sup>) optimiert. Die Optimierung wurde mit den Datensätzen der kristallisierten Liganden, sowie aktiven und inaktiven ChEMBL Verbindungen durchgeführt (Gesamt: 1.425 LTA4H Verbindungen, 2.476 sEH Verbindungen). Als erstes wurde das Teilungsschema (Teilung des Datensatzes in Trainings- und Testdatensatz) für jedes Zielprotein und jeden Fingerabdruck (PLIF|MOE, PLIF|PLANTS, AtomPair, Morgan, FeatMorgan, MACCS) mit jedem der vier verwendeten ML Algorithmen optimiert. Daraus resultieren 48 individuelle Modell-Optimierungen. Die Modelle wurden mittels 10-facher Überkreuzvalidierung evaluiert. Der Grad der Genauigkeit als primärer Leistungsmesser wurde durch die Modell-Optimierung maximiert. Als zweites wurden individuelle Parameter der verwendeten ML Algorithmen optimiert. Eine Rastersuche wurde durchgeführt, um eine maximale Leistungsfähigkeit der Modelle zu erreichen. Allgemein kann festgestellt werden, dass unabhängig von Fingerabdruck und ML Algorithmen, bessere Ergebnisse für das Zielprotein LTA4H erzielt werden.

Weiterhin zeigen die Resultate, dass in 20-30 % der Fälle, 2D-Fingerabdrücke besser zwischen aktiven und inaktiven Verbindungen unterscheiden können. Zusammenfassend lässt sich sagen, dass es einen signifikanten Unterschied in der Vorhersagegenauigkeit zwischen 2D- und 3D-Fingerabdrücken gibt, jedoch keinen signifikanten Unterschied unter den verschiedenen ML Algorithmen.

Anhand der Fingerabdrücke wurden die Modelle trainiert aktive Verbindungen von inaktiven zu unterscheiden. Die optimierten Teilungsschemata und Parameter wurden für die Vorhersage neuer möglicher aktiver Verbindungen genutzt. Dazu wurden die neu designten Verbindungen der kombinatorischen Bibliothek genutzt. Für die zwei Zielproteine, die fünf Fingerabdrücke, die zwei Docking-Programme und die vier ML Algorithmen wurde jeweils ein Modell aufgebaut (48 unterschiedliche Modelle).

Das Ziel, dual aktive Verbindungen vorherzusagen, wurde erreicht, indem die Sets von aktiv vorhergesagten Verbindungen für LTA4H und sEH miteinander verglichen wurden. Für den 3D-Fingerabdruck PLIF (MOE Docking) wurde der ML Algorithmus Random Forest ausgewählt. Von den dual aktiv vorhergesagten Verbindungen (insgesamt 115 Verbindungen) wurden sechs Verbindungen für die Synthese und biologische Testung ausgewählt. Die Kriterien der Auswahl waren: Durchführbarkeit der Synthese, abgeschätzte Löslichkeit und Einzigartigkeit der Verbindung. Die sechs ausgewählten Verbindungen sind strukturell heterogen. Alle Verbindungen verfügen jedoch über einen aromatischen Ring, welcher direkt mit dem Amid verbunden ist. Eine der Verbindungen konnte nicht vom Reaktionsgemisch isoliert werden, für die restlichen fünf Verbindungen wurden inhibitorische Aktivitäten gemessen. Zwei dieser Verbindungen (Verbindung **3** und **6**) zeigen auf beiden Zielproteinen eine sehr gute/moderate inhibitorische Aktivität (Table 12). Zudem enthält Verbindung **3** ein Oxazol-Motiv, ein bis dato unbekanntes Strukturelement in der Inhibition von LTA4H und sEH. Von den 2D-Fingerabdrücken wurde der AtomPair-Fingerabdruck in Kombination mit dem ML Algorithmus Random

**Table 132: Inhibitory activity values of dual active compounds.** Shown are the compound number, the compound structure and the IC50 values on LTA4H and sEH.

Compound nr.	R	LTA4H (IC50)	sEH (IC50)
Compound 3		0.57 ± 0.08 μM	0.32 ± 0.01 μM
Compound 6		4.7 ± 0.9 μM	4.2 ± 0.8 μM

Forest ausgewählt. Von den dual aktiv vorhergesagten Verbindungen (insgesamt 116 Verbindungen) wurden acht Verbindungen für Synthese und biologische Testung ausgewählt. Zwei der Verbindungen konnten nicht synthetisiert werden (Verbindung **8** und **10**). Alle verbleibenden Verbindungen beinhalten ein N-substituiertes Piperidin oder Pyrrolidin-Motiv, welche bekannte Elemente in verschiedenen LTA4H und sEH Inhibitoren sind. Einer der synthetisierten Verbindungen (Verbindung **12**) zeigt gute inhibitorische Aktivität auf LTA4H und sEH. Die drei neu vorhergesagten dual aktiven Verbindungen wurden anhand der generierten Docking-Posen in den Zielproteinen LTA4H und sEH genauer analysiert. Alle Verbindungen formen starke Interaktionen zwischen der Amid-Gruppe und unterschiedlichen Aminosäuren innerhalb der Zielproteine. Weitere geformte Interaktionen stellen sich wie folgt dar:

- Verbindung **3** formt eine starke Pi Interaktion zwischen einem der aromatischen Ringe des Schlüsselfragment C1 und der Aminosäure Ala137 (LTA4H) (**Figure 32 A**),
- Verbindung **6** formt eine mittelstarke Pi Interaktion zu Aminosäure Trp336 (sEH) (**Figure 33 B**),
- Verbindung **12** formt eine starke Interaktion vom Amid Carbonyl zu dem Zink Ion in der Proteinstruktur LTA4H (**Figure 34 A**).

