Master Thesis

Large Scale Activity Profile Induction for Small Molecules

Kaswara Saleh Kraibooj

06.05.2013

Albert-Ludwigs-University Freiburg im Breisgau

Technical Faculty

Institute for Computer Science

Chair for Bioinformatics

Eingereichte Masterarbeit gemäß den Bestimmungen der Prüfungsordnung der Albert-Ludwigs-Universität Freiburg für den Studiengang Master of Science (M. Sc.) Informatik vom 06. May 2013.

Bearbeitungszeitraum

24. 09. 2012 – 06. 05. 2013

Gutachter

Prof. Dr. Rolf Backofen

Prof. Dr. Stefan Guenther

Betreuer

Dr. Fabrizio Costa

Declaration

Statement of authenticity I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I hereby also declare, that my thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Erklärung Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Ort, Datum

Unterschrift

Inhaltsverzeichnis

Zusammenfassung

Virtuelle Screeningtechniken (VS) sind zu einem unentbehrlichen Instrument der Arzneimittelforschung- und entwicklung geworden. Vorhersage von Bioaktivität ist eine der wichtigsten Felder der VS. Viele Modelle wurden für dieses Ziel entwickelt. Jedoch sind die verfügbaren Modelle meist abhängig von der Beziehung zwischen Struktur und Aktivität (Ähnlichkeitsprinzip). Es ist bewiesen, dass dieses Prinzip Fehlfunktionen aufweist und somit in vielen Fällen nicht korrekt ist. Aus diesem Grund wurde das Aktivitäts-Aktivitäts Prinzip eingeführt. Aufgrund des Mangels an experimentellen Daten sind viele Moleküle nicht oder nur unzureichend repräsentiert.

In dieser Arbeit werden die virtuellen Aktivitätsprofile als eine neue Form der Repräsentation von Molekülen vorgestellt, die eine Alternative zu experimentellen Aktivitätsprofilen bieten. Darüber hinaus verbessern wir die Aktivitätsprofile, indem wir die Zahl der beitragenden Assays erhöhen. Dies wird erreicht, indem semi-überwachte Algorithmen eingesetzt werden. Sie werden verwendet, um neue Moleküle zu Assays hinzuzufügen, die über eine zu geringe Anzahl an experimentell verifizierten Molekülen verfügen. Dadurch verfügen diese Assays über genug Moleküle, um gute Modelle zu erhalten. Wir erstellten die gesamte benötigte Infrastruktur, sowohl Hardware als auch Software. Wir verwendeten Daten aus dem öffentlichen Repository PubChem. Die resultierenden Modelle der virtuellen Aktivitätsprofile

zeigten eine vergleichbare Qualität zu den strukturellen Profilen. Deshalb kombinierten wir beide. Die neuen Modelle waren leicht, jedoch nicht signifikant besser als die strukturellen. Vielversprechende Ergebnisse ergaben sich aus einigen Experimenten, nachdem der semi-überwachte Algorithmus miteinbezogen wurde. Jedoch ist weitere Forschung nötig, um die Richtigkeit dieser Ergebnisse zu überprüfen. Als Ergebnis der vorgelegten Arbeit lässt sich festhalten, dass eine gute Repräsentation für fast alle Komponenten des Aktivitätsraumes gefunden werden kann. Es wird erwartet, dass dies uns zu vielfältigen neuen Vorhersageergebnissen in Bezug auf Aktivität führen wird.

Abstract

Virtual screening techniques (VS) have become an essential part of drug discovery research. Biological activity prediction is one of the most important fields of VS. Many models has been generated for this goal. However, usually, the available models depend on the structure-activity relationship (similarity principle). It is proven that this principle has malfunctions and is not correct in many cases. Therefore, the activity-activity principle has been introduced. This requires molecular representation in the activity space. Many molecules are without representation or with no sufficient representation because of the lack in the experimental data.

Here, we introduce virtual activity profiles as a new representation of the molecules which are alternatives to the experimental activity profiles. Moreover, we improve the activity profiles by increasing the number of contributing assays. This is done by employing a semi supervised algorithm to add new molecules to the assays of too small number of molecules which are experimentally verified. In this way these assays have sufficient number of molecules to get good models. We prepared the required infrastructure (hardware and software). Our data were brought from the public repository PubChem. The resulting models of the virtual activity profiles have shown worse quality than the structural ones, but the difference is not considerably significant. Therefore, we combined both of them. The new models are shown to be slightly, but still not significantly better than the structural ones. Promising

results have been shown by few experiments after incorporating the semi supervised algorithm. However, further research is needed to investigate the correctness of these results. As a result of the presented work, we can come up with a good representation for almost every compound in the activity space. This is expected to lead us to diverse new activity prediction results.

Chapter 1

Introduction

Motivation Despite the large amount of work and effort expended in the process of drug discovery and development, this process is still not efficient. It is too costly in money and in time. The overall costs of developing one new drug is about 1.8 billion US$ [40]. In about 16 years one molecule out of 9000 to 10,1000 reaches the market [12]. Figure 1.1 shows a diagram of the very general stages of the process and the average time required for each stage. Many in vitro and in silico methods have been developed for this goal. This work belongs to the in silico methods or what is called virtual screening (VS) methods. These methods reduce the number of drug candidates at different stages of the process. For a specific target, many molecules can be excluded because they do not show the expected activity. This process is very important, since it reduces the time and the costs required by High Throughput Screening.

QSAR methods are one of the most important methods in VS. In QSAR, available experimental data are used to train models. The resulting models are then used to predict new data (not experimented yet). An essential issue of this method is how to get good models. The quality of the models depends on several factors. Two of them are very important. The first one is the way in which the molecules are represented,

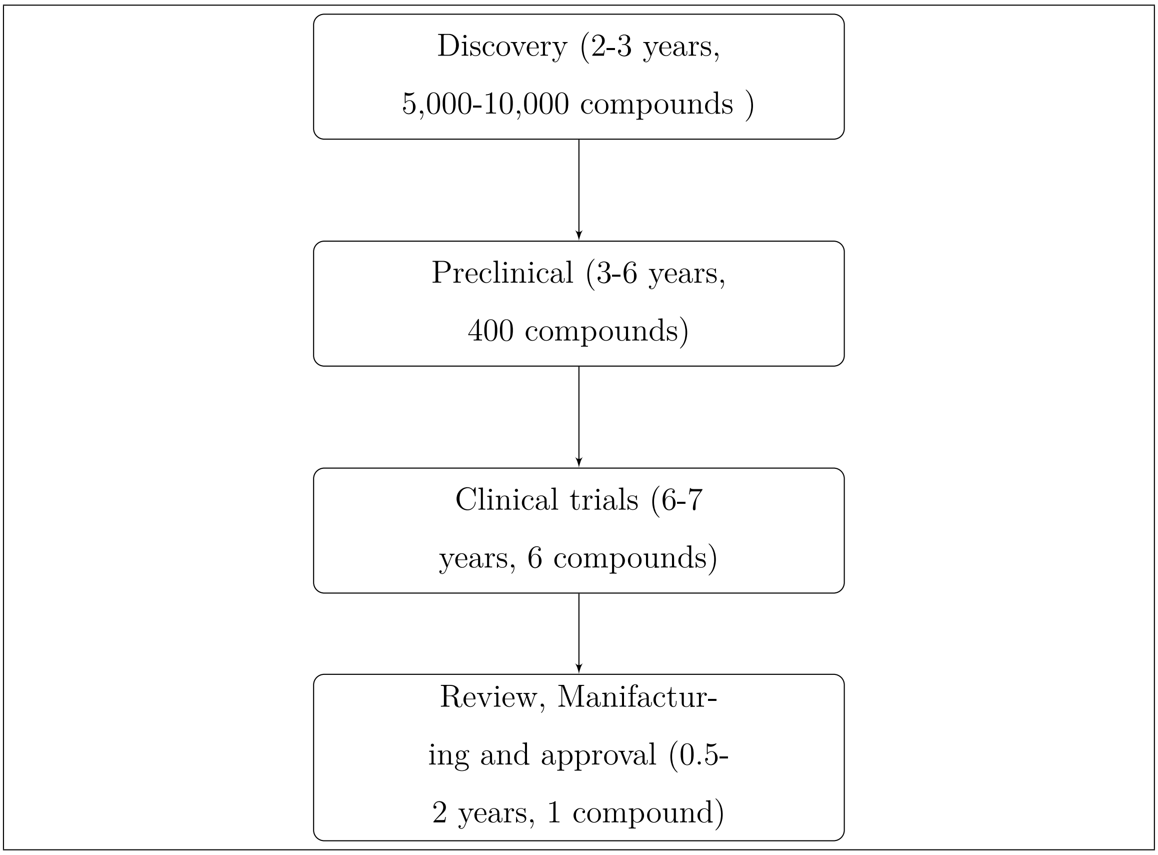

Discovery (2-3 years, 5,000-10,000 compounds)
Preclinical (3-6 years, 400 compounds)
Clinical trials (6-7 years, 6 compounds)
Review, Manifacturing and approval (0.5-2 years, 1 compound)

Figure 1.1.: The general process of drug discovery and development with the approximate time required to finish each step.

and the second one is the number of molecules used to train the models. This is of course without forgetting the importance of the learning algorithm by which we get the models.

Previous related work QSAR methods depend on the similarity hypothesis: structurally similar compounds are likely to have similar activity [18]. However, it is known that in some cases a small change on the structure of an active compound changes the activity of the compound to be even inactive [26]. This problem and others raised the need for a new solution. Therefore, in the few past years new approaches have introduced to circumvent this problem. One of them is replacing the structural descriptors by bioactivity ones. In early work at the institute NCI

(1991), activity fingerprints of 60 bits (60 cancer cell lines) have been generated and then used to discover the similarity of compounds. It was shown that similarity in activity patterns often indicates similarity in mechanism of action, mode of resistance, and molecular structure [55]. The tool HTS-FB in [41] uses fingerprints of 195 bits (Novartis assay panel) for ~ 1.5 million compounds to study the relationship between the molecules (and other applications). Using this tool, new bioactive compounds were detected starting from dissimilar compounds in structure. In [52] a chemical similarity searching exploits the tool HTS-FB to include the not yet tested molecules in the existing set of molecules. This is done by finding a similar reference compound, and then replacing it by the compound under study and then continuing with HTS-FB.

Contribution The main obstacles of this type of methods is that they don not work when we have a small amount of experimented data. To tackle this problem, we suggest two approaches. First, we replace biological activity profiles by virtual activity profiles. Second, computationally, we increase the number of molecules contributing to one assay. The first approach gives every molecule a representation in the activity space and the second approach extends the activity profile by new predicted activity values which assumingly increases the power of the descriptors.

Method overview For this we prepared the required infrastructure. A database of required tables are created. Data are collected from the public repository PubChem [5] and stored in the database. Models are generated by the machine learning technique SVM which is integrated with the kernel NSPDK [7]. The validation of the models are done by the 10 cross validation method. Different models are then compared by measuring their performance. The mean value of APR and ROC measurements are used for comparison. The significance of the results are measured by standard deviation.

Thesis structure overview Chapter 2 introduces basic conceptual terms and an overview of the drug discovery field. Chapter 3 contains terms and concepts required to understand the methodology of the work. Chapter 4 explains all the data and material and methods used in the project. Chapter 5 shows the results of the experiments and the evaluation of the results. In chapter 6 we conclude the results and propose possible extension of the work as a future work.

Chapter 2

Fundamentals

This chapter provides all definitions and concepts necessary to understand this work. Section 2.1 contains basic definitions of chemistry, biochemistry and pharmacology which helps to build a good knowledge base required for understanding the rest of the sections. Section 2.2 explains in more detail the drug and related concepts such as the concept of the drug activity, which is an essential element in this work. Section 2.3 contains an overview about drug discovery and the most important methods in this field in addition to other relevant issues.

2.1. Basic definitions

A *molecule* [8, 30] is a specific set of atoms connected in a specific way by attractive forces. The terms *biological molecule* or *biomolecule* are used to refer to molecules inside a living organism such as proteins, nucleic acids and small molecules. A *compound* [8] is a molecule such that it has at least two different atoms. *Hint:* In this thesis I will use the terms molecule, compound, chemical to refer to the same entity. This entity itself is not our concern in this work. We use the entity as an input to achieve the main goal of improving the predicted models. A *small molecule* is a molecule with a molecular weight of less than 800 Daltons such as metabolites, see [50]. An *assay* is a very accurate method for measuring properties of a system or object and for interpreting this measurement, see [58]. A *bioassay* is defined as an assay that measures biological activity by measuring the response of a biological test system to a test substance, see [58]. *Biological activity* or *bioactivity* is the effect of a substance on a biological test system, see [58]. A *receptor* is a molecule on the surface of a cell that binds to other molecules receiving chemical signals that effect the cell [13]. A *ligand* is the molecule that binds to receptors [13]. A ligand may fit or may not fit to a receptor as shown in figure 2.1.

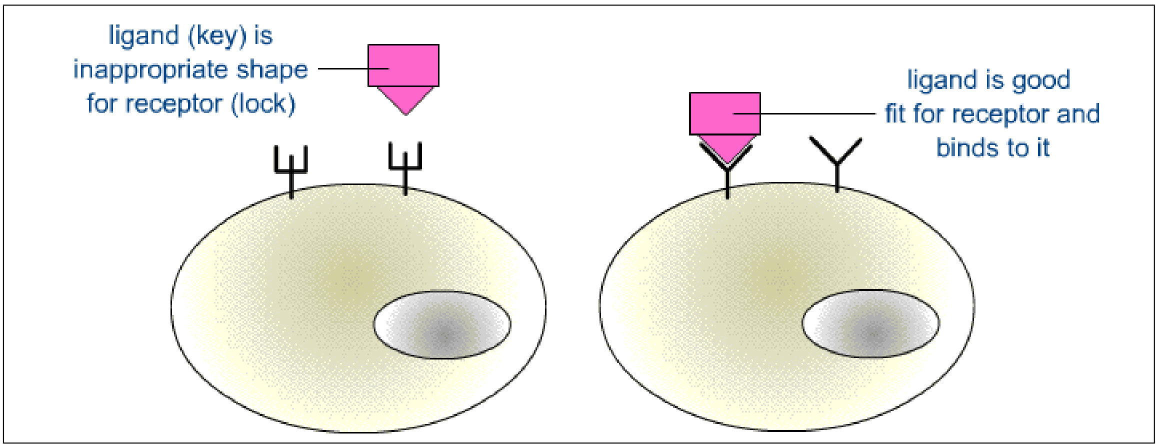


Figure 2.1.: Ligand and receptor [49]. On the left hand side, the ligand can not bind to the receptor. On the right hand side, the ligand can bind to the receptor.

2.2. The concept of drug

In pharmacology, a *drug* is any chemical substance which can react with a body and causes a physiological effect on it [34]. Generally, a drug may not provide a therapeutic effect [47]. A drug is called medicine if it causes a therapeutic effect. Two main types of drugs can be distinguished according to the type of their action:

- An *agonist*: A chemical compound which binds to a receptor and provoke some response in the corresponding cell [38].
- An *antagonist*: A chemical compound which inhibits a response when it binds to a cell's receptor [38].

Small molecules are easier to be absorbed by the cell membrane than the large ones. Therefore, most of available drugs are small molecules. The molecules in this work are small ones.

Drug target In pharmacology, a *biological target* is a biological molecule whose function can be affected by an external stimulus [43]. Drug targets are in general proteins (like G-protein-coupled receptors (or GPCRs) and protein kinases), ion channels (like ligand-gated ion channels and voltage-gated ion channels) and nucleic acids [22, 38]. Target-based methods in drug discovery are the most common ones. For this aim, several databases were established like TTD (Therapeutic Targets Database). This work is not a target-based method. However, what we need to know about the target is just the bioactivity value (*IC50* see below) resulted from the reaction between targets and molecules.

