Department of Computer Science
Algorithmic Bioinformatics

Universitätsstr. 1 D–40225 Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Identifying significant targets of molecular fragments

Jonas Weber

Identifikation signifikanter Targets von molekularen Fragmenten

Submission:	24.08.2018
Supervisor:	Prof. Dr. G. Klau
Second Assessor:	Prof. Dr. H. Schwender
Advisor:	Dr. M. Engler

Declaration

I hereby confirm that this thesis is my own work. I have documented all sources an tools used. Any direct or indirect quote has been marked as such clearly with specification of the source.

Düsseldorf, August 24, 2018

Jonas Weber

Abstract

Molecular fragments are connected substructures of molecules. Studying common molecular fragments of molecules is an important aspect of cheminformatical research. This research yields new insight on drug candidates. The fragments can be linked to the molecules which contain them, and to the targets of these molecules using chemical databases. In this bachelor thesis we identify significant fragments and their associated targets from a given set of fragments.

The studied fragments originate from the Fragment Database (FDB). We obtained the related molecules and targets by employing the Automated Topology Builder (ATB) and the ChEMBL database, respectively. 11388 fragments, 107927 molecules and 6950 targets of these molecules in total were gathered.

To identify the significant fragments, and therefore the targets, we used a linear regression and a bootstrap algorithm to find a cut-off value. Fragments below this value were deemed significant. We implemented this using Python 3.5.5. 141 fragments were identified as significant within this bachelor thesis.

These fragments and targets were found to be relevant in chemical research and possibly in medical applications.

Contents

1 Introduction 1

1.1 Objective . . . 1

2 Background information 1

2.1 Chemical background . . . 1

2.1.1 Molecular fragment . . . 1

2.1.2 Target . . . 2

2.2 Used databases . . . 3

2.2.1 Fragment Database . . . 3

2.2.2 Automated Topology Builder . . . 3

2.2.3 ChEMBL . . . 3

2.3 Mathematical background . . . 3

2.3.1 Regression . . . 3

2.3.2 Bootstrap method . . . 4

3 Gathering data 4

3.1 Obtaining molecules . . . 4

3.2 Obtaining targets . . . 6

3.3 Discussion . . . 6

4 Finding significant fragments 7

4.1 Linear regression on data . . . 7

4.2 Bootstrap algorithm . . . 7

4.3 Discussion . . . 9

5 Implementation 11

5.1 Conda environment . . . 11

5.2 Workflow . . . 11

5.2.1 Data collection . . . 11

5.2.2 Regression . . . 11

6 Results 12

6.1 Molecular fragment 57302 . . . 12

6.1.1 Related molecules and target CHEMBL203 . . . 13

6.2 Molecular fragment 60336 . . . 14

6.2.1 Related molecules and targets . . . 14

7 Discussion and outlook 15

7.1 Relevance of molecular fragments . . . 15

7.1.1 Relevance of fragment 57302 . . . 15

7.1.2 Relevance of fragment 60336 . . . 16

7.2 Summary and outlook . . . 17

8 Acknowledgments 17

9 Bibliography 18

A Source code 20

1 Introduction

Using pharmaceutical drugs as medication for disease has long been a standard in modern medicine. A great amount of resources is put into the research on known compounds and the development of new pharmaceutical compounds. Constantly, researchers are learning more about their structure, function and relation to other molecules. Biological and chemical databases containing information about these molecules are growing. Because of these large amounts of data, scientists often use algorithms to study the properties of drugs. An important aspect of this cheminformatical research is to discover shared molecular fragments of these molecules. Many molecular fragments have already been described and saved in the Fragment Database (FDB). Due to the amount of information of molecules and targets saved in various databases, it is possible to link the fragments to the molecules they are contained in, and therefore the functions of these molecules, via an algorithmic approach.

1.1 Objective

For this bachelor thesis we aim to find molecular fragments which are significantly linked to functions. To achieve this, we collected the required data from the databases FDB, ATB and ChEMBL, and identified the significant fragments and targets using linear regression and a bootstrap algorithm. These steps will be described in individual chapters and in more detail later. Finally, we will show the results and discuss their relevance for further research.

2 Background information

For this thesis we define and describe the following background information to share the same terminology and knowledge with our readers.

2.1 Chemical background

2.1.1 Molecular fragment

We define a molecular fragment as introduced by Engler et al. *Enumerating common molecular substructures* [1]. This paper describes the process of enumerating the fragments used for this thesis. The definition is based on molecules being represented by a molecular graph. A molecular graph is a simple graph $G = (V, E)$ whose nodes and edges correspond to atoms and bonds, respectively. In the named paper, nodes are labeled by their partial charge $w : V \to R$ and their atom type $t : V \to \Sigma$ where Σ is the set of all atom types.

Definition 1 (k-neighborhood)**.**

The k-neighborhood of a node $u \in V$ is defined recursively as

$$N^k(u) = \begin{cases} \{u\}, & \text{if } k = 0, \\ N^{k-1}(u) \cup \{w | (v, w) \in E, v \in N^{k-1}(u)\}, & \text{if } k \geq 1. \end{cases}$$

For a subset $V' \subseteq V$, we define $N^k(V')$ to be the set $\bigcup_{v \in V'} N^k(v) \setminus V'$.

Note that $|N^1(u)| = deg(u) + 1$, where $deg(u)$ denotes the degree of node u. We denote a subgraph of a graph $G = (V, E)$ induced by $V' \subseteq V$ as $G[V']$. Given molecules $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ with atom types $t_1 : V_1 \to \Sigma$ and $t_2 : V_2 \to \Sigma$ and $k \in \mathbb{N}$, we define a k-common fragment and its shell as follows.

Definition 2 (k-common fragment).
Given $k \in \mathbb{N}$, a k-common fragment is a triple (V_1', V_2', h) with $V_1' \subseteq V_1$ and $V_2' \subseteq V_2$ and a bijection $h : V_1' \cup N^k(V_1') \to V_2' \cup N^k(V_2')$ such that

1. *$G_1[V_1']$ and $G_2[V_2']$ are connected,*
2. *(u, v) is an edge in $G_1[V_1' \cup N^k(V_1')]$ if and only if $(h(u), h(v))$ is an edge in $G_2[V_2' \cup N^k(V_2')]$,*
3. *$t_1(v) = t_2(h(v))$ for all $v \in V_1' \cup N^k(V_1')$.*

Definition 3 (shell).
The shell of a k-common fragment (V_1', V_2', h) is given by $(N^k(V_1'), N^k(V_2'))$.