Durch diese Arbeit war es möglich Vor- und Nachteile der 2D- und 3D-Fingerabdrücke in Kombination mit ML Algorithmen aufzuzeigen. Beide Strategien (2D: Ligand-basiert, 3D: strukturbasiert) resultieren in der Vorhersage von neuartigen dual aktiven Verbindungen mit moderater bis sehr guter inhibitorischer Aktivität (Verbindung **3**, **6** und **12**) auf den Zielproteinen (LTA4H und sEH). 2D-Fingerabdrücke haben den Vorteil, dass nur Ligand-Strukturen benötigt werden und eine schnelle Berechnung möglich ist. Auf der anderen Seite ist die Ligand-basierte Strategie voreingenommen gegenüber bestimmter Chemotypen und erzeugte keine neuartigen Strukturmerkmale in den vorhergesagten Verbindungen. Das Verwenden von 3D-Fingerabdrücken ist sehr zeitaufwändig, da ein Docking zur Generierung von Docking-Posen durchgeführt werden muss. Jedoch resultieren aus den 3D-Fingerabdrücken diverse dual aktive Verbindungen mit neuen Strukturelementen und sehr guter inhibitorischer Aktivität auf beiden Zielproteinen (Verbindung **3**).

Die unterschiedlichen ML Algorithmen zeigen kaum Unterschiede in der Genauigkeit der Vorhersagen, jedoch hatten sie große Differenzen in Bezug auf die Anzahl an aktiv vorhergesagten Verbindungen (Variation zwischen 0 Verbindungen und 20.000 Verbindungen). Random Forest, als ein sehr einfacher ML Algorithmus generierte eine manuell verarbeitbare Anzahl an aktiv vorhergesagten Verbindungen und resultierte in der Vorhersage von drei neuartigen dual aktiven Verbindungen. Allgemein profitieren alle ML Algorithmen von großen Datensätzen. Eine zu geringe Datenmenge führt wahrscheinlich zu einem Versagen der ML Algorithmen.

Die in dieser Arbeit verwendeten Zielproteine interagieren mit ähnlichen natürlichen Liganden, was zu einer ähnlichen Bindetasche führt (in Bezug auf hydrophobe Muster in der Bindetasche). Ob die hier vorgestellte Methode auch mit sehr unterschiedlichen Zielproteinen funktioniert lag außerhalb des Umfangs dieser Arbeit.

Die hier entwickelte *in silico* Methode ist in der Lage dual aktive Verbindungen mit sehr guter inhibitorischer Aktivität vorherzusagen. Auch zuvor unbekannte Strukturelemente in Bezug auf die Inhibition von LTA4H und sEH konnten generiert werden. Dieser Beitrag zu *in silico* Wirkstoffdesign ist vielversprechend und kann für die weiteren Vorhersagen dual aktiver Verbindungen genutzt werden. Die hier vorhergesagten Verbindungen können in weiteren Studien in Bezug auf ihre Bindungsaffinität, Löslichkeit und weiterer pharmakologischer und physikalisch-chemischer Eigenschaften optimiert werden.

## 7 Appendix

### 7.1 Python code for ML partitioning scheme optimization

```
#Open data set with active and inactive compounds (LTA4H: 1,425 compounds;  
sEH: 2,476 compounds)
```

```
data = pd.read_csv(target_fp.csv)
```

```
#Split data set into training and test set. x_train contains fingerprints of training data  
set; x_test contains fingerprints of test data set; y_train contains class labels of  
training data set; y_test contains class labels of test data set; test_size specifies size  
of test data set to be between 5% and 25%
```

```
x_train, x_test, y_train, y_test = train_test_split(data,  
test_size=0.05-0.25)
```

```
#mla (machine learning algorithm) specifies the machine learning algorithm used to train  
the model
```

```
mla=SVC()/RandomForestClassifier()/XGBClassifier()/AdaBoostClassifier()
```

```
#Model is build using the training data set (containing fingerprints and class labels)
```

```
mla.fit(x_train,y_train)
```

```
#Build model uses fingerprints of test set (x_test) to make class label prediction
```

```
predict = mla.predict(x_test)
```

```
#Model accuracy is calculated. Class label predictions (predict) are compared with true  
class labels of the test data set (y_test)
```

```
acc_mla = accuracy_score(y_test, predict)
```

## 7.2 Python code for ML parameter optimization

### *Random Forest parameter optimization:*

Number of estimators corresponds to the number of trees in the forest.

