Structure-Based Virtual Screening

New Methods and Applications in Infectious Diseases

MICAELE JACOBSSON
Dissertation presented at Uppsala University to be publicly examined in B41, BMC, Husargatan 3, Uppsala, Friday, October 31, 2008 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

A drug discovery project typically starts with a pharmacological hypothesis: that the modulation of a specific molecular biological mechanism would be beneficial in the treatment of the targeted disease. In a small-molecule project, the next step is to identify hits, i.e. molecules that can affect this modulation. These hits are subsequently expanded into hit series, which are optimised with respect to pharmacodynamic and pharmacokinetic properties, through medicinal chemistry. Finally, a drug candidate is clinically developed into a new drug. This thesis concerns the use of structure-based virtual screening in the hit identification phase of drug discovery.

Structure-based virtual screening involves using the known 3D structure of a target protein to predict binders, through the process of docking and scoring. Docking is the prediction of potential binding poses, and scoring is the prediction of the free energy of binding from those poses. Two new methodologies, based on post-processing of scoring results, were developed and evaluated using model systems. Both methods significantly increased the enrichment of true positives. Furthermore, correlation was observed between scores and simple molecular properties, and identified as a source of false positives in structure-based virtual screening.

Two target proteins, *Mycobacterium tuberculosis* ribose-5-phosphate isomerase, a potential drug target in tuberculosis, and *Plasmodium falciparum* spermidine synthase, a potential drug target in malaria, were subjected to docking and virtual screening. Docking of substrates and products of ribose-5-phosphate isomerase led to hypotheses on the role of individual residues in the active site. Additionally, virtual screening was used to predict 48 potential inhibitors, but none was confirmed as an inhibitor or binder to the target enzyme. For spermidine synthase, structure-based virtual screening was used to predict 32 potential active-site binders. Seven of these were confirmed to bind in the active site.

Keywords: drug discovery, docking, scoring, virtual screening, malaria, tuberculosis

Michael Jacobsson, Department of Medicinal Chemistry, Box 574, Uppsala University, SE-75123 Uppsala, Sweden

© Michael Jacobsson 2008

ISSN 1651-6192
urn:nbn:se:uu:diva-9302 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-9302)
“...wealth and greatness are mere trinkets of frivolous utility. [...] It is this deception which rouses and keeps in continual motion the industry of mankind.”

Adam Smith
This thesis is based on the following papers, which are referred to in the text by their Roman numerals:


Reprints made with permission from the publishers.
Contents

1 Introduction .................................................................................................................. 11
  1.1 Drug discovery ........................................................................................................... 11
  1.2 Structure-based virtual screening ............................................................................. 12
    1.2.1 Docking .............................................................................................................. 12
    1.2.2 Scoring .............................................................................................................. 14
  1.3 Malaria ....................................................................................................................... 16
  1.4 Tuberculosis ............................................................................................................. 19

2 Aims of the present study ........................................................................................... 22

3 New methodology in structure-based virtual screening ............................................. 23
  3.1 Background .............................................................................................................. 23
  3.2 Supervised learning of scoring data (Paper I) ..................................................... 23
    3.2.1 Generation of data sets ..................................................................................... 25
    3.2.2 Multivariate classifiers ...................................................................................... 26
    3.2.3 Results and conclusions ................................................................................... 28
    3.2.4 Related studies ................................................................................................... 30
  3.3 Transforming scoring data without a training set (Paper II) ............................. 31
    3.3.1 Test sets .......................................................................................................... 31
    3.3.2 Transforms ...................................................................................................... 35
    3.3.3 Results .............................................................................................................. 36
    3.3.4 Conclusions ...................................................................................................... 41

4 Structure-based virtual screening applications ............................................................. 42
  4.1 Background .............................................................................................................. 42
  4.2 Mycobacterium tuberculosis ribose-5-phosphate isomerase (Paper III) ............. 42
    4.2.1 Protein structure ............................................................................................... 43
    4.2.2 Docking studies ............................................................................................... 44
    4.2.3 Structure-based virtual screening ..................................................................... 47
    4.2.4 Experimental testing of predicted inhibitors ..................................................... 50
    4.2.5 Discussion ......................................................................................................... 54
    4.2.6 Subsequent results ............................................................................................ 55
  4.3 Plasmodium falciparum spermidine synthase (Paper IV) ..................................... 56
4.3.1 Structure-based virtual screening ...........................................57
4.3.2 Experimental testing of predicted binders .........................61
4.3.3 Results .................................................................................65
4.4 Discussion ..............................................................................66

5 Concluding remarks .....................................................................67

Acknowledgements ........................................................................68
References .....................................................................................72

Appendix: Experimental procedures for biophysical evaluation of predicted binders to mtRpiB .................................................................85
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn</td>
<td>Asparagine</td>
</tr>
<tr>
<td>4MCHA</td>
<td>4-methylcyclohexylamine</td>
</tr>
<tr>
<td>4PE</td>
<td>4-phospho-D-erythronocarboxylic acid</td>
</tr>
<tr>
<td>4PEH</td>
<td>4-phospho-D-erythronohydroxamic acid</td>
</tr>
<tr>
<td>AChE</td>
<td>Acetylcholine esterase</td>
</tr>
<tr>
<td>ACT</td>
<td>Artemisinin-based combination therapy</td>
</tr>
<tr>
<td>AdoDATO</td>
<td>S-adenosyl-1,8-diamino-3-thiooctane</td>
</tr>
<tr>
<td>AdoMet</td>
<td>S-Adenosylmethionine</td>
</tr>
<tr>
<td>AdoMetDC</td>
<td>S-Adenosylmethionine decarboxylase</td>
</tr>
<tr>
<td>CDK2</td>
<td>Cyclin dependent kinase 2</td>
</tr>
<tr>
<td>CMR</td>
<td>Calculated molecular refractivity</td>
</tr>
<tr>
<td>COX2</td>
<td>Cycloxygenase 2</td>
</tr>
<tr>
<td>Cys</td>
<td>Cysteine</td>
</tr>
<tr>
<td>dcAdoMet</td>
<td>Decarboxylated S-adenosylmethionine</td>
</tr>
<tr>
<td>DHFR</td>
<td>Dihydrofolate reductase</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethylsulphoxide</td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Commission number</td>
</tr>
<tr>
<td>ERα</td>
<td>Oestrogen receptor α</td>
</tr>
<tr>
<td>ESI-MS</td>
<td>Electrospray ionisation mass spectrometry</td>
</tr>
<tr>
<td>E. coli</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>fXa</td>
<td>Factor Xa</td>
</tr>
<tr>
<td>Glu</td>
<td>Glutamic acid</td>
</tr>
<tr>
<td>HAC</td>
<td>Heavy atom count</td>
</tr>
<tr>
<td>His</td>
<td>Histidine</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>HTS</td>
<td>High throughput screening</td>
</tr>
<tr>
<td>MASC</td>
<td>Multiple active site correction</td>
</tr>
<tr>
<td>MMP-3</td>
<td>Matrix metalloprotease 3</td>
</tr>
<tr>
<td>MTA</td>
<td>Methylthioadenosine</td>
</tr>
<tr>
<td>MTB</td>
<td><em>Mycobacterium tuberculosis</em></td>
</tr>
<tr>
<td>mtRpiB</td>
<td><em>Mycobacterium tuberculosis</em> Ribose-5-phosphate isomerase B</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>NA</td>
<td>Neuraminidase</td>
</tr>
<tr>
<td>NADPH</td>
<td>Nicotinamide adenine dinucleotide phosphate</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>ODC</td>
<td>Ornithine decarboxylase</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>OPLS-AA</td>
<td>Optimized potential for liquid simulations – all atom</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>Pf</td>
<td><em>Plasmodium falciparum</em></td>
</tr>
<tr>
<td>PLS</td>
<td>Partial least-squares projections to latent structures</td>
</tr>
<tr>
<td>PLS-DA</td>
<td>PLS discriminant analysis</td>
</tr>
<tr>
<td>PMF</td>
<td>Potentials of mean force</td>
</tr>
<tr>
<td>R5P</td>
<td>Ribose-5-phosphate</td>
</tr>
<tr>
<td>RAPID</td>
<td>Rational approach to pathogen inhibitor discovery</td>
</tr>
<tr>
<td>RpiB</td>
<td>Ribose-5-phosphate isomerase B</td>
</tr>
<tr>
<td>SBVS</td>
<td>Structure-based virtual screening</td>
</tr>
<tr>
<td>SpdSyn</td>
<td>Spermidine synthase</td>
</tr>
<tr>
<td>SPL</td>
<td>Sybyl programming language</td>
</tr>
<tr>
<td>STD</td>
<td>Saturation transfer difference</td>
</tr>
<tr>
<td>TB</td>
<td>Tuberculosis</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Drug discovery