Definition 4 (maximal k-common fragment).
A k-common fragment (V_1', V_2', h) is maximal if there exists no k-common fragment (V_1'', V_2'', h'') such that $V_1' \subsetneq V_1''$ and $V_2' \subsetneq V_2''$.

The fragments used throughout this bachelor thesis are all maximal k-common fragments, found by the algorithm described in the previously named paper. We can see, that every considered fragment is contained in at least two molecules. For this bachelor thesis we are mostly interested in molecular fragments of metabolites. Metabolites are small molecules, that are intermediate and end products of the metabolism. They have various functions such as stimulatory and inhibitory effects on proteins, signaling or catalytic activity. We are interested in these molecules since they are primary drug candidates.

2.1.2 Target

A target is an entity inside a living being with which another entity, in our case a molecule, binds or interacts. This interaction results in a change of the structure or function of the target. Targets range from entire organisms, such as the target CHEMBL376, which is the brown rat (Rattus norvegicus), over cell lines and tissues, such as the target CHEMBL613690, which is a hepatocyte, a cell of the main parenchymal tissue of the liver, to single proteins, such as the target CHEMBL4719, which is a Nitric-oxide synthase inside the human brain. A molecule can have one target, multiple targets, or none. For this thesis all targets were identified using their ChEMBL ID.

2.2 Used databases

To collect the fragments, their molecules and their targets, the FDB, ATB and ChEMBL databases were used, respectively.

2.2.1 Fragment Database

The Fragment Database (FDB) consists of molecular fragments and their properties. These fragments were computed by finding all common fragments inside molecules found in the ATB database [1]. All molecular fragments used in this thesis originate from a snapshot of the FDB.

2.2.2 Automated Topology Builder

The Automated Topology Builder (ATB) is a Web-accessible server, that provides topologies and parameters for a wide range of molecules. Next to generating forcefield descriptions of novel molecules, one of its primary functions is to act as a repository for molecules, that have been parametrized as part of the GROMOS family of force fields [2]. The ATB contains currently about 250000 molecules, of which most are metabolites. An API and a public Python API client for the ATB API [3] are also provided by the ATB. The ATB is relevant for this bachelor thesis, since the fragments of the FDB were computed from an ATB snapshot [1].

2.2.3 ChEMBL

ChEMBL is an Open Data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds [4]. It also contains 2-D structures, calculated properties and abstracted bioactivities. ChEMBL also provides an API and a Python client for accessing the ChEMBL API [5]. In this thesis the targets of molecules are accessed through the ChEMBL API.

2.3 Mathematical background

To find the significant fragments and targets from the collected data, we employed a linear regression and a bootstrap algorithm.

2.3.1 Regression

Regression is a set of statistical methods for fitting a curve to a given set of points using a goodness of fit criterion. Regression analysis can be used, among many other purposes, for finding relations between data, prediction and classification. Here we used the linear regression method, a type of regression, where the fitted curve is linear.

2.3.1.1 Standard Model of Simple Linear Regression

Let the data be $D = (y_i, x_i), i = 1, ..., n$ with continuous variables x and y. The model is in the following form:

$$y_i = \beta_0 + \beta_1\ x_1 + \epsilon_i, i = 1, ..., n$$

The errors $\epsilon_1, ..., \epsilon_n$ are independent and identically distributed (i.i.d.) with

$$E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$$

We can interpret the estimated regression line $\hat{f}(x) = \hat{\beta_0} + \hat{\beta_1}x$ as an estimate $E(\hat{y}|x)$ for the conditional expected value of y given the covariate value x. We can, thus, predict y through $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x$ [12].

2.3.2 Bootstrap method

Bootstrap methods are computer-intensive methods of statistical analysis, which use simulation to calculate standard errors, confidence intervals, and significance tests. The methods apply for any level of modelling, and can therefore be used for fully parametric, semiparametric, and completely nonparametric analysis. Applications include stratified data, finite populations, censored and missing data, linear, nonlinear, and smooth regression models, classification, time series and spatial problems [13]. A bootstrap algorithm employs random resampling with or without replacement to compute an estimator. The bootstrap algorithm used during this bachelor thesis samples the given data 10000 times with replacement. From each sample a cut-off value is computed by fitting a regression line to the sample. Finally, a single cut-off value is estimated from the 10000 cut-off values.

3 Gathering data

For each molecular fragment, the molecules containing the fragments, and the targets of those molecules were gathered as the first step of this thesis. The data was organized as shown in Figure 1 for each molecular fragment. The obtained data consists of 11388 fragments, 107927 molecules and 6950 targets of those molecules. We will now go into detail on how this data was collected.

3.1 Obtaining molecules

As mentioned previously, all molecular fragments of the FDB were computed from an ATB snapshot [1]. The data corresponding to a molecular fragment of the FDB contains a molecular ID (molid), which is an ID assigned to every existing molecule in the ATB. Thus, we obtained most molecules by searching their molecular ID in the ATB.

