DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
114

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
114

KALEV TAKKIS

Virtual screening of chemical databases for bioactive molecules

TARTU UNIVERSITY PRESS

Institute of Chemistry, University of Tartu, Estonia

This Dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Molecular Design in 27 February 2012, by the Doctoral Committee of the Department of Chemistry, University of Tartu.

Supervisor: Dr. Sulev Sild, University of Tartu

Opponent: Prof. Marjana Novič, National Institute of Chemistry, Ljubljana, Slovenia

Commencement: 27 April 2012 at 14A Ravila Str., room 1021, 11:00 h

ISSN 1406-0299
ISBN 978-9949-19-960-0 (trükis)
ISBN 978-9949-19-961-7 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 106

Contents

List of original publications

This thesis is based on four publications, listed below. All papers are denoted in the text by Roman numerals I-IV.

1. "QSAR Modeling of HIV-1 Protease Inhibition on Six- and Seven-membered Cyclic Ureas"
   Takkis, K.; Sild, S. *QSAR Comb. Sci.* **2009**, *28*, 52-58

2. "The QSAR Modeling of Cytotoxicity on Anthraquinones"
   Takkis, K.; Sild, S.; Maran, U. *QSAR Comb. Sci.* **2009**, *28*, 829–833

3. "Combined Approach Using Ligand Efficiency, Cross-Docking, and Antitarget Hits for Wild-Type and Drug-Resistant Y181C HIV-1 Reverse Transcriptase"
   García-Sosa, A. T.; Sild, S.; Takkis, K.; Maran, U. *J. Chem. Inf. Model.* **2011**, *51*, 2595-2611

4. "A Novel Structure-based Virtual Screening Method Finds Active Ligands through the Process of 'Topological Docking"'
   Takkis, K.; García-Sosa, A. T.; Sild, S.

Author's contribution

Publication I: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

Publication II: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

Publication III: The author was involved in preparing the data set.

Publication IV: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

List of abbreviations

2D	2-dimensional
3D	3-dimensional
ADME	Absorption, distribution, metabolism, excretion
ANN	Artificial neural network
AUC	Area under the curve
BMLR	Best multi-linear regression
DUD	Directory of useful decoys
EF	Enrichment factor
HIV	Human immunodeficiency virus
HIVRT	Human immunodeficiency virus reverse transcriptase
HTS	High-throughput screening
LBVS	Ligand-based virtual screening
logD	Logarithm of water-octanol distribution coefficient
logP	Logarithm of water-octanol partition coefficient
MLR	Multiple linear regression
PCA	Principal component analysis
QSAR	Quantitative structure-activity relationships
QSPR	Quantitative structure-property relationships
ROC	Receiver-operator characteristic
SBVS	Structure-based virtual screening
SVM	Support vector machine
VS	Virtual screening

1 Introduction

Drug design is a process of developing new marketable drugs for the treatment of diseases humans are plagued with. Usually it starts by identifying the target, be it a receptor, enzyme, ion channel, DNA or something else, which a drug is supposed to activate, inhibit or regulate in order to achieve a desired medicinal effect. Target found, begins a search for a lead compound, a molecule that in assay tests exerts activity towards the target, but is not suitable for a drug due to any number of reasons, such as excessive toxicity, poor selectivity and, most likely, low activity. However, lead compound can be tweaked to give it a desired pharmacokinetic profile and enhance its potency. This process ends up with a drug candidate, a compound that exhibits the desired properties of the drug and is therefore subjected to thorough biological and pharmacological testing. Final step in drug design process is clinical drug, a drug candidate that has passed all the required conditions regarding toxicity, metabolism, side effects etc., and is ready for clinical trials.[1,2] The process is often iterative, with several steps of redesign and retesting, as exemplified on some existing drugs.[3]

Drug design is expensive and time consuming, therefore not surprisingly, options of transferring it into computers as extensively as possible are intensively investigated. There is no way to avoid the clinical trials in foreseeable future,[4] however, it is a different story for less stringent steps of drug development such as the lead discovery. This part has always heavily relied on random screening and in cases where there is no knowledge about the target, or no known natural lead compound is known, it is essentially the only option.[2] Today this takes a form of high throughput screening (HTS) – hundreds of parallel experiments conducted simultaneously in robotised manner. Since HTS is intended for scanning millions of compounds, of which only a tiny fraction are even remotely active, it is often compared to finding a needle in a haystack. The idea then is not to accurately measure several million activities, but rather to find a few interesting compounds to continue with. This is feasible for computational methods and therefore lead and drug candidate discovery have on numerous occasions been successfully complemented with or even replaced by *in silico* methods collectively known as virtual screening (VS). For that reason VS is often seen simply as an *in silico* analogue of HTS, and while it can be seen in section 2.3, that its ambition goes a bit further than that, they do work and gain results in a rather similar manner. Like HTS, VS does not deliver a finished product. Instead, the objective is to concentrate the data set in terms of interesting compounds for further development, eliminate those that are obviously unsuitable or just pick out a handful of most prospective ones.[5]

This thesis presents a research on the topic of virtual screening. Overview of the current state, problems, perspectives as well as most important methods is given in chapter 2. Chapter 2.3 summarises the methodology used in the study, and the results of the original research are presented in chapter 4.

2 Literature overview

2.1 Methods of virtual screening

While the objective of all the VS methods is the same, means of achieving it can be considerably different. For that reason it is often observed, that different methods may have considerably different behaviour. Some methods can work better on some systems, they can be complementary, or even select completely different compounds. Understanding of the methods quirks can therefore be insightful when choosing the one to use.

The most important means of classification is distinguishing between ligand- and structure-based methods. Structure-based virtual screening (SBVS) requires a receptor structure to be known beforehand, derived either form from X-ray diffraction, nuclear magnetic resonance or homology modelling. This structure is then used to describe ligand-receptor interactions that are necessary for good binding. Ligand-based virtual screening (LBVS) does not require a target structure, only one or several known ligands are needed and the new ones are found based on similarities to the old ones.

The second important difference is based on the structural information methods rely on, namely 2D and 3D methods. 2D refers to the structure of molecular graph, atom connectivity and whatever information can be derived from this. 3D is the molecule's spatial structure, atomic coordinates in space. The dimensionality axis can be elongated in both directions, there are 1D methods, that require only molecular composition, and 4D methods which also consider molecule's flexibility; but the difference between 2D and 3D is the most significant, and these two are the most common in VS.

Due to plethora of methods, their unique classification is rather hopeless and various authors have used different schemes in their papers. For that reason, the following list does not follow any particular classification scheme, it is a selection of most important VS methods by the names used commonly in literature.

2.1.1 Molecular docking

The method that most closely resembles an actual experiment, to the extent that it can be called virtual experiment, is molecular docking. The ligand molecule is quite literally placed into the binding site of the target and their affinity towards each other is evaluated. Its straightforward approach to the problem attributes to its popularity – a recent study by Ripphausen *et al.*[6] where almost 500 papers were examined, shows that docking is the single most popular VS method out there. Combined with the fact that it has been around for a while,[7,8] it has become almost a synonyme of virtual screening.

While the general idea is intuitive and easily graspable, internals of the process are dauntingly complicated. It can be seen as a two step process, whereas the first step includes finding a correct pose for the ligand and placing it into the binding site. While earliest adaptations of docking techniques treated both ligand and protein as rigid bodies, today most docking programs employ flexible ligand-rigid receptor paradigm; and with the increasingly powerful computer hardware, combined with the emerging understanding of importance of protein flexibility[9–11],

fully flexible systems are emerging. Ligand flexibility can be accounted through pre-generated conformational ensemble or in-place conformational sampling. For protein flexibility, there are more options: soft docking, which allows simply some overlap of van der Waals radii in ligand and protein atoms; side chain flexibility, where protein backbone is kept fixed whereas side chains are allowed to move; molecular relaxation, which allows also some backbone flexibility; and, similarly to ligand flexibility modelling, an ensemble of protein structures.[12]

After the ligand is fitted, its affinity towards protein is evaluated through some scoring function. They fall into three classes: force-field-based, empirical, and knowledge-based.[12,13] The former takes advantage of some existing force field and uses its parameters to evaluate individual interaction terms such as van der Waals energies, electrostatic energies, bond stretching/bending/torsional energies, etc. separately. It is simple and straightforward, but its problem areas are accounting for entropic and solvent effects.[12] Empirical and knowledge-based scoring functions derive their predictive capacity from a set of known protein-ligand interactions, but handle the data in training set a bit differently. Empirical scoring functions evaluate binding energy by summing the contributions of individual components, commonly hydrogen bonds, entropic terms, ionic, and hydrophobic interactions. These are evaluated as a geometric function of ligand and receptor coordinates which are derived form empirical data using least-square fitting. They are more efficient to calculate, but depend heavily on the quality of training set. Knowledge-based scoring functions make their decision by summing all the interaction terms of all protein-ligand atom pairs, which are derived from the frequency that a particular distance is observed in a database. It is based on the concept of potential mean force, defined by inverse Boltzmann relation.[14,15] The large number of complexes used for training is expected to bias the values towards the existence of specific contacts and absence of repulsive interactions. Knowledge-based functions are a good compromise between force-field based and empiric scoring functions in terms of both computational efficiency and accuracy, while being relatively robust and general.[16–18]

One use for docking is the calculation of the binding affinities between ligand and receptor. There is a number of docking programs available, such as Autodock[19], FlexX[20], Gold[21], Glide[22,23], DOCK[24], to name just a few of the most popular ones. Different programs, employing different geometry handling procedures and different scoring functions give different results, which is not all that surprising. However, no single program regrettably performs consistently better than the others. Studies comparing different programs have pointed out that certain programs work better on certain target classes or active site types, but not uniformly across target space.[25–29] Some limited success has been noted on homogeneous series,[30] but otherwise it has to be concluded that to date, binding affinity prediction remains the toughest assignment of the docking, despite intense studies.[31–33] In fact, the correlation between experimental and calculated affinities is so short of perfect, docking programs lack even the capacity to correctly rank the molecules. Results of docking studies are therefore often evaluated using enrichment[34] – number of active ligands in top x% of the top-scoring compounds. Docking is also used to find a ligand's binding mode, and when applied to this problem, the results are much better. The same studies quoted earlier, that concluded

poor performance on binding energies, report greater success on pose prediction and finding natural binding poses, to the extent that on certain cases, a single best performing program could be pointed out.[29]

The difficulties with affinity predictions stem from the condensed phases in which biology occurs, and the many degrees of freedom of biomolecules.[35] All contributing factors, such as structure of the host, ionization states, structure of the complex, internal degrees of freedom, solvation states, and the host-ligand energetics, are potential error sources. As Tirado-Rives and Jorgensen pointed out, there is only a small activity window for a good hit; it is highly unlikely to find a library compound with the activity above ~50 nM, and below 100 µM there are too many difficulties with experimental determination and lead optimisation to consider the molecule further. That difference corresponds to free energy difference of 4.5 kcal/mol and the free energy contributions from conformational factors alone for typical drug-like ligands (which are usually neglected in most scoring functions) can be as large as this.[36] The authors conclude that predicting affinities reliably for large and diverse molecular libraries, is currently beyond reach.

Despite of the current shortcomings, docking is, as mentioned earlier, the most popular VS method and it holds the greatest potential for the future. It is actively developed,[33,37] and has publicly available libraries ready for docking such as the ZINC database[38]; and validation, such as the Directory of useful decoys (DUD).[39,40]

2.1.2 Fingerprints and similarity search

While docking is clearly a structure-based method, similarity search in majority of cases refers to ligand-based approach. The idea is to describe a molecule using a collection of parameters and search for new ligands with similar values. The result then depends on both how the molecules are characterised and how the conformity of parameters is assessed.

A fingerprint is in essence a vector, in a simplest case a binary vector where each bit corresponds to a predefined structural feature which is searched from the molecule and marked down as 1 if found and 0 if not. This representation is quite intuitive and even human-readable, although to carry a meaningful amount of information, structural key libraries must be rather large, which has a negative effect on extraction time and comparison efficiency. Additionally, they would contain mostly zeroes, because in order to be representative, a library must encompass a large number of structural features and one cannot possibly expect to find them all in a single drug-like molecule.

There are two solutions to the sparse vector problem, one is to use some compressing technique or hashing function which loses negligible amount of information, but results with shorter and denser fingerprint. Another is to extract information directly from the molecule itself, to describe each atom, its environment, connectivity to neighbours and paths connecting to other atoms. Since this type of information depends on the molecule's size, doing it in a manner where each individual bit in the fingerprint corresponds to a certain feature, would result in fingerprints of variable length, a problem when the aim is to compare them. Instead, the fingerprint is set to a fixed length and the information is added using

logical "OR" operation. The advantages over predefined libraries include (besides not having to compose one), faster extraction times and avoiding the problem of poorly constructed library.