The research and development process often used in the pharmaceutical industry to develop new drugs is illustrated in Figure 1. Early drug discovery efforts are aimed at identifying a candidate drug. This is subsequently developed and evaluated in the clinic, before it is eventually approved by regulatory authorities. Today’s drug discovery process\(^1\) is well illustrated by the discovery of sorafenib\(^2\) and imatinib.\(^3\)

![Diagram of the drug discovery process](image)

**Figure 1.** Schematic illustration of the drug discovery process.

Target selection means choosing a specific target, which is hypothesised to be pivotal in treating the disease targeted by the drug discovery project. The target is often a receptor, an enzyme or a similar substance. Hit identification is the discovery of molecules having the desired effect in one or more primary assays. These assays are intended to be predictive for the ultimately desired pharmacological effect. Hit identification is often carried out using high throughput screening (HTS), where large compound libraries, \(10^5–10^6\) molecules, are tested experimentally. Validated hits from this phase serve as starting points for medicinal chemistry expansion and exploration, in the lead generation and optimisation stage. As this process progresses, more and more information on the pharmacodynamic and pharmacokinetic properties of the various compounds is discovered, and intellectual property is created. Validation of the initial pharmacological hypothesis, posed in the target selection phase, is often a parallel process. Here, various disease models, together with compounds generated during the drug discovery process, are
often used. Ultimately, a single compound, the drug candidate, is chosen for clinical testing.

This thesis concerns the use of molecular modelling in the early drug discovery process, specifically in hit identification. In general, molecular modelling is used in drug discovery to predict molecular properties and experimental results, to guide in planning experiments and making decisions. The use of molecular modelling in hit identification is referred to as virtual screening. This is used both to design sets of compounds to screen in HTS and to predict smaller sets of compounds to test in lower throughput assays.

1.2 Structure-based virtual screening

Structure-based virtual screening (SBVS) is the prediction of binders to target proteins through computational methods, using the known 3D structure of these targets. The basic approach in SBVS is to predict the binding pose of each small molecule in a test library (docking), and from that predict the free energy of binding of that molecule (scoring). The set of hits is then predicted by sorting all compounds in the test library by this score, and deciding on a threshold score. Compounds scoring better than this threshold are regarded as hits, and are evaluated further. This is analogous to experimental HTS, where the percent inhibition obtained from HTS serve the same role as the score in SBVS.

Structure-based virtual screening has evolved over the past decade, and many different variations of the basic methodology have been proposed. Literally hundreds of reviews have been published on virtual screening. One of the more recent ones, from March 2008, covers the current state of the art and many of the practical aspects of SBVS.

1.2.1 Docking

In the context of molecular modelling, docking means predicting the bioactive conformation of a molecule in the binding site of a target structure. In essence, this is equivalent to finding the global free energy minimum of the system consisting of the ligand and the target. Docking is used as a tool in structure-based drug design as well as in SBVS.

The first algorithm developed to dock small molecules into the binding pocket of a macromolecule, the DOCK algorithm, was published in 1982 by Kuntz et al. In a review from 2007, more than sixty published docking
programs and thirty scoring functions were listed. However, the earliest and most widely used docking programs over the past years are probably DOCK, AutoDOCK, GOLD and FlexX, and in recent years also e.g. Glide, ICM, FRED and Surflex.

All docking programs contain two components, a fitness or scoring function, whose global minimum is intended to coincide with the global free energy minimum of the target-ligand system, and a search method, used to sample the search space in which the scoring function is optimised. This search space can be very large, combining all ligand positions and rotations with all possible conformations of the ligand and possibly also the target protein.

In DOCK the ligand was initially treated as a rigid body, as was the protein. An incremental construction algorithm has since been adopted to include ligand flexibility. In this version, the ligand is partitioned into rigid fragments placed incrementally in the active site. The fitness function is the sum of the van der Waals and Coulomb interactions between the ligand and the protein atoms. Ligand positions and orientations are sampled through matching of spheres describing the active site and the ligand, using geometrical methods. The fitness function is evaluated using a pre-calculated grid covering the active site, to reduce the CPU time required to dock each ligand.

AutoDOCK also treats the target as a rigid body, and uses a pre-calculated grid to evaluate the fitness function. This function is again force-field based, but also includes intramolecular interactions of the ligand. The ligand conformations, orientations and positions were initially sampled by simulated annealing, but today genetic algorithms are also used.

GOLD is based on genetic algorithms. Here, candidate conformations and placements of the ligands are generated using site points, representing optimal placement of hydrogen-bond donors and acceptors as well as metal-ion interactions. A sample pose is generated by matching site points from the ligand to site points from the protein. The protein is not represented by a pre-computed grid, and flexibility of terminal hydrogen-bond donors and acceptors in the protein side chains is integral to the algorithm.

FlexX is based on incremental construction of the ligand in the active site. The protein is described using a set of interaction groups of different types, as is each ligand fragment. The scoring function is calculated by matching interaction groups and summing their contributions. The entire docking

* http://www.eyesopen.com/docs/html/fred/
procedure is described as a search tree, where each level of the tree corresponds to one ligand fragment.