Drug activity Several properties refer to the activity of a drug: *Potency*: A measure of the amount of a drug required to produce a desired response [39]. *Efficacy or intrinsic activity*: A measure refers to the potential maximum therapeutic response that a drug can produce [46]. *Affinity*: A measure of the strength of the bind between a drug and a receptor. The higher the affinity, the lower the side effects [42]. Moreover, these properties are related to each other. For example, the higher the potency, the higher the affinity.

Measurement of drug activity A quantitative measurement of these properties can be IC50 and EC50. *The half maximal inhibitory concentration (IC50)* measures the concentration of a drug to inhibit a biological process [16]. IC50 is used in PubChem [5], which is the source of our data. IC50 measures the potency of an antagonist.

The half maximal effective concentration (EC50) measures the concentration of a drug to trigger a biological process [16]. EC50 is used to measure the potency of an agonist.

2.3. Drug discovery

A more detailed view of the process of drug discovery is necessary to understand what this work is about. The field of drug discovery is related to several other fields, medicine, pharmacology, biology, chemistry, biotechnology and others. The process is too long, too costly and tedious despite the big progress in the field [3]. Throughout the history, the principle of developing a new drug candidate has changed. Currently, the work in this field is a collaboration between the dry lab efforts and the wet lab efforts. The methods in the wet lab are called real screening whereas the methods in the dry lab (computational machines) are called the virtual screening (VS). In the following sections I explain these methods in summary.

2.3.1. High-Throughput Screening

HTS (biological screening or real screening) is the process in which a large number of chemical compounds are (in vitro) tested against a target for the purpose of finding hits. More than 100,000 compounds can be screened per day [51]. HTS has several advantages such as rapidness, simplicity and automation. Additionally, HTS could reduce the costs of drug development in several types of assays [10]. However, HTS is still not efficient. Therefore it was necessary to make use of the computational power in software and hardware. This led to the emergence of virtual screening methods.

2.3.2. Virtual Screening

VS is screening by computational machines. The increasing availability of 1) very powerful computational machines and 2) chemical compounds, has increased the importance of VS. VS does not substitute HTS, rather than VS integrates HTS [12]. There are two types of virtual screening, the *ligand-based techniques* and the *target--based techniques* [29]. In the ligand-based techniques, active compounds (ligands) and their targets are available. Using these compounds, new compounds of similar activity are searched. The new compounds should be of different structure [4]. In target-based techniques (structure-based techniques), accurate 3D descriptors of the studied targets should be available [12]. Docking algorithms are used to find the well binding ligands [25] and then scoring functions are employed to estimate the strength of the bind. The two techniques can be combined to increase the probability to detect new hits [20].

2.3.2.1. Molecular descriptors

Molecule properties Molecule comparison is essential in ligand-based VS. This requires a way to computationally represent the molecules using their properties. This is the reason why molecular descriptors were made [12]. Properties of molecules are characterized as numerical values to form the so-called molecular descriptors [23]. Generally, the properties used in descriptors are classified to molecular structure, physiochemical properties and pharmacophore ones. There are thousands of different molecular descriptors [48]. The diversity of descriptors is due to the diversity of objectives for which these descriptors are made [59].

However, two general important parameters distinguish descriptors from another. The first parameter is the computational efficiency in calculating and using them. The second parameter is the amount of information they encode. Moreover, There is a relation between these two parameters. The more efficient in computation the descriptors are, the less content of information they have.

The simplest descriptors are those which depend on a small number of features such as molecular weight, hydrogen bonds donors and others. These descriptors are not sufficient in isolating molecules from each other. Therefore, usually they are merged with other descriptors.

Dimensionality Several principles are considered when classifying descriptors. Usually, descriptors are classified according to their dimensionality, 1D, 2D and 3D.

1D descriptors are the simplest ones and the most efficient in computation, however, they do not hold a lot of information. Therefore, usually they are combined with other descriptors.

2D descriptors have the advantage of being a trade-off between efficiency and information content. Examples of 2D descriptors are those which exploit physio-chemical properties like the ones used in fragment-based methods. These methods break the molecule into fragments and give each fragment a value. The most common and used type of physio-chemical methods are 2D fingerprints. A fingerprint is a boolean array or a sequence of bits such that (1) means the presence and (0) means the absence of a particular fragment of the molecule [12]. 2D fingerprints are good at molecular similarity searching [23]. There exist two types of 2D fingerprints, the dictionary-based fingerprints and the hashed fingerprints. The first type has the advantage of that each bit represents one substructure which helps in interpreting the results. The second type has the advantage that it does not need a already defined dictionary. This feature lets this type of fingerprints applicable to any molecular structure [24].

3D descriptors are the richest ones. They represent more properties of molecules such as the electronic properties of the molecules. However, they are computationally expensive. Examples of them are 3D fragment screens and pharmacophore keys. Selecting the descriptor which fits the goal of a study is not a simple task. For this, automated selection tools were created like in [14].

SMILES The simplified molecular-input line-entry system (SMILES) is a linear ASCII string for representing the structure of chemical molecules. This string is a computer readable format. It can be converted into 2D and 3D descriptors. SMILES allows users to annotate any chemical structure [2, 53, 54].

The main advantage of SMILES that it is easy and efficient to be processed by computers. This notation was invented by Arthur Weininger and David Weininger in 1987. Each structure has a unique SMILE. There are several rules to write the syntax of a SMILE and several various algorithms to generate the SMILE. This work is not the right place for these rules and algorithms. However, we can mention basics to give a general idea of what these rules are and how the algorithms work. For example, a five-rules system is used to generate the syntax of a SMILE of a two-dimensional structure of a chemical. These rules consider the following basic chemical structures: atoms and bonds, simple chains, branches, rings, and charged Atoms. Figure 2.2 shows a SMILE and a general algorithm to generate it.

Chemical (Molecular) Graph ”The graph with differently labeled (colored) vertices (chromatic graph) which represent different kinds of atoms and differently labeled (colored) edges related to different types of bonds. Within the topological electron distribution theory, a complete network of the bond paths for a given nuclear configuration” [32].

In this way, a molecules becomes a mathematical element ready to be theoretically studied using all available methods. Molecular graphs can be generated from

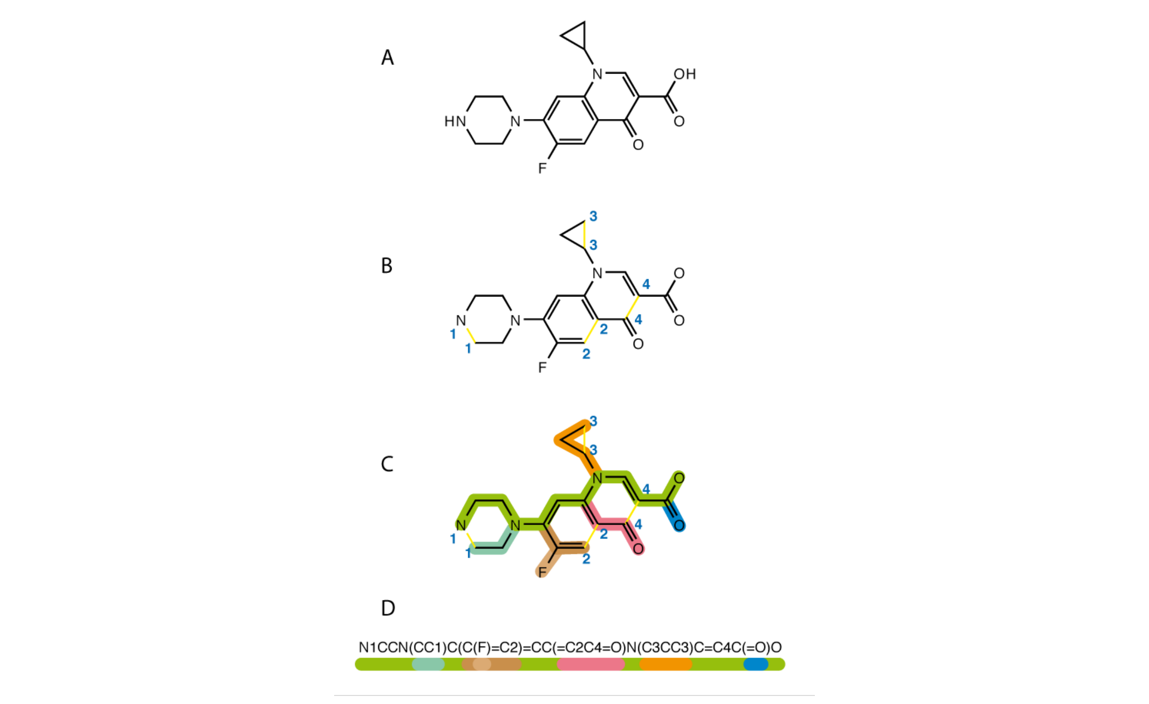


Figure 2.2.: A chemical structure and its SMILE. Generation of SMILES: Break cycles, then write as branches off a main backbone [57].

SMILES.

2.3.2.2. Molecular Similarity

Similarity principle Similarity methods depend on the *similarity principle* which states that similar molecules in structure have similar properties [18]. Being merely an assumption means that the principle is not always correct [21]. However, this concept has been widely applied in chemoinformatics [18, 27, 35] and VS (similarity-based VS techniques [44]) and has showed positive results [28].

In similarity searching methods a database is searched for structurally similar compounds of a query compound of known activity [23]. The goal of studying the similarity is to explore properties of new molecules. Figure 2.3 shows examples of some similar compounds. To decide the similarity of molecules we need two requirements: 1) A suitable representation of molecules and 2) An efficient comparison method.

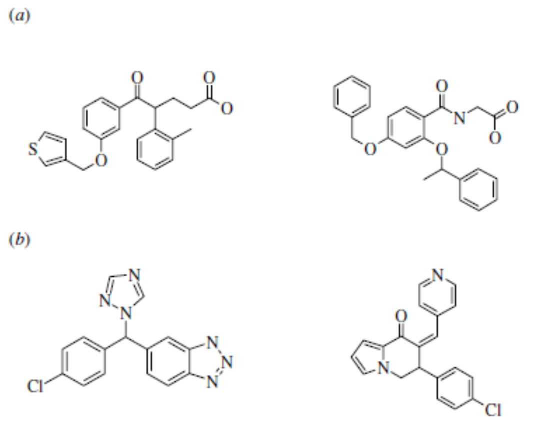


Figure 2.3.: 2 examples of similar compounds [23]. Similarity relationships that were detected using similarity methods: (a) endothelin A antagonists, (b) aromatase inhibitors.

Similarity metrics Table 2.1 shows a list of the most frequently used similarity metrics. For example, the Tanimoto similarity which enumerates the number of common fragments between two molecules.

Metric name	Formula
Tanimoto coeficient	$Tc = n_{ij}/(n_i + n_j - n_{ij})$
Dice coeficient	$Dc = n_{ij}/(n_i + n_j)/2$
Cosine coeffiecient	$Cc = n_{ij}/(n_i n_j)^{1/2}$
Hamming distance	$HD_{ij} = (n_i + n_j - 2n_{ij})$
Euclidean distance	$ED_{ij} = (n_i + n_j - 2n_{ij})^{1/2}$

Table 2.1.: The most common similarity metrics [12].

2.3.2.3. QSAR

Quantitative structure-activity relationship (QSAR) is the process in which the relation between the molecule properties and the molecule activity is explored

[12]. QSAR methods can be employed for hit identification and lead optimization. Classical QSAR methods assume that this relation is a linear one [11].

$$\text{Activity} = f(\text{physicochemical properties and/or structural properties}) + \text{Error} \quad (2.1)$$

QSAR methods may be classified in accord with the molecule descriptors. Accordingly, we will have 1D-QSAR; 2D-QSAR, 3D-QSAR and 4D-QSAR. As earlier mentioned, for efficiency reasons, 2D descriptors are preferred when screening large databases of molecules. Given a set of input data, QSAR models these data and then predicts new data of the same type. The resulting models from QSAR need to be validated. The quality of a QSAR models depend on several factors such as the type of used descriptors, the quality of data, the modeling methods and even on the validation. QSAR is used in VS. In this context, QSAR takes the chemical descriptors as input and produces the activity of new chemicals.

Chapter 3

Scientific Background

3.1. Basic definitions in machine learning

The following definitions are based on the book ”foundations of machine learning” [33].

Learning: It is the process of understanding existing knowledge and using this knowledge to predict not existing knowledge of the same type. *Machine learning*: It is a computational algorithm for learning. It should be efficient and accurate. *Examples*: Instances of data used for learning. *Features*: Set of properties which represent an example as vectors. *Labels*: Values given to examples for categorizing. *Training sample*: A set of examples represented by their features. This set is used by learning algorithm to learn from. *Validation sample*: A set of examples used to empirically find the optimal values of the parameters of a learning algorithm. *Test sample*: A set of examples used to test the performance of a learning algorithm. *Loss function*: A function used to measure the difference between a predicted and true label. The true labels and predicted labels of a test set can be input of this function. *Supervised learning*: All training data are labeled. *Unsupervised learning*: All training data are unlabeled. *Semi supervised learning*: Training data are a combination of labeled and unlabeled data. *Classification*: Assign a label to each example. *Clustering*: Partition examples into homogeneous classes.

Support vector machine (SVM) One of the most effective methods in the field of machine learning. It belongs to supervised learning algorithms. It was introduced first in 1995 by Vladimir N. Vapnik [6]. SVM is used in classification and regression. Non-linear classification can efficiently done by SVM. This is due to the kernel trick [1]. Given the features and the lables of a training sample, SVM tries to construct a hyperplane which is used to classify the examples to their labels. The best hyperplane is the one which separates the examples by the largest margin. Figure 3.1 shows two hyperplanes for the same training set. The right hand side hyperplane separates the set with a bigger margin than the one of the left hand side.

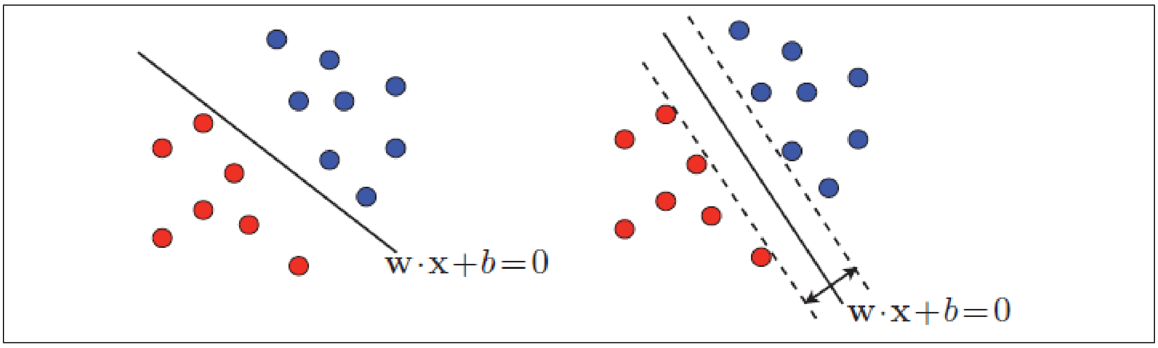

$\mathbf{w}\cdot\mathbf{x}+b=0$
$\mathbf{w}\cdot\mathbf{x}+b=0$

Figure 3.1.: Two hyperplanes of one training set [33]

3.2. K nearest neighbors (Knn)

It is one of the simplest machine learning algorithms for classification [9]. The input of the algorithm is a set of vectors (molecules) of known classes and another set of vectors of unknown classes and the question is how to predict the class of these vectors. The core idea of the algorithm is that for one new vector k nearest neighbors are searched. A majority voting of the k neighbors decide the class of the vector. Selecting k and the distance metric between the vectors are two essential matters. Both depend on the data sets and the size of these data.