- (1) #Dictionary is generated
- (2) #Grid search range for number of estimators is defined starting from 10 to 1000 in steps of 10
- (3) #Random Forest model is trained using training data set
- (4) #Model is used to predict class labels of test data set
- (5) #Model accuracy is calculated
- (6) #Model accuracy is written into dictionary

```
(1) dic = {}
(2) for n in range(10,1001,10):
    rfc = RandomForestClassifier(n_estimators = n)
(3)     rfc.fit(x_train, y_train)
(4)     predict = rfc.predict(x_test)
(5)     acc_rfc = accuracy_score(y_test, predict)
(6)     dic[(n)] = acc_rfc
```

### *XGBoost parameter optimization:*

The maximum depth limits the number of nodes in the tree. Number of estimators defines the number of boosting stages to perform.

- (1) #Dictionary is generated
- (2) #Grid search range for maximum depth limits is defined starting from 10 to 200 in steps of 10. Grid search range for number of estimators is defined starting from 100 to 1000 in steps of 100. Learning rate values were set to 0.01 and 0.001. Alpha regulates the weights, values were set to 0.0 (default) and 0.005.
- (3) #XGBoost model is trained using training data set
- (4) #Model is used to predict class labels of test data set
- (5) #Model accuracy is calculated
- (6) #Model accuracy is written into dictionary

```
(1) dic = {}
(2) for n in range(10,201,10):
    for esti in range(100,1001,100):
        xgb = XGBClassifier(max_depth = n,
            learning_rate=0.01/0.001, n_estimators=esti,
            reg_alpha=0.0/0.005)
(3)     xgb.fit(X_train, y_train)
(4)     predict = xgb.predict(x_test)
(5)     acc_xgb = accuracy_score(y_test, predict)
(6)     dic[(n)] = acc_xgb
```



### *AdaBoost parameter optimization:*

Number of estimators defines when boosting is terminated.

- (1) #Dictionary is generated
- (2) #Grid search range for number of estimators is defined starting from 10 to 1000 in steps of 10
- (3) #Random Forest model is trained using training data set
- (4) #Model is used to predict class labels of test data set
- (5) #Model accuracy is calculated
- (6) #Model accuracy is written into dictionary

```
(1) dic = {}
(2) for n in range(10,1001,10):
      ada = AdaBoostClassifier(n_estimators = n)
(3)     ada.fit(x_train, y_train)
(4)     predict = ada.predict(x_test)
(5)     acc_ada = accuracy_score(y_test, predict)
(6)     dic[(n)] = acc_ada
```

### 7.3 Python code for ML prediction

#Models are trained using the training set (optimized parameters), predictions are made on test set and accuracy is calculated.

```
m1a=SVC()/RandomForestClassifier()/XGBClassifier()/AdaBoostClassifier()
m1a.fit(x_train,y_train)
pred = m1a.predict(x_test)
acc_m1a = accuracy_score(y_test, pred)
```

#10-fold cross validation is conducted on trained models. Mean and standard deviation is calculated on 10-fold cross validation.

```
cv_m1a = cross_val_score(m1a, x_train, y_train, cv=10,
scoring="accuracy")
mean_cv_m1a = cv_score_m1a.mean()
sd_cv_m1a = cv_score_m1a.std()
```

#Trained models are used to predict class labels on combinatorial library compounds. Predictions are labeled active and inactive and converted into a panda DataFrame.

```
predict = pickle_model.predict(combi_lib)
predict.pred_activity[predict.pred_activity == 1] = 'inactive'
```

```
predict.pred_activity[predict.pred_activity == 0] = 'active'  
predicted_mla = pd.DataFrame(predict)  
predict_mla.to_excel(path)
```

Table A1: MOE docking validation results LTA4H.

Scoring function combination	RMSD [Å]	%	RMSD [Å]	%
ASE_ASE	<1	41.9	<2	67.4
ASE_Affinity dG	<1	32.6	<2	55.8
ASE_Alpha HB	<1	41.9	<2	67.4
ASE_GBVI/WSA dG	<1	46.5	<2	69.8
ASE_London dG	<1	46.5	<2	69.8
Affinity dG_ASE	<1	34.9	<2	62.8
AffinitydG_Affinity dG	<1	20.9	<2	46.5
Affinity dG_Alpha HB	<1	32.6	<2	62.8
Affinity dG_GBVI/WSA dG	<1	37.2	<2	60.5
Affinity dG_London dG	<1	37.2	<2	60.5
Alpha HB_ASE	<1	32.6	<2	60.5
Alpha HB_Affinity dG	<1	23.3	<2	44.2
Alpha HB_Alpha HB	<1	34.9	<2	62.8
Alpha HB_GBVI/WSA dG	<1	37.2	<2	60.5
Alpha HB_London dG	<1	37.2	<2	62.8
GBVI/WSA dG_ASE	<1	41.9	<2	65.1
GBVI/WSA dG_Affinity dG	<1	27.9	<2	48.8
GBVI/WSA dG_Alpha HB	<1	39.5	<2	67.4
GBVI/WSA dG_GBVI/WSA dG	<1	44.2	<2	65.1
GBVI/WSA dG_London dG	<1	46.5	<2	60.5
London dG_ASE	<1	41.9	<2	72.1
London dG_Affinity dG	<1	32.6	<2	55.8
London dG_Alpha HB	<1	44.2	<2	72.1
London dG_GBVI/WSA dG	<1	46.5	<2	69.8
London dG_London dG	<1	44.2	<2	65.1

Table A2: MOE docking validation results sEH.