The docking calculations described in this thesis have mostly been performed using Glide. In Glide, the protein is again represented by a set of pre-computed grids. The grids are calculated using the non-bonded part of the OPLS-AA force field\textsuperscript{37} and the empirical ChemScore\textsuperscript{38} scoring function. The docking algorithm aims to mimic an exhaustive search of all possible ligand conformations in all possible positions in the docking region. For this purpose, the ligand is represented by a set of conformers. This conformational ensemble is described as a set of core conformations, often less than 500, together with terminal rotamers. The most coarse-grained description of the core conformations consists of a midpoint and a histogram of binned distances along predefined directions, from the midpoint to the ligand surface. The midpoint is placed at candidate positions in the docking region, by evaluating the overlap of the histogram and a similar histogram derived from the distance between the candidate position and the protein surface. For each candidate position, the core conformations are orientated along orientations not giving rise to too many steric clashes, and rotated around each such orientation. The all-atom placements of the ligand are scored using the pre-computed ChemScore grids, and placements with sufficiently good scores are refined through local rigid-body translations and rotations. Each remaining candidate pose is subsequently minimised in the OPLS-AA force field, allowing the rotatable bonds in the ligand to vary. The lowest energy poses are finally subjected to a Monte Carlo search of nearby torsional minima. The resulting poses are ranked according to the EModel scoring function. This function combines the ChemScore-based GlideScore, the OPLS-AA intermolecular interaction energy and the internal energy of the ligand.

1.2.2 Scoring

In the present context, scoring is the computationally efficient prediction of the free energy of binding of a small molecule, using only the binding pose. In SBVS, scoring is also used to predict whether a compound binds at all to a given target structure, by introducing a threshold score. Scoring functions are also used in the docking procedure, as described in the previous section. The scoring functions used to rank candidate poses of the same compound are not necessarily the same scoring functions that are used to predict the free energy of binding and to rank diverse compounds. This is, for example, the case in Glide, where one scoring function, EModel, is used to rank poses in docking, and another, GlideScore, is used to rank compounds in SBVS.
Scoring functions are generally divided into three categories: force-field-based, empirical and knowledge-based. Force-field-based scoring functions are evaluated by calculating the intermolecular interaction energy between the ligand and the target protein. This is done by summing the van der Waals and Coulomb interactions from a force field (e.g. DOCK). The intra-molecular energy of the ligand is often included as well (e.g. AutoDOCK). The scoring function used in GOLD, another force-field-based function, also includes a term for intermolecular hydrogen bonds, as does the AutoDOCK scoring function. Recently, the authors of AutoDOCK introduced an alternative, empirical scoring function for their program.18

Empirical scoring functions are based on fitting linear combinations of terms, assumed to be important for ligand–protein binding, to known affinities of compounds with known binding poses. These terms, which can be called interaction descriptors, may include the number of hydrogen bonds, the number of rotatable bonds of the ligand, etc. The first empirical scoring function intended for SBVS was published by Böhm in 1994, and further developed in 1998. In 1997, the aforementioned ChemScore function was published. Another empirical scoring function package is X-Score, which is one of the scoring function packages used in the study described in Paper II.

Since empirical scoring functions are based on fitting parameters to experimental data, their performance and range of applicability could be expected to improve with more training data. This has been addressed by the “Scoring Function Consortium”, a collaboration between academia and the pharmaceutical industry, where publicly available data have been augmented with data from industrial collaborators to create large training sets. The first results for the family of scoring functions created in this consortium, SFCscore, were published in April 2008. When evaluated on a large test set, the best-scoring function still gives a mean error of more than one order of magnitude, when used to predict $K_i$ values. This may indicate that this is the limit on the accuracy of empirical scoring functions of this type, due to their being based on only one binding pose, and the inherent experimental variability of the training data.

Finally, knowledge-based scoring functions work by calculating distance-dependent interaction energies between ligand and protein atom pairs from experimentally determined complexes. The energies are called potentials of mean force (PMF). This is done by deriving radial distribution functions for different pairwise combinations of atom types, and assuming that the experimentally observed complexes represent the global free energy minima of the systems. The resulting interaction energy functions are not based on physical principles, but are intended to implicitly describe all kinds of
interactions important for ligand-protein binding. The first such scoring function for SBVS was reported by Muegge and Martin in 1999,44 and further developed in 2006.45 Other implementations are DrugScore46,47 and M-Score.48

In Glide, two different scoring functions are implemented, Glide SP and Glide XP. Both are empirical scoring functions that also incorporate force-field-based terms. The Glide XP score involved more terms, and is more computationally expensive than Glide SP. Glide XP rewards compounds whose predicted binding poses lead to replacement of highly strained water molecules. Also, the formation of so-called correlated hydrogen bonds is emphasised. In this, the scoring function tries to describe hydrogen bonding and hydrophobic interactions in an involved and fine-grained manner.32 The two scoring functions are intended to be run in sequence. The Glide SP score is used first, in large-scale SBVS, after which the Glide XP score is applied to a subset of compounds obtaining high scores with the SP function.49

1.3 Malaria

As part of the work described in this thesis, SBVS was applied in projects within the area of infectious diseases. One enzyme, proposed as a target for developing new drugs to treat malaria, and another enzyme believed to be important in tuberculosis, were specifically targeted.

Malaria is caused by protozoan parasites of the genus Plasmodium.50 Four different species infect humans: Plasmodium falciparum (Pf), Plasmodium vivax, Plasmodium ovale and Plasmodium malarie. Of these, Pf is the most dangerous, followed by Plasmodium vivax. Plasmodium vivax and Plasmodium ovale can exist in the liver in a dormant stage called hypnozoites, for months, and Plasmodium malarie can lie dormant in the blood for decades.

These single-cell parasites have a complex life cycle, involving two different hosts: humans and female mosquitoes of the genus Anopheles. They enter the human host via the blood stream after a mosquito bite, and infect the liver. In the liver the parasites replicate non-sexually and are released back into the blood stream through so-called merozoites. The infected hepatocytes eventually die. In the blood stream the parasites infect red blood cells, in which they undergo meiosis. The resulting gametocytes are transferred back to the mosquito upon a subsequent bite. In the mosquito, the life cycle is completed through sexual reproduction.51 The entire genome of Pf was published in 2002.52
Malaria is usually classified as a mild or severe form. General symptoms are fever, vomiting and coughing. Severe malaria often manifests itself differently in children and adults. In adults, severe malaria often leads to failure of the kidneys and other organs, while children often show extreme weakness (prostration), respiratory problems, anaemia and/or cerebral malaria. The latter is a condition in which the patient falls into a coma, which is not attributable to any other condition, such as meningitis or hypoglycaemia. It is believed to be caused, somehow, by the sequestration of parasites in the capillaries of the brain.

The global incidence of malaria is shown in Figure 2. There are between 350 and 500 million clinical malaria episodes each year, with 0.7 to 2.7 million deaths. Most deaths are African children below the age of five.

Figure 2. Per-country estimated annual incidence of clinical malaria episodes, caused by any vector, in 2004. Reproduced with permission from the WHO, “World Malaria Report 2005”.

The first successful synthetic chemotherapy against malaria was chloroquine, first synthesised in 1934. Two other basic quinolines, quinine and quinidine, are the active ingredients in extracts from the bark of the South American cinchona tree, known for hundreds of years to possess anti-
malarial properties. Drugs currently used to treat malaria are shown in Figure 3.