3.3. Classifier performance

A binary classifier gives tow results, one or zero, yes or no, positive or negative[1]. However, these results may be true or false. Therefore, from a statistical point of view, we have four results, true active, true inactive, false active and false inactive. I will notate these four values as: TA, TI, FA, FI. Based on these four values we have the definition of the sensitivity and specificity.

Sensitivity: It measures the proportion of the number of molecules correctly identified as active to the whole number of active molecules. It is given by the equation:

$$\frac{Number\ of\ TA}{Number\ of\ TA\ +\ Number\ of\ FI} \tag{3.1}$$

Specificity: It measures the proportion of the number of molecules incorrectly identified as active to the whole number of active molecules. It is given in the equation:

$$\frac{Number\ of\ TI}{Number\ of\ TI\ +\ Number\ of\ FA} \tag{3.2}$$

ROC It stands for Receiver Operating Characteristic curve or Relative Operating Characteristic curve [31, 61]. It is one of the best accuracy measurements of diagnostic tests. It has several applications. One of them is to measure the accuracy of binary classification as in this work. A ROC curve represents the relation between

[1] in this work our data (molecules) are classified into active and inactive. Therefore we will use these two values as classifier results

the sensitivity and specificity. The area under this curve represents the value of ROC. The bigger this area the more accurate the measured test.

3.4. Graphs

This section contains basic definitions of graph theory. The content depends on the book ”Handbook of Graph Theory” [15] and the article [7].
A Graph $G = (V, E)$ consists of two sets V, E. The elements of V are called *nodes* or *vertices*. The elements of E are called *edges*. The notations $V(G), E(G)$ are used when the graph G is not the only one under consideration. Each edge has a set of one or two vertices associated to it. This set is called *endpoints*. A node u is *adjacent* to node v if they are joined by an edge. Each two adjacent nodes are called *neighbors*. A walk in graph G is an alternating sequence of edges and vertices,

$$W = v_0, e_1, v_1, ..., e_n, v_n$$

such that for $j = 1, ..., n$, the vertices v_{j-1} and v_{j-1} are the endpoints of the edge e_j. The *distance* between two nodes v, u, denoted $\mathcal{D}(v, u)$ is the length of the shortest walk between them. A *connected graph* is a graph which has a walk between every pair of vertecis. The eccentricity of a vertex v in a connected graph is the distance of a vertex farthest from v. The *radius* r of a connected graph is its minimum eccentricity. The *neighborhood* of radius r of vertex v is the set of vertecis of distance D from v such that $\mathcal{D} \leq r$ and is denoted by $N_r(v)$. In graph G, the *induced subgraph* on a set of vertices $W = w_1, ..., w_k$, denoted by $G(W)$, has W as it vertex-set, and it contains every edge of G whose endpoints are in W. The *neigborhood subgraph* of radius r of vertex v is the subgraph induced by the neighborhood of radius r of v and is denoted by $\mathcal{N}_r^v$. A *labeled graph* is a graph whose vertices and/or edges are labeled. The mapping function from a vertex/edge to a label is denoted as $\mathcal{L}$. Tow graphs $G_1 = (V_1, E1)$ and $G_2 = (V_2, E_2)$ are *isomorphic* ($G_1 \simeq G_2$), if there is a bijection $\phi : V_1 \to V_2$, such that for any two vertices $u, v \in V_1$, there is an edge vu *iff* there is an edge $\phi(u)\phi(v)$ in G_2.

3.4.1. Kernels

Kernel methods depend on the concept *kernel*. These methods are widely used in machine learning. The advantage of kernel methods that they can compute inner products efficiently [33]. Therefore, whenever a dot product problem exists, a kernel function is used. This is called the kernel trick [1]. When typical classification methods fail to find a hyperplane, kernels come into play. Kernels solve the problem by projecting the data into a space of higher dimensionality. In other words, a kernel function performs a mapping $x \to (x, x^2)$. Combining kernels with other techniques like SVM results in a powerful learning methods. Kernels are often used as similarity measurement. Many kernels are available. Selecting the best kernel strongly depend on the problem at hand. Tuning the parameters of the selected kernel is another important issue.

Definition and notation The following formal definition and notation is based on [7,17]. Given a set X and a function $K : X \times X \to \mathbb{R}$, K is a kernel on $X \times X$ if K is symmetric. That is, for any x and $y \in X$, $K(x, y) = K(y, x)$ and if K is a *positive--semidefinite.* K is positive-semidefinitive if for any $N \geq 1$ and any $x_1, ..., x_N \in X$ the matrix defined by $K_{ij} = K(x_i, x_j)$ is positive-semidefinitve, that is $\sum\limits_{ij} c_i c_j K_{ij} \geq 0$ for all $c_1, ..., c_N \in \mathbb{R}$. If each $x \in X$ can be represented as $\phi(x) = \{\phi(x)\}_{n \geq 1}$ such that K is the ordinary l_2 dot product $K(x, y) = \langle \phi(x), \phi(y) \rangle = \sum\limits_{n} \phi_n(x)\phi_n(y)$ then K is a kernel. The vector space induced by ϕ is called the *feature space.*

Chapter 4

Methods

This chapter includes an explanation of the data and the methods used to accomplish the work. In the first section we describe the data, how to store, organize and retrieve them. In the next section, we describe the models, how to get, validate and measure their performance. Finally, a semi-supervised algorithm is explained. Figure 4.1 shows a simplified and general diagram of all steps used in the work.

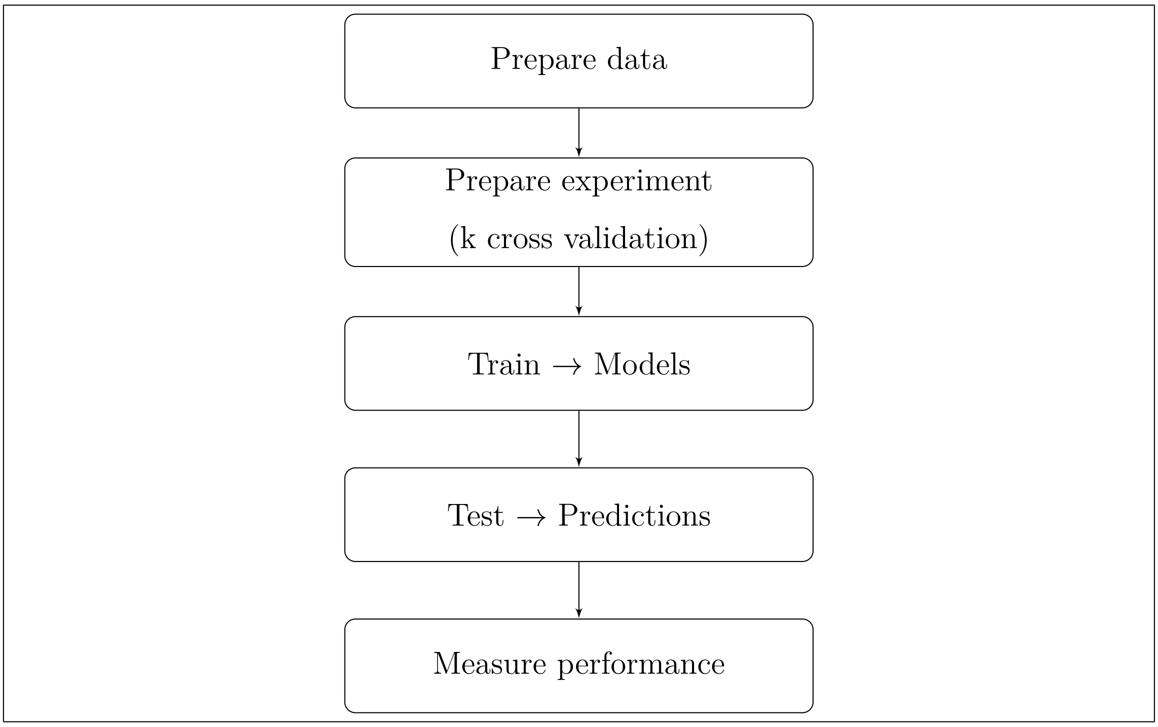


Figure 4.1.: All steps of the work and their order.

4.1. Infrastructure

The code of the application is written by Perl 5.10 in addition to bash Shell in Linux Fedora.

4.1.1. Data collection

The needed data for this work is taken from the free online public repository PubChem [5]. Specifically, our data are downloaded from the FTP site (`ftp://ftp.ncbi.nih.gov/pubchem/`). Molecules are stored in zipped SDF files. The data we need for each molecule are the identifier of the molecule and its SMILE. Assays are stored as zipped CSV files. The data we need for each assay is its identifier, molecule identifiers and their activity values against this assay. The application automatically downloads the files and extracts the required information. Downloading is essentially done by the tool *GNU wget* [36].

4.1.2. Data organization

A MySQL database is established for storing the data. The database mainly consists of four tables. One table is for storing the compounds (cid, SMILE). Another table is for storing the assays (aid). Another table is for storing the experimental activity values *IC50*. It has three fields (aid, cid, activity). The fourth table has in each row a molecule identifier and a descriptor of the molecule(cid, feature). A descriptor represents a feature generated by NSPDK (see paragraph 4.2). Figure 4.2 shows the four main tables and their relations.

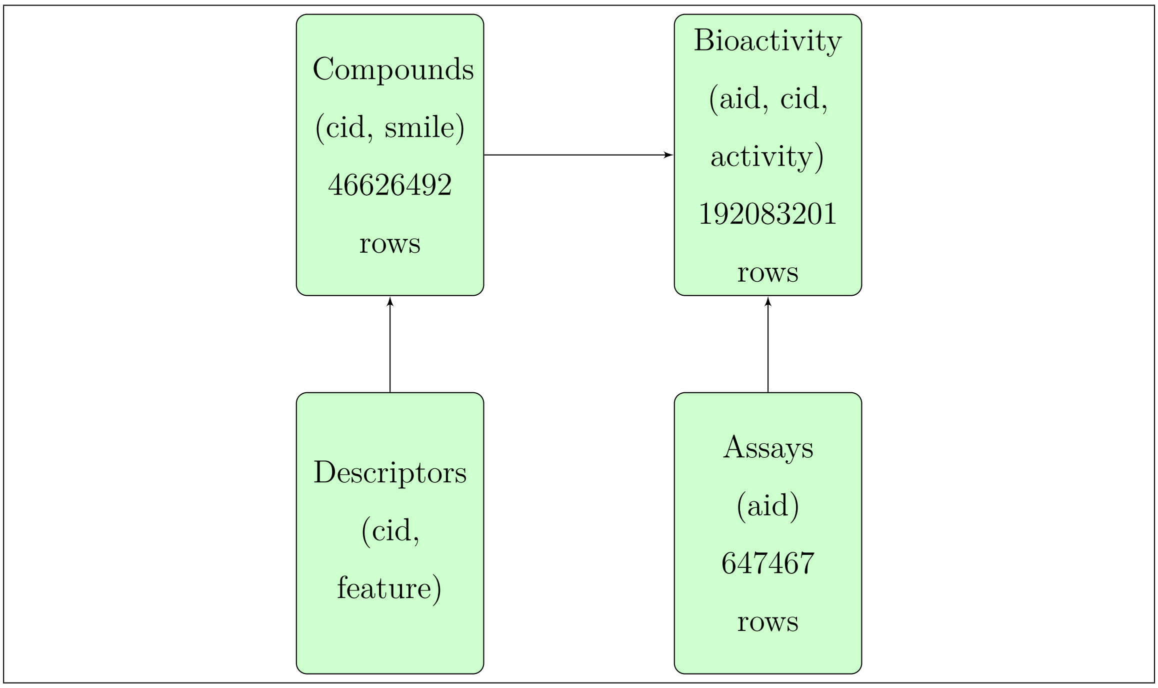

Compounds (cid, smile) 46626492 rows
Bioactivity (aid, cid, activity) 192083201 rows
Descriptors (cid, feature)
Assays (aid) 647467 rows

Figure 4.2.: The main four tables of the database, the number of entries of each of them, and their relations.

4.1.3. Data retrieval

Having the data organized in this way we can now simply retrieve the data we need. For example, given a number of assays we can retrieve all compounds related to it, i.e. all compounds which are already experimented on it. Moreover, in case we want to work on a specific data set, for example cancer, the application can do it by taking advantage of the available APIs of Pubchem. For this goal, we integrated our application with an API called *Ebot* [56] in order to fetch all assay identifiers which are related to a search text like cancer or HIV or breast cancer, etc.

4.1.4. Hardware

The application has been run on a machine whose CPU is of the model Intel(R) Core(TM)2 Duo CPU E6550 @ 2.33GHz and its main memory size is 4GB. Parallelized parts of the script has been run on a cluster of 100 node. The CPU of each node is of the model Dual Core AMD Opteron(tm) Processor 875. The size o the main memory of each node is 32GB.

4.2. Models

The most important part of this work are the models. Our models are predictive ones. A model represents the relation between molecules and their activity values. The model is expressed by the equation:

$$m = Xw \tag{4.1}$$

Where X is the feature of a molecular descriptor and w is the predicted weight. Figure 4.3 shows the flow diagram of the process of generating a model. In the next lines we show how to get these models, validate them and measure their performance.

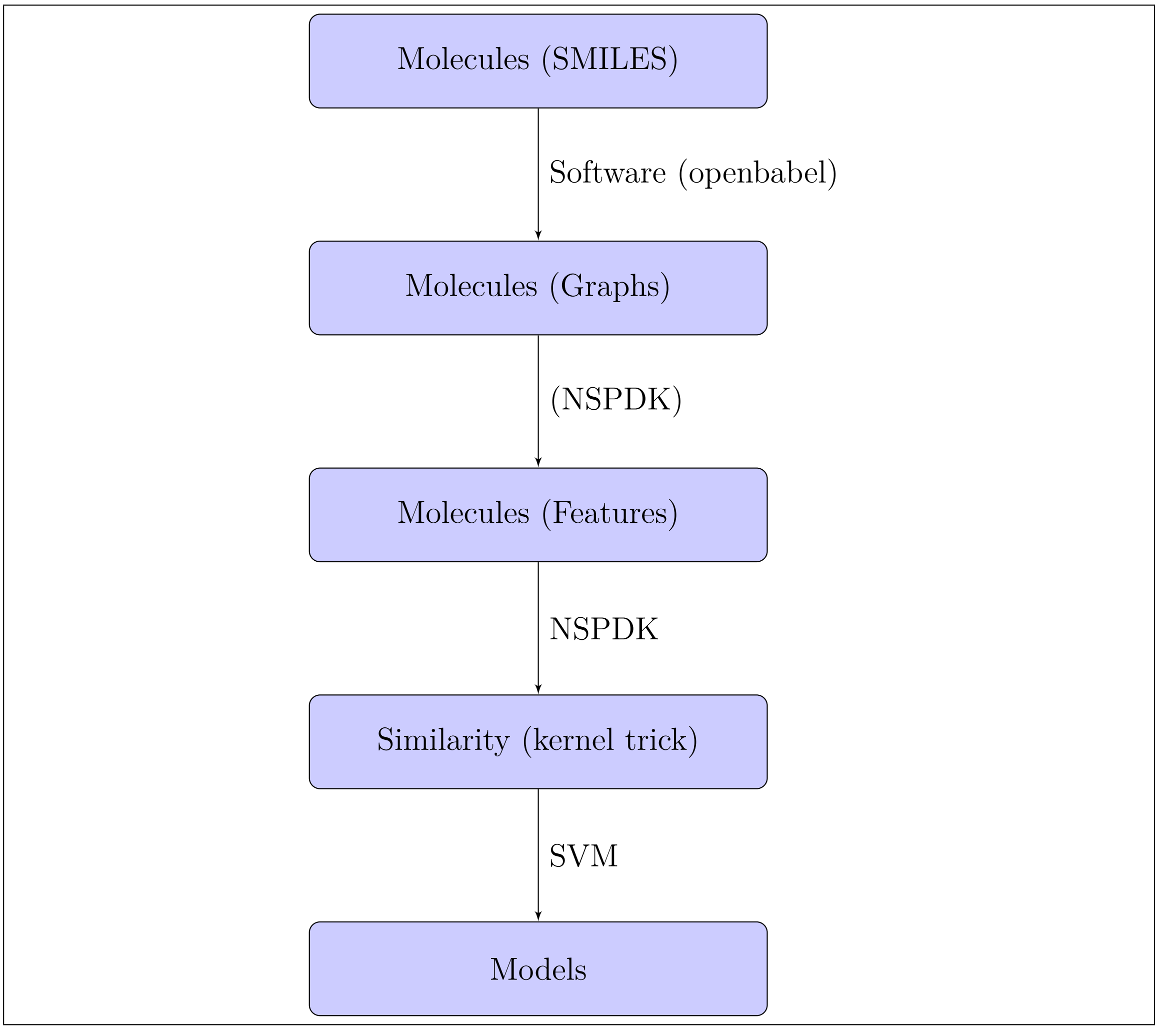


Figure 4.3.: The general procedure for getting models.