Scoring function combination	RMSD [Å]	%	RMSD [Å]	%
ASE_ASE	<1	48.9	<2	71.1
ASE_Affinity dG	<1	43.5	<2	67.4
ASE_Alpha HB	<1	46.7	<2	70.7
ASE_GBVI/WSA dG	<1	50	<2	77.2
ASE_London dG	<1	52.2	<2	72.8
Affinity dG_ASE	<1	45.7	<2	66.3
AffinitydG_Affinity dG	<1	41.3	<2	65.2
Affinity dG_Alpha HB	<1	42.4	<2	60.9
Affinity dG_GBVI/WSA dG	<1	50	<2	73.9
Affinity dG_London dG	<1	48.9	<2	69.6
Alpha HB_ASE	<1	46.7	<2	65.2
Alpha HB_Affinity dG	<1	41.3	<2	58.7
Alpha HB_Alpha HB	<1	46.7	<2	64.1
Alpha HB_GBVI/WSA dG	<1	45.7	<2	64.1
Alpha HB_London dG	<1	48.9	<2	66.3
GBVI/WSA dG_ASE	<1	47.8	<2	70.7
GBVI/WSA dG_Affinity dG	<1	44.6	<2	64.1
GBVI/WSA dG_Alpha HB	<1	50	<2	70.7
GBVI/WSA dG_GBVI/WSA dG	<1	52.2	<2	73.9
GBVI/WSA dG_London dG	<1	52.2	<2	72.8
London dG_ASE	<1	53.3	<2	73.9
London dG_Affinity dG	<1	47.8	<2	72.8
London dG_Alpha HB	<1	51.1	<2	71.7
London dG_GBVI/WSA dG	<1	55.4	<2	79.3
London dG_London dG	<1	56.5	<2	73.9

Table A3: PLANTS docking validation result LTA4H.

Scoring function	Radius [Å]	RMSD [Å]	%	RMSD [Å]	%
<b>CHEMPLP</b>	10	≤ 1	9.3	≤ 2	27.9
	15	≤ 1	48.8	≤ 2	90.7
	20	≤ 1	44.2	≤ 2	83.7
	25	≤ 1	30.2	≤ 2	69.8
<b>PLP</b>	10	≤ 1	7.0	≤ 2	20.9
	15	≤ 1	44.2	≤ 2	86.1
	20	≤ 1	30.2	≤ 2	83.7
	25	≤ 1	27.9	≤ 2	83.1
<b>PLP95</b>	10	≤ 1	9.3	≤ 2	25.6
	15	≤ 1	30.2	≤ 2	86.1
	20	≤ 1	25.6	≤ 2	93.0
	25	≤ 1	30.2	≤ 2	88.4

Table A4: PLANTS docking validation result sEH.

Scoring function	Radius [Å]	RMSD [Å]	%	RMSD [Å]	%
<b>CHEMPLP</b>	10	≤ 1	64.1	≤ 2	85.9
	15	≤ 1	77.2	≤ 2	95.7
	20	≤ 1	73.9	≤ 2	95.7
	25	≤ 1	76.1	≤ 2	95.7
<b>PLP</b>	10	≤ 1	67.4	≤ 2	84.8
	15	≤ 1	75.0	≤ 2	94.6
	20	≤ 1	76.1	≤ 2	94.6
	25	≤ 1	75.0	≤ 2	94.6
<b>PLP95</b>	10	≤ 1	60.9	≤ 2	84.8
	15	≤ 1	68.5	≤ 2	91.3
	20	≤ 1	65.2	≤ 2	90.2
	25	≤ 1	66.3	≤ 2	91.3

## 8 Bibliography

1. Huang, S.-Y. & Zou, X. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.* **11**, 3016–3034 (2010).
2. Lionta, E., Spyrou, G., Vassilatis, D. & Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **14**, 1923–1938 (2014).
3. Baldwin, J. *et al.* High-throughput screening for potent and selective inhibitors of Plasmodium falciparum dihydroorotate dehydrogenase. *J. Biol. Chem.* **280**, 21847–21853 (2005).
4. Polgar, T. & M. Keseru, G. Integration of Virtual and High Throughput Screening in Lead Discovery Settings. *Comb. Chem. High Throughput Screen.* **14**, 889–897 (2011).
5. Siles, S. A., Srinivasan, A., Pierce, C. G., Lopez-Ribot, J. L. & Ramasubramanian, A. K. High-Throughput Screening of a Collection of Known Pharmacologically Active Small Compounds for Identification of Candida albicans Biofilm Inhibitors. *Antimicrob. Agents Chemother.* **57**, 3681–3687 (2013).
6. Brideau, C., Gunter, B., Pikounis, B. & Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8**, 634–647 (2003).
7. Tanrikulu, Y., Krüger, B. & Proschak, E. The holistic integration of virtual screening in drug discovery. *Drug Discov. Today* **18**, 358–364 (2013).
8. Ferreira, L., dos Santos, R., Oliva, G. & Andricopulo, A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **20**, 13384–13421 (2015).
9. Keseru, G. M. & Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **11**, 741–748 (2006).
10. Patani, G. A. & LaVoie, E. J. Bioisosterism: A rational approach in drug design. *Chem. Rev.* **96**, 3147–3176 (1996).
11. Friedman, H. Influence of Isosteric Replacements upon Biological Activity. *Symp. Chem. Correl.* **206**, 295–358 (1951).
12. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–42 (2000).
13. Kuhnert, M. & Diederich, W. Structure-Based Drug Design in Medicinal Chemistry: The Devil is in the Detail. *Synlett* **27**, 641–649 (2016).
14. Kleywegt, G. J. & Jones, T. A. Model building and refinement practice. *Methods Enzymol.* **277**, 208–230 (1997).