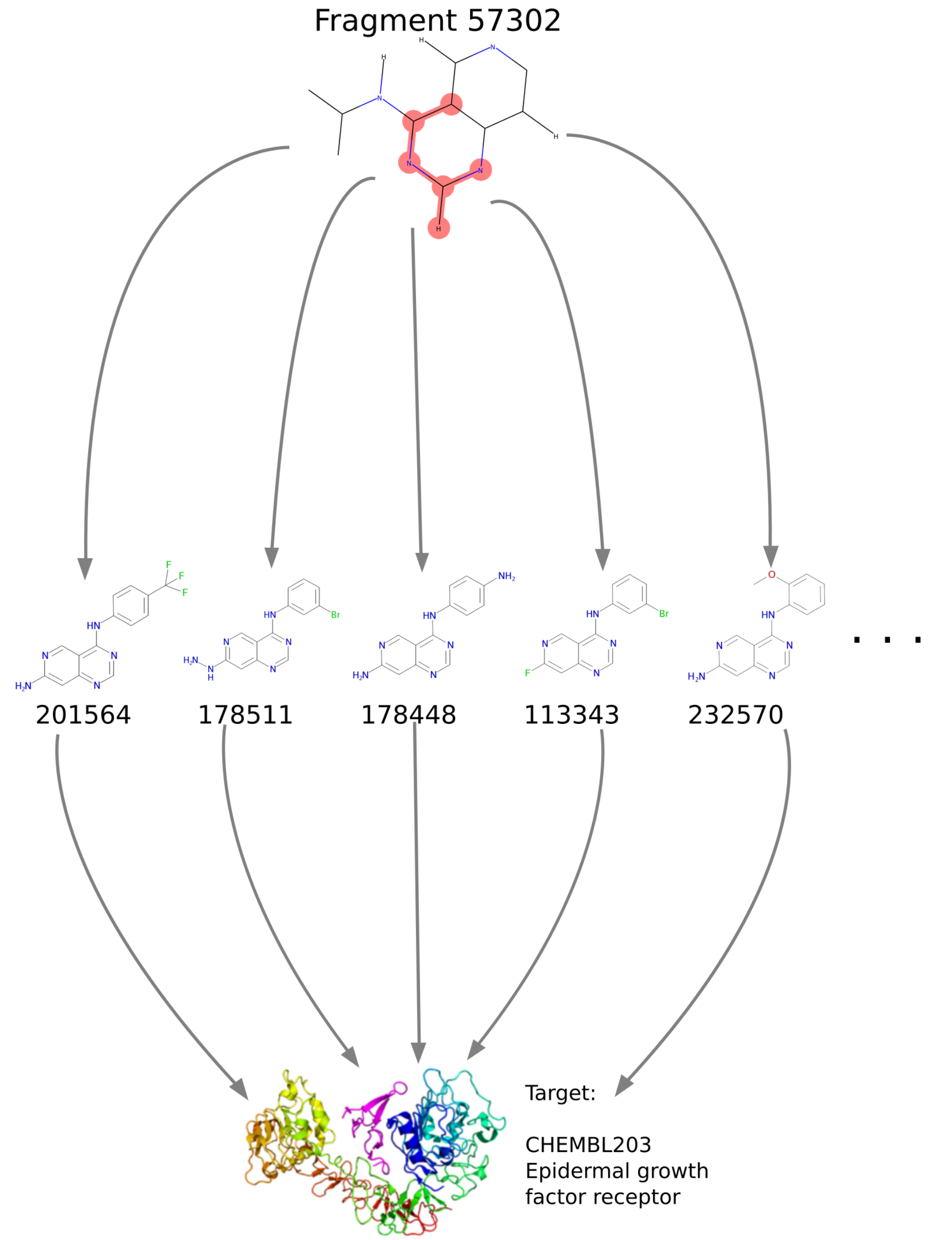


Figure 1: Example of the data structure displaying fragment 57302 and five of its molecules with molids 201564, 178511, 178448, 113343, 232570, and the single target of those with ChEMBL ID 203. The red framed atoms of the fragment are its core atoms

3.2 Obtaining targets

Besides the molecular id, most molecules from the ATB also are assigned a unique ChEMBL ID. This ChEMBL ID can be used to identify known molecules in the ChEMBL database. The data corresponding to molecules of the ChEMBL database often incorporates the targets of these molecules. Thus, we found most of the obtained molecules of the ATB in the ChEMBL database, and thereby obtained the targets of these molecules.

3.3 Discussion

While obtaining the data, a few interesting cases arose, which we now will discuss.

255 molecules could not be found in the ATB by searching their molecular ID. This is caused by changes of the ATB database. The FDB was generated through a snapshot of the ATB database. Hence changes in the ATB database cannot be reflected in the snapshot of the FDB, which is used for this bachelor thesis. These molecules will not be regarded during this thesis. An example for this case is the molecule with molecular ID 23883. Its molecular ID cannot be found in the ATB database.

Some molecules from the ATB were assigned to the same ChEMBL ID or are the same molecule but with different molecular IDs. This is caused by the molecular ID not being unique, meaning that a molecule can have multiple molecular ids. The ATB database contains an entry for every different molecular id, despite some describing the same molecule. These molecular ids of duplicate molecules can be identified by comparing the InChI key of all molecules. The International Chemical Identifier (InChI) key is another unique property of a molecule saved in the ATB. It encodes molecular information in textual form. The ChEMBL ID is unique, meaning only a single ChEMBL id can be assigned to a molecule. If molecules with different molecular ids are isomorphic, only one molecular ID will be regarded during this bachelor thesis. An example for this case are the molecular ids 217310 and 220226 which both correspond to the same ChEMBL ID CHEMBL2144822.

We discovered that 5167 molecules from the ChEMBL database do not have targets stored in their properties. This can be caused by two reasons. Either a molecule has no targets, and therefore there are no targets to be named in the database, or a molecule has a target, but it has yet to be described and added to the database. Molecules with no targets remain in the obtained data. An example for a molecule with no targets is the molecule with the ChEMBL ID CHEMBL2010316.

For a few molecules of the ChEMBL database, the search yielded, among other targets, duplicate targets. Duplicates of the same target were discarded, keeping only one of each found target. An example for such a molecule can be found in the molecule with the ChEMBL ID CHEMBL1333203. Searching this molecule in the ChEMBL database yielded, among others, the target with ChEMBL ID CHEMBL1293231 twice.

After obtaining the molecules and the targets of all fragments, it was discovered that 4511 fragments were not contained in any molecule of the ATB database. These fragments were discarded from the obtained data.

In addition, the size of fragments and molecules, meaning the number of atoms contained in the fragment or molecule, were also collected. This data can be used for future analysis of the fragments.

4 Finding significant fragments

We used the collected data about the fragments for the identification of the significant fragments, and thus the significant targets. We searched for fragments, which are contained in many molecules, but do not have many targets. To do so, we fit a linear regression on the collected data and subsequently used the bootstrap algorithm to identify a cut-off, below which a fragment is deemed significant. In the following sections the two steps will be described and discussed individually.

4.1 Linear regression on data

The first step of computing a linear regression is to transform your data into the correct form, in which it can be used for the regression. We transformed the data into the form $D = (m_i, t_i), i = 1, ..., n$, where m_i is the logarithmic molecule count of a fragment, t_i is the logarithmic target count of those molecules, and n is the number of fragments. In our case n is 11388. Using the logarithmic counts limits the impact of the larger values. We fit the regression line to the two-dimensional data, predicting y through $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$. This yielded the regression line $reg(D)$, which is shown in Figure 2. The distances of every fragment to the regression line were computed by calculating $dy_i = t_i - ry_i, i = 1, ..., n$, where dy_i is the calculated distance, and ry_i is the y value of the regression line at the x value of m_i. The significant fragments are those with the lowest values of dy_i, meaning the points furthest below the regression line.

4.2 Bootstrap algorithm

After fitting the regression line on the data points and calculating the distances, a cut-off had to be found in order to discriminate between significant and non-significant fragments. This was implemented with the following bootstrap algorithm.