Figure 3. Drugs currently used to treat malaria.
Chloroquine acts by inhibiting the detoxification of free heme in the parasite.\textsuperscript{58} Other drugs acting through this mechanism are amodiaquine, mefloquine, halofantrine and lumefantrine.\textsuperscript{59} Quinoline-based-drug resistant strains of Pf are now widespread. Drugs that act on specific target enzymes are dapsone (acting on dihydropteroate synthase), sulfadoxine (acting on dihydropteroate synthase), pyrimethamine (acting on dihydrofolate reductase), cycloguanil (acting on dihydrofolate reductase)\textsuperscript{60} and atovaquone (acting on the mitochondrial $bc_1$ complex).\textsuperscript{61} The currently recommended first-line therapy for treating malaria in endemic areas is a combination therapy, employing artemisinin or one of its analogues together with another drug (ACT, artemisinin-based combination therapy).\textsuperscript{62} Artemisinin is a natural product extracted from the leaves and flowers of the traditional Chinese medicinal plant *Artemisia annua*.\textsuperscript{63} It is believed to act by inhibiting the sarco-endoplasmic reticulum Ca\textsuperscript{2+}-ATPase (SERCA) of Pf, PfATP6.\textsuperscript{64}

Due to the development of drug-resistant strains of *Plasmodium*, new drugs to treat malaria, preferably acting through novel mechanisms of action, are urgently needed.

## 1.4 Tuberculosis

Tuberculosis (TB) is caused by the bacterium *Mycobacterium tuberculosis* (MTB). Infection normally occurs through inhaling bacteria-containing droplets emitted from infected individuals when coughing, sneezing or even speaking. Inhaled bacteria are phagocytosed by macrophages in the lungs. To combat the infection, these macrophages aggregate, and recruit various T-cells and granulocytes to form granuloma, so-called tubercles. The bacterium can lie dormant in these granuloma for decades.\textsuperscript{65,66} Actually, only approximately 10\% of all MTB infections ever give rise to the disease in the infected individual. Triggering of TB is associated with collapse of the tubercles and resumed bacterial replication. In situations where the immune system is weakened, as in individuals infected with HIV, the lifetime risk of developing TB for an infected individual increases to 50\%.\textsuperscript{67}

Pulmonary TB is the most common form of the disease in non-HIV infected patients. In conjunction with HIV, MTB spreads more readily to other parts of the body, and non-pulmonary disease is as common as the pulmonary form. The non-pulmonary forms include infections of the meninges and the central nervous system, the genitourinary system and the lymphatic system, as well as a disseminated form. The bacteria cause necrosis and tissue remodelling in the infected organs, with clinical effect depending on the site of action. In the lungs, TB initially causes intense coughing and sometimes
pneumothorax with associated pain, and eventually dyspnoea and severe respiratory failure. Systemic symptoms include weight loss and fever.\textsuperscript{68}

Globally, MTB infection and TB is an enormous problem. The number of people infected with MTB is estimated to be a staggering two billion, a third of the world’s population. The incidence and prevalence of the disease were estimated in 1997 and 2000.\textsuperscript{69,70} In the former, the number of people suffering from TB was estimated to be approximately 16 million, with almost 6 million new cases each year, and a TB death rate of 23%. In the latter, the incidence of new TB cases had increased to 8.3 million per year. According to the World Health Organization (WHO), the incidence in 2006 was 9.2 million new cases each year, the prevalence of TB was 14.4 million and the number of annual TB deaths was 1.7 million, corresponding to a mortality rate of 11.5\%.' The global spread of TB in 2006 is shown in Figure 4.

\begin{center}
\textbf{Figure 4.} Estimated number of TB cases by country in 2006. Reproduced with permission from the WHO, “Global Tuberculosis Control: Surveillance, Planning, Financing” (2008).
\end{center}

The MTB bacterium was discovered as the causative agent in TB by Robert Koch in 1882. For this and other findings in the field of microbiology, he received the Nobel Prize in 1905.\textsuperscript{71} The MTB bacterium has evolved to survive within macrophages, and is characterised, together with other mycobacteria, by extraordinary slow growth and a thick, complex cell wall

\textsuperscript{*} http://www.who.int/tb/publications/global_report/2008/xls/annex3_global.xls
with a unique composition. The complete genome of MTB was published in 1998, and is currently believed to contain 3995 genes, of which 376 encode proteins with no homology to known proteins.

The thick cell wall of MTB makes it immune to most common anti-bacterial chemotherapies. There are currently only five first-line anti-tuberculosis drugs (see Figure 5): isoniazid, rifampicin, pyrazinamide, ethambutol and streptomycin. These are usually given in combination. Isoniazid was discovered in 1952, and is an irreversible enzyme inhibitor, acting on the cell wall synthesis of MTB. It is still the most effective TB drug, and is included in all recommended first-line treatments. Ethambutol is another cell wall synthesis inhibitor, rifampicin inhibits transcription, streptomycin acts on the ribosome to inhibit translation and pyrazinamide inhibits the transport of nutrients over the bacterial cell membrane. Even though there are a number of less effective second-line antibiotics, development of multi-drug-resistant strains of MTB underlines the need for effective novel drugs to treat TB.

![Figure 5. First-line drugs currently used to treat TB.](image-url)
2 Aims of the present study

This study was part of a multi-disciplinary project: Rational Approach to Pathogen Inhibitor Discovery (RAPID), aimed at discovering new drugs to treat infectious diseases through rational structure-based drug design.

The specific objectives of this study were:

- To explore and develop new methodology in structure-based virtual screening, and

- To use structure-based virtual screening to identify novel inhibitors of *Mycobacterium tuberculosis* ribose-5-phosphate isomerase and *Plasmodium falciparum* spermidine synthase, target proteins relevant in tuberculosis and malaria.
3 New methodology in structure-based virtual screening

3.1 Background

In SBVS, as in all kinds of screening, the goal is to predict a set of putative hits for further evaluation or testing. Screening methods, as opposed to for example de novo design methods, are based on the evaluation of a pre-defined library of compounds. In the work presented in this thesis, libraries of existing screening compounds are always used. This facilitates experimental verification of the predictions, without having to consider the synthetic tractability or availability of the predicted hits.

In virtual screening, enrichment is a central measure. Basically, it describes how much better than random selection a specific virtual screening method is. Numerous papers have been published on ways in which this can be evaluated and measured, and even more comparing the performance of different docking and scoring algorithms with respect to enrichment. We have developed methodologies to try to increase the enrichment of existing docking and scoring tools. The observation or assumption behind these methodologies was that predicting a few active compounds from a large population of largely inactive compounds is quite different from correctly ranking a set of known active compounds. Different empirical transformations of scoring data were explored. The goal was to increase the discriminatory power when discriminating active compounds from inactive ones. In doing so, the physical interpretability of the scoring data as free energies of binding was sacrificed.