Once the fingerprints are composed, they can be compared using certain coefficients. It is not an unique problem, several disciplines require such comparisons, therefore numerous similarity indices have been developed.[41] Some of them are strongly correlated or monotonous (ordering the molecules in the same way), others again entirely unrelated, suggesting that they capture completely different aspects of the object.[41,42] A similarity coefficient can directly measure similarity between two compounds, like Tanimoto or Cosine index or dissimilarity, distance in chemical space like Euclidean or Hamming distance. Some of them see a common absence of a feature a similarity, others do not. Tanimoto index, for example, does not, therefore it could be slightly size-biased, making small molecules, that have less bits set in their fingerprints, appear less similar.[43] Despite that it has become a *de facto* standard in measuring similarity of compounds. Most common similarity measures in chemoinformatics have been discussed by Willett *et al.*[43]

Since fingerprints are simple mathematical constructs, their usage in chemoinformatics as a molecular descriptor is rather universal. Fingerprints described here so far are extracted from the 2D molecular graphs, but really anything can be encoded using the same principle, including 3D structures,[44] protein-ligand complexes,[45–47] and even spatial shapes of molecules.[48] Due to a wide range of application, they are often used and found to be advantageous in many ways. Their calculation, storage and comparison is computationally efficient, they are based on simple concepts, but despite that, they are often found to be comparably effective to much more elaborate methods.[49] Compared to docking, which attempts to model the exact physical interaction between the ligand and receptor, similarity search using fingerprints, might seem a bit sketchy and unreliable, without real physical background. This is exactly the way as it was seen earlier – in 1998 Willet *et al.* comment on the similarity search as "a very crude way of accessing a structural database" and suggest that when there is more than one molecule available, other LBVS methods should be applied and ultimately, when receptor structure is available, SBVS should be used.[43] Since then, the attitude has changed considerably,[50,51] and as of today, it is an established VS method with many success stories.[6,52]

2.1.3 Pharmacophores

A pharmacophore describes the portion of the ligand that is responsible for triggering the desired biological response during interaction with the macromolecule. It is not an exact description of the ligand, it is rather a common denominator of several ligands that follow the same binding mode. Functional groups contributing into binding, usually hydrophobic, aromatic, charges, hydrogen bond donors and acceptors, are identified and described as points in three-dimensional space. This pattern of groups can then be used to find new ligands, based on the placement of the same type of functional groups in the molecule, and distances between them. This is indeed the most straightforward usage of the pharmacophore, but being such an abstract and universal concept it is, similarly to fingerprints, embedded

deeply into virtual screening, and can be encountered in many different methods.[53]

This flexibility is what makes pharmacophores such a powerful approach. It can be used as both ligand-based and structure-based method, meaning that pharmacophoric pattern can be extracted from active ligands or receptor structure.[54] It can manage the flexibility of ligands and proteins through series of conformations.[55] When activity values of ligands are added to the equation, the most favourable groups can be identified. When derived from the receptor structure, the shape of the binding site can be taken into account by defining forbidden areas for the ligand, effectively avoiding clashes with the receptor.[56] Since only the groups responsible for the interaction are defined and no assumption is made about the molecule between them, finding novel scaffolds is easily achievable, both in theory and practice.[57] All that considered, it is not surprising that pharmacophores have many success stories. Due being faster than docking, it is often used as a pre-screening tool, although studies have demonstrated its comparable efficiency.

Pharmacophores are usually depicted separately from similarity search, but the information extracted in this way can easily be encoded into fingerprint notation and used as such.[44,53,58]

2.1.4 Shape similarity

Considering the close complementarity of receptor and a bound ligand, observable in an X-ray structure, the importance of molecular shape is deeply rooted in medicinal chemistry.[59] The "lock-and-key" metaphor for ligand-receptor binding allows to raise an hypothesis that active ligands should have similar shape and volume. This concept is exploited in shape matching techniques, which attempt to find molecules with similar volumetric arrangement in three-dimensional space.

The first task in the process is to describe the shape in some mathematically approachable way. Popular methods have been described by Putta and Beroza in their excellent review.[48] To summarise, they fall into four categories: moment-based, gnomic-based, volume-based, and surface-based methods. Moment-based methods describe the molecule as the set of multipole moments of inertia;[60,61] they are efficient to calculate and are often used as preliminary alignment. Gnomic methods map the molecule onto a simple surface and the points of that surface are encoded with additional information, such as distance to closest pharmacophoric groups.[62,63] Volume-based shape representation is currently most common, it is used in leading shape matching programs, ROCS[64,65] and SW/SQW[66]. Atoms in molecule are described as intersecting hard spheres with van der Waals radius,[67] or Gaussian functions.[68] The latter is advantageous due to more efficient calculation and simplified mathematical operations. Steric grid can also be used in volume-based representations; each grid point receives a value describing its relation to underlying molecular shape. Probably the most accurate approach in shape description is the surface-based representation. The surface can be described as a shell of finite width[69] or set of patches on the surface of the molecule's shape.[70,71] This is less common due to complicated calculations.

The shape alone is not enough to determine if a ligand binds to a macromolecule. Interactions such as hydrogen bonding, ionic, hydrophobic, and van der Waals forces among others, also have an important role,[72] therefore shape similar-

ity methods also look for an additional information in the molecules. Depending on the method used for shape description, this information can be encoded as atom types,[66,73,74] coloured grid points[75] or surface points.[62,76] Interestingly, too detailed description may not be the best solution, the findings of Sastry *et al.* suggest that simple pharmacophoric colouring is preferable to exact atom mapping.[77]

Similarly to pharmacophores, shape derived from the molecule can be used directly to search for matches with other molecules, but it is also possible to encode it in the form of molecular descriptors, such as fingerprints.[78,79] Following a similar derivation route, 3D shape fingerprints are composed from a library of predefined shapes. Once prepared, they can be used as conventional fingerprints, described in section 2.1.2. Same approach of comparison applies also to a method described by Zauhar *et al.*, where a histogram is composed from the lengths of rays reflected inside the surface of the molecule.[80]

The strength of the descriptor-based approach is that it avoids the most challenging aspect of shape matching – the alignment of structures. Alignment attempts to find one or more three-dimensional overlays of one molecule onto the other. This problem can be solved in numerable ways and the details of individual programs vary, but a distinction of global and local shape alignment can be made – global methods seek to match entire molecules, local methods identify smaller fragments and try to match them individually. This raises the problem of combinatorics, as many local-to-local combinations have to be examined, making the local approach more time consuming. It is useful however, when query and target molecule follow different binding modes, or only a portion of query molecule is responsible for binding interactions with the receptor.

Comparative studies with docking have revealed that shape similarity methods perform as well or even better than docking,[65,81] yet given the simpler approach, are about an order of magnitude faster.[48] While described here as 3D method, it has been demonstrated to work encompassing only 2D structures of molecules with no significant loss in results, but strong beneficial effect on speed.[82] Considering its performance in active ligand retrieval and scaffold hopping it has been suggested to be the best option in lead discovery stage.[59]

2.1.5 QSAR and data mining methods

Methods described in this section share a common trait of using conventional statistical techniques on virtual screening. They cannot work on their own though, their usage is always preceded by some parameterisation of the molecules, i.e. calculation of molecular descriptors.[83–85] Over the years, the number of these has grown almost incomprehensibly large and is way beyond the scope of this thesis, or even a single book; only a very brief summary of those most important to virtual screening, is therefore given here.

There is a special interest in VS towards descriptors that can be calculated quickly and capture biological properties that are essential for a drug molecule, such as absorption in gastrointestinal tract; distribution and metabolism in organism; and excretion; collectively denoted as ADME. Parameters that are often used to describe them, are water-octanol partition coefficient (logP),[86–89] which is used to model the cell permeability; and water solubility,[90–92] describing the overall chances

of drug reaching the right place. The acid dissociation constant (pKa),[93,94] has also received a lot of attention. It is necessary to describe the charge distribution of the molecule and serves therefore as the basis for calculating, for example the distribution coefficient (logD) or tautomerisation states. When it comes to speed of calculation, simpler is obviously better, and 1D fragment-based descriptors, such as number of acceptor or donor sites, or number of rotatable bonds, have been used extensively.[95–97] Topological and other 2D information-based descriptors have become common in VS.[85] They are simple to calculate, and have also been proven to be a reasonable replacement for logP.[98]

It should be emphasised, that the data mining techniques themselves have no problem making use of any kind of molecular descriptors. Once they are calculated, what follows is the application of common statistical methods, that are applicable in any discipline to clusterise, classify or rank molecules. These methods are vastly numerous and have been reviewed in context of virtual screening repeatedly.[99–101] One of the most popular appears to be regression analysis, which in VS, and chemoinformatics in general, is known as quantitative structure-activity relationships (QSAR) or quantitative structure-property analysis. These methods attempt to establish a relationship between dependent, and one or more independent variables through statistical techniques such as multiple linear regression (MLR), partial least squares (or projection to latent structures) (PLS), principal component regression (PCR) or artificial neural networks (ANN). Statistical models resulting from the data mining methods allow to analyse the data set, and a good understanding of the data allows to predict properties or activities of new compounds. Regression analysis is generally a simple and reliable method, but it requires the user to have a good understanding about the property or mechanism, and the proper validation of the results. Misconceptions about this principle have lead to publication of many inadequate models, those resulting from chance correlations and containing meaningless variables;[102,103] and this has had an effect on the credibility of the whole research area.[104,105] The problem has been phrased as "Kubinyi paradox – the models that fit the best retrospectively tend to predict the worst prospectively".[106] This has resulted in adopting more stringent development and validation methods.[107–110]

When the existing data does not allow to create a strict regression model, or it is not required, classification can be used instead, a common case in VS being separation of active ligands from the inactive ones. A well-established method in VS is, for example, a decision tree, where classification occurs in a successive evaluation of descriptors, often depicted in a tree-like manner, where each leaf node denotes a different class. They are fast and efficient to calculate,[111] but have a common problem of overtraining - it is possible to perfectly classify a training set, but in the process, the capability of generalisation is lost i.e. the method just learns the data. Another popular classification technique is principal component analysis (PCA).[112] In this method, the information in descriptor pool is transformed into principal components, linear combinations of input descriptors. Since principal components are extracted in such way to maximise the information content of the descriptor, just a handful of them are needed to capture almost all the information in descriptor pool. It is used therefore for one, as a data reduction method. Clustering is performed using a first few principal components, since they capture the most of

the information contained in descriptors.

Classification can also be done by using the support vector machines (SVM).[113] In this method, a hyperplane is constructed in descriptor space maximising the distance to the nearest training data points belonging to different classes (active and inactive for example). In its original form, SVM is a linear binary classifier, but being an attractive method, several modifications have been proposed, which allow the method to overcome its original limitations, and enable its usage as a multi-class classifier,[114] non-linear modeller,[115] or regression method.[116]

A family of methods stemming from probability theory have become increasingly popular in VS and chemistry in general. They are based on Bayes theorem[117] and therefore referred to as Bayesian methods.[118,119] In VS, they are generally used for classification,[120,121] such as in naïve Bayesian classifiers[122–125] or binary kernel discrimination,[126,127] but in principle, they can also be used for ranking and regression.

2.1.6 Filters

The underlying mechanisms of filter methods do not represent any conceptually new technologies. Their purpose is, however, slightly different from other VS methods, which is why they receive a separate section in this overview. The best known among them is Lipinski's rule-of-five. In 1997 Lipinski *et al.* studied a set of known drugs and noticed that they tend to lie between certain values of some simple physicochemical parameters.[128] This enabled them to formulate a set of rules, which state that the compound is likely to have poor absorption if at least two of the four parameters are true: the number of hydrogen bond acceptors $>$ 10, the number of hydrogen bond donors $>$ 5, logP $>$ 5 and molecular weight $>$ 500. This rather crude effort was immediately picked up by other authors[129] and further enhanced by adding other parameters. Among the most popular ones were the number of rotatable bonds, to account for molecule's flexibility, and polar surface area to better describe hydrogen bonding properties.[130] But it did not stop there, many other parameters were tried, such as molar refractivity, net charge, number of heavy atoms and many others.[96,97,130–133] Sets of toxic and reactive fragments have been composed,[134–136] allowing a molecule containing any of these to be removed. Statistical models predicting ADME properties,[137–139] toxicity[140] and brain-blood barrier[141–143] have been developed and used, as well as solubility in water[90–92] and of course endless takes on logP.[86–89,144]

It appears that filters exploit the concepts of "drug-likeness",[130,132,133] or "lead-likeness"[133,145,146] and attempt to focus the database to more prospective molecules by eliminating molecules that are too large, poorly soluble, overly toxic or otherwise unsuitable. It makes sense not to waste time on them,[147] but in the light of lead optimisation, the technique of improving quality of a hit molecule found in (virtual) screening, there is another, more pragmatic reason. Given the size of molecular databases, containing millions of entities, it is often desirable to restrict the initial size of the library as early, as strongly, and with as little effort as possible. Filters provide an intuitive and easily understandable means of achieving that. This helps to define their scope in virtual screening. They are not specific to a single target, they are mainly concerned with pharmacokinetic, rather than pharmacodynamic

properties, and they are aimed to be fast, which as a result makes them quite crude, a point succinctly illustrated by the name of one of the programs used for filtering: REOS, or Rapid Elimination Of Swill.[134]

Today, filters have become common and their usage in VS is widely accepted, thus they do not receive a lot of attention in the papers any more. It is not even necessary to apply them, because virtual screening databases, such as ZINC[38], offer pre-filtered sets of drug-like[148] or lead-like[149] molecules; PubChem[150] offers filtering search results according to rule-of-five; and ChEMBL provides info about compounds regarding rule-of-five and rule-of-three.