The algorithm is given the data points and the regression line fitted by a linear regression. The first step of this algorithm is to sample the data points 10000 times and calculate the confidence intervals of each sample. One sampling is done by selecting a point from the uniformly distributed data points n times, without removing it, n being the amount of data points. In our case this means the selection happens 11388 times. For each sample $D'_i, i = 1, ..., 10000$ of those sampled data sets the algorithm fits an individual regression line $reg(D'_i)$ and

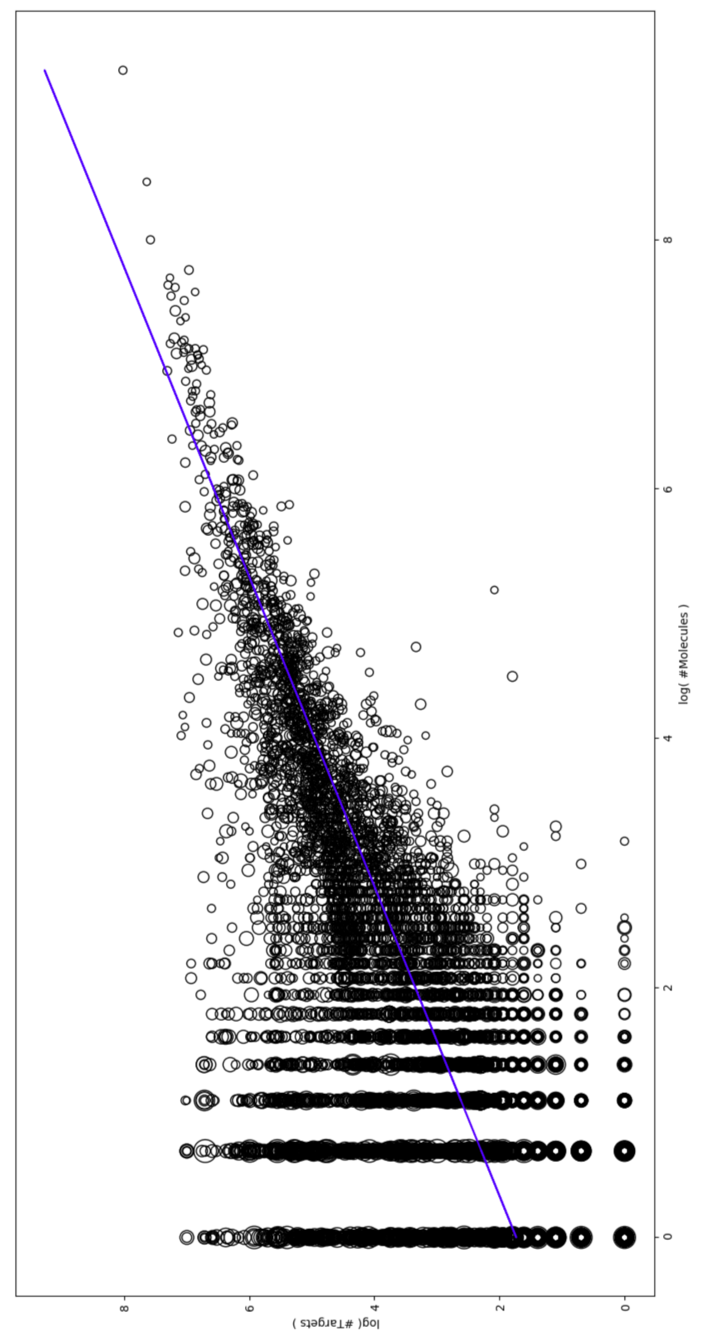


Figure 2: Plot showing all data points (*black*) and the regression line calculated on those points (*blue*). The size of a data point is determined by the size of its fragment, meaning the number of atoms.

calculates the distances of the points to the regression line, as described before. Now it computes a value c_1 and c_2 for each sample, such that 95% of the data points are contained between the two lines $g_1 = reg(D'_i) + c_1$ and $g_2 = reg(D'_i) + c_2$, with $c_1 > 0 > c_2$. This can be calculated by setting c_1 to the 97.5% quantil of the distances dy_i and c_2 to the 2.5% quantil of the distances dy_i. From the values c_1 and c_2 of the samples we can calculate the values $\hat{c_1}$ and $\hat{c_2}$ of the original data points. This can be done by taking the 97.5% quantil of the 10000 c_1 values as $\hat{c_1}$, and the 2.5% quantil of the 10000 c_2 values as $\hat{c_2}$. The regression line with the corresponding lines $\hat{g_1} = reg(D) + \hat{c1}$ and $\hat{g_2} = reg(D) + \hat{c2}$ of the original data set is shown in Figure 4.

We get the values $\hat{c_1}$ and $\hat{c_2}$ from the algorithm. The value $\hat{c_2}$ now functions as the cut-off to determine the significant fragments. This means every fragment with a distance $dy < \hat{c_2}$ is a significant fragment.

4.3 Discussion

The bootstrap algorithm was used since the distances dy between the fragments and the regression line are not normally distributed. This was determined by a statistical test, with the null hypothesis being that a sample comes from a normal distribution. The null hypothesis was discarded after running the test with our data set. Thus, the cut-off cannot be determined by fitting the distances dy to a normal distribution and taking a specific quantil as the value. The bootstrap algorithm is a possibility to solve this problem. The difference between the distribution of our data points and the normal distribution is displayed in the qqplots shown in Figure 3.

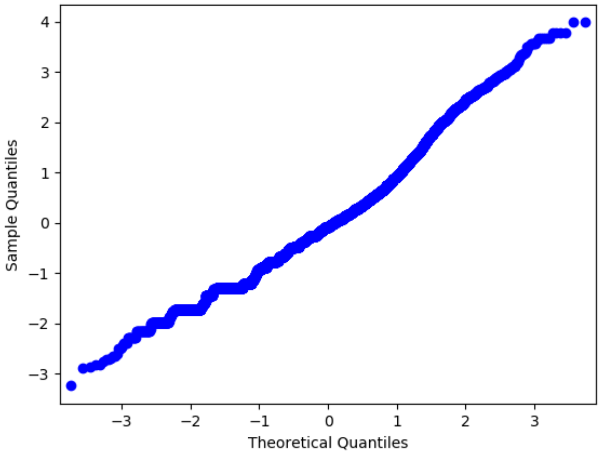


Figure 3: qqplot of the distances of the data points to the regression line. This plot shows the difference between the distribution of the distances and the normal distribution.