3.2 Supervised learning of scoring data (Paper I)

Charifson et al. introduced consensus scoring as a way of increasing enrichment in SBVS. This involves the use of several scoring functions on the same test library, and combining the sets of predicted hits from each scoring function to achieve a higher enrichment. Wang and Wang showed...
that consensus scoring works. However, they made the radical assumptions that there were no systematic errors in scoring functions, and that the random errors of different scoring functions were independent. In Paper I, we built on the approach of applying multiple scoring functions to generate a scoring vector for each compound. The resulting sets of vectors were analysed with multivariate statistical methods. Ultimately, the scoring data was used to construct multivariate classifiers, able to distinguish actives from inactives. The classifiers were developed using supervised learning, as is schematically shown in Figure 6.

**Figure 6.** Schematic representation of the methodology presented in Paper I.
3.2.1 Generation of data sets

To be able to draw general conclusions, four different targets were used, together with sets of known actives collected from the literature. The targets were oestrogen receptor α (ERα),\textsuperscript{107-109} acetylcholine esterase (AChE),\textsuperscript{110,111} matrix metalloprotease 3 (MMP-3)\textsuperscript{112,113} and factor Xa (fXa).\textsuperscript{114,115} In total, 389 known actives were used (see Table 1). 999 diverse compounds, selected from the MDL Drug Data Report (Symyx Technologies, Inc.) database, were used as inactives.

Table 1. Number of active compounds for each included target (Paper I).

<table>
<thead>
<tr>
<th>Target</th>
<th>PDB\textsuperscript{116} ID</th>
<th>Number of binders</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERα</td>
<td>1ERE</td>
<td>146</td>
</tr>
<tr>
<td>AChE</td>
<td>1EVE</td>
<td>54</td>
</tr>
<tr>
<td>MMP-3</td>
<td>1HY7</td>
<td>60</td>
</tr>
<tr>
<td>fXa</td>
<td>1G21</td>
<td>129</td>
</tr>
</tbody>
</table>

Initially, three different docking programs were used to generate poses: GOLD\textsuperscript{21,22} (Cambridge Crystallographic Data Centre), Glide\textsuperscript{30,31} (Schrödinger LLC) and ICM\textsuperscript{9,33,34} (Molsoft LLC). However, the choice of docking program used to generate the poses did not affect the results to any large extent, in terms of the performance of the multivariate classifiers. The continued analysis was therefore based only on poses generated with ICM.

The compounds were docked to each target structure, storing the best scoring pose per compound and target, according to the ICM scoring function. Prior to docking, target structures were prepared in ICM. The compound structures were prepared using a Sybyl Programming Language (SPL) script to ionise ionisable groups, and CORINA\textsuperscript{117} (Molecular Networks GmbH) to create starting 3D conformations. The docked poses were scored using seven different scoring functions: two scoring functions implemented in ICM and five scoring functions implemented in CScore\textsuperscript{105} (Tripos Inc.). The five scoring functions in CScore were FlexX score,\textsuperscript{25} DOCK score,\textsuperscript{10} PMF score,\textsuperscript{44} GOLD score\textsuperscript{22} and ChemScore.\textsuperscript{38}

The final data sets consisted of four matrices of scores, one for each target. Each matrix had seven columns and slightly more than one thousand rows. The columns corresponded to scoring functions, and the rows to active and inactive compounds. Note that the actual number of rows was smaller than the maximum for each target, since not all inactive compounds docked to all target structures.
3.2.2 Multivariate classifiers

The four different data sets were partitioned into training and validation sets, by sorting actives by potency and inactives by molecular weight. Every third compound was set apart for the validation sets. In this way, the training sets and validation sets had similar distributions of activities and molecular weights. Three separate multivariate methods were used to construct classifiers, namely Bayesian classification, partial least-squares projections to latent structures (PLS) discriminant analysis (PLS-DA) and rule-based methods as implemented in Rule Discovery System (Compumine AB).

Naïve Bayesian classifiers

The naïve Bayesian classifiers were implemented in MatLab (The MathWorks, Inc.), according to Equation 1, which is derived from Bayes’ theorem and the multivariate normal distribution.

\[
g_i(x) = P(i) \left( -\frac{1}{2} \ln |C_i| + \mu_i^T C_i^{-1} x - \frac{1}{2} \mu_i^T C_i^{-1} \mu_i - \frac{1}{2} x^T C_i^{-1} x \right)
\]

Here \( i \) is the class (active or inactive), \( x \) is the scoring vector for the compound being classified, \( C \) is the covariance matrix and \( \mu \) is the mean vector. \( C \) and \( \mu \) are calculated from the training set. The compound represented by the vector \( x \) is classified as belonging to the class \( i \) with the highest value of the discriminant function \( g(i) \).

\( P(i) \) is the a priori probability of class \( i \). By increasing the a priori probability of a compound being active, the resulting classifier would predict more active compounds. This would result in fewer false negatives, at the cost of predicting more false positives. Accordingly, the precision and enrichment of a Bayesian classifier can be tuned by changing \( P(i) \). In each specific case, \( P(i) \) can be set to give a suitable number of predicted actives, determined, for example, by the assay capacity.

PLS-DA

The PLS-DA classifiers were implemented by constructing PLS models from the training set, with the scoring data as independent variables, and the class, coded as -1 or +1, as the dependent variable. In standard PLS-DA classification, the class is predicted to be ‘active’ if the value predicted by the PLS model is above zero, and predicted to be ‘inactive’ if the value predicted by the PLS model is below zero. In the present case, however, the cross-validation results, from the training data, indicated that the threshold value used to classify a compound as active should be shifted to a value
above zero. This is similar to varying the a priori probability in Bayesian classification.

**Rule-based classification**

Rule-based models are sets of rules, where each rule is a condition on one or more of the attributes (scores) of the objects (compounds) being classified. In this study, models were induced by either recursive partitioning\(^{120}\) or covering\(^{121}\), sometimes together with so-called bagging\(^{122}\). The models were evaluated using cross-validation with the training set, and different kinds of models were selected for different targets, based on the cross-validation results.

**Evaluation of the results**

For each target, three multivariate classifiers were constructed, one for each method described above. These were applied to the validation sets, and the number of true positives (\(tp\)), true negatives (\(tn\)), false positives (\(fp\)) and false negatives (\(fn\)), were counted.

To compare the performance to that of already existing methods, two other classifiers were used. Firstly, the ICM score was used directly. The score threshold was set based on the scores of the actives in the training set. Secondly, consensus scoring, as proposed by Clark et al.\(^{105}\) and implemented in CScore (Tripos Inc.), was used. This consensus score was calculated by giving each scoring function one vote, and setting the threshold of each scoring function for voting ‘active’ to half the range of the scores from that scoring function, when applied to the set of compounds being classified. A consensus score threshold of 4 or 5 was used in this study.

Four different performance measures were calculated to compare the classifiers, accuracy, precision, recall and the enrichment factor (\(EF\) in Figure 7). The precision and recall were calculated with respect to the ‘active’ class. The measures are defined in Figure 7.

\[
\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \quad \text{Precision} = \frac{tp}{tp + fp} \quad \text{Recall} = \frac{tp}{tp + fn}
\]

\[
EF = \frac{\text{Precision}}{(tp + fn)/(tp + fp + tn + fn)}
\]

**Figure 7.** Definitions of the measures calculated to evaluate the multivariate classifiers.
3.2.3 Results and conclusions

The performance of the five classifiers for the validation sets is shown in Figure 8.