2.2 Comparing virtual screening methods

The number of current virtual screening methods raises a question how make a choice between them. Comparison of methods and evaluation of their advantages, drawbacks, performance, cost and results is therefore of highest interest. The obvious way to compare methods is based on their ability to separate active compounds from inactive ones by classification or ranking. Several common statistical methods can be used, such as analysis of variance (ANOVA),[151] the Z-score[152] or Matthews correlation coefficient[153] among others.[154] Surpassing these by popularity is enrichment factor (EF),[23,155] which is calculated as in eq. 1. $HITS$ is the number of known actives and N is the number of compounds. The $sampled-set$ commonly refers to 1% or 5% of the best scoring compounds.

$$EF = \frac{HITS_{sampled-set}/N_{sampled-set}}{HITS_{total-database}/N_{total-database}} \tag{1}$$

To write it out in human-readable form, EF simply shows how much more actives are contained in a set of selected N compounds compared to randomly chosen N compounds. Completely random selection will give $EF = 1$; values above that indicate concentrating the set in terms of active ligands, values below 1 mean worse than random selection.

Another frequently used method in virtual screening are the receiver-operator characteristic (ROC) curves.[156,157] They are in essence a 2D representation of cost-benefit analysis; in virtual screening terms it says how many false positives (compounds falsely predicted to be active) accompany true positives in a given classifier settings. They provide a powerful and thorough insight into the method's performance, but being in graphical representation, are somewhat cumbersome to compare. To yield a more comfortable, single scalar value, area under the ROC-curve (AUC) is often used, which is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.[157]

While EF and AUC are both popular methods of assessing success of virtual screening, they both have the 'early recognition' problem. What it means, is that in practical matter, VS is expected to preselect compounds form database for experimental testing, and the successful method is expected to significantly reduce the amount of experiments. This sets a requirement to a VS method, to rank active compounds the best. But EF doesn't rank compounds within threshold limit, and AUC can evaluate curve with more suitable shape the same as worse

one. This can be overcome by using logarithmic transformation on the curve[158] or using exponential weighting scheme.[154,159] Indices have been developed to tackle this problem, for example robust initial enhancement (RIE)[159] and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC).[154]

Since the current level of virtual screening cannot guarantee to find all possible active compounds, it is therefore of highest interest, that the set of compounds selected, also contains novel chemotypes. These are structures with new scaffolds that can explore wider areas in chemical space and are not covered by existing patents. This so-called "scaffold hopping" ability is another criteria determining the success of virtual screening. It is generally assumed that on this area, structure-based methods have the advantage over ligand-based ones, because they do not set any requirements on ligand's structure, but only its function. That belief is not without a reason, a recent comprehensive study by Ripphausen *et al.* demonstrated, that novel chemotypes were more frequently discovered using SBVS.[6] However, LBVS had a surprising advantage: while SBVS was more successful in scaffold hopping, LBVS often found ligands with higher experimental activity.[6] Ligand-based methods rely on the concept of molecular similarity and assume the presence of known active, and preferably also inactive ligands. The level of the performance of the method is therefore inherently connected to the limited amount of information provided by training compounds, and how well they sample the chemical space, especially around activity cliffs, areas in chemical space, where small changes in structure are accompanied by large changes in activity.[160,161] LBVS methods can be therefore likened to local models. The binding site structure, on the other hand, contributes much more information to the process, it defines the shape of the ligand, as well as how it interacts with the target protein, but does it without setting rules on how exactly atoms in the molecule should be arranged, keeping open the possibility of fitting it with a new ligand, which looks completely different, but has similar activity. It defines the complete possible chemical space for a potential ligand and can therefore be seen as a global model. This is where the difference emerges: while global models have the capacity to describe the wider area, they are less accurate on separate entities, and that's where local models with their closer focus have an edge.

For purely practical reasons, method's resource requirement is also a point of interest, especially in the case when comparing methods with otherwise similar capacity. It can be generally said, that ligand-based methods are faster than structure-based methods, because they have to process less information. The same holds true also along the other important divide in VS methods, 3D methods with their need to generate spatial geometry or to perform a conformational search, tend to be more time-consuming than 2D methods. But when it comes to comparing results, things are not so clear. Intuitively, putting more information into an equation should give more accurate results, but 3D methods are not significantly better than 2D methods, and similarly, using the receptor structure doesn't guarantee an advantage over ligand-based methods.

2.3 Current trends and problems in virtual screening

Virtual screening has been around now for a few decades. It has been widely accepted as a useful technique in drug design, but despite the popularity and active development, it has some critical problems, casting a shadow to the whole area. This section summarises the most acute difficulties of the VS, as well as the directions it is evolving.

Starting with the problems, this section can pick up where the last one left off, because validation or evaluation of methods is one of the biggest concerns of VS today. Given the number of methods in existence, it is unrealistic to compare them all in a single study, the lack of universal standards, however, makes the comparison across different articles precarious. While there are popular evaluation methods, that can often be found, such as enrichment factor and ROC-curves, the outcome depends also on which target and data sets are used. Different nature of binding sites ensures that methods do not to work equally well on all targets, and data sets are prone to bias, thereby favouring different methods. The reasons why data sets are biased are purely pragmatic – when a suitable scaffold is discovered, it is usually put through a lead modification process, meaning that a series of compounds with the same scaffold but different substituents, are synthesized and tested. As a result, the number of known ligands could seem large enough, but the underlying chemical space is poorly sampled and does not provide enough information for virtual screening[158] The popular validation database, DUD, was therefore cleaned from structural redundancy, which has in many cases significantly reduced the individual ligand and decoy sets.[40]

Another critical problem, also suggested in previous section, stems from the observation that complex and advanced methods are not always better than simple ones. An interesting study by Bender and Glen,[161] where several types of fingerprints were compared, demonstrated that a simple atom count vectors outperformed in some cases much more sophisticated fingerprints. This is truly remarkable, because atom count vectors are essentially molecular formula, 1D descriptor, which is not really able to distinguish molecules due to high redundancy. This raises some disturbing questions as what do we exactly know about virtual screening. Unpredictable compound retrieval and level of accuracy when using different methods with varying complexity, clearly indicates our poor apprehension of molecular descriptors and their interpretation.[162] Despite the work done in this area,[163] the results of virtual screening are still inconsistent. An obvious and rather crude solution at the moment would be to simply use several methods simultaneously, and select compounds based on consensus or average.[51,162,164,165] This may have a hindering effect, when one method is significantly worse than the other, but usually covering different methods and thereby different chemical descriptions allows to reduce false positives and negatives.

So far in the present thesis, virtual screening has been depicted as a method to find ligands for a single target. However, VS is more versatile than that and it is a growing trend to employ many targets in a single project. The exact opposite of the usual paradigm is called "target fishing" and as the name suggests, it is finding a suitable target for a known ligand.[166–168] The idea of this approach is drug re-purposing[169] – existing drugs may have undiscovered beneficial effects and since they have already undergone medicinal testing and approval, the process of

bringing them to market becomes significantly faster and cheaper. But this is not the only reason to use many targets. When a drug is introduced into the organism, it has a theoretical possibility to interact with many other proteins or nucleic acids, which may cause severe side effects. For example, the Ether-à-go-go Related Gene (hERG),[170] blocking of which is related to cardiac arrhythmia; the cytochrome P450 superfamily,[171] that play crucial roles in the metabolism and biosynthesis of endogenous compounds, and the metabolism of drugs and non-drug xenobiotics; pregnane X receptor (PXR),[172] mediator of expression of several proteins, including the P450 enzymes; and transporter proteins,[173,174] among others.[175] Identifying these 'off-targets' or 'anti-targets', and testing whether they have significant affinity towards the ligand by using virtual screening, has therefore great potential to reduce drug candidate fail-rate in clinical trials.

Besides the targets that the drugs are not supposed to hit, there may also be more than one target the drug is required to interact with, in order to work effectively. This approach, called polypharmacology or network pharmacology,[176,177] comes from the observation that designing a perfectly selective ligand for a target may not be enough for effective treatment of a disease.[178] The reasons lie in cell biology – a hugely complex structure a cell is, there is functional redundancy, meaning that like network connection, there is a possibility to find alternative routes to bypass the deletion of individual nodes.[179–181] That means disabling one of two genes has no significant effect on cell's viability, but simultaneous inhibition leads to impaired functionality or death.[182–185] For example, it has been demonstrated that synergistic effect of compound targeting two kinases is greater than the additive sum of it acting on each kinase individually.[186] Four targets have been identified to be necessary to inhibit to stop metastatic progression of breast cancer in mouse model.[187,188] Effective antibiotics often target several rather than single proteins,[189,190] and similar observations have been made also about central nervous system disorder drugs.[191]

For drug discovery, the obvious implications are, that instead of searching for a single "disease-causing" protein, polypharmacology suggests targeting the appropriate subgraph in the network.[192,193] In virtual screening, this can either mean parallel screens for different protein structures using docking or pharmacophore models, or fragment based approaches based on selected fragments known to bind several required targets separately.[194,195]

As a rather fresh approach in VS, multi-target screening has, as of yet, been used sparsely, and only a few examples can be brought here. Huang *et al.*, in a search of a cure for Alzheimer's disease, used first a pharmacophore model of -secretase 1 (BACE 1) inhibitors, and then filtered the results using molecular docking to acetylcholinesterase (AChE) structure.[196] The same group also reports two multi-target cancer studies, targeting several kinases using support vector machines and docking;[197,198] and also a study on anticoagulant activities, targeting factor Xa and thrombin in human clotting cascade.[199] Prado-Prado *et al.* have developed ligand-based classification approach based QSAR models, trained on known antibacterial drugs, active against several lines.[200,201] More examples can be brought from a simpler case, where both mutant and wild-type structures of the same target are used. This is more common, because using multiple structures is a well known technique in docking to model protein flexibility,[202] and incorporating also mutant

protein structures, making this effectively a multi-target approach, is just a logical step further. In here the predominant interest has towards HIV and malaria.[203]

To sum up the literature overview part of this thesis, VS, its methods and paradigms undergo an extremely active development. Originally seen as simply an *in silico* analogue for high-throughput screening, it has now widened its territory outside of its traditional habitat and has become universal tool in drug design. Most of the papers published on the subject do not simply exploit some existing technique, but rather complement them with something new, from additional validation scheme to entirely new method. And as it can be seen in next chapter, this holds true also for the articles this thesis is based on.

3 Methods

Virtual screening as a process can rarely manage with just one step. In addition to the deployment of actual screening with any of the methods described in section 2.1, the data also needs to be prepared, analysed and validated. Several methods can be used, for example a fast filter method before the more demanding one. Virtual screening workflows as found in literature, can therefore grow to be rather elaborate. This chapter takes a closer look on the schemes used in individual articles this thesis is based on. Examples of the two important paradigms in virtual screening – SBVS and LBVS can both be found. Articles I and II employ the ligand-based approach by following the classical QSAR methodology. A series of known ligands is analysed with linear regression and as a result, a model is developed. Article III uses docking as its main method, and a new structure-based VS method is developed in the fourth one.

3.1 QSAR and data mining

Articles I and II use a similar set of methods which are described in section 2.1.5. In both cases, a multiple linear regression model is developed on a set of training compounds with a goal of analysing the existing data and predicting properties of new molecules.

The first step in the workflow was the conformational search with MacroModel,[204] followed by geometry optimisations and quantum-chemical calculations with MOPAC,[205] and calculation of several topological, geometric, charge distibution-related and quantum-chemical molecular descriptors with CODESSA software.[206,207] After calculation, descriptors were subjected to nonlinear transformations, namely inverse, square, square-root and logarithmic. This is due to the observation, that the relationship between property and parameters is often nonlinear, therefore these transformations help to fit the property more accurately.