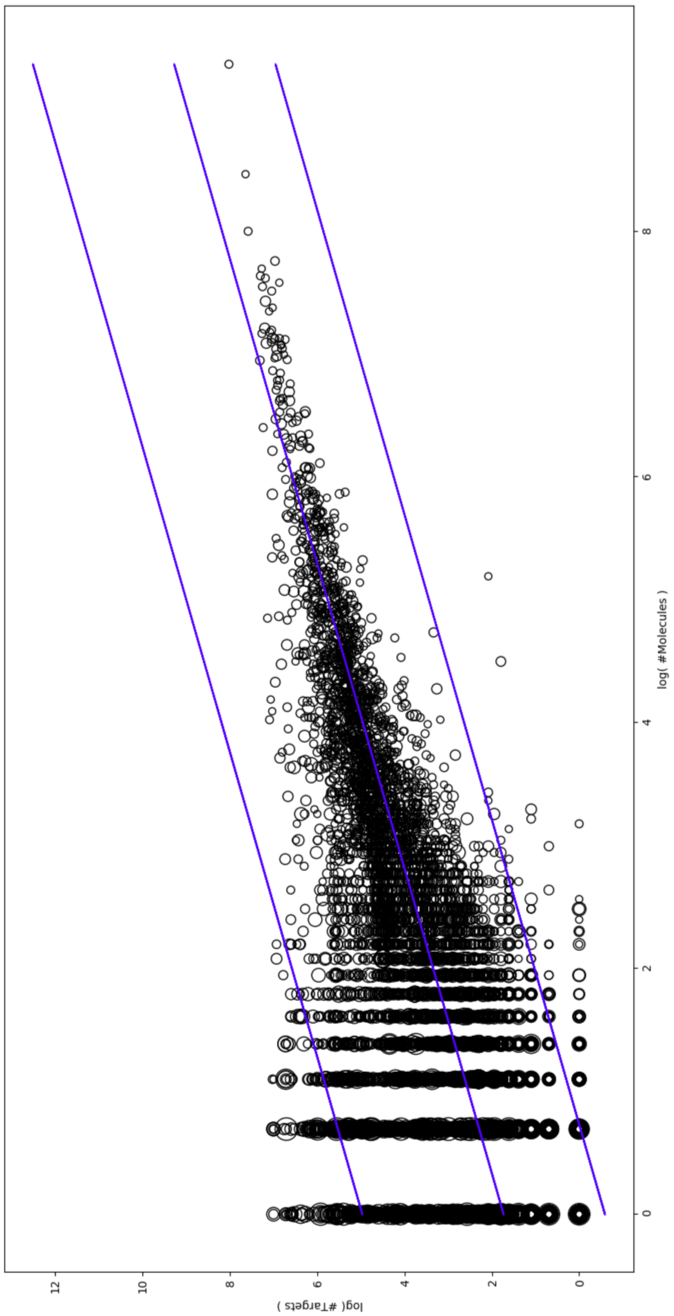


Figure 4: Plot showing all data points (*black*), the regression line calculated on those points (*middle blue line*) and the lines of c_1 and c_2 computed by the bootstrap algorithm (*left and right blue line*). The size of a data point is determined by the size of its fragment.

5 Implementation

The code of this bachelor thesis is written in Python 3.5.5 and can be accessed via this work's GitLab repository found in Appendix A.

5.1 Conda environment

The scripts run in a Conda environment. Conda is an open source package and environment manager, contained in the Anaconda Distribution [6]. The environment.yaml file, found in the GitLab repository, can be used to recreate the Conda environment used for this bachelor thesis. The environment contains all packages used for this bachelor thesis, such as scipy, the ATP API Python client, and the Python client for accessing ChEMBL API.

5.2 Workflow

The python scripts used within this bachelor thesis are managed by a Snakemake workflow. The Snakemake workflow management system [7] is a tool to describe workflows using a Python based language. By defining rules, which depend on other rules, a workflow is created.

The Snakemake workflow of this bachelor thesis can be divided into two sections, the data collection and the regression sections, as shown in Figure 5.

5.2.1 Data collection

This section of the workflow controls the implementation of Chapter 3 ”Gathering data”. Therefore this section of the workflow manages the collection of the needed data from the FDB snapshot, from the ATB and from the ChEMBL database. This is accomplished by using the Python API client for the ATB API [3] and the Python client for accessing ChEMBL API [5]. The collected data is organized in json files created from python dictionaries. This section can be run using the rule all_data.

5.2.2 Regression

The second section of the workflow regulates the implementation of the Chapter 4 ”Finding significant fragments” of this bachelor thesis. The regression and the bootstrap algorithm are calculated, and the significant fragments and targets are identified by finding the cut-off value. Furthermore, plots of the regression and the result of the bootstrap algorithm are generated. The linear Regression is computed using the sklearn library [8]. The bootstrap algorithm is calculated using the bootstrap function from the library seaborn [9]. The plots are generated by the libraries matplotlib [10], statsmodels [11] and seaborn [9]. Calling the rule all_regression runs both sections of the workflow.

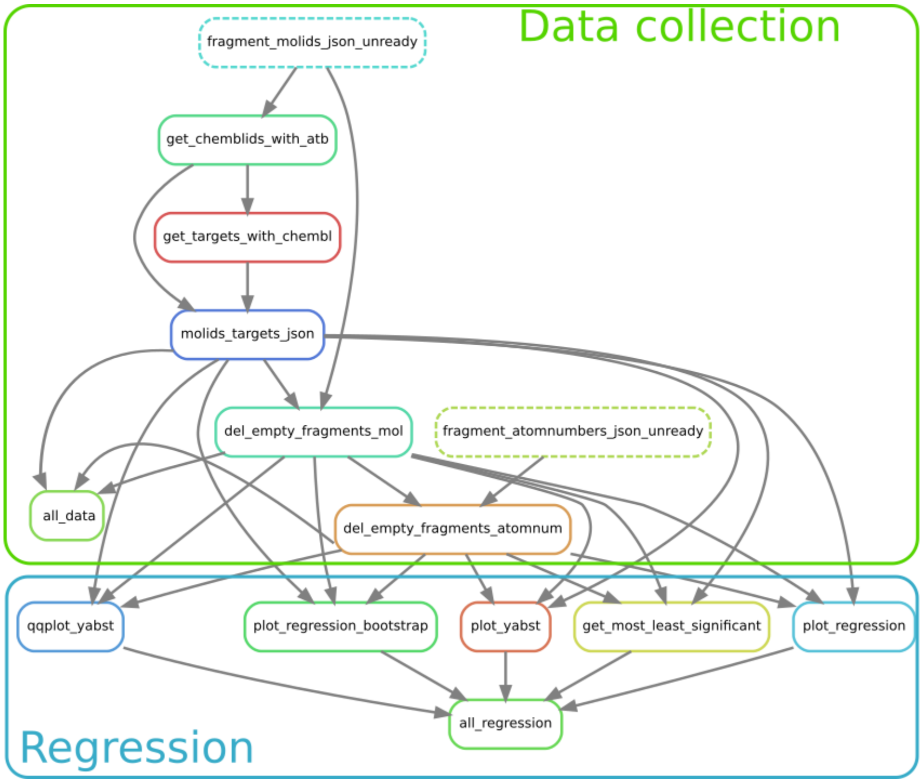


Figure 5: DAG showing the Snakemake workflow of this bachelor thesis. The workflow is separated into data collection and regression.