![Precision](image1)

**Figure 8.** Performance of the five different classifiers when applied to the validations sets, evaluated for each target, and in the case of ERα, for each set of known actives.
In the case of ERα, two separate sets of actives were used. One set, referred to as ERα-mimics, was a set of thirty-six structurally rather homogenous oestrogen mimics, and the other set, referred to as ERα-toxins, was a set of 110 more diverse compounds. These were treated separately, as is shown in Figure 8, but also in a combined and cross-wise manner. That is, on one hand the combined training sets were used to construct classifiers that were applied to the combined validation sets, and on the other hand, the training set from one set of compounds was used to create the classifiers applied to the validation set of the other set of compounds. The results are given in Table 2.

Table 2. Performance of the different multivariate classifiers when applied to the two different sets of actives for ERα. ‘Actives’ is the combination of the two sets, ‘Mimics’ and ‘Toxins’ are the sets applied to themselves and the last two groups of results are for the cross-wise classifiers.

<table>
<thead>
<tr>
<th>Data set combination</th>
<th>No. compounds (No. actives)</th>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>381 (48)</td>
<td>Rule-based</td>
<td>0.979</td>
<td>0.955</td>
<td>0.875</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayes</td>
<td>0.895</td>
<td>0.551</td>
<td>0.896</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS-DA</td>
<td>0.961</td>
<td>0.811</td>
<td>0.896</td>
<td>6.4</td>
</tr>
<tr>
<td>Mimics</td>
<td>345 (12)</td>
<td>Rule-based</td>
<td>0.994</td>
<td>1.000</td>
<td>0.833</td>
<td>28.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayes</td>
<td>0.980</td>
<td>0.692</td>
<td>0.750</td>
<td>19.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS-DA</td>
<td>0.991</td>
<td>0.909</td>
<td>0.833</td>
<td>26.1</td>
</tr>
<tr>
<td>Toxins</td>
<td>369 (36)</td>
<td>Rule-based</td>
<td>0.976</td>
<td>0.935</td>
<td>0.806</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayes</td>
<td>0.905</td>
<td>0.508</td>
<td>0.861</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS-DA</td>
<td>0.959</td>
<td>0.784</td>
<td>0.806</td>
<td>8.0</td>
</tr>
<tr>
<td>Mimics to toxins</td>
<td>443 (110)</td>
<td>Rule-based</td>
<td>0.887</td>
<td>0.969</td>
<td>0.564</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayes</td>
<td>0.935</td>
<td>0.945</td>
<td>0.782</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS-DA</td>
<td>0.932</td>
<td>0.955</td>
<td>0.764</td>
<td>3.8</td>
</tr>
<tr>
<td>Toxins to mimics</td>
<td>369 (36)</td>
<td>Rule-based</td>
<td>0.986</td>
<td>0.919</td>
<td>0.944</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayes</td>
<td>0.911</td>
<td>0.522</td>
<td>1.000</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS-DA</td>
<td>0.965</td>
<td>0.745</td>
<td>0.972</td>
<td>7.6</td>
</tr>
</tbody>
</table>

In summary, it was found that multivariate classification of multidimensional scoring data significantly increased the performance in SBVS, compared with using a single scoring function. This was also true when using chemically quite diverse compounds to generate the scoring data used for training the classifiers. It was also observed that consensus scoring as implemented in CScore did not work very well. The enrichment was always worse than the best-scoring function for each target (but better than the worst).

Two important considerations led on to the study described in Paper II. Firstly, the methods described in Paper I all require a training set, i.e. a set of
known actives. It would be very useful to develop methods not requiring previously known binders, but still increasing the enrichment compared to any single scoring function. Secondly, why did these multivariate classifiers actually work? What information was encoded in the scoring vectors that enabled the multivariate classifiers to perform better than any of the scores on their own, in almost all cases?

3.2.4 Related studies

Terp et al. applied principal component analysis (PCA) and PLS to the same kind of data matrices as those described in Paper I. They started with eight scoring functions, but ended up using only five of them. Their study was aimed at ranking actives in terms of the free energy of binding. Accordingly, regression models were constructed, instead of classifiers. The resulting predictions showed good correlations with measured potencies of known binders.

Teramoto and Fukunishi applied a rule-based algorithm, ‘Random Forest’, to construct classifiers used to rank predicted poses of the same compound, and also applied this in SBVS.

Klon et al. took a slightly different approach. The compounds were ranked using only one scoring function, but a Bayesian classifier was constructed with the predicted actives as the training set, and molecular fingerprints, not multiple scores, as descriptors. The actual SBVS predictor was then the Bayesian classifier, which would identify high-scoring compounds according to the single scoring function, and compounds structurally similar to those compounds. This method was also combined with classical consensus scoring.

Prathipati and Saxena applied a different multivariate method, binary QSAR, to the same data set as in the current work (Paper I), but with a different docking program and different scoring functions. They also used all terms of the scoring functions instead of the actual scores, and obtained results comparable to or better those reported in Paper I, in terms of enrichment.
3.3 Transforming scoring data without a training set (Paper II)

Structure-based virtual screening is most valuable when it can be used to identify novel binders to target protein with no previously known binders. The method described in Paper I does not approach this problem, since a set of known actives is needed to train the multivariate classifiers. In 2004, Vigers and Rizzi suggested one such method; Multiple Active Site Correction (MASC).\textsuperscript{131} Basically, their idea was to dock and score each compound against the target structure, and a set of unrelated protein structures. The score obtained against the target structure was then ‘corrected’ using the scores obtained against all protein structures.

Another method was suggested by Pan et al.\textsuperscript{132} They observed that the sets of hits predicted by the scoring functions they evaluated were skewed towards higher molecular weight compounds compared with the starting test library. To obtain a predicted set of hits with a molecular weight distribution more similar to that of the starting distribution, they suggested a transformation in which the scores were divided by some power of the number of heavy atoms or the molecular weight of each compound.

In the search for methods that would also improve SBVS enrichments when there were no actives known beforehand, the methods introduced by Vigers and Rizzi and Pan et al. were evaluated and expanded. In the process of this, significant correlations were also observed between scores and the physio-chemical properties of the compounds being scored. This correlation is a computational artefact of the docking and scoring calculations. Such a relation between the physiochemical properties and potency is not observed in experimental screening of large, diverse sets of compounds.