Prior to the model development, an additional clustering step was used in article II, where the data set was broken down into smaller pieces and individual QSAR models were derived on these subsets. Cluster analysis was performed with the PCA methodology and the classification was done based on score plot of first two principal components, which were found to be correlated to size and hydrophobicity.

Descriptor selection and model development was carried out using CODESSA's Best Multi-Linear Regression (BMLR) method,[207] which is essentially a forward selection method, where descriptors are successively added to existing models, and retained if significant improvement in statistics is observed. All models were validated internally using statistical tests – cross-validated correlation coefficients, F-test and t-test; external validation using a set of test compounds was carried out where possible.

When used in virtual screening, the applicability domain of the models must be considered. As a similarity-based method, the models do not perform reliably on compounds significantly different form those in training set, therefore it would not make sense to apply it on entire ZINC library for example. Instead, a combinatorial library can be created, using a set of possible substituents (analysis of outliers in the papers can be helpful at this point) and adding them to the proper locations in core structure.

3.2 Docking

Molecular docking was used as the predominant method in article III, and as an additional validation method in article IV. Due to problems with exact activity prediction and ranking, as pointed out in section 2.1.1, a common practice is to compare the results with known good, and if possible, bad binders. The score values of good binders, such as existing drugs or other known ligands, are used to set the threshold value for new ligands. If some known bad binders are available, they can be used to make sure, that the algorithm can tell them apart from good ligands. This approach was also used in article IV.

Article III however, employs a much more comprehensive and stringent selection process. The schematic depiction of the workflow is brought in figure 1. Ligands were chosen from the ZINC database, which at first was filtered to manageable size using a set of pharmacokinetic and structural rules. The filter parameters were set based on literature[96,97,128,130,131,208] and are depicted in table 1. The approach used was conservative, while rule-of-five allows some rules to be broken, this time all compounds had to comply exactly. The objective was to retain only a small subset, so that the following docking would not take excessively long time. After applying all filters, the initial database of 5,627,809 compounds was reduced to just 65,035.

Several X-ray structures of the receptor were used, covering wild-type and mutant structures, different hydration states and some conformational flexibility. A set of five anti-targets were used, human sulfotransferase 1A3, pregnane X receptor, and three cytochrome P450 enzymes. They are some of the proteins that the drug is likely to interact with in the organism, and could therefore alter it's potency. Two docking programs were used in the study, Autodock and Glide, to get two opinions on ligands, but the consensus was used differently with targets and anti-targets. With targets, to reduce the occurrence of false positives, both programs had to give high rating to the ligand, higher than the set of known ligands used for reference. With anti-targets, as a safety measure, high score from only one sufficed.

After docking, additional ranking was performed according to the efficiency indices, and the number of hydrogen bonds formed with the protein backbone.

	Min	Max
Water-octanol partition coefficient $logP$	-3	5
Molecular weight (g/mol)	300	650
Number of rotatable bonds	5	12
Topological polar surface area[209] $(\AA^2)$	25	180
Number of hydrogen bond acceptors	2	
Number of hydrogen bond donors	3	
Solubility in water $logS$	-5	
Number of rings		6
Number of fused rings		6
Ring size		7

Table 1: Filtering parameters.

Protein chain backbone atoms are less susceptible to mutations than side chain atoms, they are often found to be more stable and easier to detect from X-ray studies, which makes their position more reliable. Therefore, re-ranking ligands according to hydrogen bonds with protein backbone, gives preference to ligands that are less affected by mutations. Ligand efficiency indices are free energy of binding normalised with molecular weight or any other available parameters, including and not limited to molecular surface area, polar surface area, the number of heavy atoms or the number of rotatable bonds; and can thereby indicate effective binding per atom or other pharmacokinetically relevant parameter.[210,211] It has been detected that they are able to improve correlations between experimental and calculated binding parameters[211] and separate drugs from non-drugs. In this study, they were calculated using the average binding energy across different programs and target structures. As a result of these additional steps, the ligands selected are more likely to be pharmacologically relevant.

3.3 Topological docking

Topological docking is a novel structure-based virtual screening method, which development and deployment is described in article IV. The method starts with transforming the 3D structure of the binding site into the 2D distance matrix, in order for it to be directly comparable with the molecular graph. Hydrophobic, hydrogen bond donor and hydrogen bond acceptor subsites are identified and a point in three-dimensional space is assigned to each, usually coinciding with the centre of the area. To complete the transformation, distances between all points are calculated and stored in a distance matrix. Distances are measured along the shortest path that does not intersect with the protein surface that ensures more accurate description of the binding site, because it implicitly defines forbidden areas for the ligand.

Principally, the topological representation of the molecule is already in the form

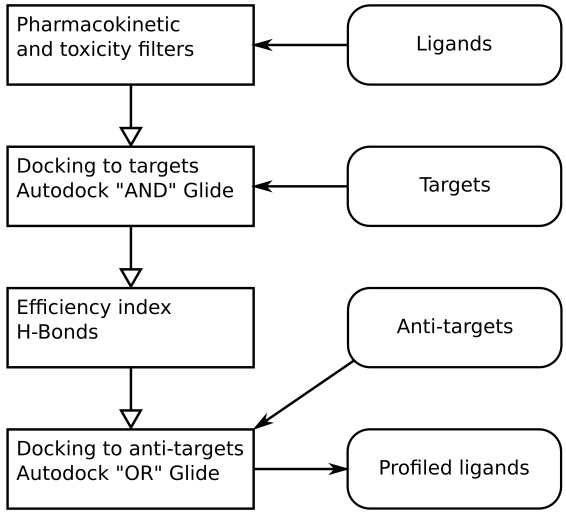

Figure 1: The workflow in article III.

of the distance matrix, but some small modifications are still necessary. Hydrogen bond donors and acceptors reside on one atom, but since hydrophobic areas in the molecule are comprised of several atoms, they need to be described using an additional dummy atom. Distance matrix is composed now using the donors, acceptors and dummy atoms marking the hydrophobic centre. Distances this time are not distances in three-dimensional space but rather along the shortest path in the molecular graph. This approach ensures a certain amount of flexibility of both protein and ligand to be accounted for, because distances along the path in molecular graph should almost always exceed the corresponding distances in 3D space, unless the molecular fragment is perfectly rigid and linear, an extremely rare case when more than 3 atoms are involved.

Testing whether a ligand fits into the binding site is achieved by a process which can be viewed as comparing the corresponding distance matrices. Two matrices, overlaid onto each other, depict a pose of the ligand, i.e. the way the ligand's functional groups are assigned into the subsites in the binding pocket. This is analogous to an actual three-dimensional arrangement of the ligand in the binding site, as in docking, ensuring a clear interpretation of the results. When the distances in ligand's matrix are greater than or equal to those in binding site's matrix, the pose is considered a match, otherwise it is rejected and a new pose is evaluated. Generating a new pose is as simple as shuffling the columns in one of the matrices. Besides the pose, size of the ligand is also assessed, whether the molecule can actually fit into the binding site, or is it too big for this. Altogether this makes the process quite similar to docking, hence the method was called 'topological docking'.

Testing and development was done on six target systems: poly(ADP-ribose) polymerase, P38 mitogen activated protein kinase, phosphodiesterase 5A, platelet

derived growth factor receptor kinase, thrombin and HIV-1 protease. Data for the tests, active ligands and non-active decoys were acquired from the DUD database, and a subset of around 7,000 randomly selected compounds were used from the ZINC database. A full-scale screening for HIV-1 protease inhibitors was performed on the entire ZINC database, about 14 million compounds, and the results were further validated with docking.

4 Summary of original publications

4.1 QSAR Modeling of HIV-1 Protease Inhibition on Six- and Seven-membered Cyclic Ureas

Inhibiting HIV-1 protease is a potential pathway of blocking the reproduction of virus in organism, and as such has been an object of interest in drug design since 1988[212]. Several known drugs are exploiting that mechanism, such as Saquinavir or Atazanavir among others. Protease inhibitors usually form hydrogen bonds not only directly with the protein itself but also to a water molecule in the active site of protease. The cyclic urea-based inhibitors (figure 2, a), developed by DuPont Merck[213] are therefore an unconventional class of inhibitors because the carbonyl atom in the core structure effectively replaces the water molecule, giving the inhibitors of this type higher entropic efficiency.

Article I describes the development of linear QSAR model that can predict the activities of cyclic urea-based HIV-1 protease inhibitors. The model is based on large and diverse data set, and according to statistical parameters fits the training data well ($R^2 = 0.82$; $s = 0.46$). The standard error of estimation was close to experimental error of measurement found in the source articles, indicating that further improvement begins to compromise the generalisation and lead to overfitting. The model was validated both internally (using statistical tests) and externally (with a set of compounds not used in model training). The model contained four parameters, one topological descriptor, describing the size and shape of the molecule; and three charge distribution-related descriptors, which characterised the hydrogen bonding and electrostatic properties of ligand molecules. None of the descriptors require laborious quantum-mechanical calculations, they are all simple and fast to calculate, making the model especially suitable for virtual screening. Some care must be taken about conformational analysis because these types of inhibitors are quite flexible. The analysis of the model's applicability domain revealed that due to the diverse training set it successfully functions over a wide range of substituents. Problematic areas were detected around structures with rare functional groups and also isomery.

4.2 The QSAR Modeling of Cytotoxicity on Anthraquinones

Cytotoxicity in virtual screening is often related to cancer treatment studies. Article II examines the case of anthraquinone derivatives (figure 2, b), and their cytotoxicity on human hepatoma G2 cell line. Similarly to article I, the initial plan was to compose a large and diverse data set, and develop a fairly general model to predict anthraquinones cytotoxic activity. The problem, however, turned out to

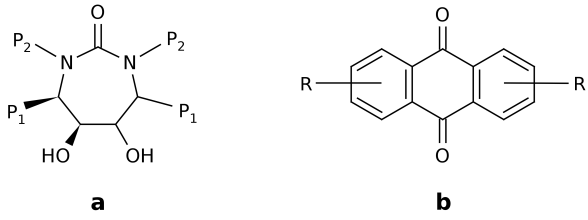

Figure 2: Core structures used in article I (a) and article II (b).

be more difficult, and instead of a single model, a classification scheme and three separate models for each class was required for successful modelling.

The data set was divided into three classes based on first two principal components of the PCA analysis, which were related to size and hydrophobicity. The largest of these, 47 compounds, contained the most hydrophobic ones, and the parameters in the resulting model describe the molecule's polarity, electrostatic-, and other charge distribution-related properties. Second group contains compounds with least hydrophobicity, most compounds which were protonated under physiological condition, were placed in this group by the PCA. Descriptors in the model are appropriate, all three describe the charge distribution. Final PCA cluster was the smallest (20 compounds) and their characteristic properties were small size and low hydrophobicity. Besides the size and hydrophobicity, the descriptors in the model were also related to reactivity.

Why this data set, which was not particularly big or terribly diverse did not yield any meaningful results on the whole and needed to be broken into smaller sets, remains somewhat unclear. It can be speculated to be caused by the complicated property. The toxicity of anthraquinones is attributed to their intercalation with the DNA. Their three-ring core structure is planar, aromatic and suitably sized to fit between the base pairs of the DNA strand, disrupting thereby it's transcription, replication and repair processes. Other unknown toxicity pathways, though less likely, can also not be excluded. Additionally, the property measured was not directly the binding to the DNA, but rather the 50% inhibition of cellular growth (IC_{50}), so pharmacokinetic in addition to pharmacodynamic properties (such as solubility and membrane permeation), come into play. Since the property seemed to contain too many unknowns, the best we could do was to provide a clustering scheme and derive focused models for different classes which were statistically more coherent.

4.3 Combined Approach Using Ligand Efficiency, Cross-Docking, and Antitarget Hits for Wild-Type and Drug-Resistant Y181C HIV-1 Reverse Transcriptase

HIV-1 reverse transcriptase (HIVRT) is an attractive target in AIDS treatment. Its function, producing a DNA sequence from viral RNA, is unprecedented in humans, allowing theoretically the design of highly selective inhibitors. It's attractiveness is however somewhat hindered by it's fast mutation rate, which enables the virus to

quickly develop resistance to drugs.[214] Despite several drugs already in existence, the search for new HIVRT inhibitors actively continues. This is also the aim in article III, which is set out to find novel ligands, that are effective against both wild type and Y181C mutant structures, a troublesome mutation which modifies the binding site to the extent of causing drug resistance.[215]

This paper brings an example of how virtual screening can be used in a more thorough manner than just finding some potential hits. A comprehensive profile was compiled of the potential ligands, considering not only their estimated activity, but also sensitivity to a mutation, adaptability to HIVRT's rather flexible binding site, and the activity against anti-targets, proteins that the drug is likely to come across in the organism while travelling to the site of action. As a result, a number of small and structurally diverse compounds were identified as a potential inhibitors of both wild-type and mutant HIVRT, which having passed all the selection criteria, are expected to have higher likelihood of surviving in the following steps of the drug design process.