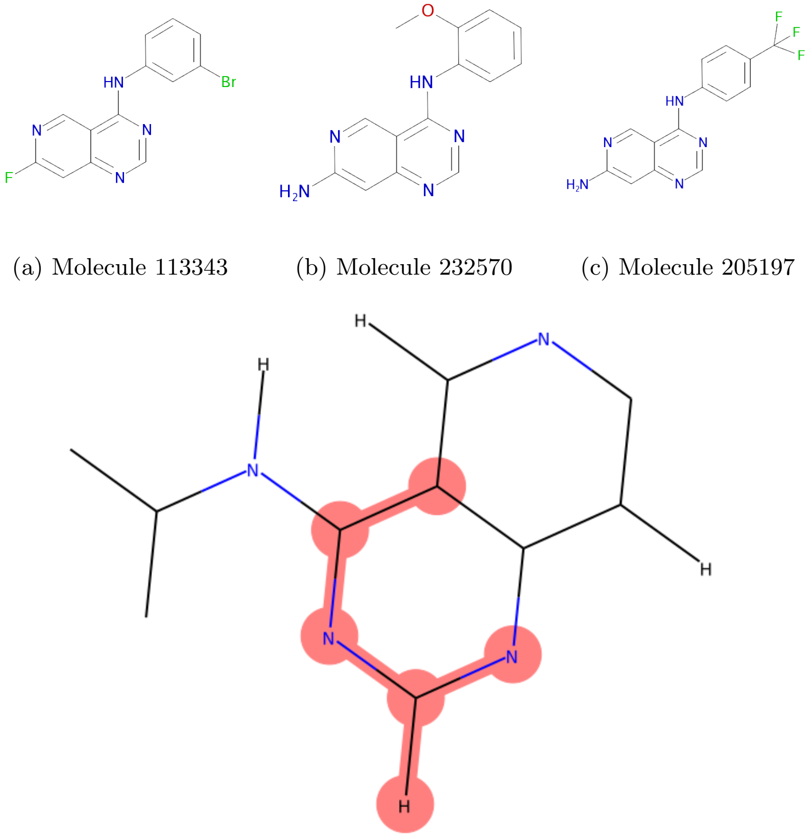
6 Results

The bootstrap algorithm yielded the values $\hat{c_1} = 3.233$ and $\hat{c_2} = -2.139$. Therefore, every fragment, which is lower than the found $\hat{c_2}$ value, is considered significant. Within the 11388 considered molecular fragments, 141 significant fragments were found using this cut-off value. These are the fragments located below the line generated by the $\hat{c_2}$ value in Figure 4.

We will look at the two significant molecular fragments with molids 57302 and 60336 specifically. These two fragments are located furthest below the regression line, meaning the most significant targets.

6.1 Molecular fragment 57302

The fragment 57302, which is shown in the Figure 6, has a distance of $dy = -4.289$ to the regression line. It is contained in 24 molecules, which have one common target with the ChEMBL ID CHEMBL203. The core of the molecular fragment contains six molecules, of which two are nitrogen atoms, three are carbon atoms and one is a hydrogen atom. The molecular structure of this fragment includes two rings.



(a) Molecule 113343

(b) Molecule 232570

(c) Molecule 205197

(d) Molecular structure of the molecular fragment 57302. Red framed atoms belong to the core of the fragment.

Figure 6: The molecular fragment 57302 (d) and three molecules in which it is contained (a-c)

6.1.1 Related molecules and target CHEMBL203

According to the ChEMBL database, the common target of the molecules in which the fragment 57302 is contained, is a single protein functioning as an epidermal growth factor receptor inside the human organism named erbB1. It is also called EGFR, ERBB, ERBB1, HER1, Proto-oncogene c-ErbB-1 and Receptor tyrosine-protein kinase erbB-1. The epidermal growth factor receptor is a transmembrane protein, a protein that is embedded in the membrane of human cells passing through on both sides. An epidermal growth factor is a protein that, when binding to the EGFR, stimulates the growth and differentiation of cells.

The molecules, in which the fragment 57302 is contained, were found to be of similar size and structure, all containing an additional ring structure besides the contained fragment.

6.2 Molecular fragment 60336

The fragment 60336, which is shown in the Figure 6, has a distance $dy = -4.289$ to the regression line. It is contained in 180 molecules, which have eight common targets. The core of the molecular fragment contains 5 molecules, of which one is a nitrogen atom, one is a carbon atom, one is an oxygen atom and two are hydrogen atoms. The molecular structure of this fragment includes a ring.

6.2.1 Related molecules and targets

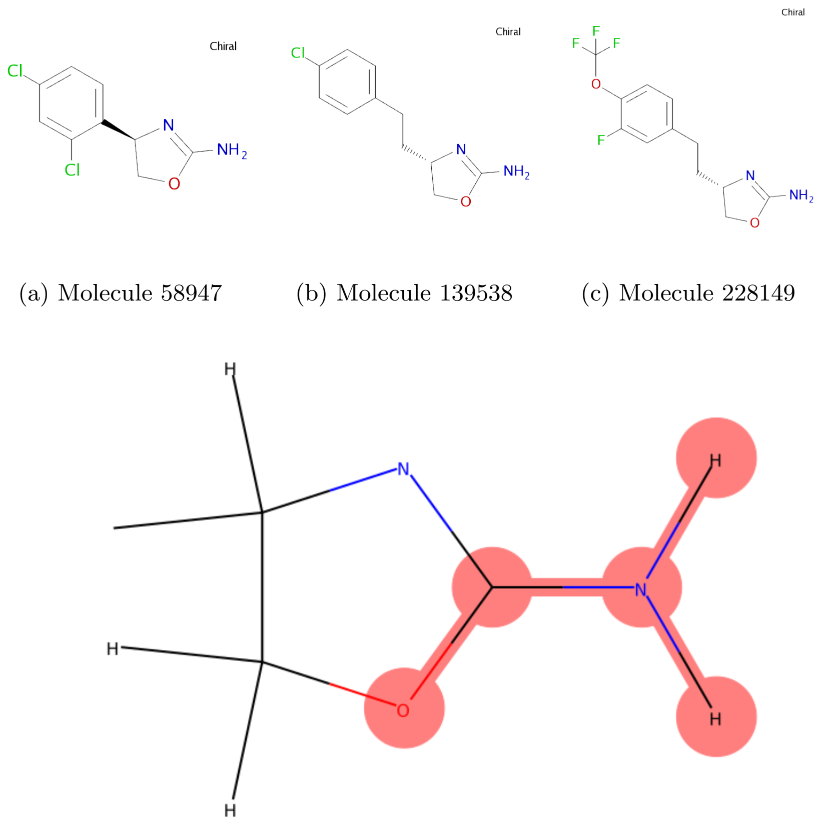
Three molecules, in which the fragment 60336 is contained, are shown in Figure 7. The eight targets can be divided into three groups.