Molecular graphs In order to be able to generate models we need a mathematical representation of our data. SMILES are strings. Therefore, SMILES are converted into Graphs see paragraph 2.3.2.1. We used the software *Openbabel* [37] for conversion.

Features At this point, the molecules are graphs. Thought, we do not use the graphs to be the descriptors of molecules. But instead we compose new descriptors. The new descriptor is a vector of features. Each index of the vector represents a subgraph (feature) and the corresponding value represents the number of times this subgraph exists in the graph. The enumeration is done by **NSPDK**.

Similarity The similarity between two feature vectors then is decided by computing the inner product of the resulting vectors. The higher the value of the inner product, the more similar the relevant molecules. For this, we used the kernel **NSPDK**. The similarity metric is the Dice coefficient metric.

NSPDK It stands for Neighborhood Subgraph Pairwise Distance Kernel. It is a state of the art tool [7]. The kernel is employed to generate the feature vectors of the molecular graphs. Moreover it is applied in SVM to generate the models.
NSPDK works through two steps (the same notations used in the method chapter for graphs and kernels):

1. It decomposes a molecular graph into all its subgraphs. This is done by defining the relation R as:
   A_v, B_u are rooted graphs and have the relation $R_{r,d}(A_v, B_u, G)$ which is true when $v, u \in V(G)$ and $d = D(u, v)$.

2. It calculates the number of identical pairs of subgraphs of radius r at distance d. This is done by a decomposition kernel $\kappa_{r,d}$ given by the equation Equation 4.2

$$\kappa_{r,d}(G, G') = \sum_{\substack{A_v, B_u \in R^{-1}_{r,d}(G) \\ A'_{v'}, B'_{u'} \in R^{-1}_{r,d}(G')}} \delta(A_v, A'_{v'})\delta(B_u, B'_{u'}) \tag{4.2}$$

where G, G' are two rooted graphs. such that:

$$\delta(x,y) = \begin{cases} 1 & \text{if } x \simeq y \quad \text{(x, y are isomorphic graphs).} \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

The final equation which enumerates the number of identical subgraph pairs of increasing radius r at distance d is given by the next equation Equation 4.4:

$$K(G,G') = \sum_r \sum_d \kappa_{r,d}(G,G') \tag{4.4}$$

Model generation Having the descriptors (features, activity profiles, combined features) and their targets, SVM can train the corresponding models.

4.3. Generating new descriptors

Activity profile Activity profile of a molecule is a string of activity values against each assay under study. Given a number of assays and a number of molecules we can generate the activity profile of each molecule like the following:

1. Prepare train and test sets for each assay.
2. Generate models for each assay.
3. Predict the activity for each molecule using the models.
4. For one molecule concatenate the activity values in one string to be the activity profile of the molecule.

An activity profile looks like in Figure 4.4.

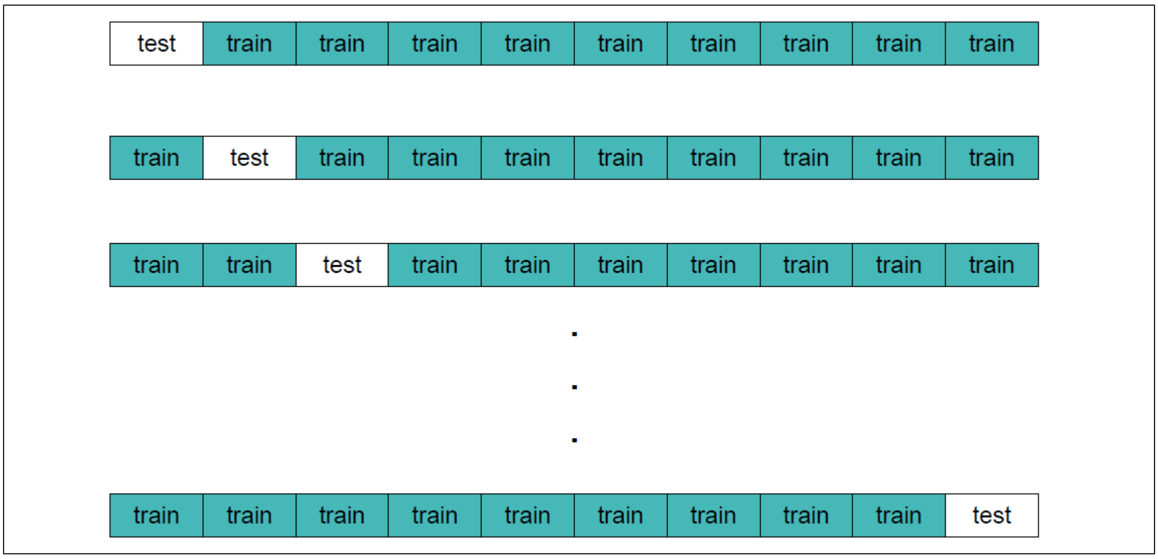
```
1:+1 3:-1 4:-1 20:+1 25:+1 101:-1 90:+1 3:-1 9:+1 5:+1 33:-1 43:+1 871:+1
```

Figure 4.4.: An activity profile.

Combined descriptors A new descriptor can be generated by combining both descriptors. To combine the both we need to normalize each vector. Additionally, we use importance factor $\alpha \in [0, 1]$ which specifies the importance of the activity profiles relatively to the structural features which have the importance $(1 - \alpha)$. Each value of the vector is multiplied by its importance factor. Afterwards, we change the index of each vector such that we avoid any conflict. Finally, we concatenate the resulting two vectors to get the new combined descriptor.

4.4. Model performance

K cross validation K cross validation is one of the best techniques to measure the generalization of a model on a set of data which does not belong to the train set which is used to get the models [19, 33, 45]. In this technique, k is an integer number (typically k=10). The data set under study is randomly divided into k equal sets (folds). Each fold represents a test set whereas the remained k-1 folds represent the corresponding train set. As a result of this splitting we get k test sets and k corresponding train sets. For each molecules we will have k predictions. An average of these predictions is calculated. The average represents the final prediction value. The advantage of this technique is that the resulting prediction value is a reliable one. Figure 4.5 shows a 10 fold cross validation.



test	train	train	train	train	train	train	train	train	train
train	test	train	train	train	train	train	train	train	train
train	train	test	train	train	train	train	train	train	train
train	train	train	train	train	train	train	train	train	test

Figure 4.5.: 10-folds cross validation.

Performance measurement Each fold of each assay results in a model. This model is tested on the corresponding test set. The resulting predictions are compared to the true values. ROC and APR are used to measure how accurate these predictions are. For one assay we will have k of ROC values and k of APR values. The mean value of ROC or APR is considered as an evaluation of the models performance. The error is estimated by standard deviation. The mean value μ and the standard deviation σ are given by the following relations respectively:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.5}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4.6}$$

4.5. Semi-supervised algorithm

The lack of the number of molecules experimented against some assay makes activity prediction a very difficult task. This is due to the fact that good predictive models require a sufficient size of training set. To overcome this problem, we need to increase the number of molecules of known activity for some assay. This is can be done computationally. For this we proposed exploiting a semi-supervised algorithm. Given a set of molecules and their experimental activity, an efficient algorithm can classifies a set of new molecules to active and inactive. In this way, the number of total molecules of known activity becomes bigger and sufficient for training good models.

For simplicity reasons we chose the algorithm in [60] to be employed in our work. The algorithm exploits the intrinsic manifold structure of the given set of data. The input data are required to be represented in the Euclidean space. The algorithm is summarized by three main steps:

1. Construct a graph using k nearest neighbor algorithm, see section 3.2. The nodes x_i are vectors of molecules features.
2. Weight the edges of the resulting graph using equation Equation 4.2 which represents the similarity value between two molecules x_i, x_j.
3. Normalize the weight matrix:
4. Iterate calculating the classification function $f(t)$ until convergence: $f(t+1) = \alpha S f(t) + (1-\alpha)y$; α is a parameter in $[0,1]$, y is the target vector whose values are either +1 or -1 or 0 in case the molecule is of unknown value.

In words: Given a set of molecules and their targets. First, we represent them as indirected graph. Construction of the graph is done by k nearest neighbor algorithm. Then, the edges are weighted. Second, the weights are normalized for convergence reasons. Third, the classification function is iterated until convergence. During iteration the classification score of each node gets effected by the classification score of its neighbors. The amount of contribution of classification score of other is expressed by α. The result is a vector of classification values of input molecules. These values are the activity values. The classification values are either positive real numbers or negative real numbers. We consider the negative numbers as inactive values (-1) and the positive numbers as active values (+1).

1

Chapter 5

Results

This chapter shows the results of the solution hypothesis, and then the result are evaluated. As already mentioned, the proposed solution (for improving the activity models) is based on two main points. First, using novel descriptors. Second, increasing the number of contributing molecules per assay. Therefore we designed all possible[1] experiments necessary to check whether our proposals achieve our goal. The next sections describe the specifications of one experiment. Afterwards, we compare models resulting from using different descriptors. Finally, in section 4.5 we discover the effect of employing the semi supervised algorithm described in the methods chapters on the quality of models.

5.1. Experiment

Before we introduce the results of the experiments, it is necessary to precisely specify what are the specifications of the experiment conducted here. An *experiment* is a

[1] We restricted ourselves to the available number of assays and molecules in our database (all Pubchem data) so that resulting models stay of good quality.

number of repetitions of k-fold cross validation of models resulting from a set of molecules which belong to a set of assays which represents some data set (here disease). Accordingly, an experiment can be specified by six parameters. First, number of folds k. Second, number of molecules (active and inactive). Third, number of assays. Fourth, the data set. Fifth, type of the molecular descriptor. Sixth, number of repetitions. For all experiments we fixed the number of folds (k=10) and the number of repetitions (=1). All other parameters were varied to serve studying the results from different points of view. The size of the experiment is specified by the first two parameters (see table 5.2). The fourth parameter, the disease name, allows us to increase the diversity of our experiments. We worked on cancer, HIV, and flu. The fifth parameter, the descriptors, is described separately in section 4.3. Table 5.1 shows three examples of experiments of various parameters.

Table 5.1.: Three variations of experiments using different values of the parameters.

# Assays	# Mol	Disease	Descriptor
small	small	cancer	structure
large	medium	HIV	activity profiles
medium	large	Influenza	combined

Experiment size Theoretically, we want to have specific values of the parameters of the experiments (see table 5.2) to conduct the required experiments. Unfortunately, the actual available number of molecules per assay is small. Therefore, we restrict the experiment size to the available number of molecules for each assay. For this, we defined the *available experiment size*. In the available experiment size, the upper value of the number of molecules is the maximal number of molecules per assay such that the number of assays does not exceed a lower bound (=10). In the same way, the upper value of the number of assays is the maximal number of assays which has a number of molecules does not exceed a lower bound (=300 active and 300 inactive molecules). As a result of this definition the experiment size in HIV case becomes as displayed in table 5.3.

Table 5.2.: Theoretical experiment size.

size	# Assays	# Mol per assay
small	10	10-100
medium	100	1000 -5000
large	1000	10000-100000

Table 5.3.: Available experiment size of HIV.

size	# Assays	# Mol per assay
small	12	300
medium	30	1000
large	47	1500

5.2. Descriptors

As described in the methods chapter, we have three descriptors. Our aim is to find out which descriptor leads us into better models. To make this comparison we have designed the necessary experiments. For this, 3 data sets (cancer, HIV, flu) were experimented. And for each set, we studied the effect of the number of assays as well as the number of molecules on the quality of the resulting models for each of the three descriptors. Moreover, we observed the effect of the experiment size on the models of each descriptor individually. For clarity and simplicity we used the plots to show the results of the experiments. Detailed results are presented in chapter appendix A. Models quality is measured by two measurements APR, and ROC. However, only ROC values are used in the plots. The values of APR are put in Appendix A.

5.2.1. Effect of the importance factor

Before comparing the descriptors it is important to highlight the fact that the normalization factor (α) affects the quality of the combined descriptor models. The size of the activity profile descriptors is equal to the number of the corresponding assays. The size of the activity profiles increases or decreases the importance of the activity profile in the combined descriptors. A profile of size 5, for example, can not be of higher importance than the corresponding structural features. To understand the behavior of the models in relation to α we tested 3 values of the activity profile size and measured the quality of the models for 9 values of α. The plot in Figure 5.1 shows that the quality of models is almost the same for $\alpha \in [0.1, 0.3]$ for sizes 10, 50, and 150. Then, for bigger values of α, the quality of models gradually decreases. This implies that the activity profiles are of lesser importance than the features at relatively small sizes (10, 50, 150). For better understanding of α effect we drew another plot at 3 points (0.1, 0.5, 0.9) for the same dataset (cancer) but this time at a smaller number of molecules (300). Using a smaller number of molecules we avoid the negative effect of larger number of molecules which reduces the quality of combined models. Therefore, in this plot we see that the ROC value at $\alpha = 0.5$ is bigger than this at $\alpha = 0.1$. This implies that activity profiles become more important than those in the first plot for smaller number of molecules and for relatively small number of assays (activity profile size).