15. Gramaccioni, C. M. *et al.* Atomic Displacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallogr. Sect. A Found. Crystallogr.* **52**, 770–781 (2002).
16. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
17. RCSB-PDB. PDB-101: Guide to Understanding PDB Data. Available at: <https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction>.
18. Davis, A. M., Teague, S. J. & Kleywegt, G. J. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design. *Angew. Chemie - Int. Ed.* **42**, 2718–2736 (2003).
19. Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A. & Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov. Today* **17**, 1270–1281 (2012).
20. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. A review of protein-small molecule docking methods. *J. of Computer-Aided Mol. Des.* **16**, 151–166 (2002).
21. Leach, A. R., Shoichet, B. K. & Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **49**, 5851–5855 (2006).
22. Keserü, G. M. & Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* **8**, 203–212 (2009).
23. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: II. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **54**, 1717–1736 (2014).
24. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079–93 (2009).
25. Kalinowsky, L., Weber, J., Balasubramanian, S., Baumann, K. & Proschak, E. A Diverse Benchmark Based on 3D Matched Molecular Pairs for Validating Scoring Functions. *ACS Omega* **3**, 5704–5714 (2018).
26. Open-Source Cheminformatics Software. RDKit2018.09.1 documentation. (2018). Available at: <https://www.rdkit.org/docs/index.html>.
27. MDL Information Systems/Symyx. *MACCS-II*. (1984).
28. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. **71**, 58–63 (2015).
29. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**,

- 742–754 (2010).
30. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).
  31. Hu, Y., Lounkine, E. & Bajorath, J. Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density- Dependent Similarity Function. *ChemMedChem* **4**, 540–548 (2009).
  32. Xue, L. & Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. High Throughput Screen.* **3**, 363–372 (2000).
  33. Lo, Y., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
  34. Tanimoto, T. T. An elementary mathematical theory of classification and prediction. in (International Business Machines Corporation, 1958).
  35. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **29**, 157–170 (2010).
  36. Sliwoski, G., Mendenhall, J. & Meiler, J. Autocorrelation descriptor improvements for QSAR : 2DA \_ Sign and 3DA \_ Sign. *J. Comput. Aided. Mol. Des.* **30**, 209–217 (2016).
  37. Bajorath, J. Selected Concepts and Investigations in Compound Classification , Molecular Descriptor Analysis , and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **41**, 233–245 (2001).
  38. Verma, J., Khedkar, V. M. & Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **10**, 95–115 (2010).
  39. Awale, M., Jin, X. & Reymond, J. L. Stereoselective virtual screening of the ZINC database using atom pair 3D-fingerprints. *J. Cheminform.* **7**, 1–15 (2015).
  40. Brownlee, J. How Machine Learning Algorithms Work (they learn a mapping of input to output). (2016). Available at: <https://machinelearningmastery.com/how-machine-learning-algorithms-work/>.
  41. Difference between Supervised and Unsupervised Learning. Available at: <https://www.geeksforgeeks.org/difference-between-supervised-and-unsupervised-learning/>.



42. Regression and Classification | Supervised Machine Learning. Available at: <https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/>.
43. Fumo, D. Classification Versus Regression — Intro To Machine Learning #5. (2017). Available at: <https://medium.com/simple-ai/classification-versus-regression-intro-to-machine-learning-5-5566efd4cb83>.
44. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today* **20**, 318–331 (2015).
45. Vapnik, V., Golowich, S. E. & Smola, A. *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing*. (1997).
46. Vapnik, V. *The Nature of Statistical Learning Theory*. (Springer, 1995).
47. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **36**, 1171–1220 (2008).
48. Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A. & Rückert, D. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage* **65C**, 167–175 (2013).
49. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
50. Anyanwu, M. N. & Shiva, S. G. Comparative Analysis of Serial Decision Tree Classification Algorithms. *Int. J. Comput. Sci. Secur.* **3**, 230–240 (2009).
51. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **31**, 249.268 (2007).
52. Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
53. Parr, T. & Howard, J. How to explain gradient boosting. Available at: <https://explained.ai/gradient-boosting/index.html>.
54. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407 (2002).
55. Freund, Y. & Schapire, R. E. A Short Introduction to Boosting. *Comptes rendus l'Academie des Sci. Ser. III, Sci. la vie* **14**, 771–780 (1999).
56. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).
57. Schapire, R. E. & Freund, Y. A Decision-Theoretic Generalization of On-Line

- Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
58. Piomelli, D. Arachidonic acid in cell signaling. *Curr. Opin. Cell Biol.* **2**, 274–280 (1993).
  59. Zeldin, D. C. Epoxygenase Pathways of Arachidonic Acid Metabolism \*. *J. Biol. Chem.* **276**, 36059–36062 (2001).
  60. Glover, S., Timothy, B., Jonas, M., Chi, E. & Gelb, M. H. Translocation of the 85-kDa phospholipase A2 from cytosol to the nuclear envelope in rat basophilic leukemia cells stimulated with calcium ionophore or IgE/antigen. *J. Biol. Chem.* **270**, 15359–15367 (1995).
  61. Haeggström, J. Z., Wetterholm, A., Shapiro, R., Vallee, B. L. & Samuelsson, B. Leukotriene A4 hydrolase: A zinc metalloenzyme. *Biochem. Biophys. Res. Commun.* **172**, 965–970 (1990).
  62. Imig, J. D. & Hammock, B. D. Soluble Epoxide Hydrolase as a Therapeutic Target for Cardopvascular Diseases. *Nat Rev Drug Discov.* **8**, 794–805 (2009).
  63. Haeggström, J. Z. & Funk, C. D. Lipoxygenase and Leukotriene Pathways : Biochemistry , Biology , and Roles in Disease. *Chem. Rev.* **111**, 5866–5898 (2011).
  64. Wetterholm A., Blomster M., H. J. Z. Leukotriene A4 Hydrolase: A Key Enzyme in the Biosynthesis of Leukotriene B4. in *Eicosanoids* 1–12 (Springer, 1996). doi:DOI [https://doi.org/10.1007/978-1-4899-0200-9\\_1](https://doi.org/10.1007/978-1-4899-0200-9_1)
  65. Snelgrove, R. J. *et al.* A critical role for LTA4 H in limiting chronic pulmonary neutrophilic inflammation. *Science (80-. )*. **330**, 90–94 (2011).
  66. Shen, H. C. & Hammock, B. D. Discovery of Inhibitors of Soluble Epoxide Hydrolase : A Target with Multiple Potential Therapeutic Indications. *J. Med. Chem.* **55**, 1789–1808 (2012).
  67. Newman, J. W., Morisseau, C. & Hammock, B. D. Epoxide hydrolases : their roles and interactions with lipid metabolism. *Prog. Lipid Res.* **44**, 1–51 (2005).
  68. Amin, A. R., Attur, M. G., Pillinger, M. & Abramson, S. B. The pleiotropic functions of aspirin: Mechanisms of action. *Cell. Mol. Life Sci.* **56**, 305–312 (1999).
  69. Varjabedian, L., Bourji, M., Pourafkari, L. & Nader, N. D. Cardioprotection by Metformin: Beneficial Effects Beyond Glucose Reduction. *Am. J. Cardiovasc. Drugs* **18**, 181–193 (2018).
  70. Rogliani, P., Ora, J., Di Daniele, N. & Lauro, D. Pleiotropic effects of hypoglycemic agents: implications in asthma and COPD. *Curr. Opin. Pharmacol.* **40**, 34–38 (2018).

71. Kaur, G. & Silakari, O. Multiple target-centric strategy to tame inflammation. *Future Med. Chem.* **9**, 1361–1376 (2017).
72. Hwang, S. H., Wecksler, A. T., Wagner, K. & Hammock, B. D. Rationally Designed Multitarget Agents Against Inflammation and Pain. *Curr. Med. Chem.* **20**, 1783–1799 (2013).
73. Asako, H. *et al.* Indomethacin-induced leukocyte adhesion in mesenteric venules: Role of lipoxygenase products. *Am. J. Physiol. - Gastrointest. Liver Physiol.* **262**, G903–G908 (1992).
74. Gilroy, D. W., Tomlinson, A. & Willoughby, D. A. Differential effects of inhibitors of cyclooxygenase (cyclooxygenase 1 and cyclooxygenase 2) in acute inflammation. *Eur. J. Pharmacol.* **355**, 211–217 (1998).
75. Knapp, H. R., Sladek, K. & Fitzgerald, G. A. Increased excretion of leukotriene E4 during aspirin-induced asthma. *J. Lab. Clin. Med.* **119**, 48–51 (1992).
76. J.-Y., L. *et al.* Inhibition of soluble epoxide hydrolase enhances the anti-inflammatory effects of aspirin and 5-lipoxygenase activation protein inhibitor in a murine model. *Biochem. Pharmacol.* **79**, 880–887 (2010).
77. Liu, J. Y. *et al.* Inhibition of soluble epoxide hydrolase enhances the anti-inflammatory effects of aspirin and 5-lipoxygenase activation protein inhibitor in a murine model. *Biochem. Pharmacol.* **79**, 880–887 (2010).
78. Sreedhar, D., Subramanian, G. & Udupa, N. Combination drugs: Are they rational? *Curr. Sci.* **91**, 406 (2006).
79. Frantz, S. The trouble with making combination drugs - Drug compound interactions in a tablet are still difficult to predict Simon. *Nat. Rev. drug Discov.* **5**, 881–882 (2006).
80. Korb, O., Stütze, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.* **1**, 115–134 (2007).
81. Chemical Computing Group Inc. Molecular Operating Environment (MOE), 2018.01. *Molecular Operating Environment (MOE), 2018.01* (2018).
82. Korb, O., Stütze, T. & Exner, T. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. *Lect. Notes Comput. Sci.* **4150**, 247–258 (2006).
83. Edelsbrunner, H. *Weighted alpha shapes.* (1992).
84. Liang, J., Edelsbrunner, H. & Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein*