3.3.1 Test sets

Eight different targets were used, with a total of 945 known binders and 10,000 random diverse compounds as decoys, and ten different scoring functions, to construct the data sets used in the evaluation. The targets were the same as in the study described in Paper I, together with dihydrofolate reductase (DHFR), cyclooxygenase 2 (COX2), cyclin dependent kinase 2 (CDK2) and neuraminidase (NA). This large set of data was used to be able to draw generally applicable conclusions from the results. The known actives were collected from the Protein Data Bank (PDB) and from the literature, see Table 3.
The 10 000 decoys were chosen from a database of commercially available screening compounds. The database was filtered using the ‘Rule of 5’ prior to the random selection. The compounds were selected using the OptiSim algorithm as implemented in ChemEnlighten (Tripos Inc.). LigPrep (Schrödinger LLC) was used to calculate 3D conformations of both known actives and decoys. LigPrep calculated different ionisation states, stereo-isomers and tautomers for each compound. Thus, each compound was represented by one or more structures. The calculation of 3D coordinates failed for ten of the decoy compounds, so the final number of decoys was 9 990, represented by 18 851 3D structures. The 945 known actives were represented by 2 649 3D structures.

Table 3. Sources and number of known actives for each target used (Paper II).

<table>
<thead>
<tr>
<th>Target</th>
<th>PDB ID</th>
<th>Reference</th>
<th>Number of compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHFR</td>
<td>1RX3</td>
<td>Otzen,135 Zolli-Juran,137</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>4</td>
</tr>
<tr>
<td>MMP3</td>
<td>1G49</td>
<td>Ha113</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>14</td>
</tr>
<tr>
<td>AchE</td>
<td>1EVE</td>
<td>Sippl, Contreras</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>9</td>
</tr>
<tr>
<td>ERα</td>
<td>3ERD</td>
<td>Sippl108</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shi109</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>5</td>
</tr>
<tr>
<td>tXa</td>
<td>1IQE</td>
<td>Matter115</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>26</td>
</tr>
<tr>
<td>COX2</td>
<td>1PXX</td>
<td>Clark141</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>5</td>
</tr>
<tr>
<td>NA</td>
<td>1L7F</td>
<td>Y1143</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>10</td>
</tr>
<tr>
<td>CDK2</td>
<td>1K7</td>
<td>Tominaga145</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDB</td>
<td>44</td>
</tr>
</tbody>
</table>

All compound structures were docked to all target structures using Glide. One pose was stored per combination of structure and target; the highest scoring pose according to the Glide EMModel scoring function. All poses were re-scored, using all ten scoring functions. The highest score was stored for each compound, according to each scoring function, for each target, respectively. The resulting data set was a three-dimensional matrix with approximately 875 000 scoring function results (see Figure 9). Note that this was not a full matrix, as all compound structures could not be docked against all targets.
Ten scoring functions were used in this study. Firstly, the same five CScore scoring functions as in the study described Paper I, secondly GlideScore and EModel from Glide, and finally the three scoring functions HMScore, HPScore and HSScore, implemented in XScore. The SBVS set-up used in this study was similar to a real case, but with one or two orders of magnitude fewer compounds in the test library. The number of known actives per target was quite reasonable. The known actives were, however, not very diverse, compared, for example, to a set of validated primary hits from HTS. This was due to their being partly collected from papers describing structure–activity relationships and the expansion of chemical series. The diversity of the compound sets is illustrated in Figure 10. In this figure, the compounds are described by six molecular descriptors: heavy atoms count (HAC), molecular weight (MW), calculated molecular refractivity (CMR), molecular volume, number of heteroatoms and calculated logP (ClogP). The resulting data set was analysed by PCA. From this analysis it was clear that the actives for a particular target were physiochemically similar. For example, the ERα actives were small and hydrophobic while the fXa actives were large and hydrophilic. It was also clear that the decoys covered the same physiochemical space as the known actives.
Figure 10. PCA score plots of the known actives and the decoys. The loading plot (a) shows that the first principal component describes size and the second describes polarity. The score plots are: (b) CDK2, (c) COX2, (d) DHFR, (e) ERα, (f) tXa, (g) NA, (h) MMP3 and (i) decoys.
3.3.2 Transforms

The scoring data were manipulated using three different kinds of transforms; firstly the MASC transform with different variations, secondly by dividing by HAC or MW or a power thereof and thirdly a new transform named PLS MASC.

**MASC**

This transformation was defined by Vigers and Rizzi\textsuperscript{131} as:

\[
S_{i,j}^* = \frac{(S_{i,j} - \mu_i)}{\sigma_i}
\]  \hspace{1cm} (2)

Where \( S_{i,j} \) is the score for compound \( i \) against target \( j \), \( \mu_i \) is the mean of the scores for compound \( i \) over all targets, and \( \sigma_i \) is the standard deviation of the scores for compound \( i \) over all targets. Different variations of this transform were tested, such as using the median instead of the mean, excluding the score against the intended target when calculating the means and standard deviations, and rank-transforming the scores before calculating means and standard deviations. This did not improve the resulting enrichments to any significant extent. However, excluding the standard deviation from the transform, and normalising the scores to unit variance and zero mean, against each target separately, before calculating the MASC-transformed scores did improve the enrichment in most cases. This modified MASC transform is given in Equations 3 and 4.

\[
S_{i,j}'' = \frac{(S_{i,j} - \mu_j)}{\sigma_j}
\]  \hspace{1cm} (3)

\[
S_{i,j}^* = S_{i,j}'' - \mu_i
\]  \hspace{1cm} (4)

Here, \( \mu_j \) and \( \sigma_j \) are the mean and standard deviation of the scores from all compounds against target \( j \).

**Size normalisation**

Based on the suggestions of Pan et al.,\textsuperscript{132} the scores were also divided by the molecular weight, the number of heavy atoms or a power of these molecular properties prior to ranking the compounds.
PLS MASC

In the analysis of the scoring data, a rather strong linear correlation was observed between the HAC and the scores for most scoring functions. The worst scoring function in this respect, the DOCK function from CScore, had an r-square ranging from 0.33 (DHFR) to 0.89 (COX2), when fitting the scores for all compounds to the HAC. These observations, together with the sometimes rather remarkable improvement in enrichment following size normalisation, led to the proposal of PLS MASC.

The basic idea was that the scores of a large set of diverse compounds should not correlate with simple physiochemical properties of the compounds. The vast majority of the compounds are not true binders of the target protein, and hence their free energy of binding is positive. The scores are intended to be interpreted as free energies of binding. Only if the compounds being scored are all true binders, can a correlation be expected between for example size and the free energy of binding. Hence, the success of size normalisation and MASC may be due to correcting the scoring functions employed in SBVS for this correlation, or ligand bias. If so, it may be possible to describe the ligand-bias part of a particular scoring function in a PLS model, correlating molecular properties to mean scores, as calculated in the MASC transform. This PLS model could then be used to calculate the MASC factors that would be subtracted from the actual scores.

PLS models were constructed using up to 192 different molecular descriptors. However, the best results were achieved using only up to six descriptors, the same descriptors as in the PCA illustrated in Figure 10. Tables 4-10 give the results for four different PLS models. PLS1 is a least squares fit to MW, PLS2 a least squares fit to HAC, PLS3 a four-component PLS model correlating all six descriptors to the mean scores, and PLS4 a one-component PLS model correlating MW, HAC and molecular volume to the mean scores.