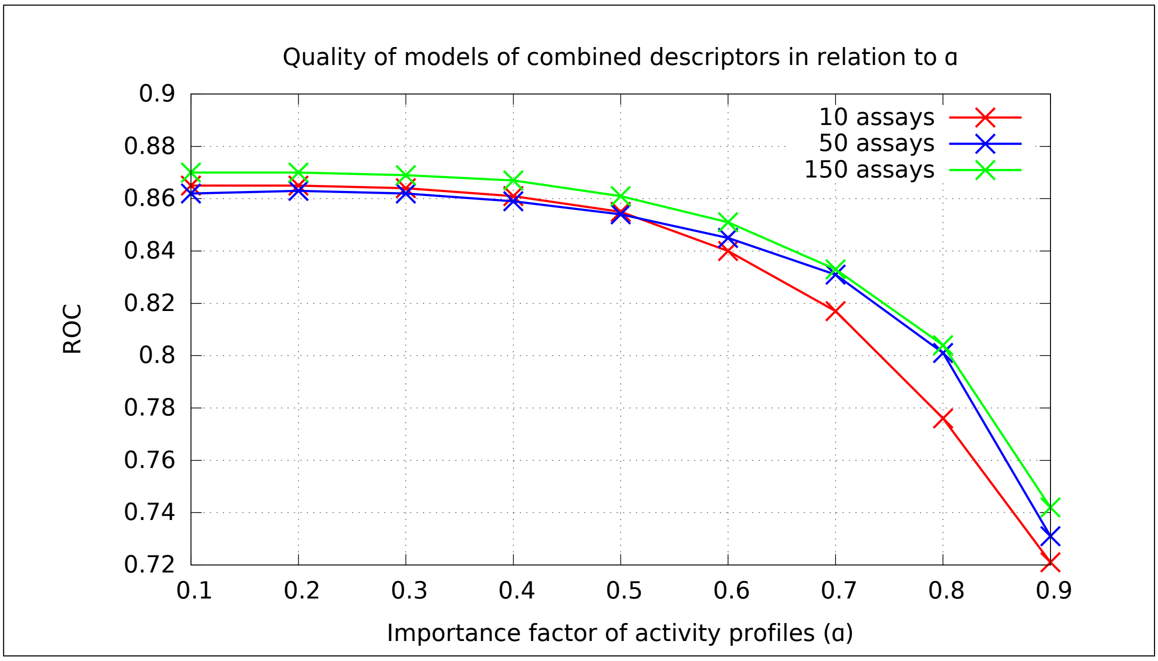


Figure 5.1.: The relation between ROC and the importance of the activity profiles (α) in the combined descriptors at 3 different sizes. Number of molecules is 1400.

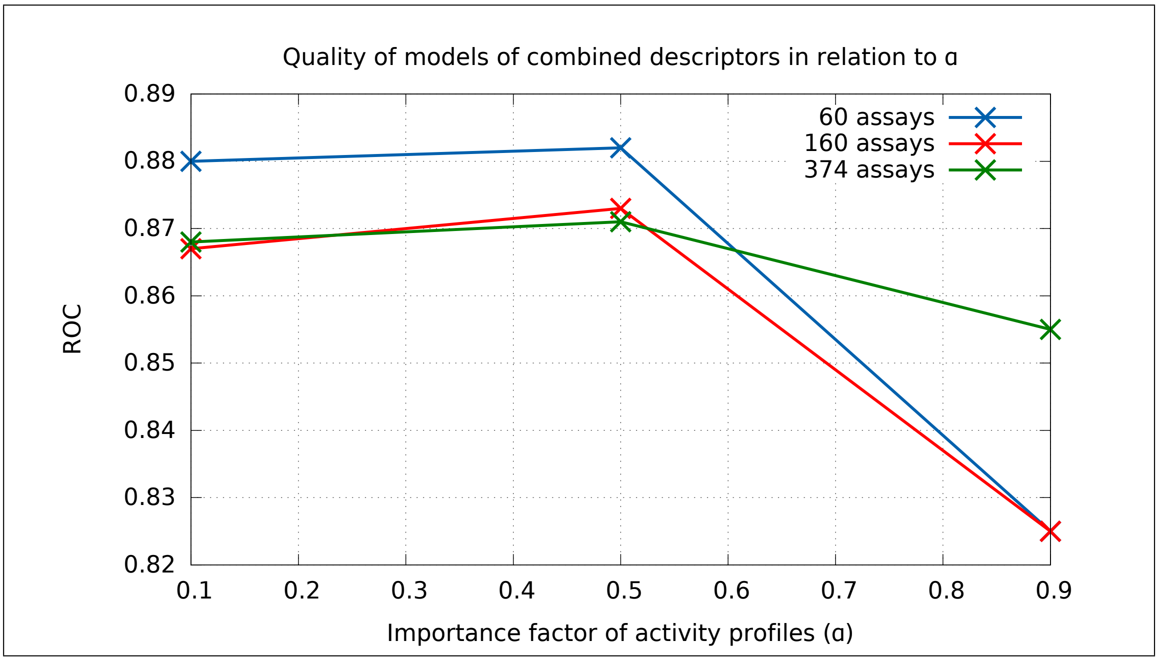


Figure 5.2.: The relation between ROC and the importance of the activity profiles in the combined descriptors at 3 different sizes. The combined descriptors.

5.2.2. Effect of the number of assays

Descriptors comparison To study the effect of the number of assays on the quality of the models, we fix the number of molecules and vary the number of assays. Figure 5.3, figure 5.4, and figure 5.5 show the mean value and standard deviation of the measurement ROC using three different number of assays in each of them. Number of molecules is 300 for all. There are two common observations for the three figures. First, the structure bars and the combined bars totally overlap in all cases. Second, the activity profile bars partially overlap with the other bars in all cases except for one (cancer, 60 assays, 300 molecules). These two observations imply that the structure models and the combined models are of the same quality. Additionally, the activity profile models are of lesser equality but it is still comparable to the two other models.

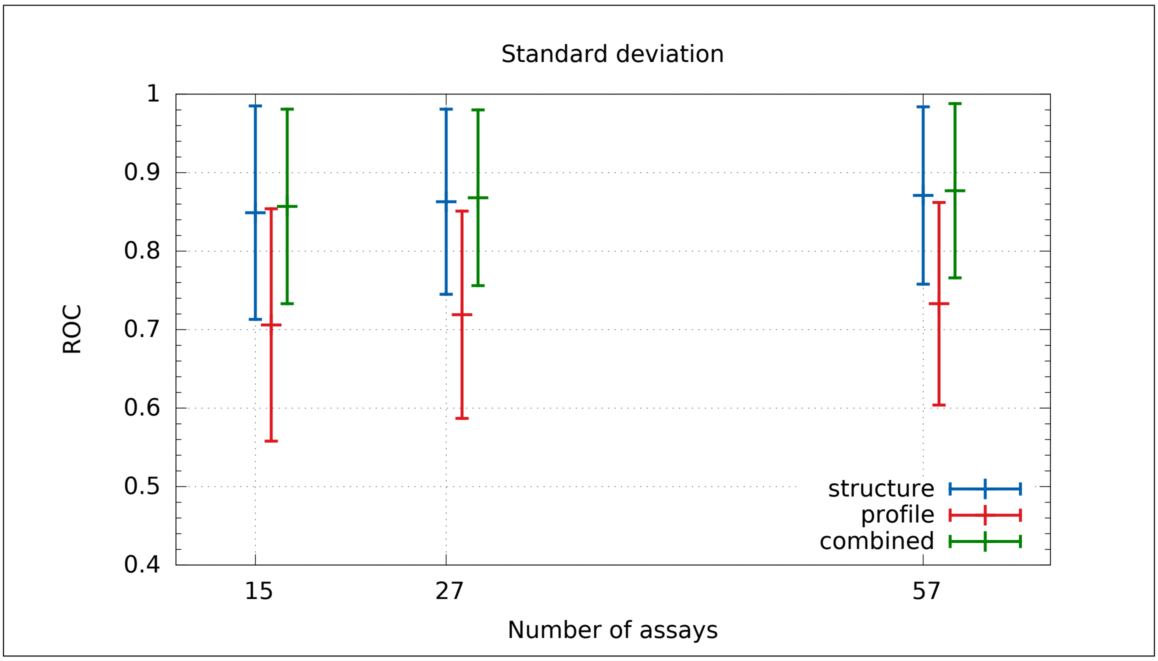


Figure 5.3.: Standard deviation of ROC values. 300 Molecules. 3 values of the number of assays 15, 27, 57. Flu dataset.

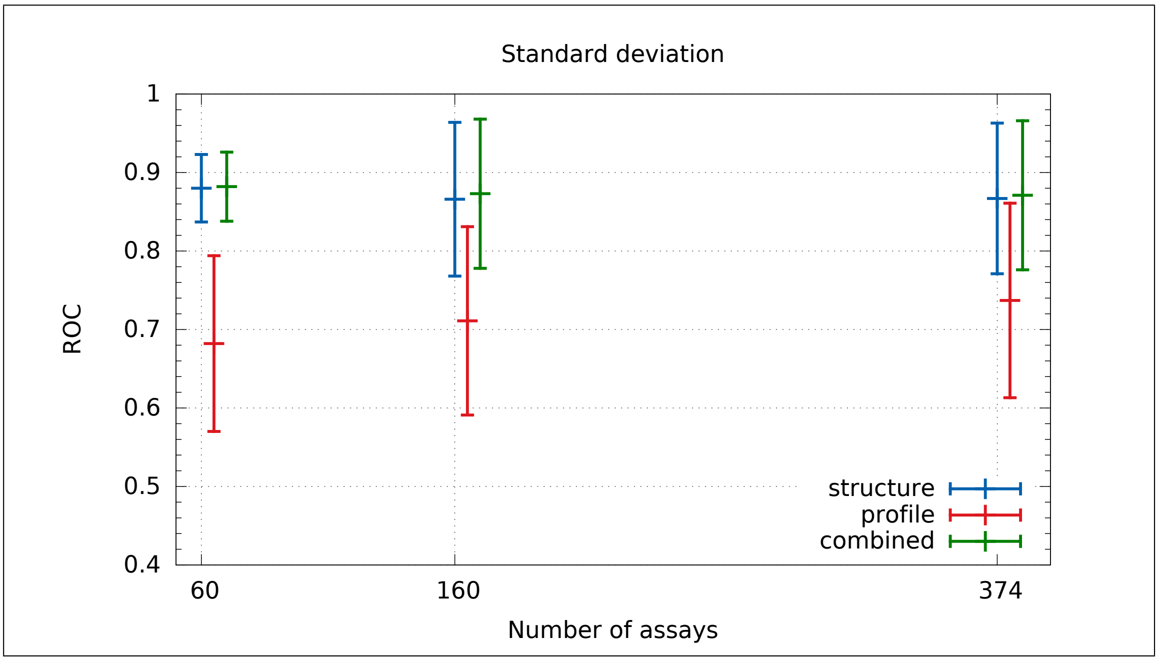


Figure 5.4.: Standard deviation of ROC values. 300 Molecules. 3 values of the number of assays 60, 160, 374. Cancer dataset.

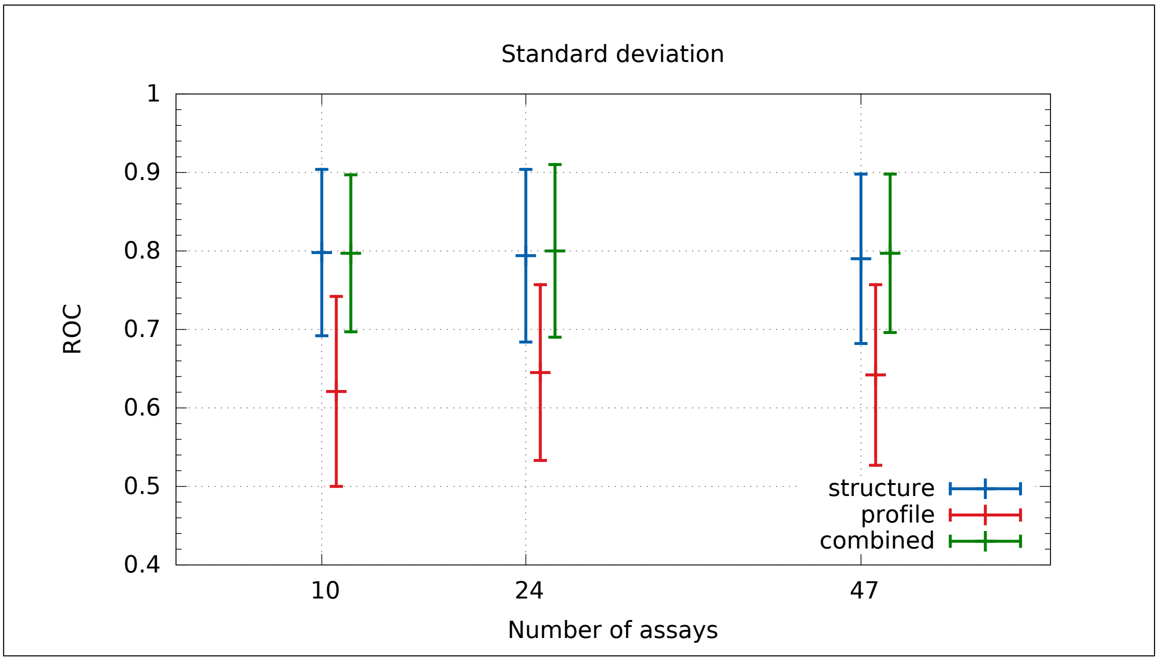


Figure 5.5.: Standard deviation of ROC values. 300 Molecules. 3 values of the number of assays 10, 24, 47. HIV dataset.

Models of each descriptor The same figures can be exploited to show the change on the models quality with increasing number of assays for the 3 descriptors. Apparently, the combined and the structural descriptor do not significantly improve with increasing number of assays. However, the models of the activity profiles show increase in the quality with increasing number of assays. As attempt to estimate to which extent the ROC value increases in the number of assays we draw a fit curve Figure 5.6 for the three points we have in the figure 5.6. This fit curve tells that for a reasonably big number of assays (1000 assays), a significant improvement on the quality of models will not occur.

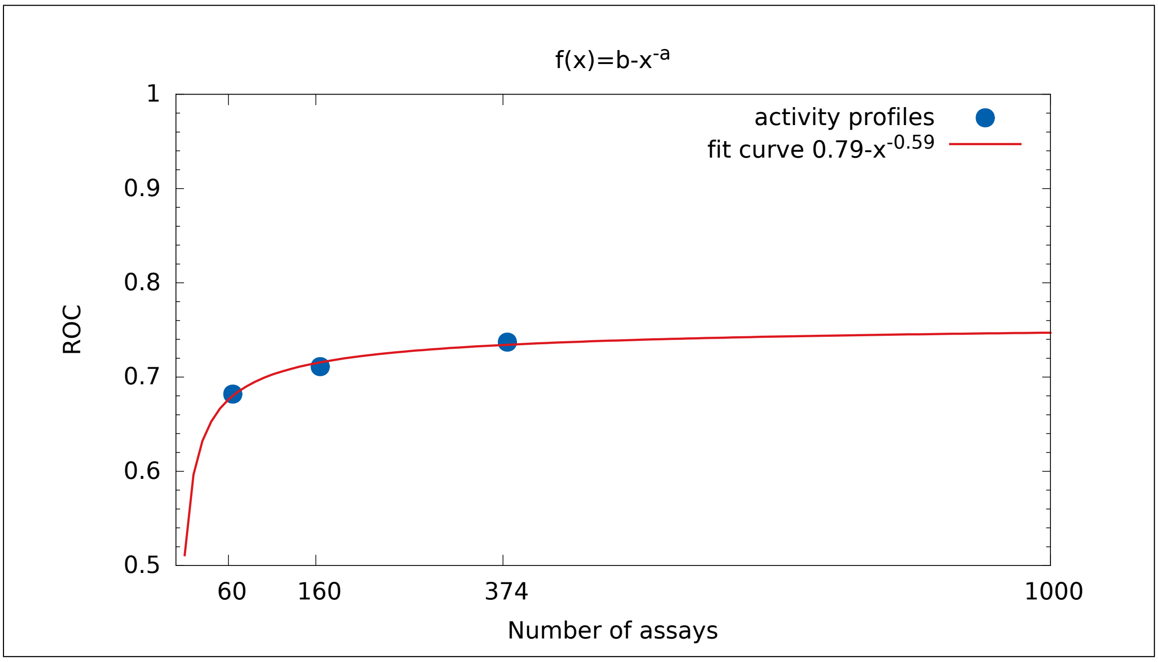


Figure 5.6.: The fit curve of the mean values of ROC in the number of assays. Descriptor is activity profile. Number of molecules is 300. Cancer dataset.