Post-screening research revealed that some of the proposed ligands also appeared to be active against other targets, exhibiting anti-tumor and anti-influenza properties. This confirms the sound composition principles of the data set used in docking. The pharmacokinetic rules applied onto the set concentrated the database in terms of pharmacologically relevant compounds and thus allow it to be used as a starting point in virtual screening workflow. The versatility of this set is also demonstrated by the fact that it has been successfully employed in other projects besides the current article. Inhibitors have been found for E3 ubiquitin ligase neuregulin receptor degrading protein 1 (Nrdp1), Ras-related C3 botulinum toxin substrate 1 (Rac1) (both currently unpublished), and avian influenza H5N1 neuraminidase.[216]

4.4 A Novel Structure-based Virtual Screening Method Finds Active Ligands through the Process of 'Topological Docking'

The fourth article describes the development of a new virtual screening method, which attempts to combine the generalisation power of structure-based methods with the ease of computation allowed by encompassing only 2D information of the ligand molecules. Previous work on environmental toxicology,[98] had revealed that the topological representation is a thorough and powerful way to describe a molecule, especially in the case of a drug-like molecules, which are preferably quite rigid and limited in size. For that reason, abandoning 3D information does not result in a critical loss of information, but gains an advantage in speed, therefore the method aims to be a quick and accurate scanning method for large databases.

The main result of the article, the method itself, is described in section 3.3, so just a few points are left to emphasise here. As it is common in VS, the method's behaviour was different on different targets. The results ranged from unsuccessful in the case of poly(ADP-ribose) polymerase, where the method was unable to discriminate active ligands from decoys, to great in the case of HIV protease, where almost 1,000-fold reduction of the input data was observed. Subsequent docking of HIV protease results found several ligands with known experimental activity,

confirming method's capability to recognise active ligands.

The analysis of these results revealed, that the current problem areas are mainly concentrated around binding site description, especially small and hydrophobic binding pockets. Currently the method appears to be too coarse and has too low resolution to achieve sufficiently detailed description of such binding sites. This is the reason behind the good results on the HIV protease, its binding site is large and well-defined, giving the reason to believe, that the method can be improved with more sophisticated handling of the binding sites.

5 Summary

Virtual screening is currently under great interest and active development, thanks to its potential to significantly reduce the time and cost of the drug development process. It is a vast and prospective research area, comprising of several paradigms and numerous individual methods. The present thesis takes a closer look on some of them, as an example of ligand-based methods, a classical QSAR approach is tried, and from structure-based methods, docking was the primary choice. As a result of this work, QSAR models describing the activity of cyclic urea-based HIV-1 protease inhibitors and anthraquinone-based cytotoxic compounds were found; a focused library of drug-like molecules for virtual screening was composed; a few prospective ligands for HIV-1 protease and HIV-1 reverse transcriptase were discovered; and as a summarisation of the accumulated knowledge over the PhD studies, a new virtual screening method was developed.

References

[1] Eglen, R. M.; Schneider, G.; Bohm, H.-J. In *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH Verlag GmbH, 2000; Vol. 10, Chapter 1, pp 1–14.

[2] Silverman, R. B. *The Organic Chemistry of Drug Design and Drug Action*, 2nd ed.; Elsevier Academic Press, 2004; pp 7–120.

[3] Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Curr. Top. Med. Chem.* **2010**, *10*, 127–141.

[4] Teague, S. J. Learning lessons from drugs that have recently entered the market. *Drug Discovery Today* **2011**, *16*, 398–411.

[5] Shoichet, B. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

[6] Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.

[7] Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

[8] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269 – 288.

[9] Gunasekaran, K.; Nussinov, R. How Different are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J. Mol. Biol.* **2007**, *365*, 257 – 273.

[10] Boström, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion? *J. Med. Chem.* **2006**, *49*, 6716–6725.

[11] Kontoyianni, M.; Madhav, P.; Suchanek, E.; Seibel, W. Theoretical and Practical Considerations in Virtual Screening: A Beaten Field? *Curr. Med. Chem.* **2008**, *15*, 107–116.

[12] Huang, S.-Y.; Zou, X. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.* **2010**, *11*, 3016–3034.

[13] Stahl, M. In *Virtual Screening for Bioactive Molecules*; Wiley-VCH Verlag GmbH, 2000; Chapter 11, pp 229–264.

[14] Sippl, M. J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229 – 235.

[15] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337 – 356.

[16] Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.

[17] Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875.

[18] Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876–1882.

[19] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

[20] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470 – 489.

[21] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727 – 748.

[22] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

[23] Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.

[24] Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.* **1998**, *7*, 938–950.

[25] Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225–242.

[26] Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

[27] Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962–976.

[28] Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.

[29] Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742–755.

[30] Moitessier, N.; Therrien, E.; Hanessian, S. A Method for Induced-Fit Docking, Scoring, and Ranking of Flexible Ligands. Application to Peptidic and Pseudopeptidic -secretase (BACE 1) Inhibitors. *J. Med. Chem.* **2006**, *49*, 5885–5894.

[31] Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

[32] Chang, C.-E.; Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host−Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

[33] Steinbrecher, T.; Labahn, A. Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies. *Curr. Med. Chem.* **2010**, *17*, 767–785.

[34] Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

[35] van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.

[36] Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein−Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.

[37] Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149–164.

[38] Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182, http://zinc.docking.org/.

[39] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

[40] Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 14.

[41] Hubálek, Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews* **1982**, *57*, 669–689.

[42] David Ellis, J. F.-H.; Willett, P. Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management* **1994**, *3*, 128–149.

[43] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[44] Khedkar, S. A.; Malde, A. K.; Coutinho, E. C.; Srivastava, S. Pharmacophore Modeling in Drug Discovery and Development: An Overview. *Med. Chem.* **2007**, *3*, 187–197.

[45] Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

[46] Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

[47] Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-Target Interaction-Based Weighting of Substructures for Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.

[48] Putta, S.; Beroza, P. Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1514–1524.

[49] Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719–729.

[50] Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

[51] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046 – 1053.

[52] Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372–376.

[53] Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Predicting Molecular Interactions in silico: I. A Guide to Pharmacophore Identification and its Applications to Drug Design. *Curr. Med. Chem.* **2004**, *11*, 71–90.

[54] Löwer, M.; Proschak, E. Structure-Based Pharmacophores for Virtual Screening. *Mol. Inf.* **2011**, *30*, 398–404.

[55] Amaro, R. E.; Li, W. W. Emerging Methods for Ensemble-Based Virtual Screening. *Curr. Top. Med. Chem.* **2010**, *10*, 3–13.

[56] Sotriffer, C.; Klebe, G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Il Farmaco* **2002**, *57*, 243 – 251.

[57] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.

[58] Sun, H. Pharmacophore-Based Virtual Screening. *Curr. Med. Chem.* **2008**, *15*, 1018–1024.

[59] Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.

[60] Zyrianov, Y. Distribution-Based Descriptors of the Molecular Shape. *J. Chem. Inf. Model.* **2005**, *45*, 657–672.

[61] Mansfield, M. L.; Covell, D. G.; Jernigan, R. L. A New Class of Molecular Shape Descriptors. 1. Theory and Properties. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 259–273.

[62] Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.

[63] Perry, N. C.; Van Geerestein, V. J. Database searching on the basis of three-dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607–616.

[64] Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

[65] Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.

[66] Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.

[67] Connolly, M. L. Computation of molecular volume. *J. Am. Chem. Soc.* **1985**, *107*, 1118–1124.

[68] Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

[69] Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular shape comparison of angiotensin II receptor antagonists. *J. Med. Chem.* **1993**, *36*, 1230–1238.

[70] Goldman, B. B.; Wipke, W. T. Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 644–658.

[71] Cosgrove, D.; Bayada, D.; Johnson, A. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573–591.

[72] Silverman, R. B. *The Organic Chemistry of Drug Design and Drug Action*, 2nd ed.; Elsevier Academic Press, 2004; pp 122–172.

[73] Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 357–368.

[74] Robinson, D. D.; Lyne, P. D.; Richards, W. G. Partial Molecular Alignment via Local Structure Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 503–512.

[75] Pitman, M. C.; Huber, W. K.; Horn, H.; Krämer, A.; Rice, J. E.; Swope, W. C. FLASHFLOOD: A 3D Field-based similarity search and alignment method for flexible molecules. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 587–612.

[76] Hahn, M. Receptor Surface Models. 1. Definition and Construction. *J. Med. Chem.* **1995**, *38*, 2080–2090.

[77] Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455–2466.

[78] Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673–684.

[79] Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A Novel Shape-Feature Based Approach to Virtual Library Screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230–1240.

[80] Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.* **2003**, *46*, 5674–5690.

[81] McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

[82] Ebalunode, J. O.; Zheng, W. Unconventional 2D Shape Similarity Method Affords Comparable Enrichment as a 3D Shape Method in Virtual Screening Experiments. *J. Chem. Inf. Model.* **2009**, *49*, 1313–1320.

[83] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons, 2000.

[84] *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: New York, 1999.

[85] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH, 2009; Vol. 41.

[86] Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule approaches for calculating log P. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337–354.

[87] Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47–66.

[88] Viswanadhan, V.; Ghose, A.; Wendoloski, J. Estimating aqueous solvation and lipophilicity of small organic molecules: A comparative overview of atom/group contribution methods. *Perspect. Drug Discovery Des.* **2000**, *19*, 85–98.

[89] Petrauskas, A.; Kolovanov, E. ACD/Log P method description. *Perspect. Drug Discovery Des.* **2000**, *19*, 99–116.

[90] Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355 – 366.

[91] Faller, B.; Ertl, P. Computational approaches to determine drug solubility. *Adv. Drug Delivery Rev.* **2007**, *59*, 533 – 545.

[92] Duchowicz, P. R.; Castro, E. A. QSPR Studies on Aqueous Solubilities of Drug-Like Compounds. *Int. J. Mol. Sci.* **2009**, *10*, 2558–2577.

[93] Ho, J.; Coote, M. L. pKa Calculation of Some Biologically Important Carbon Acids - An Assessment of Contemporary Theoretical Procedures. *J. Chem. Theory Comput.* **2009**, *5*, 295–306.

[94] Riccardi, D.; Schaefer, P.; Cui, Q. pKa Calculations in Solution and Proteins with QM/MM Free Energy Perturbation Simulations: A Quantitative Test of QM/MM Protocols. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

[95] Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382–384.

[96] Walters, W. P.; Murcko, M. A. In *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH Verlag GmbH, 2000; Chapter 2, pp 15–32.

[97] Muegge, I. Selection Criteria for Drug-Like Compounds. *Med. Res. Rev.* **2003**, *23*, 302–321.

[98] Maran, U.; Sild, S.; Tulp, I.; Takkis, K.; Moosus, M. In *In Silico Toxicology*; Cronin, M., Madden, J., Eds.; Issues in Toxicology; The Royal Society of Chemistry: Cambridge, UK, 2010; Chapter 6, pp 148–192.

[99] Ekins, S.; Shimada, J.; Chang, C. Application of data mining approaches to drug delivery. *Adv. Drug Delivery Rev.* **2006**, *58*, 1409–1430.

[100] Weaver, D. C. Applying data mining techniques to library design, lead generation and lead optimization. *Curr. Opin. Chem. Biol.* **2004**, *8*, 264–270.

[101] Guha, R.; Gilbert, K.; Fox, G.; Pierce, M.; Wild, D.; Yuan, H. Advances in Cheminformatics Methodologies and Infrastructure to Support the Data Mining of Large, Heterogeneous Chemical Datasets. *Curr. Comput.-Aided Drug Des.* **2010**, *6*, 50–67.

[102] Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

[103] Mager, P. P. A random number experiment to simulate resample model evaluations. *J. Chemom.* **1996**, *10*, 221–240.

[104] Doweyko, A. QSAR: dead or alive? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81–89.

[105] Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241–266.

[106] van Drie, J. Computer-aided drug design: the next 20 years. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 591–601.

[107] Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

[108] Scior, T.; Medina-Franco, J. L.; Do, Q.-T.; Martinez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297–4313.

[109] Gonzalez, M. P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable Selection Methods in QSAR: An Overview. *Curr. Top. Med. Chem.* **2008**, *8*, 1606–1627.

[110] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.

[111] Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning1. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–1026.

[112] Eriksson, L.; Andersson, P.; Johansson, E.; Tysklind, M. Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Mol. Diversity* **2006**, *10*, 169–186.