The three targets with the ChEMBL IDs CHEMBL5857, CHEMBL4908 and CHEMBL3833 are single proteins. These proteins are similar in structure and function. Their preferred name is trace amine-associated receptor 1, but they are also known as TAAR1, Ta1, TaR-1, Taar1, Tar1, Trace amine receptor 1 and Trar1. The target CHEMBL5857 can be found in the human body, whereas the target CHEMBL4908 can be found in the house mouse (*Mus musculus*), and the target CHEMBL3833 can be found in the brown rat (*Rattus norvegicus*). TAAR1 is an intracellular amine-activated receptor, which is primarily found in internal organs, such as stomach and duodenum, and specific cells, for example neurons.

The two targets with the ChEMBL ID CHEMBL4719 and CHEMBL3464 are also single protein enzymes, named nitric oxide synthase. Nitric oxide synthases are enzymes, which catalyze the synthesis of nitric oxide (NO) from L-arginine. The target CHEMBL4719 can be found in the brain of the house mouse (*Mus musculus*). It is also known as constitutive NOS, N-NOS, NC-NOS, NOS type I, Neuronal NOS, Nos1, Peptidyl-cysteine S-nitrosylase NOS1, bNOS and nNOS. The other target CHEMBL3464 can also be found in the house mouse. A specific area of activity is not stated in the ChEMBL database. This target is also known as inducible NO synthase, Inducible NOS, Inosl, MAC-NOS, Macrophage NOS, NOS type II, Nos2, Peptidyl-cysteine S-nitrosylase NOS2 and iNOS.

The target with the ChEMBL ID CHEMBL613690 is a cell line, known as hepatocyte. This target can be found in the human. Hepatocytes are the main tissue of the human liver, making up 70-85% of the liver's mass. The target with the ChEMBL ID CHEMBL2367379 is a tissue named liver microsome, found within the human liver. Liver microsomes are vesicles of the hepatocyte endoplasmic reticulum. They contain proteins.

The target with ChEMBL ID CHEMBL612545 is unchecked in the ChEMBL database and therefore cannot be regarded.



(a) Molecule 58947 (b) Molecule 139538 (c) Molecule 228149

(d) Molecular structure of the molecular fragment 60336. Red framed atoms belong to the core of the fragment.

Figure 7: The molecular fragment 60336 (d) and three molecules in which it is contained (a-c)

7 Discussion and outlook

7.1 Relevance of molecular fragments

We will now look further into the two described fragments 57032 and 60336 and their targets, and discuss their relevance for chemical research.

7.1.1 Relevance of fragment 57302

The only target of the molecular fragment 57302 is a single protein functioning as an epidermal growth factor receptor inside the human organism, as described in Chapter 6.1.

Andrew M. Thompson et al. [14] describe the function of the molecules which contain this fragment as inhibitors of the tyrosine kinase activity of the epidermal growth factor receptor, the found target. Within this article the molecules were evaluated according to their ability to inhibit the tyrosine-phosphorylating action of the EGF receptor enzyme and for their inhibition of autophosphorylation of the EGF receptor in human epidermoid carcinoma cells. The authors state, that two of the described compounds produced in vivo tumor growth delays of 13-21 days against advanced stage in nude mice, when administered twice per day on days 7-21 posttumor implant. Treated tumors did not increase in size during therapy and resumed growth at the termination of therapy, indicating an apparent cytostatic effect for these compounds under these treatment conditions. The data suggest that continuous long-term therapy with these compounds may result in substantial tumor growth inhibition.

As we can observe from the named article, the molecules and target of the found fragment 57302 are relevant in chemical research and possibly in medical applications.

7.1.2 Relevance of fragment 60336

The four targets of the molecular fragment 60336 are the single protein known as trace amine-associated receptor 1, the single protein enzymes named nitric oxide synthase, the cell line hepatocyte found in the human liver, and the tissue named liver microsome, as described in Chaper 6.2.

Miller G. M. et al. [15] describe the trace amine associated receptor 1 (TAAR1), its functional role in the regulation of brain monoamines, and the mediation of action of amphetamine-like psychostimulants. The authors state that research on TAAR1 opens the door to a new avenue of approach for medications development to treat drug addiction, as well as the spectrum of neuropsychiatric disorders hallmarked by aberrant regulation of brain monoamines. Furthermore, they state that the research on TAAR1 will likely promote the development of a new generation of therapeutics. Additionally, Jing L et al. [16] state that TAAR 1 knockout mice demonstrate increased sensitivity to dopaminergic activation, while TAAR 1 agonists reduce the neurochemical effects of cocaine and amphetamines, attenuate abuse- and addiction-related behavioral effects of cocaine and methamphetamine. The authors conclude that TAAR 1 agonists appear to be promising pharmacotherapies against psychostimulant addiction.

The regulation and function of the nitric oxide synthases (NOS) is described by Förstermann U. et al. [17]. Nitric oxide (NO) is an unorthodox messenger molecule, which has numerous molecular targets. In mammals, NO can be generated by three different isoforms of the enzyme NOS. All three NOS isozymes have regulatory functions in the cardiovascular system. Inducible NOS is found expressed in atherosclerotic plaque and is an important mediator of the fall in blood pressure in septic shock.

Liver microsomes are by far the most widely used in vitro model, providing an affordable way to give a good indication of the CYP and UDP-glucuronyltransferases (UGT) involved in the metabolism of a drug [18].