5.2.3. Effect of the number of molecules

In the last results, the number of molecules were fixed, but the number of assays were variable. Now we try to find out the effect of the number of molecules per assay. Therefore, we fix the number of assays and vary the number of molecules. Figure 5.7, figure 5.8, and figure 5.9 show three plots. Each plot is for one data set. The number of assays is fixed in each plot. The number of molecules in each plot is represented by three different values. Each plot represents the mean and standard deviation of the values of the measurement ROC.

Descriptors comparison The standard deviation bars of the models of the structure and the combined descriptors substantially overlap at each point. The standard deviation bars in the case of the activity profile descriptors partially overlap with the other bars at a small amount of molecules, but it does not overlap when the number is reasonably big. What we can conclude from these plots that the quality of models in case of structural and combined descriptors are very similar, whereas it is lower in the case of activity profiles.

The models of each descriptor The quality of the models in both cases the structural descriptors and the combined ones explicitly increases with increasing number of molecules. This is intuitive since the learning algorithm performs better with more true examples. However, in contrast, increasing the number of molecules does not improve the models of activity profiles. This is due to the prediction error, since the activity profiles are artificially synthesized by the prediction values using structural descriptors.

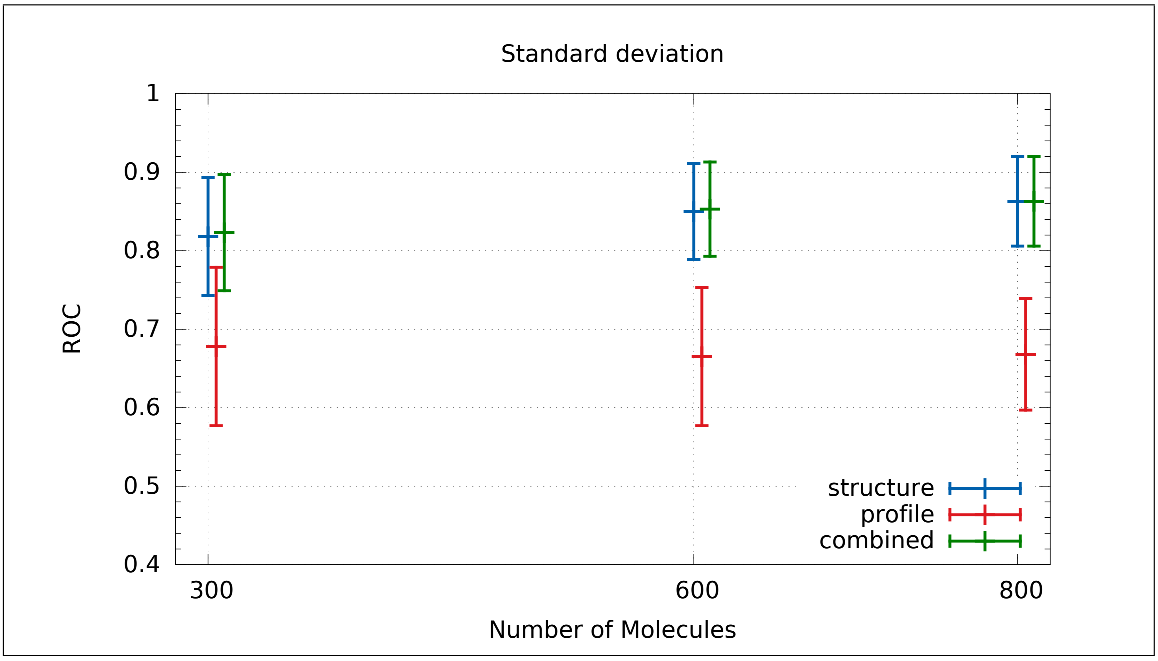


Figure 5.7.: Standard deviation of ROC values. 30 Assays. 3 values of the number of molecules 300, 600, 800. Flu dataset.

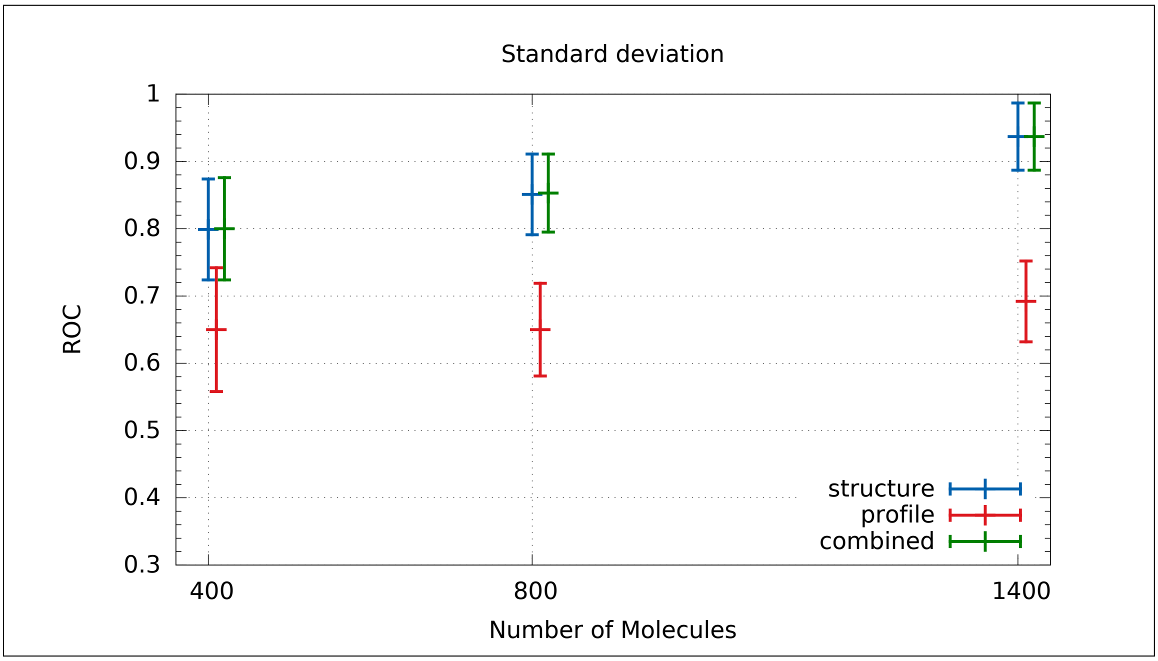


Figure 5.8.: Standard deviation of ROC values. 150 assays. 3 values of the number of molecules 400, 800, 1400. Cancer dataset.

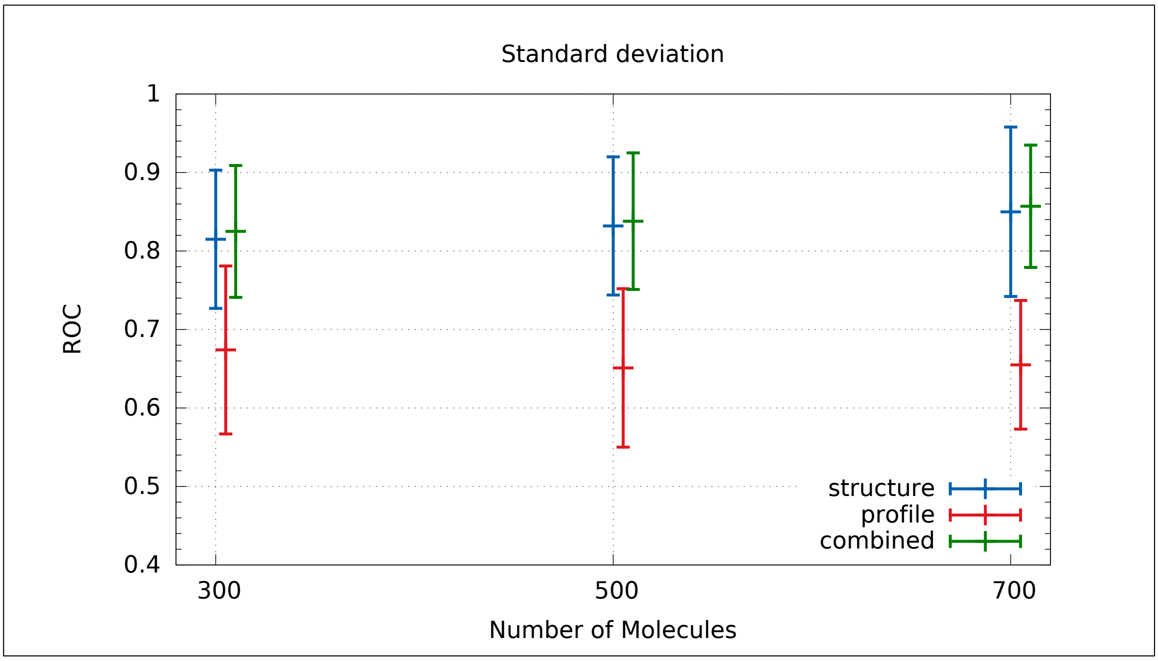


Figure 5.9.: Standard deviation of ROC values. 30 assays. 3 values of the number of molecules 300, 500, 700. HIV dataset.

5.2.4. Run time

The run time of the application for one experiment is the sum of the run time for preparing the descriptors + the run time for preparing files of 10 cross validation experiment + the run time required for training the models + the run time for testing the models. Here, we study the time of the application in the number of assays, the total number of unique molecules of all assays, and the type of the descriptor of one experiment. The plots in figures 5.10 and 5.11 show that the time is linear in the number of molecules for different number of assays and different descriptor. However, the plot in figure 5.12 shows a divergence in the time line. This is because our applications works in parallel on a cluster which sometimes results in waiting time till at least a node of the cluster becomes free. The application has been run on the machines mentioned in the methods chapter. The linearity of the application is due to several reasons: 1) efficiency of the kernel NSPDK, 2) efficient database queries and 3) parallelism. The detailed values of the time are displayed in the chapter Appendix A.

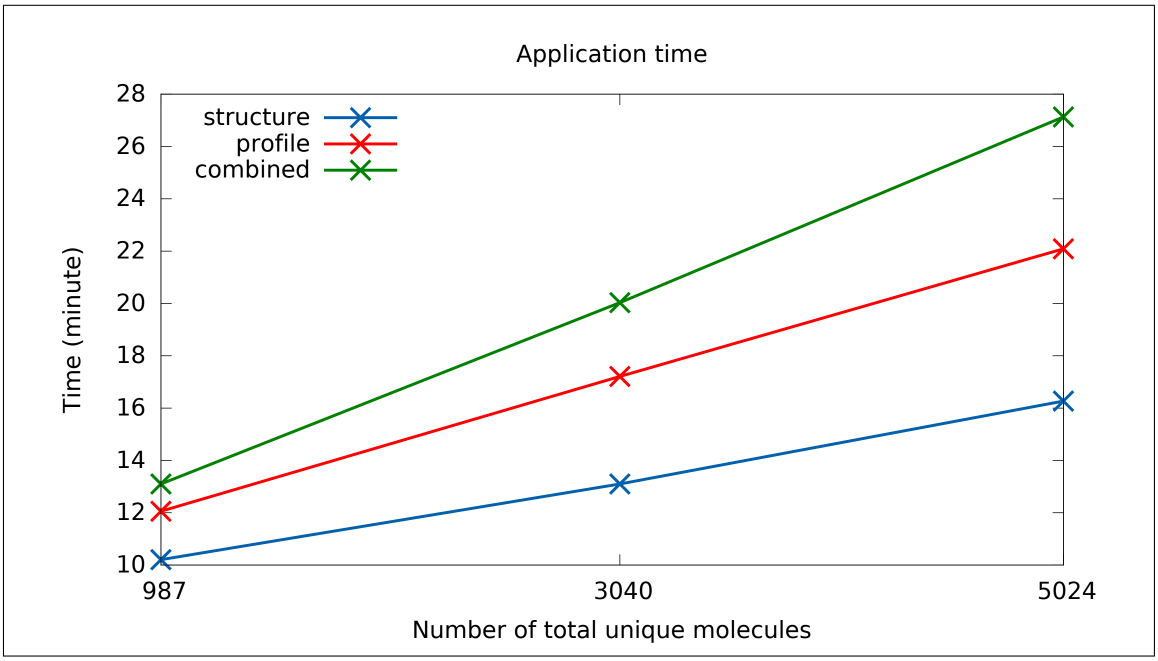


Figure 5.10.: 10 assays. 3 different numbers of the total number of unique molecules. The 3 Descriptors.

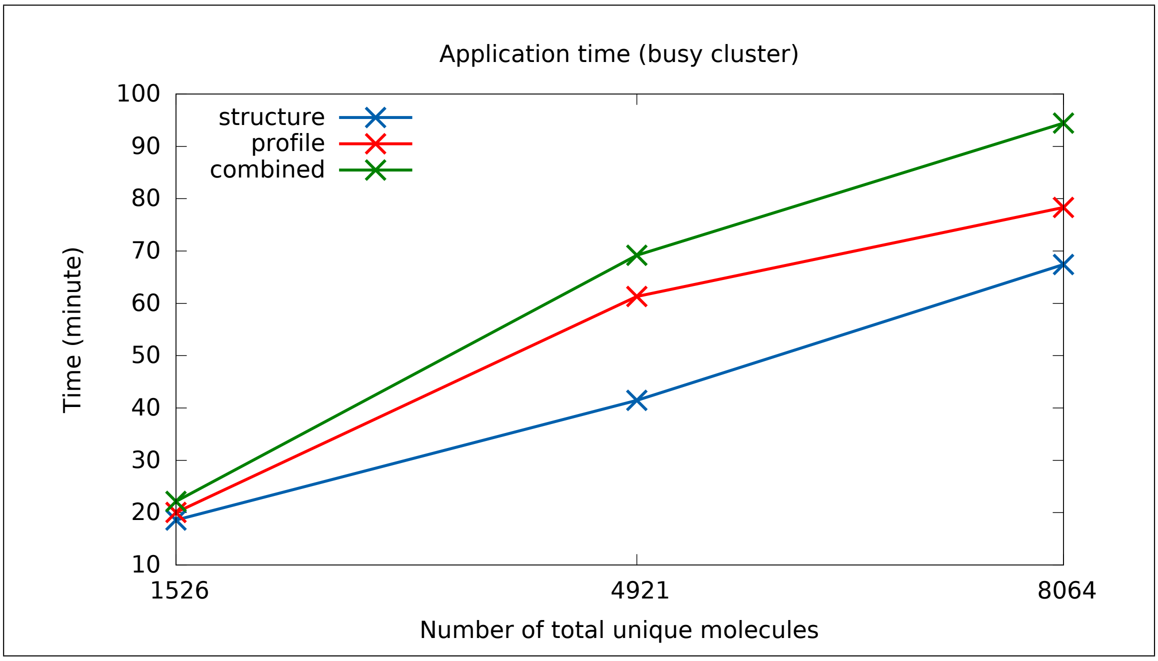


Figure 5.11.: 50 assays. 3 different numbers of the total number of unique molecules. The 3 Descriptors.