[113] Yap, C. W.; Xue, Y.; Chen, Y. Z. Application of Support Vector Machines to In Silico Prediction of Cytochrome P450 Enzyme Substrates and Inhibitors. *Curr. Top. Med. Chem.* **2006**, *6*, 1593–1607.

[114] Crammer, K.; Singer, Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research* **2001**, *2*, 265–292.

[115] Azencott, C.-A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J. Chem. Inf. Model.* **2007**, *47*, 965–974.

[116] Drucker, H.; Burges, C.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Advances in neural information processing systems 9: Proceedings of the 1996 conference*, Denver, Co., 1997; pp 155–161.

[117] Bayes, T. An essay towards solving a Problem in the Doctrine of Chances. *Philos. Trans. R. Soc. London* **1763**, *53*, 370–418.

[118] Armstrong, N.; Hibbert, D. B. An introduction to Bayesian methods for analyzing chemistry data: Part 1: An introduction to Bayesian theory and methods. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 194–210.

[119] Hibbert, D. B.; Armstrong, N. An introduction to Bayesian methods for analyzing chemistry data: Part II: A review of applications of Bayesian methods in chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 211–220.

[120] Næs, T.; Indahl, U. A unified description of classical classification methods for multicollinear data. *J. Chemom.* **1998**, *12*, 205–220.

[121] Mello, K. L.; Brown, S. D. Novel 'hybrid' classification method employing Bayesian networks. *J. Chemom.* **1999**, *13*, 579–590.

[122] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

[123] Klon, A. E.; Diller, D. J. Library Fingerprints: A Novel Approach to the Screening of Virtual Libraries. *J. Chem. Inf. Model.* **2007**, *47*, 1354–1365.

[124] Watson, P. Naive Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166–178.

[125] Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325.

[126] Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295–1300.

[127] Willett, P.; Wilton, D.; Hartzoulakis, B.; Tang, R.; Ford, J.; Madge, D. Prediction of Ion Channel Activity Using Binary Kernel Discrimination. *J. Chem. Inf. Model.* **2007**, *47*, 1961–1966.

[128] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3 – 25.

[129] Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68.

[130] Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86–96.

[131] Xu, J.; Stevenson, J. Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.

[132] Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

[133] Oprea, T.; Allu, T.; Fara, D.; Rad, R.; Ostopovici, L.; Bologa, C. Lead-like, drug-like or "Pub-like": how different are they? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113–119.

[134] Walters, W.; Stahl, M. T.; Murcko, M. A. Virtual screening–an overview. *Drug Discovery Today* **1998**, *3*, 160–178.

[135] Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.

[136] McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.

[137] Huynh, L.; Masereeuw, R.; Friedberg, T.; Ingelman-Sundberg, M.; Manivet, P. In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part I: beyond the reduction of animal model use. *Drug Discovery Today* **2009**, *14*, 401 – 405.

[138] Jacob, A.; Pratuangdejkul, J.; Buffet, S.; Launay, J.-M.; Manivet, P. In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part II: the body in a Hilbertian space. *Drug Discovery Today* **2009**, *14*, 406 – 412.

[139] Kharkar, P. S. Two-Dimensional (2D) In Silico Models for Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME/T) in Drug Discovery. *Curr. Top. Med. Chem.* **2010**, *10*, 116–126.

[140] Jr., L. G. V. In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.* **2009**, *241*, 356 – 370.

[141] Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood–brain distribution. *Adv. Drug Delivery Rev.* **2002**, *54*, 291 – 313.

[142] Mensch, J.; Oyarzabal, J.; Mackie, C.; Augustijns, P. In vivo, in vitro and in silico methods for small molecule transfer across the BBB. *J. Pharm. Sci.* **2009**, *98*, 4429–4468.

[143] Mehdipour, A. R.; Hamidi, M. Brain drug targeting: a computational approach for overcoming blood–brain barrier. *Drug Discovery Today* **2009**, *14*, 1030 – 1036.

[144] Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

[145] Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876 – 877.

[146] Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255 – 263.

[147] Oprea, T. I. Current trends in lead discovery: Are we looking for the appropriate properties? *Mol. Diversity* **2000**, *5*, 199–208.

[148] Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235 – 249.

[149] Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743–3748.

[150] Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: National Institutes of Health, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, MD 20894, USA, 2008; Vol. 4, Chapter 12, pp 217 – 241, http://pubchem.ncbi.nlm.nih.gov/.

[151] Seifert, M. H. J. Assessing the Discriminatory Power of Scoring Functions for Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456–1465.

[152] Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

[153] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424.

[154] Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

[155] Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand−Protein Interactions Using OWFEG Free Energy Grids. *J. Med. Chem.* **2001**, *44*, 502–511.

[156] Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

[157] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **2006**, *27*, 861–874.

[158] Clark, R. D.; Shepphird, J. K.; Holliday, J. The effect of structural redundancy in validation sets on virtual screening performance. *J. Chemom.* **2009**, *23*, 471–478.

[159] Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395–1406.

[160] Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010–1024.

[161] Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

[162] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

[163] Nisius, B.; Vogt, M.; Bajorath, J. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback-Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.

[164] Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903 – 911.

[165] Muegge, I. Synergies of Virtual Screening Approaches. *Mini-Rev. Med. Chem.* **2008**, *8*, 927–933.

[166] Petsko, G. When failure should be the option. *BMC Biology* **2010**, *8*, 61.

[167] Keiser, M. J. et al. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.

[168] Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inf.* **2010**, *29*, 176–187.

[169] Palmeira, A.; Rodrigues, F.; Sousa, E.; Pinto, M.; Vasconcelos, M. H.; Fernandes, M. X. New Uses for Old Drugs: Pharmacophore-Based Screening for the Discovery of P-Glycoprotein Inhibitors. *Chem. Biol. Drug Des.* **2011**, *78*, 57–72.

[170] Aronov, A. M. Predictive in silico modeling for hERG channel blockers. *Drug Discovery Today* **2005**, *10*, 149–155.

[171] Schuster, D.; Laggner, C.; Steindl, T. M.; Langer, T. Development and Validation of an In Silico P450 Profiler Based on Pharmacophore Models. *Curr. Drug Discovery Technol.* **2006**, *3*, 1–48.

[172] Schuster, D.; Langer, T. The Identification of Ligand Features Essential for PXR Activation by Pharmacophore Modeling†. *J. Chem. Inf. Model.* **2005**, *45*, 431–439.

[173] Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L.-B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Three-Dimensional Quantitative Structure-Activity Relationships of Inhibitors of P-Glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964–973.

[174] Ekins, S.; Johnston, J. S.; Bahadduri, P.; D'Souza, V. M.; Ray, A.; Chang, C.; Swaan, P. W. *In vitro* and Pharmacophore-Based Discovery of Novel hPEPT1 Inhibitors. *Pharm. Res.* **2005**, *22*, 512–517.

[175] Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing Drug Discovery Through In Silico Screening: Strategies to Increase True Positives Retrieval Rates. *Curr. Med. Chem.* **2008**, *15*, 2040–2053.

[176] Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127 – 136.

[177] Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

[178] Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discovery Today* **2005**, *10*, 139–147.

[179] Hartman, J. L.; Garvik, B.; Hartwell, L. Principles for the Buffering of Genetic Variation. *Science* **2001**, *291*, 1001–1004.

[180] Barabási, A.-L.; Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113.

[181] Kitano, H. Towards a theory of biological robustness. *Mol. Syst. Biol.* **2007**, *3*, 137.

[182] Kaelin, W. G. The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nat. Rev. Cancer* **2005**, *5*, 689–698.

[183] Gu, Z.; Steinmetz, L. M.; Gu, X.; Scharfe, C.; Davis, R. W.; Li, W.-H. Role of duplicate genes in genetic robustness against null mutations. *Nature* **2003**, *421*, 63–66.

[184] Tong, A. H. Y. et al. Global Mapping of the Yeast Genetic Interaction Network. *Science* **2004**, *303*, 808–813.

[185] Cowen, L. E.; Lindquist, S. Hsp90 Potentiates the Rapid Evolution of New Traits: Drug Resistance in Diverse Fungi. *Science* **2005**, *309*, 2185–2189.

[186] Kung, C.; Kenski, D. M.; Dickerson, S. H.; Howson, R. W.; Kuyper, L. F.; Madhani, H. D.; Shokat, K. M. Chemical genomic profiling to identify intracellular targets of a multiplex kinase inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 3587–3592.

[187] Gupta, G. P.; Nguyen, D. X.; Chiang, A. C.; Bos, P. D.; Kim, J. Y.; Nadal, C.; Gomis, R. R.; Manova-Todorova, K.; Massague, J. Mediators of vascular remodelling co-opted for sequential steps in lung metastasis. *Nature* **2007**, *446*, 765–770.

[188] Eltarhouny, S. A.; Elsawy, W. H.; Radpour, R.; Hahn, S.; Holzgreve, W.; Zhong, X. Y. Genes controlling spread of breast cancer to lung "gang of 4". *Exp. Oncol.* **2008**, *30*, 91–95.

[189] Lange, R. P.; Locher, H. H.; Wyss, P. C.; Then, R. L. The Targets of Currently Used Antibacterial Agents: Lessons for Drug Discovery. *Curr. Pharm. Des.* **2007**, *13*, 3140–3154.

[190] Janoir, C.; Zeller, V.; Kitzis, M. D.; Moreau, N. J.; Gutmann, L. High-level fluoroquinolone resistance in Streptococcus pneumoniae requires mutations in parC and gyrA. *Antimicrob. Agents Chemother.* **1996**, *40*, 2760–2764.

[191] Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–359.

[192] Csermely, P.; Ágoston, V.; Pongor, S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* **2005**, *26*, 178–182.

[193] Chen, Y. et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* **2008**, *452*, 429–435.

[194] Morphy, R.; Kay, C.; Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discovery Today* **2004**, *9*, 641–651.

[195] Ma, X.; Shi, Z.; Tan, C.; Jiang, Y.; Go, M.; Low, B.; Chen, Y. *In-Silico* Approaches to Multi-target Drug Discovery. *Pharm. Res.* **2010**, *27*, 739–749.

[196] Huang, W.; Tang, L.; Shi, Y.; Huang, S.; Xu, L.; Sheng, R.; Wu, P.; Li, J.; Zhou, N.; Hu, Y. Searching for the Multi-Target-Directed Ligands against Alzheimer's disease: Discovery of quinoxaline-based hybrid compounds with AChE, H3R and BACE 1 inhibitory activities. *Bioorg. Med. Chem.* **2011**, *19*, 7158–7167.

[197] Li, Y.; Tan, C.; Gao, C.; Zhang, C.; Luan, X.; Chen, X.; Liu, H.; Chen, Y.; Jiang, Y. Discovery of benzimidazole derivatives as novel multi-target EGFR, VEGFR-2 and PDGFR kinase inhibitors. *Bioorg. Med. Chem.* **2011**, *19*, 4529–4535.

[198] Luan, X.; Gao, C.; Zhang, N.; Chen, Y.; Sun, Q.; Tan, C.; Liu, H.; Jin, Y.; Jiang, Y. Exploration of acridine scaffold as a potentially interesting scaffold for discovering novel multi-target VEGFR-2 and Src kinase inhibitors. *Bioorg. Med. Chem.* **2011**, *19*, 3312–3319.

[199] Li, Q.; Xudong, L.; Canghai, L.; Lirong, C.; Jun, S.; Yalin, T.; Xiaojie, X. A Network-Based Multi-Target Computational Estimation Scheme for Anticoagulant Activities of Compounds. *PLoS ONE* **2011**, *6*, 14774.

[200] Prado-Prado, F. J.; Uriarte, E.; Borges, F.; González-Díaz, H. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. *Eur. J. Med. Chem.* **2009**, *44*, 4516–4521.

[201] Prado-Prado, F. J.; Vega, O. M. d. l.; Uriarte, E.; Ubeira, F. M.; Chou, K.-C.; González-Díaz, H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug–drug complex networks. *Bioorg. Med. Chem.* **2009**, *17*, 569–575.

[202] Sheridan, R.; McGaughey, G.; Cornell, W. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 257–265.

[203] Jenwitheesuk, E.; Horst, J. A.; Rivas, K. L.; Voorhis, W. C. V.; Samudrala, R. Novel paradigms for drug discovery: computational multitarget screening. *Trends Pharmacol. Sci.* **2008**, *29*, 62–71.

[204] Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macromodel-an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440–467.

[205] Stewart, J. J. P. *MOPAC Program Package*, 1989.

[206] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.

[207] *CODESSA PRO User's Manual*; University of Florida, 2005.

[208] Charifson, P.; Walters, W. Filtering databases and chemical libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 311–323.

[209] Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.

[210] Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.