- Sci.* **7**, 1884–1897 (1998).
85. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, (2009).
  86. Corbeil, C. R., Williams, C. I. & Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput. Aided. Mol. Des.* **26**, 775–786 (2012).
  87. Goto, J., Kataoka, R., Muta, H. & Hirayama, N. ASEDock-docking based on alpha spheres and excluded volumes. *J. Chem. Inf. Model.* **48**, 583–590 (2008).
  88. Jain, A. N. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **7**, 407–420 (2006).
  89. Muegge, I. PMF scoring revisited. *J. Med. Chem.* **49**, 5895–5902 (2006).
  90. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
  91. Weiner, P. K. & Kollman, P. A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem.* **2**, 287–303 (1981).
  92. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
  93. Naim, M., Bhat, S. & Rankin, K. Solvated interaction energy (SIE) for scoring protein-ligand binding affinities. 1. Exploring the parameter space. *J. Chem. Inf. Model.* **47**, 122–133 (2007).
  94. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
  95. Ponder, J. W. & Case, D. A. Force Fields For Protein Simulations. *Adv. Protein Chem.* **66**, 27–85 (2003).
  96. Wojciechowski, M. & Lesyng, B. Generalized Born model: Analysis, refinement, and applications to proteins. *J. Phys. Chem. B* **108**, 18368–18376 (2004).
  97. Korb, O., Stützle, T. & Exner, T. E. *PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design.* (2006).
  98. Daniel K Gehlhaar, Gennady M Verkhivker, Paul A Rejto, Christopher J Sherman, David B Fogel, Lawrence J Fogel, S. T. F. Molecular recognition of the inhibitor AC-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **2**, 317–324 (1995).
  99. Ervø, J. *et al.* Improved Protein–Ligand Docking Using GOLD. **52**, 609–623 (2003).

100. Clark, M., Cramer, R. D. & Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **10**, 982–1012 (1989).
101. Verkhivker, G. M. Computational analysis of ligand binding dynamics at the intermolecular hot spots with the aid of simulated tempering and binding free energy calculations. *J. Mol. Graph. Model.* **22**, 335–348 (2004).
102. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Bioinforma.* **52**, 609–623 (2003).
103. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
104. Berthold, M. R. *et al.* *KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (Springer, 2007). doi:10.1007/978-3-540-78246-9
105. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
106. Weber, L. Current status of virtual combinatorial library design. *QSAR Comb. Sci.* **24**, 809–823 (2005).
107. Amano, Y., Tanabe, E. & Yamaguchi, T. Identification of N-ethylmethylamine as a novel scaffold for inhibitors of soluble epoxide hydrolase by crystallographic fragment screening. *Bioorganic Med. Chem.* **23**, 2310–2317 (2015).
108. Irwin, J. J. & Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model* **45**, 177–182 (2005).
109. Willett, P. Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **672**, 133–158 (2011).
110. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
111. Khan, A. U. Descriptors and their selection methods in QSAR analysis : paradigm for drug design. *Drug Discov. Today* **21**, 1291–1302 (2016).
112. Eckert, H. & Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **12**, 225–233 (2007).
113. Smith, D. H., Carhart, R. E. & Venkataraghavan, R. Atom Pairs as Molecular Features

- in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
114. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).
  115. Jelínek, J., Škoda, P. & Hoksza, D. Utilizing knowledge base of amino acids structural neighborhoods to predict protein-protein interaction sites. *BMC Bioinformatics* **18**, (2017).
  116. Deng, Z., Chuaqui, C. & Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* **49**, 490–500 (2006).
  117. Deng, Z., Chuaqui, C. & Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions. *J. Med. Chem.* **47**, 337–344 (2004).
  118. Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, J. D. T. *Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016). doi:10.3233/978-1-61499-649-1-87
  119. Pedregosa, F., Weiss, R. & Brucher, M. Scikitlearn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
  120. Orning, L., Gierse, J. K. & Fitzpatrick, F. A. The bifunctional enzyme leukotriene-A4 hydrolase is an arginine aminopeptidase of high efficiency and specificity. *J. Biol. Chem.* **269**, 11269–11273 (1994).
  121. Wittmann, S. K. *et al.* Thermodynamic properties of leukotriene A4 hydrolase inhibitors. *Bioorg. Med. Chem.* **24**, 5243–5248 (2016).
  122. Lukin, A. *et al.* Discovery of polar spirocyclic orally bioavailable urea inhibitors of soluble epoxide hydrolase. *Bioorg. Chem.* **80**, 655–667 (2018).
  123. Hahn, S. *et al.* Complementary Screening Techniques Yielded Fragments that Inhibit the Phosphatase Activity of Soluble Epoxide Hydrolase. *ChemMedChem* **6**, 2146–2149 (2011).
  124. Perola, E., Walters, W. P. & Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins Struct.*