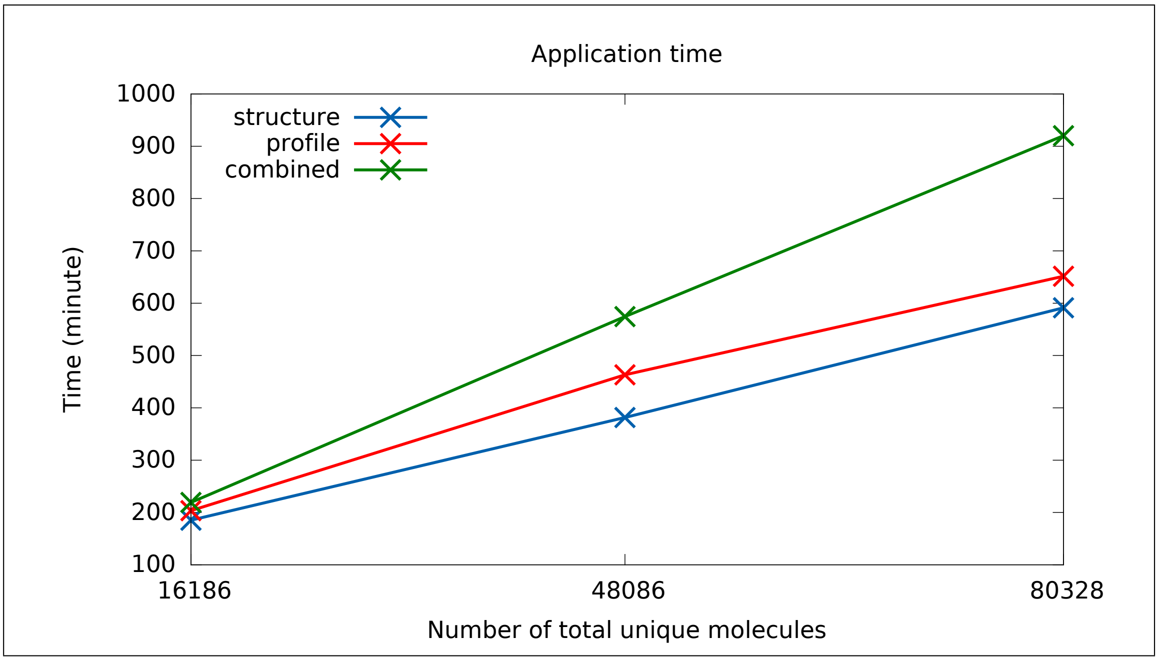


Figure 5.12.: 150 assays. 3 different numbers of the total number of unique molecules. The 3 Descriptors.

5.3. Semi supervised algorithm

As we saw in the beginning of this chapter, the number of assays with sufficient number of molecules is limited in reality. In this section we study the models resulting from the application after employing the semi supervised algorithm introduced in the methods chapter. The algorithm help in increasing the number of molecules per assay. Having a number of molecules and their activity values per assay, the algorithm classifies a number of new molecules into active and inactive. Consequently, now we can contribute new assays which very limited number of experimented molecules. In this section, we show the performance of the algorithm. Afterwards, we study the effect of the number of neighbors. Finally, we introduce one example to present the effect of increasing the number of assays on the activity profile models.

5.3.1. Performance of the algorithm

The input of the algorithm are two files, the molecules (represented by some descriptor) and the activity values of them. Part of the molecules of known activity values (+1,-1) and the rest of unknown activity value (0). Here, we vary the number of molecules and measure the time and the precision of the algorithm. As table 5.4 shows, the time is linear in the number of molecules and the accuracy is very high even at a big number of molecules. In this test, we fixed the number of neighbors at 10.

Table 5.4.: Performance and accuracy of the semi-supervised algorithm for $\alpha = 0.001$ and 10 neighbors.

# of mol (known + unknown)	CPU time (second)	precision(%)
10+100	1	100
30+300	4	98.2
40+400	7	100
50+500	9	96
70+700	13	97.7
100+1000	35	94
300+3000	42	94.3

5.3.2. Effect of the number of neighbors

As we mentioned in the methods chapter we use the algorithm ”k nearest neighbor” when building our the graphs of molecule vectors. Now, we want to check whether the number of the neighbors has an influence on the quality of the models. To do this we varied the number of neighbors for the same number of assays and the same number of molecules per assay. Figure 5.13 shows that the number of neighbors does not substantially change the quality of the models. However, for a relatively big number of molecules (here 50) the ROC value becomes 0.5 which refers random prediction results. This result is a logical one since the more neighbors the lesser precise results, since the algorithm will consider molecules of small similarity value as neighbors of the molecule.

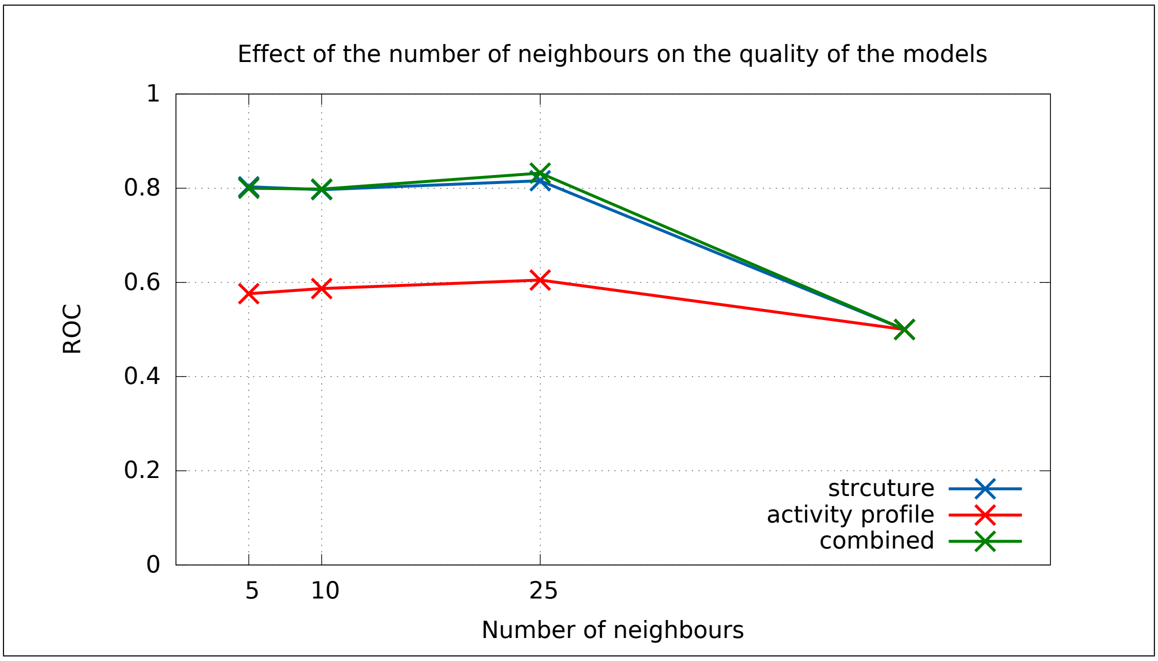


Figure 5.13.: 3 different number of neighbors. 300 Molecules. 10 assays. HIV dataset.

5.3.3. Effect of the number of new assays

As already mentioned, the algorithm can be exploited to increase the number of contributing assays in one experiment. This helps to increase the size of the activity profiles. Figure 5.14 is a plot of ROC values in the number of assays per one experiment. One value, the blue one, is obtained from running normal experiment (without semi-supervised algorithm). The red line is obtained by ROC values of four numbers of assays after applying the semi supervised algorithm. As we can notice, increasing the number of contributing assays improves the quality of obtained models of activity profiles. For theoretically estimating how the models behave with increasing number of new assays, we drew a fit curve plot in figure 5.15. The fit curve shows that for about 1000 new assays, the ROC value reaches 0.8.

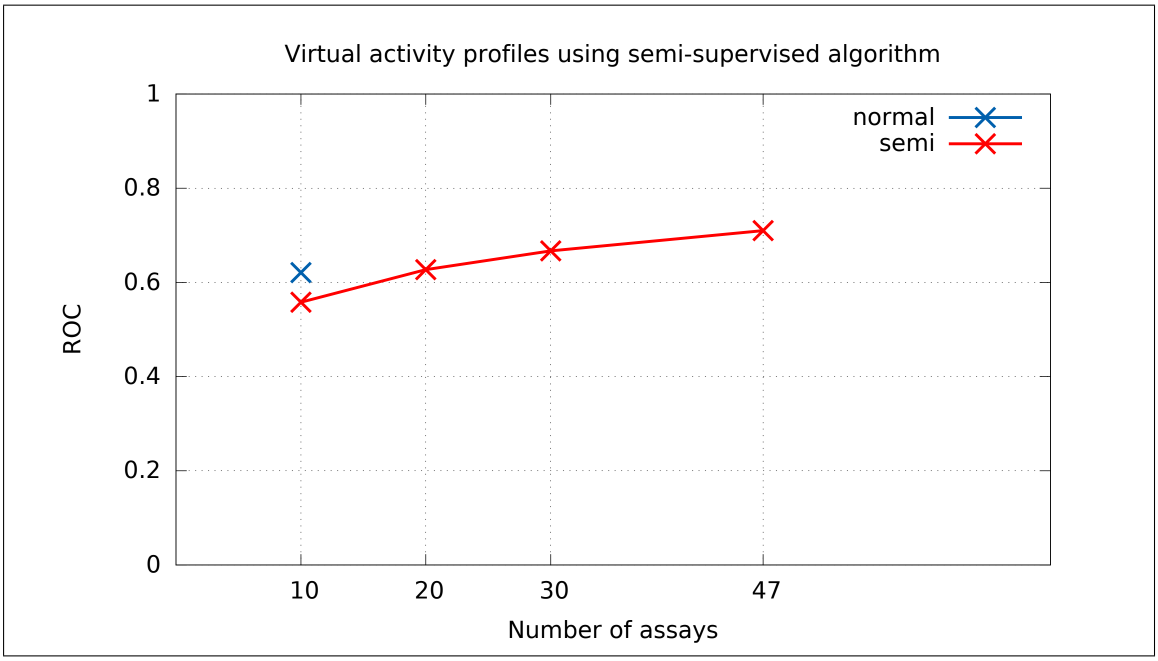


Figure 5.14.: Prediction by activity profiles after adding new assays by employing the semi supervised algorithm. 3 different numbers of assays. 300 molecules. HIV dataset.

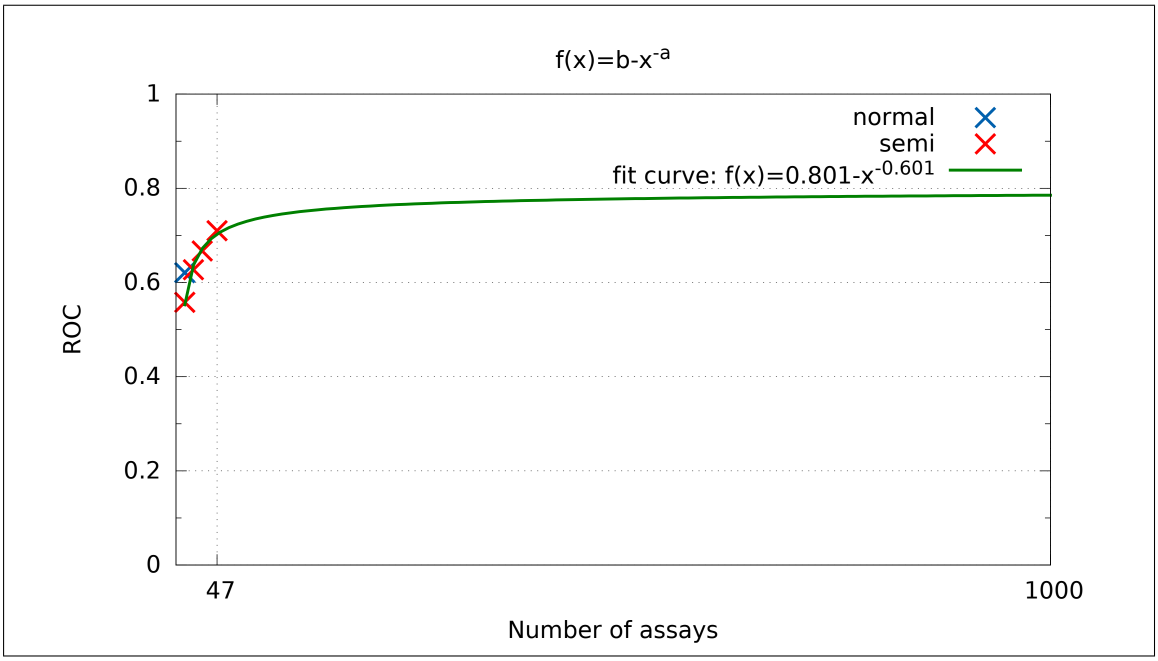


Figure 5.15.: A fit curve of four points of ROC values in the number of assays. 3 different number of assays (10, 20, 30, 40). 300 molecules. HIV dataset.

Chapter 6

Conclusions

6.1. Discussion

The main goal of this work was to come up with new computational predictive models for predicting the biological activity of small molecules against new assays. To achieve this goal, we proposed to generate virtual descriptors so that they lead to good models, and to employ a semi-supervised algorithm so that it allows us to include new molecules to existing models which, in turn, from one side improves the predictive power of models for one assay, and from another side increases the number of contributing assays in some experiment. We prepared the required infrastructure and wrote an efficient script. All the required experiments were designed. However, not all experiments have been conducted due to time limitations. Therefore, we could not completely show the power of our method. Nevertheless, the conducted experiments were sufficient to tell us many things. Two new descriptors allowed us to generate new predictive models. A comparison between the models of these descriptors and the models of structural descriptors has shown that our new models have a similar quality to the structural ones. Two main factors were varied to make this comparison, number of assays and number of molecules per assay. Increasing the number of assays has slightly improved the activity profiles models, but did not improve the two others. The reason might be that the number of molecules directly impacts the activity profiles, since the activity profile size equals the number of assays. On the other side, the number of assays does not impact the models of the structural and combined descriptors, because the number of molecules per assay stays the same which means trained models stay the same. Increasing the number of molecules (for a fixed number of assays) slightly decreases the quality of activity profile models, and significantly increases the quality of the two others. The reason behind this might be that in activity profile case, we have a prediction error which increases with more new molecules, but the new molecules in the other two cases would help the learning algorithm with more new true examples.

The second approach, which integrates the first one, tries to increase the contributing molecules per assay. After presenting the power of the semi supervised algorithm we showed that increasing the number of molecules helps in adding new assays to some experiment. This, in turn, increases the quality of activity profiles models. The impact of the semi-supervised algorithm needs further investigations. Although the new descriptors did not outperform the quality of the structure-based models, the new models are of comparable quality. Moreover, the virtual activity profiles have the advantage that they can be used to represent molecules not yet tested against an assay under study. This advantage is very beneficial because it allows us to test compounds on many new assays in a too short time in comparison to the biological screening.

6.2. Future work

Extensions of the work are possible in several ways. First, we can add new types of descriptors to the database. This would give us the possibility to select the best descriptor for generating better activity profiles. Second, we can add the available experimental information to the virtual activity profiles. This is expected to increase the quality of the models. Third, a statistical study may be conducted to discover the diversity of the truly predicted hits by the new models. Are the hits of the new models different from the ones of other models? What kind of molecules do the new models predict?. Fourth, instead of randomly choosing the new contributing molecules to some experiment in the semi supervised algorithm, we can choose them such that they are similar to the molecules of the corresponding assay. For this, for example, we can design a library of molecules related to cancer data set. The library has molecules which have never tested against cancer, but they are similar in structure (or activity or both) to the molecules already tested against cancer. Fifth, we can consult chemists or pharmacologists to ask them how we can extend the assays of some experiment by new assays such that they enrich the experiment by new information. Sixth, new molecules and assays can be added to the database. This would help to get better models of all descriptors.

Acknowledgments

It is my pleasure to thank everyone who contributed to this work directly or indirectly or in any way,...or even did not contribute or even hampered ...., actually thanks everyone :-).