[211] García-Sosa, A. T.; Hetényi, C.; Maran, U. Drug efficiency indices for improvement of molecular docking scoring functions. *J. Comput. Chem.* **2010**, *31*, 174–184.

[212] Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. Active Human Immunodeficiency Virus Protease is Required for Viral Infectivity. *Proc. Natl. Acad. Sci.* **1988**, *85*, 4686–4690.

[213] Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **1994**, *263*, 380–384.

[214] Preston, B. D.; Poiesz, B. J.; Loeb, L. A. Fidelity of HIV-1 reverse transcriptase. *Science* **1988**, *242*, 1168–1171.

[215] Nichols, S. E.; Domaoal, R. A.; Thakur, V. V.; Tirado-Rives, J.; Anderson, K. S.; Jorgensen, W. L. Discovery of Wild-Type and Y181C Mutant Nonnucleoside HIV-1 Reverse Transcriptase Inhibitors Using Virtual Screening with Multiple Protein Structures. *J. Chem. Inf. Model.* **2009**, *49*, 1272–1279.

[216] García-Sosa, A. T.; Sild, S.; Maran, U. Design of Multi-Binding-Site Inhibitors, Ligand Efficiency, and Consensus Screening of Avian Influenza H5N1 Wild-Type Neuraminidase and of the Oseltamivir-Resistant H274Y Variant. *J. Chem. Inf. Model.* **2008**, *48*, 2074–2080.

Summary in Estonian

Sobiva omaduste profiiliga ühendite tuvastamine keemiliste struktuuride andmekogudest

Keemiliste ühendite digitaalsete andmebaaside kasutuselevõtuga kaasneb vajadus leida neist arvutuslikke vahendeid kasutades sobivate omadustega molekule. Probleem on eriti huvipakkuv ravimitööstuses, kus aja- ja ressursimahukate katsete asendamine arvutustega, võimaldab märkimisväärset säästu. Ravimitele esitatavate rangete ohutusnõuete tõttu ei ole lähemas tulevikus kindlasti võimalik kogu ravimidisaini protsessi algusest lõpuni arvutitesse ümber kolida. Isegi täpseimad kasutada olevad meetodid ei suuda järjekindla usaldusväärsusega ennustada kuidas käituks mingi aine ravimina, kas ta on sobival määral efektiivne, kas tal on mürgiseid või muidu põnevaid kõrvaltoimeid, ning palju muud potentsiaalsele patsiendile huvipakkuvat informatsiooni. Kuid lugu on teine, kui vaadelda suuri andmekogusid. Arvutusmeetod, mis töötab teadaoleva statistilise vea piires, visates välja mõne sobiva ühendi ja lugedes mõni ekslikult aktiivseks, tihendab lõppkokkuvõttes andmekomplekti tuntaval määral huvitavate ühendite suhtes. Seetõttu on ravimiarenduse lihtsamate ja vähenõudlikkumade etappide puhul, nagu juhtühendite või ravimikandidaatide leidmine, edukalt võimalik rakendada arvutuslikke vahendeid.

Selline tegevus on tuntud virtuaalsõelumisena ning seda ongi käesolevas doktoritöös uuritud. Ülevaate temaatikast võib leida töö sissejuhatavast osast peatükist 2, kokkuvõtte töö aluseks olevates artiklites kasutatud metoodikast ning tulemustest vastavalt peatükkidest 3 ja 4. Töö põhineb neljal artiklil, millest kahes esimeses kasutatakse ligandipõhist lähenemist s.t. kasutades ainult teadmisi senituntud ligandidest, püütakse leida ühendeid, mis on neile struktuurilt sarnased ning võiks eeldatavasti olla seda ka aktiivsuselt. Mõlemal juhtumil koostatakse teadaolevate aktiivsustega ühendite andmekomplekti põhjal lineaarne regressioonimudel, mis võimaldab ennustada uute samalaadsete ühendite aktiivsusi. Artiklis I on uurimisobjektiks HIV-1 proteaasi tsüklilisel uureal põhinevad inhibiitorid. Tulemusena leitakse lihtne ning kiirelt arvutatavate parameetritega mudel, mis on vägagi sobiv suuremahuliseks sõelumiseks. Molekulaardsekriptorid mudelis kirjeldavad molekuli suurust ja kuju, elektrostaatilisi omadusi, ning vesiniksideme andmise võimet. Artiklis II analüüsitakse antrakinoonide tsütotoksilisust. Keerulisema andmestiku tõttu jagatakse ühendid eelnevalt struktuurselt erinevatesse klassidesse, seejärel sarnaselt esimese artikliga lineaarset regressiooni kasutades leitakse mudelid kirjeldamaks erinevatesse klassidesse kuuluvate ühendite tsütotoksilisust. Mudelites leiavad kajastust erinevad suurust, hüdrofoobsust ja laengujaotust kirjeldavad deskriptorid. Nii uureate kui antrakinoonide puhul on deskriptorite valik bioloogilise mehhanismi kontekstis põhjendatud, ning mudelid on sisemiselt, võimaluse korral ka väliselt, valideeritud.

Artiklis III kasutatakse retseptoripõhist lähenemist. Sel puhul on vajalik teada retseptori, kuhu ligand toimet avaldab, kolmemõõtmelist struktuuri. Antud töös on selleks HIV-1 pöördtranskriptaas, ning tulemusena leitakse ligandid, mis hinnan-

guliselt on aktiivsed nii loodusliku kui ka muteerunud pöördtranskriptaasi vastu, kuid samas ei toimi mõningate teiste oluliste organismis leiduvate ensüümide vastu. Töö tulemuste hulgas on ka eeltöödeldud andmekomplekt, mis sisaldab suurema tõenäousega juhtühendiks või ravimikandidaadiks sobivaid molekule kui juhuslik valik, ning mis on end juba mitmes töös heast küljest näidanud.

Viimases artiklis kirjeldatakse uut retseptoripõhist sõelumismeetodit, mis sarnaneb veidi dokkimisele kuna piltlikult sobitab ligandi retseptori aktiivtsentrisse, kuid erinevalt dokkimisest ei vaja ühendite kolmemõõtmelisi geomeetriaid, vaid piirdub molekuli graafiga. Kahemõõtmelise molekuli kujutusviisiga piirdumine ei kaota ühendi kohta kuigi palju infot, kuna ravimitena huvipakkuvad ühendid on piiratud suurusega ning suhteliselt jäigad, kuid annab eeliseks suurema töökiiruse. Kõrvaltulemuseks on meetodi valideerimise käigus leitud potentsiaalselt aktiivsed ligandid HIV-1 proteaasile.

Virtuaalsõelumine on avar ning praegusel hetkel hoogsalt arenev uurimisteema, mille lai haare teeb selle käsitlemise ühe töö mahus keeruliseks. Käesolevas doktoritöös on tehtud valik mõningate suundade kasuks, ning uuritud nende võimekust ja tulemuslikkust erinevate projektide raames. Töö tulemusena on valminud mõned kasutusvalmis mudelid, eeltöödeldud andmekomplekt – mugav lähtepositsioon edasisteks töödeks; ning omandatud teadmistepagas virtuaalsõelumise vallas on leidnud väljundi uue meetodi väljatöötamises.

Acknowledgements

I would like to thank my supervisor Dr. Sulev Sild, who has always provided valuable advice and counseling during my research and studies. I also thank my colleagues and all the co-authors of my publications for their collaboration and also for their contribution to my education.

My special gratitude goes to my family for their support during my studies, Kai and Hedi for linguistic consultations, and Geven and Maikki for creating excellent office atmosphere.

This work was supported by the Estonian Science Foundation (grants 7153 and 7709), Ministry of Science and Education (SF0140031Bs09), EU 6FP program CardioWorkBench – "Drug Design for Cardiovascular Diseases: Integration of in Silico and in Vitro Analysis" (LSHB-CT-2005-018671) and graduate schools UTTP and „Functional materials and technologies" (European Social Fund project 1.2.0401.09-0079)."

Publications

Curriculum Vitae

KALEV TAKKIS

Personal particulars

Born:	11.01.1981, Tartu, Estonia
Citizenship:	Estonian
Address:	Riia 65-3, Tartu, Estonia
Phone:	+372 555 89 226
e-mail:	kalev.takkis@ut.ee

Education

2007–	Ph.D. student of Molecular Engineering, University of Tartu
2004–2007	M.Sc. in Molecular Engineering, University of Tartu
1999–2004	B.Sc. in Chemistry, University of Tartu

Work experience

2012–	analyst, State Agency of Medicines
2006–2009	chemist, University of Tartu, Institute of Chemistry

Elulookirjeldus

KALEV TAKKIS

Delikaatsed isikuandmed

Sünniaeg:	11.01.1981, Tartu, Eesti
Kodakondsus:	Eesti
Aadress:	Riia 65-3, Tartu, Eesti
Tel:	+372 555 89 226
e-mail:	kalev.takkis@ut.ee

Haridus

2007–	molekulaartehnoloogia doktorant, Tartu Ülikool
2004–2007	M.Sc. molekulaartehnoloogias, Tartu Ülikool
1999–2004	B.Sc. keemias, Tartu Ülikool

Erialane töökogemus

2012–	analüütik, Ravimiamet
2006–2009	keemik, Tartu Ülikool, Keemia Instituut

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS

1. **Toomas Tamm.** Quantum-chemical simulation of solvent effects. Tartu, 1993, 110 p.
2. **Peeter Burk.** Theoretical study of gas-phase acid-base equilibria. Tartu, 1994, 96 p.
3. **Victor Lobanov.** Quantitative structure-property relationships in large descriptor spaces. Tartu, 1995, 135 p.
4. **Vahur Mäemets.** The ^{17}O and ^{1}H nuclear magnetic resonance study of H_2O in individual solvents and its charged clusters in aqueous solutions of electrolytes. Tartu, 1997, 140 p.
5. **Andrus Metsala.** Microcanonical rate constant in nonequilibrium distribution of vibrational energy and in restricted intramolecular vibrational energy redistribution on the basis of slater's theory of unimolecular reactions. Tartu, 1997, 150 p.
6. **Uko Maran.** Quantum-mechanical study of potential energy surfaces in different environments. Tartu, 1997, 137 p.
7. **Alar Jänes.** Adsorption of organic compounds on antimony, bismuth and cadmium electrodes. Tartu, 1998, 219 p.
8. **Kaido Tammeveski.** Oxygen electroreduction on thin platinum films and the electrochemical detection of superoxide anion. Tartu, 1998, 139 p.
9. **Ivo Leito.** Studies of Brønsted acid-base equilibria in water and non-aqueous media. Tartu, 1998, 101 p.
10. **Jaan Leis.** Conformational dynamics and equilibria in amides. Tartu, 1998, 131 p.
11. **Toonika Rinken.** The modelling of amperometric biosensors based on oxidoreductases. Tartu, 2000, 108 p.
12. **Dmitri Panov.** Partially solvated Grignard reagents. Tartu, 2000, 64 p.
13. **Kaja Orupõld.** Treatment and analysis of phenolic wastewater with microorganisms. Tartu, 2000, 123 p.
14. **Jüri Ivask.** Ion Chromatographic determination of major anions and cations in polar ice core. Tartu, 2000, 85 p.
15. **Lauri Vares.** Stereoselective Synthesis of Tetrahydrofuran and Tetrahydropyran Derivatives by Use of Asymmetric Horner-Wadsworth-Emmons and Ring Closure Reactions. Tartu, 2000, 184 p.
16. **Martin Lepiku.** Kinetic aspects of dopamine D_2 receptor interactions with specific ligands. Tartu, 2000, 81 p.
17. **Katrin Sak.** Some aspects of ligand specificity of P2Y receptors. Tartu, 2000, 106 p.
18. **Vello Pällin.** The role of solvation in the formation of iotsitch complexes. Tartu, 2001, 95 p.