- Funct. Genet.* **56**, 235–249 (2004).
125. Onodera, K., Satou, K. & Hirota, H. Evaluations of molecular docking programs for virtual screening. *J. Chem. Inf. Model.* **47**, 1609–1618 (2007).
  126. Vernalis. Vernalis Research. (2016). Available at: <http://www.vernalis-research.com/>.
  127. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
  128. Kirkland, T. A. *et al.* Synthesis of glutamic acid analogs as potent inhibitors of leukotriene A4hydrolase. *Bioorganic Med. Chem.* **16**, 4963–4983 (2008).
  129. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 1987).
  130. Labute, P. LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *J. Chem. Inf. Model.* **50**, 792–800 (2010).
  131. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
  132. Sandanayaka, V. *et al.* Discovery of 4-[(2S)-2-[[4-(4-Chlorophenoxy)phenoxy]methyl]-1-pyrrolidinyl] butanoic acid (DG-051) as a novel leukotriene A4 hydrolase inhibitor of leukotriene B4 biosynthesis. *J. Med. Chem.* **53**, 573–585 (2010).
  133. Shen, H. C. *et al.* Discovery of a highly potent, selective, and bioavailable soluble epoxide hydrolase inhibitor with excellent ex vivo target engagement. *J. Med. Chem.* **52**, 5009–5012 (2009).
  134. Carosati, E., Sciabola, S. & Cruciani, G. Hydrogen bonding interactions of covalently bonded fluorine atoms: From crystallographic data to a new angular function in the GRID force field. *J. Med. Chem.* **47**, 5114–5125 (2004).
  135. Kim, I. H., Park, Y. K., Nishiwaki, H., Hammock, B. D. & Nishi, K. Structure-activity relationships of amide-phosphonate derivatives as inhibitors of the human soluble epoxide hydrolase. *Bioorganic Med. Chem.* **23**, 7199–7210 (2015).
  136. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).

## 9 List of figures

<b>Figure 1: Histogram of overall growth of released 3D-structures in the PDB.....</b>	<b>2</b>
<b>Figure 2: Simple regression problem.....</b>	<b>6</b>
<b>Figure 3: Simple classification problem. ....</b>	<b>6</b>
<b>Figure 4: Optimal separation hyperplane. ....</b>	<b>7</b>
<b>Figure 5: Schematic representation of a Random Forest model. ....</b>	<b>8</b>
<b>Figure 6: Arachidonic acid cascade.....</b>	<b>11</b>
<b>Figure 7: Transformation from LTA<sub>4</sub> into LTB<sub>4</sub>.....</b>	<b>12</b>
<b>Figure 8: Crystal structure of LTA4H, PDB code 3CHP.....</b>	<b>12</b>
<b>Figure 9: Crystal structure of sEH, PDB code 4Y2T.....</b>	<b>13</b>
<b>Figure 10: Hydration of EETs into DHETs by the sEH.....</b>	<b>13</b>
<b>Figure 11: Compilation of data sets. ....</b>	<b>24</b>
<b>Figure 12: KNIME workflow for crystalized ligand preparation. ....</b>	<b>24</b>
<b>Figure 13: KNIME workflow for active ChEMBL compound preparation. ....</b>	<b>25</b>
<b>Figure 14: KNIME workflow for inactive ChEMBL compound preparation.....</b>	<b>26</b>
<b>Figure 15: Condensation reaction between C1 and C2. ....</b>	<b>27</b>
<b>Figure 16: General construction of the AtomPair fingerprint. ....</b>	<b>28</b>
<b>Figure 17: General construction of the Morgan fingerprint. ....</b>	<b>28</b>
<b>Figure 18: General construction of the MACCS fingerprint.....</b>	<b>29</b>
<b>Figure 19: General scheme of the Protein-Ligand Interaction Fingerprint.....</b>	<b>30</b>
<b>Figure 20: Schematic representation of the fluorescence based LTA4H assay. ....</b>	<b>35</b>
<b>Figure 21: Schematic representation of the fluorescence based sEH assay.....</b>	<b>36</b>
<b>Figure 22: Co-crystalized ligands of LTA4H and sEH.....</b>	<b>38</b>
<b>Figure 23: KNIME workflow describing the MOE docking validation.....</b>	<b>38</b>
<b>Figure 24: MOE docking validation for LTA4H and sEH.....</b>	<b>40</b>
<b>Figure 25: PLANTS docking validation for LTA4H and sEH.....</b>	<b>41</b>
<b>Figure 26: Binding site shape of LTA4H and sEH.....</b>	<b>43</b>
<b>Figure 27: Scheme of the models built and used for prediction. ....</b>	<b>46</b>
<b>Figure 28: Selection of six compounds for chemical synthesis using PLIF fingerprint and Random Forest.....</b>	<b>53</b>
<b>Figure 29: Binding poses (MOE docking) of synthesized compounds using PLIF and Random Forest in the receptor structure LTA4H.....</b>	<b>55</b>



<b>Figure 30: Binding poses (MOE docking) of synthesized compounds using PLIF and Random Forest in the receptor structure sEH. ....</b>	<b>56</b>
<b>Figure 31: Selection of eight compounds for chemical synthesis using AtomPair fingerprint and Random Forest. ....</b>	<b>61</b>
<b>Figure 32: Compound 3 docked in target structures LTA4H and sEH. ....</b>	<b>65</b>
<b>Figure 33: Compound 6 docked in target structures LTA4H and sEH. ....</b>	<b>66</b>
<b>Figure 34: Compound 12 docked in target structures LTA4H and sEH. ....</b>	<b>67</b>