Many thanks to Prof. Dr. Rolf Backofen and Prof. Dr. Stefan Günther who offered me the opportunity to accomplish this work. Many thanks to Dr. Fabrizio Costa, my direct supervisor who did not hesitate in answering any question I asked. The work is his proposal. His ideas and scientific support were essential to achieve this project. A very special thank to my family in Syria, Father, Mather and my five siblings. Your support is the reason of being at this position. Many thanks to my brother Hassan and his wife Eman (Abu Taim and Um Taim) who helped me morally and financially to finish my study. Hassan! my doctor! Having a great brother like you honors me. You are great. I wish you and your family all the happiness. A very special thank to my girlfriend, Felicitas, for being more than a girlfriend. Your correction tips of my writing made my thesis at least presentable. I would like to send you a less than symbol followed by a 3 (and one wiw). I would like to thanks the administrative stuff of the technical faculty, especially Ms. Ursula Epe and Ms. Friederike Schneider for their helping role in solving student problems.

Appendix A

Results details

A.1. Cancer Results

A.1.1. Effect of the number of assays

Table A.16, table A.17 and table A.18 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of molecules for each experiment is fixed to 300 active molecules and 300 inactive molecules. The number of assays vary in $[60, 160, 374]$.

Table A.1.: Cancer, 60 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.880	0.043
Activity profile	0.682	0.112
Combined	0.882	0.044

Descriptor	*Mean*	*sd*
Structure	0.886	0.050
Activity profile	0.700	0.101
Combined	0.887	0.052

Table A.2.: Cancer, 160 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.866	0.098
Activity profile	0.711	0.120
Combined	0.873	0.095

Descriptor	*Mean*	*sd*
Structure	0.874	0.100
Activity profile	0.720	0.114
Combined	0.879	0.097

Table A.3.: Cancer, 374 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.867	0.096
Activity profile	0.737	0.124
Combined	0.871	0.095

Descriptor	*Mean*	*sd*
Structure	0.872	0.098
Activity profile	0.743	0.126
Combined	0.876	0.097

A.1.2. Effect of the number of molecules

Table A.4, table A.5 and table A.6 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of assays for each experiment is fixed to 150. The number of active molecules vary in $[400, 800, 1400]$. The number of active molecules and inactive molecules is balanced, i.e the number of inactive molecules is 300 as well. The data set is for cancer.

Table A.4.: Cancer, 150 assays, 400 molecules.

Descriptor	*Mean*	*sd*
Structure	0.799	0.075
Activity profile	0.650	0.092
Combined	0.800	0.076

Descriptor	*Mean*	*sd*
Structure	0.780	0.085
Activity profile	0.646	0.085
Combined	0.779	0.085

Table A.5.: Cancer, 150 assays, 800 molecules.

Descriptor	*Mean*	*sd*
Structure	0.851	0.060
Activity profile	0.650	0.069
Combined	0.853	0.058

Descriptor	*Mean*	*sd*
Structure	0.830	0.071
Activity profile	0.650	0.072
Combined	0.835	0.069

Table A.6.: Cancer, 150 assays, 1400 molecules.

Descriptor	*Mean*	*sd*
Structure	0.937	0.050
Activity profile	0.692	0.060
Combined	0.937	0.050

Descriptor	*Mean*	*sd*
Structure	0.940	0.053
Activity profile	0.689	0.069
Combined	0.940	0.053

A.2. Flu

A.2.1. Effect of the number of assays

Table A.7, table A.8 and table A.9 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of molecules for each experiment is fixed to 300 active molecules and 300 inactive molecules. The number of assays vary in $[15, 27, 57]$. The dataset is for flu.

Table A.7.: Flu, 15 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.849	0.136
Activity profile	0.706	0.148
Combined	0.857	0.124

Descriptor	*Mean*	*sd*
Structure	0.853	0.13
Activity profile	0.718	0.146
Combined	0.861	0.124

Table A.8.: Flu, 27 assays, 300 Molecules.

Descriptor	*Mean*	*sd*
Structure	0.863	0.118
Activity profile	0.719	0.132
Combined	0.868	0.112

Descriptor	*Mean*	*sd*
Structure	0.868	0.116
Activity profile	0.725	0.129
Combined	0.872	0.110

Table A.9.: Flu, 57 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.871	0.113
Activity profile	0.733	0.129
Combined	0.877	0.111

Descriptor	*Mean*	*sd*
Structure	0.875	0.113
Activity profile	0.737	0.130
Combined	0.881	0.109

A.2.2. Effect of the number of molecules

Table A.10, table A.11 and table A.12 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of assays for each experiment is fixed to 30. The number of active molecules vary in $[300, 600, 800]$. The number of active molecules and inactive molecules is balanced, i.e the number of inactive molecules is 300 as well. The data set is for flu.

Table A.10.: Flu, 30 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.818	0.075
Activity profile	0.678	0.101
Combined	0.823	0.074

Descriptor	*Mean*	*sd*
Structure	0.802	0.084
Activity profile	0.674	0.099
Combined	0.807	0.084

Table A.11.: Flu, 30 assays, 600 Molecules.

Descriptor	*Mean*	*sd*
Structure	0.850	0.061
Activity profile	0.665	0.088
Combined	0.853	0.060

Descriptor	*Mean*	*sd*
Structure	0.830	0.069
Activity profile	0.661	0.086
Combined	0.835	0.071

Table A.12.: Flu 30 assays, 800 molecules.

Descriptor	*Mean*	*sd*
Structure	0.863	0.057
Activity profile	0.668	0.071
Combined	0.863	0.56

Descriptor	*Mean*	*sd*
Structure	0.845	0.066
Activity profile	0.667	0.071
Combined	0.847	0.065

A.3. HIV results

A.3.1. Effect of the number of assays

Table A.13, table A.14 and table A.15 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of molecules for each experiment is fixed to 600 (300 active molecules + 300 inactive molecules). The number of assays vary in $[10, 24, 47]$. The data set is for HIV.

Table A.13.: HIV, 10 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.798	0.106
Activity profile	0.621	0.121
Combined	0.791	0.100

Descriptor	*Mean*	*sd*
Structure	0.791	0.113
Activity profile	0.641	0.117
Combined	0.792	0.105

Table A.14.: HIV, 24 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.794	0.110
Activity profile	0.645	0.112
Combined	0.800	0.110

Descriptor	*Mean*	*sd*
Structure	0.783	0.116
Activity profile	0.655	0.109
Combined	0.790	0.116

Table A.15.: HIV, 47 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.790	0.108
Activity profile	0.642	0.115
Combined	0.797	0.101

Descriptor	*Mean*	*sd*
Structure	0.778	0.110
Activity profile	0.651	0.109
Combined	0.786	0.109

A.3.2. Effect of the number of molecules

Table A.16, table A.17 and table A.18 contain values of the mean and standard deviation of the measurements ROC and APR. The right hand side table is for APR and the left hand side table is for ROC. The number of assays for each experiment is fixed to 30. The number of active molecules vary in $[300, 600, 800]$. The number of active molecules and inactive molecules is balanced, i.e the number of inactive molecules is 300 as well. The data set is for flu.

Table A.16.: HIV, 30 assays, 300 molecules.

Descriptor	*Mean*	*sd*
Structure	0.823	0.23
Activity profile	0.733	0.25
Combined	0.827	0.22

Descriptor	*Mean*	*sd*
Structure	0.803	0.096
Activity profile	0.674	0.098
Combined	0.814	0.092

Table A.17.: HIV, 30 assays, 500 molecules.

Descriptor	*Mean*	*sd*
Structure	0.792	0.18
Activity profile	0.627	0.2
Combined	0.786	0.18

Descriptor	*Mean*	*sd*
Structure	0.814	0.099
Activity profile	0.655	0.103
Combined	0.821	0.097

Table A.18.: HIV, 30 assays, 700 molecules.

Descriptor	*Mean*	*sd*
Structure	0.839	0.12
Activity profile	0.712	0.14
Combined	0.837	0.11

Descriptor	*Mean*	*sd*
Structure	0.835	0.092
Activity profile	0.660	0.086
Combined	0.840	0.089

A.4. Time complexity

Table A.19.: Structural descriptors.

# assays	# unique molecules	Time (minutes)
10	987	10.21
10	3040	13.10
10	5024	16.10
50	1526	18.57
50	4921	41.45
50	8064	67.41
150	16186	185.23
150	48086	381.51
150	80328	591.28

Table A.20.: Activity profile descriptors.

# assays	# unique molecules	Time (minutes)
10	987	12.06
10	3040	17.21
10	5024	22.09
50	1526	20.03
50	4921	61.31
50	8064	78.34
150	16186	203.45
150	48086	463.05
150	80328	651.54

Table A.21.: Combined descriptors.

# assays	# unique molecules	Time (minutes)
10	987	13.10
10	3040	20.03
10	5024	27.13
50	1526	22.15
50	4921	69.19
50	8064	94.46
150	16186	219.09
150	48086	574.44
150	80328	920.18

List of Figures

List of Tables

Bibliography

[1] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[2] E. Anderson, G.D. Veith, D. Weininger, and Minn.) Environmental Research Laboratory (Duluth. *SMILES, a Line Notation and Computerized Interpreter for Chemical Structures*. Environmental research brief. U.S. Environmental Protection Agency, Environmental Research Laboratory, 1987.

[3] D Anson Blake, Junyi Ma, and Jia-Qiang He. Identifying cardiotoxic compounds. *Genetic Engineering & Biotechnology News, TechNote (Mary Ann Liebert)*, 29(9):34–35, 2009.

[4] Jürgen Bajorath. Virtual screening in drug discovery: Methods, expectations and reality. *Curr. Drug Discov*, 2:24–28, 2002.

[5] Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, 4:217–241, 2008.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262, 2010.

[8] Steven D. Gammon Darrell D. Ebbing. *General Chemistry Ninth Edition.* Charles Hartford, Connecticut, U.S.A., 2009.

[9] Luc Devroye and Terry J Wagner. Nearest neighbor methods in discrimination. *Handbook of Statistics*, 2:193–197, 1982.

[10] Joseph A DiMasi, Ronald W Hansen, Henry G Grabowski, et al. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–186, 2003.

[11] Spencer M Free and James W Wilson. A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, 7(4):395–399, 1964.

[12] Shayne Cox Gad. *Drug discovery handbook*, volume 1. Wiley-Interscience, 2005.

[13] Hiram F Gilbert and Hiram F Gilbert. *Basic concepts in biochemistry: a student's survival guide.* McGraw-Hill, Health Professions Division USA, 2000.

[14] Jeffrey W Godden and Jürgen Bajorath. An information-theoretic approach to descriptor selection for database profiling and qsar modeling. *QSAR & Combinatorial Science*, 22(5):487–497, 2003.

[15] Jonathan L Gross and Jay Yellen. *Handbook of graph theory.* CRC press, 2003.

[16] Miles Hacker, William S Messer II, and Kenneth A Bachmann. *Pharmacology: principles and practice.* Academic Press, 2009.

[17] David Haussler. Convolution kernels on discrete structures. Technical report, Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[18] M.A. Johnson, G.M. Maggiora, and American Chemical Society. Meeting. *Concepts and applications of molecular similarity*. Wiley-Interscience Publication. Wiley, 1990.

[19] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence*, volume 14, pages 1137–1145. Lawrence Erlbaum Associates Ltd, 1995.

[20] Romano T Kroemer. Structure-based drug design: docking and scoring. *Current Protein and Peptide Science*, 8(4):312–328, 2007.

[21] Hugo Kubinyi. Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design*, 9:225–252, 1998.

[22] Yves Landry and Jean-Pierre Gies. Drugs and their molecular targets: an updated overview. *Fundamental & clinical pharmacology*, 22(1):1–18, 2008.

[23] Andrew R Leach and Valerie J Gillet. *An introduction to chemoinformatics*. Springer Verlag, 2007.

[24] Andrew R Leach and Valerie J Gillet. *An introduction to chemoinformatics*. Springer Verlag, 2007.

[25] Thomas Lengauer and Matthias Rarey. Computational methods for biomolecular docking. *Current opinion in structural biology*, 6(3):402–406, 1996.

[26] Gerald M Maggiora et al. On outliers and activity cliffs–why qsar often disap-

points. *Journal of chemical information and modeling*, 46(4):1535, 2006.

[27] Gerald M Maggiora and Veerabahu Shanmugasundaram. Molecular similarity measures. In *Chemoinformatics*, pages 1–50. Springer, 2004.

[28] Yvonne C Martin, James L Kofron, and Linda M Traphagen. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

[29] Campbell McInnes. Virtual screening strategies in drug discovery. *Current opinion in chemical biology*, 11(5):494, 2007.

[30] Alan D McNaught and Andrew Wilkinson. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997.

[31] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, pages 283–298. Elsevier, 1978.

[32] VLADIMIR I Minkin et al. Glossary of terms used in theoretical organic chemistry. *Pure and Applied Chemistry*, 71(10):1919–1981, 1999.

[33] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. The MIT Press, 2012.

[34] Paul L Munson, Robert A Mueller, and George R Breese. *Principles of pharmacology: Basic concepts and clinical applications*. Chapman & Hall New York, 1995.

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

[36] Hrvoje Nikšic. Gnu wget, 1996.

[37] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.

[38] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993–996, 2006.

[39] Clive P Page, Michael J Curtis, Morley C Sutter, Michael JA Walker, and Brian B Hoffman. *Integrated pharmacology.* Mosby Edinburgh, 2002.

[40] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[41] Paula M Petrone, Benjamin Simms, Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W Davies, Jeremy L Jenkins, and Meir Glick. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chemical Biology*, 7(8):1399–1409, 2012.

[42] Sophie Purser, Peter R Moore, Steve Swallow, and Véronique Gouverneur. Fluorine in medicinal chemistry. *Chemical Society Reviews*, 37(2):320–330, 2008.

[43] Robert B Raffa and Frank Porreca. Thermodynamic analysis of the drug-receptor interaction. *Life sciences*, 44(4):245–258, 1989.

[44] Syed Asad Rahman, Matthew Bashton, Gemma L Holliday, Rainer Schrader, and Janet M Thornton. Journal of cheminformatics. *Journal of cheminformatics*, 1:12, 2009.

[45] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and infer-*

ence, 90(2):227–244, 2000.

[46] RP Stephenson. A modification of receptor theory. *British journal of pharmacology and chemotherapy*, 11(4):379–393, 1956.

[47] Janet L Stringer. *Basic concepts in pharmacology: a student's survival guide*. McGraw-Hill, 1996.

[48] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. Wiley-Vch, 2008.

[49] Birmingham City University's. Physiology website, 2012. [Online; accessed 04-April-2013].

[50] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–2623, 2002.

[51] Hans G Vogel. *Drug discovery and evaluation: pharmacological assays*. Springer Verlag, 2002.

[52] Anne Mai Wassermann, Eugen Lounkine, and Meir Glick. Bioturbo similarity searching: Combining chemical and biological similarity to discover structurally diverse bioactive molecules. *Journal of Chemical Information and Modeling*, 2013.

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

[54] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algo-

rithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[55] John N Weinstein, Timothy G Myers, Patrick M O'Connor, Stephen H Friend, Albert J Fornace, Kurt W Kohn, Tito Fojo, Susan E Bates, Lawrence V Rubinstein, N Leigh Anderson, et al. An information-intensive approach to the molecular pharmacology of cancer. *Science*, 275(5298):343–349, 1997.

[56] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl 1):D5–D12, 2007.

[57] Wikipedia. Plagiarism — Wikipedia, the free encyclopedia, 2004. [Online; accessed 25-March-2013].

[58] Ge Wu. *Assay Development, Fundementals and Practices.* John Wiley & Sons, Inc, 2010.

[59] Ling Xue and Jurgen Bajorath. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial Chemistry & High Throughput Screening*, 3(5):363–372, 2000.

[60] Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. Ranking on data manifolds. *Advances in neural information processing systems*, 16:169–176, 2003.

[61] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.