19. **Katrin Kollist.** Interactions between polycyclic aromatic compounds and humic substances. Tartu, 2001, 93 p.
20. **Ivar Koppel.** Quantum chemical study of acidity of strong and superstrong Brønsted acids. Tartu, 2001, 104 p.
21. **Viljar Pihl.** The study of the substituent and solvent effects on the acidity of OH and CH acids. Tartu, 2001, 132 p.
22. **Natalia Palm.** Specification of the minimum, sufficient and significant set of descriptors for general description of solvent effects. Tartu, 2001, 134 p.
23. **Sulev Sild.** QSPR/QSAR approaches for complex molecular systems. Tartu, 2001, 134 p.
24. **Ruslan Petrukhin.** Industrial applications of the quantitative structure-property relationships. Tartu, 2001, 162 p.
25. **Boris V. Rogovoy.** Synthesis of (benzotriazolyl)carboximidamides and their application in relations with *N*- and *S*-nucleophyles. Tartu, 2002, 84 p.
26. **Koit Herodes.** Solvent effects on UV-vis absorption spectra of some solvatochromic substances in binary solvent mixtures: the preferential solvation model. Tartu, 2002, 102 p.
27. **Anti Perkson.** Synthesis and characterisation of nanostructured carbon. Tartu, 2002, 152 p.
28. **Ivari Kaljurand.** Self-consistent acidity scales of neutral and cationic Brønsted acids in acetonitrile and tetrahydrofuran. Tartu, 2003, 108 p.
29. **Karmen Lust.** Adsorption of anions on bismuth single crystal electrodes. Tartu, 2003, 128 p.
30. **Mare Piirsalu.** Substituent, temperature and solvent effects on the alkaline hydrolysis of substituted phenyl and alkyl esters of benzoic acid. Tartu, 2003, 156 p.
31. **Meeri Sassian.** Reactions of partially solvated Grignard reagents. Tartu, 2003, 78 p.
32. **Tarmo Tamm.** Quantum chemical modelling of polypyrrole. Tartu, 2003. 100 p.
33. **Erik Teinemaa.** The environmental fate of the particulate matter and organic pollutants from an oil shale power plant. Tartu, 2003. 102 p.
34. **Jaana Tammiku-Taul.** Quantum chemical study of the properties of Grignard reagents. Tartu, 2003. 120 p.
35. **Andre Lomaka.** Biomedical applications of predictive computational chemistry. Tartu, 2003. 132 p.
36. **Kostyantyn Kirichenko.** Benzotriazole – Mediated Carbon–Carbon Bond Formation. Tartu, 2003. 132 p.
37. **Gunnar Nurk.** Adsorption kinetics of some organic compounds on bismuth single crystal electrodes. Tartu, 2003, 170 p.
38. **Mati Arulepp.** Electrochemical characteristics of porous carbon materials and electrical double layer capacitors. Tartu, 2003, 196 p.

39. **Dan Cornel Fara.** QSPR modeling of complexation and distribution of organic compounds. Tartu, 2004, 126 p.
40. **Riina Mahlapuu.** Signalling of galanin and amyloid precursor protein through adenylate cyclase. Tartu, 2004, 124 p.
41. **Mihkel Kerikmäe.** Some luminescent materials for dosimetric applications and physical research. Tartu, 2004, 143 p.
42. **Jaanus Kruusma.** Determination of some important trace metal ions in human blood. Tartu, 2004, 115 p.
43. **Urmas Johanson.** Investigations of the electrochemical properties of polypyrrole modified electrodes. Tartu, 2004, 91 p.
44. **Kaido Sillar.** Computational study of the acid sites in zeolite ZSM-5. Tartu, 2004, 80 p.
45. **Aldo Oras.** Kinetic aspects of dATPαS interaction with $P2Y_1$ receptor. Tartu, 2004, 75 p.
46. **Erik Mölder.** Measurement of the oxygen mass transfer through the air-water interface. Tartu, 2005, 73 p.
47. **Thomas Thomberg.** The kinetics of electroreduction of peroxodisulfate anion on cadmium (0001) single crystal electrode. Tartu, 2005, 95 p.
48. **Olavi Loog.** Aspects of condensations of carbonyl compounds and their imine analogues. Tartu, 2005, 83 p.
49. **Siim Salmar.** Effect of ultrasound on ester hydrolysis in aqueous ethanol. Tartu, 2006, 73 p.
50. **Ain Uustare.** Modulation of signal transduction of heptahelical receptors by other receptors and G proteins. Tartu, 2006, 121 p.
51. **Sergei Yurchenko.** Determination of some carcinogenic contaminants in food. Tartu, 2006, 143 p.
52. **Kaido Tämm.** QSPR modeling of some properties of organic compounds. Tartu, 2006, 67 p.
53. **Olga Tšubrik.** New methods in the synthesis of multisubstituted hydrazines. Tartu. 2006, 183 p.
54. **Lilli Sooväli.** Spectrophotometric measurements and their uncertainty in chemical analysis and dissociation constant measurements. Tartu, 2006, 125 p.
55. **Eve Koort.** Uncertainty estimation of potentiometrically measured ph and pK_a values. Tartu, 2006, 139 p.
56. **Sergei Kopanchuk.** Regulation of ligand binding to melanocortin receptor subtypes. Tartu, 2006, 119 p.
57. **Silvar Kallip.** Surface structure of some bismuth and antimony single crystal electrodes. Tartu, 2006, 107 p.
58. **Kristjan Saal.** Surface silanization and its application in biomolecule coupling. Tartu, 2006, 77 p.
59. **Tanel Tätte.** High viscosity $Sn(OBu)_4$ oligomeric concentrates and their applications in technology. Tartu, 2006, 91 p.

60. **Dimitar Atanasov Dobchev**. Robust QSAR methods for the prediction of properties from molecular structure. Tartu, 2006, 118 p.
61. **Hannes Hagu**. Impact of ultrasound on hydrophobic interactions in solutions. Tartu, 2007, 81 p.
62. **Rutha Jäger.** Electroreduction of peroxodisulfate anion on bismuth electrodes. Tartu, 2007, 142 p.
63. **Kaido Viht.** Immobilizable bisubstrate-analogue inhibitors of basophilic protein kinases: development and application in biosensors. Tartu, 2007, 88 p.
64. **Eva-Ingrid Rõõm.** Acid-base equilibria in nonpolar media. Tartu, 2007, 156 p.
65. **Sven Tamp.** DFT study of the cesium cation containing complexes relevant to the cesium cation binding by the humic acids. Tartu, 2007, 102 p.
66. **Jaak Nerut.** Electroreduction of hexacyanoferrate(III) anion on Cadmium (0001) single crystal electrode. Tartu, 2007, 180 p.
67. **Lauri Jalukse.** Measurement uncertainty estimation in amperometric dissolved oxygen concentration measurement. Tartu, 2007, 112 p.
68. **Aime Lust.** Charge state of dopants and ordered clusters formation in CaF_2:Mn and CaF_2:Eu luminophors. Tartu, 2007, 100 p.
69. **Iiris Kahn**. Quantitative Structure-Activity Relationships of environmentally relevant properties. Tartu, 2007, 98 p.
70. **Mari Reinik.** Nitrates, nitrites, N-nitrosamines and polycyclic aromatic hydrocarbons in food: analytical methods, occurrence and dietary intake. Tartu, 2007, 172 p.
71. **Heili Kasuk.** Thermodynamic parameters and adsorption kinetics of organic compounds forming the compact adsorption layer at Bi single crystal electrodes. Tartu, 2007, 212 p.
72. **Erki Enkvist.** Synthesis of adenosine-peptide conjugates for biological applications. Tartu, 2007, 114 p.
73. **Svetoslav Hristov Slavov**. Biomedical applications of the QSAR approach. Tartu, 2007, 146 p.
74. **Eneli Härk.** Electroreduction of complex cations on electrochemically polished Bi(*hkl*) single crystal electrodes. Tartu, 2008, 158 p.
75. **Priit Möller.** Electrochemical characteristics of some cathodes for medium temperature solid oxide fuel cells, synthesized by solid state reaction technique. Tartu, 2008, 90 p.
76. **Signe Viggor.** Impact of biochemical parameters of genetically different pseudomonads at the degradation of phenolic compounds. Tartu, 2008, 122 p.
77. **Ave Sarapuu.** Electrochemical reduction of oxygen on quinone-modified carbon electrodes and on thin films of platinum and gold. Tartu, 2008, 134 p.
78. **Agnes Kütt.** Studies of acid-base equilibria in non-aqueous media. Tartu, 2008, 198 p.

79. **Rouvim Kadis.** Evaluation of measurement uncertainty in analytical chemistry: related concepts and some points of misinterpretation. Tartu, 2008, 118 p.
80. **Valter Reedo.** Elaboration of IVB group metal oxide structures and their possible applications. Tartu, 2008, 98 p.
81. **Aleksei Kuznetsov.** Allosteric effects in reactions catalyzed by the cAMP-dependent protein kinase catalytic subunit. Tartu, 2009, 133 p.
82. **Aleksei Bredihhin.** Use of mono- and polyanions in the synthesis of multisubstituted hydrazine derivatives. Tartu, 2009, 105 p.
83. **Anu Ploom.** Quantitative structure-reactivity analysis in organosilicon chemistry. Tartu, 2009, 99 p.
84. **Argo Vonk.** Determination of adenosine A_{2A}- and dopamine D_1 receptor-specific modulation of adenylate cyclase activity in rat striatum. Tartu, 2009, 129 p.
85. **Indrek Kivi.** Synthesis and electrochemical characterization of porous cathode materials for intermediate temperature solid oxide fuel cells. Tartu, 2009, 177 p.
86. **Jaanus Eskusson.** Synthesis and characterisation of diamond-like carbon thin films prepared by pulsed laser deposition method. Tartu, 2009, 117 p.
87. **Marko Lätt.** Carbide derived microporous carbon and electrical double layer capacitors. Tartu, 2009, 107 p.
88. **Vladimir Stepanov.** Slow conformational changes in dopamine transporter interaction with its ligands. Tartu, 2009, 103 p.
89. **Aleksander Trummal.** Computational Study of Structural and Solvent Effects on Acidities of Some Brønsted Acids. Tartu, 2009, 103 p.
90. **Eerold Vellemäe.** Applications of mischmetal in organic synthesis. Tartu, 2009, 93 p.
91. **Sven Parkel.** Ligand binding to 5-HT_{1A} receptors and its regulation by Mg^{2+} and Mn^{2+}. Tartu, 2010, 99 p.
92. **Signe Vahur.** Expanding the possibilities of ATR-FT-IR spectroscopy in determination of inorganic pigments. Tartu, 2010, 184 p.
93. **Tavo Romann**. Preparation and surface modification of bismuth thin film, porous, and microelectrodes. Tartu, 2010, 155 p.
94. **Nadežda Aleksejeva.** Electrocatalytic reduction of oxygen on carbon nanotube-based nanocomposite materials. Tartu, 2010, 147 p.
95. **Marko Kullapere.** Electrochemical properties of glassy carbon, nickel and gold electrodes modified with aryl groups. Tartu, 2010, 233 p.
96. **Liis Siinor.** Adsorption kinetics of ions at Bi single crystal planes from aqueous electrolyte solutions and room-temperature ionic liquids. Tartu, 2010, 101 p.
97. **Angela Vaasa.** Development of fluorescence-based kinetic and binding assays for characterization of protein kinases and their inhibitors. Tartu 2010, 101 p.

98. **Indrek Tulp.** Multivariate analysis of chemical and biological properties. Tartu 2010, 105 p.
99. **Aare Selberg.** Evaluation of environmental quality in Northern Estonia by the analysis of leachate. Tartu 2010, 117 p.
100. **Darja Lavõgina.** Development of protein kinase inhibitors based on adenosine analogue-oligoarginine conjugates. Tartu 2010, 248 p.
101. **Laura Herm.** Biochemistry of dopamine D_2 receptors and its association with motivated behaviour. Tartu 2010, 156 p.
102. **Terje Raudsepp.** Influence of dopant anions on the electrochemical properties of polypyrrole films. Tartu 2010, 112 p.
103. **Margus Marandi.** Electroformation of Polypyrrole Films: *In-situ* AFM and STM Study. Tartu 2011, 116 p.
104. **Kairi Kivirand.** Diamine oxidase-based biosensors: construction and working principles. Tartu, 2011, 140 p.
105. **Anneli Kruve.** Matrix effects in liquid-chromatography electrospray mass-spectrometry. Tartu, 2011, 156 p.
106. **Gary Urb.** Assessment of environmental impact of oil shale fly ash from PF and CFB combustion. Tartu, 2011, 108 p.
107. **Nikita Oskolkov.** A novel strategy for peptide-mediated cellular delivery and induction of endosomal escape. Tartu, 2011, 106 p.
108. **Dana Martin.** The QSPR/QSAR approach for the prediction of properties of fullerene derivatives. Tartu, 2011, 98 p.
109. **Säde Viirlaid.** Novel glutathione analogues and their antioxidant activity. Tartu, 2011, 106 p.
110. **Ülis Sõukand.** Simultaneous adsorption of Cd^{2+}, Ni^{2+}, and Pb^{2+} on peat. Tartu, 2011, 124 p.
111. **Lauri Lipping.** The acidity of strong and superstrong Brønsted acids, an outreach for the "limits of growth": a quantum chemical study. Tartu, 2011, 124 p.
112. **Heisi Kurig.** Electrical double-layer capacitors based on ionic liquids as electrolytes. Tartu, 2011, 146 p.
113. **Marje Kasari.** Bisubstrate luminescent probes, optical sensors and affinity adsorbents for measurement of active protein kinases in biological samples. Tartu, 2012, 126 p.