As we can observe from the named articles, the molecules and target of the found fragment 60336 are also relevant in chemical research and possibly in medical applications. Furthermore, looking into the relation of these targets may yield interesting results.

7.2 Summary and outlook

During this bachelor thesis we found significant fragments and targets of a set of fragments. This was achieved by initially collecting the needed data using the FDB, ATB and ChEMBL databases. The molecules, in which the fragments are contained, and the targets of these molecules were obtained. Subsequently the significant fragments were filtered by using a linear regression and a bootstrap algorithm to obtain a cut-off. We are interested in fragments, that are contained in many molecules, which have few targets. Fragments beyond the computed cut-off are assumed to be significant. 141 fragments were found using these methods. As discussed, the two considered fragments and their targets are relevant for chemical research and possibly medical applications.

Following this bachelor thesis, the obtained data and computed values could be expanded on, considering the fragment and molecule sizes. Calculating the cut-off value using a nonlinear regression, instead of a linear regression, and a slightly modified bootstrap algorithm, might yield better results. Furthermore, the data could be included into the FDB, additionally containing a p value for each fragment describing its significance.

The regression and bootstrap methods used for this thesis revealed new insights about the relations of fragments, molecules, and targets, and opened up new research approaches on medical drugs and their properties. As Big Data gains relevance in modern research, these methods pose promising new ways of dealing with large amount of data.

8 Acknowledgments

I would like to express my gratitude to the supervisor of this thesis Prof. Dr. Gunnar Klau for his support and for giving me the opportunity to write this bachelor thesis. In addition, I would like to thank Dr. Martin Engler for his advice and for his patience with my many questions.

I would also like to thank Prof. Dr. Holger Schwender from the Chair of Mathematical Optimization for being this work's second assessor. Additionally, I would like to thank him and Philipp Berger for the help they provided with the statistical methods used during this bachelor thesis.

9 Bibliography

References

[1] Engler MS, El-Kebir M, Mulder J, Mark AE, Geerke DP, Klau GW. (2017) **Enumerating common molecular substructures**. *PeerJ Preprints* 5:e3250v1 `https://doi.org/10.7287/peerj.preprints.3250v1`

[2] **An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0** Alpeshkumar K. Malde, Le Zuo, Matthew Breeze, Martin Stroet, David Poger, Pramod C. Nair, Chris Oostenbrink, and Alan E. Mark *Journal of Chemical Theory and Computation* 2011 7 (12), 4026-4037 DOI: 10.1021/ct200196m

[3] Bertrand Caron: **Public Python API client for the ATB API** `https://github.com/bertrand-caron/atb_api_public` (visited on August 22, 2018)

[4] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res.* 2012,40, D11001107.

[5] Micha Nowotka: **Official Python client for accessing ChEMBL API**. `https://github.com/chembl/chembl_webresource_client` (visited on August 22, 2018)

[6] **Anaconda Software Distribution**. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. `https://anaconda.com`.

[7] Köster, Johannes and Rahmann, Sven. **Snakemake - A scalable bioinformatics workflow engine**. *Bioinformatics* 2012.

[8] **Scikit-learn: Machine Learning in Python**, Pedregosa et al., *JMLR* 12, pp. 2825-2830, 2011.

[9] Michael Waskom, ., Olga Botvinnik, ., Paul Hobson, ., John B. Cole, ., Yaroslav Halchenko, ., Stephan Hoyer, ., Dan Allan, . (2014, November 14). **seaborn: v0.5.0 (November 2014)** (Version v0.5.0). *Zenodo.* `http://doi.org/10.5281/zenodo.12710`

[10] **Matplotlib: A 2D Graphics Environment** by J. D. Hunter In *Computing in Science & Engineering*, Vol. 9, No. 3. (2007), pp. 90-95

[11] Seabold, Skipper, and Josef Perktold. **Statsmodels: Econometric and statistical modeling with python.** Proceedings of the 9th Python in Science Conference. 2010.

[12] Ludwig Fahrmeir,Thomas Kneib,Stefan Lang,Brian Marx. **Regression**. *Springer-Verlag Berlin Heidelberg*, 2013. Print-ISBN 978-3-642-34332-2

[13] A. C. Davison, D. V. Hinkley. **Bootstrap Methods and their Application**. Series: *Cambridge Series in Statistical and Probabilistic Mathematics*, 1997. ISBN-13: 9780521574716 — ISBN-10: 0521574714

[14] **Tyrosine Kinase Inhibitors. 13. StructureActivity Relationships for Soluble 7-Substituted 4-[(3-Bromophenyl)amino]pyrido[4,3-d]pyrimidines Designed as Inhibitors of the Tyrosine Kinase Activity of the Epidermal Growth Factor Receptor** Andrew M. Thompson,, Donna K. Murray,, William L. Elliott,, David W. Fry,, James A. Nelson,, H. D. Hollis Showalter,, Bill J. Roberts,, Patrick W. Vincent, and, and William A. Denny*, *Journal of Medicinal Chemistry* **1997** 40 (24), 3915-3925 DOI: 10.1021/jm970366v

[15] Miller G. M. (2011). **The emerging role of trace amine-associated receptor 1 in the functional regulation of monoamine transporters and dopaminergic activity.** *J. Neurochem.* 116, 164176. 10.1111/j.1471-4159.2010.07109.x [PMC free article] [PubMed] [Cross Ref]

[16] Jing L, Li JX (August 2015). **Trace amine-associated receptor 1: A promising target for the treatment of psychostimulant addiction.** *Eur. J. Pharmacol.* 761: 345352. doi:10.1016/j.ejphar.2015.06.019. PMC 4532615Freely accessible. PMID 26092759.

[17] Förstermann U, Sessa WC. **Nitric oxide synthases: regulation and function**. *Eur Heart J.* 2012; 33:82937, 837a837d. 10.1093/eurheartj/ehr304 [PMC free article] [PubMed] [Cross Ref]

[18] Y. Parmentier, M.-J. Bossant, M. Bertrand, B. Walther, **5.10 - In Vitro Studies of Drug Metabolism**, Editor(s): John B. Taylor, David J. Triggle, *Comprehensive Medicinal Chemistry II*, Elsevier, 2007, Pages 231-257, ISBN 9780080450445, `https://doi.org/10.1016/B0-08-045044-X/00125-5`

A Source code

The following link was checked on August 24, 2017.
The source code and results of this bachelor thesis:
`https://gitlab.cs.uni-duesseldorf.de/engler/BSc-thesis-fdb-statistical-analysis`