3.3.3 Results

To evaluate the effect of the different transforms, enrichment factors were calculated in the same way as described in Paper I, at various subsetting ratios. The subsetting ratio is defined as the proportion of top-ranked compounds predicted as actives. The resulting enrichments, for all targets, scoring functions and transforms, at a subsetting ratio of 3%, are listed in Tables 4-10.
The results for AChE are not shown as the enrichments for all scoring functions for this target were close to one, meaning that the results were no better than a random selection. It was noted already in the study presented in Paper I that many of the known AChE actives docked with implausible poses, and this was also the case here, when using Glide for docking. Hence, the poor enrichments for AChE were probably due to poor docking results, and therefore not very useful when analysing scoring functions and scoring function transforms. The AChE scores were, however, used in the MASC and PLS MASC calculations.

In Tables 4-10 below, ‘Mod. MASC’ means modified MASC, as defined in Equations 3 and 4, MW and HAC means size normalisation with the molecular weight and heavy atom count, SQRT(MW) means size normalisation with the square root of the molecular weight and CRT(MW) means size normalization with the cubic root of the molecular weight.

**Table 4.** Enrichment factors at 3% subsetting for CDK2 for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Scoring function</th>
<th>Transform</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>20</td>
<td>15</td>
<td>18</td>
<td>0.68</td>
<td>0.68</td>
<td>14</td>
<td>20</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>EModel</td>
<td>13</td>
<td>13</td>
<td>19</td>
<td>10</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>FlexX</td>
<td>27</td>
<td>14</td>
<td>19</td>
<td>12</td>
<td>12</td>
<td>19</td>
<td>19</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Chem-Score</td>
<td>14</td>
<td>16</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>16</td>
<td>21</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>PMF</td>
<td>5.1</td>
<td>1.4</td>
<td>4.8</td>
<td>2.4</td>
<td>2.4</td>
<td>4.4</td>
<td>4.1</td>
<td>4.8</td>
<td>4.8</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>GOLD</td>
<td>4.1</td>
<td>9.2</td>
<td><strong>10</strong></td>
<td>1</td>
<td>0.34</td>
<td>2.7</td>
<td>3.1</td>
<td>2.7</td>
<td>2.7</td>
<td><strong>10</strong></td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>DOCK</td>
<td>2.4</td>
<td><strong>8.1</strong></td>
<td>7.8</td>
<td>0</td>
<td>0</td>
<td>1.4</td>
<td>1.4</td>
<td>2</td>
<td>2</td>
<td>2.4</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>HM-Score</td>
<td>0.99</td>
<td>9.8</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.4</td>
<td></td>
<td><strong>15</strong></td>
<td>1</td>
</tr>
<tr>
<td>HP-Score</td>
<td>0.98</td>
<td>11</td>
<td>8.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.34</td>
<td>1</td>
<td></td>
<td><strong>13</strong></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HS-Score</td>
<td>0.65</td>
<td><strong>13</strong></td>
<td>11</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>11</td>
<td>1.4</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. Enrichment factors at 3% subsetting for COX2 for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Transform</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>19</td>
<td>9.8</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3.3</td>
<td>8.8</td>
<td>13</td>
<td>13</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Emodel</td>
<td>4.7</td>
<td>0.64</td>
<td>1.6</td>
<td>0.21</td>
<td>0.21</td>
<td>0.43</td>
<td>0.86</td>
<td>0.64</td>
<td>0.54</td>
<td>0.75</td>
<td>0.64</td>
</tr>
<tr>
<td>FlexX</td>
<td>9.9</td>
<td>5.7</td>
<td>8.7</td>
<td>1.9</td>
<td>2.5</td>
<td>6.8</td>
<td>8.3</td>
<td>9.7</td>
<td>9.8</td>
<td>9.8</td>
<td>9.8</td>
</tr>
<tr>
<td>Chem-Score</td>
<td>8.8</td>
<td>10</td>
<td>5.5</td>
<td>0</td>
<td>0.43</td>
<td>3.8</td>
<td>6.1</td>
<td>5.2</td>
<td>6.3</td>
<td>9.8</td>
<td>6.3</td>
</tr>
<tr>
<td>PMF</td>
<td>5.2</td>
<td>9.7</td>
<td>11</td>
<td>0.86</td>
<td>1.3</td>
<td>2.1</td>
<td>3.2</td>
<td>3.3</td>
<td>3.4</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>GOLD</td>
<td>4.7</td>
<td>5.7</td>
<td>21</td>
<td>0.11</td>
<td>0.32</td>
<td>1.6</td>
<td>2.6</td>
<td>2.4</td>
<td>2.7</td>
<td>9</td>
<td>2.6</td>
</tr>
<tr>
<td>DOCK</td>
<td>9.8</td>
<td>2.4</td>
<td>23</td>
<td>1.2</td>
<td>2</td>
<td>5.3</td>
<td>7.8</td>
<td>4.9</td>
<td>7</td>
<td>10</td>
<td>7.6</td>
</tr>
<tr>
<td>HM-Score</td>
<td>8</td>
<td>8</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0.21</td>
<td>1.9</td>
<td>2.7</td>
<td>3.8</td>
<td>3.9</td>
<td>3.4</td>
</tr>
<tr>
<td>HP-Score</td>
<td>5.6</td>
<td>3.7</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0.11</td>
<td>1.2</td>
<td>2.3</td>
<td>4</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>HS-Score</td>
<td>4.5</td>
<td>2.8</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0.21</td>
<td>1.1</td>
<td>2</td>
<td>4.2</td>
<td>1.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Enrichment factors at 3% subsetting for DHFR for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Transform</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>5.8</td>
<td>19</td>
<td>18</td>
<td>16</td>
<td>10</td>
<td>24</td>
<td>24</td>
<td>21</td>
<td>23</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Emodel</td>
<td>3</td>
<td>23</td>
<td>8.9</td>
<td>15</td>
<td>10</td>
<td>5.9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>FlexX</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>20</td>
<td>21</td>
<td>21</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Chem-Score</td>
<td>0</td>
<td>19</td>
<td>13</td>
<td>8.9</td>
<td>5.9</td>
<td>4.4</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>3.7</td>
<td>2.2</td>
</tr>
<tr>
<td>PMF</td>
<td>20</td>
<td>18</td>
<td>13</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>28</td>
<td>27</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>GOLD</td>
<td>1.5</td>
<td>3.7</td>
<td>3.7</td>
<td>5.2</td>
<td>2.2</td>
<td>4.4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2.2</td>
<td>6.7</td>
</tr>
<tr>
<td>DOCK</td>
<td>3</td>
<td>3</td>
<td>3.7</td>
<td>3</td>
<td>0</td>
<td>4.4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>HM-Score</td>
<td>0</td>
<td>12</td>
<td>1.5</td>
<td>3.7</td>
<td>1.5</td>
<td>8.1</td>
<td>4.4</td>
<td>1.5</td>
<td>0.74</td>
<td>6.7</td>
<td>0.74</td>
</tr>
<tr>
<td>HP-Score</td>
<td>0</td>
<td>5.2</td>
<td>3.7</td>
<td>4.4</td>
<td>2.2</td>
<td>8.1</td>
<td>8.9</td>
<td>4.4</td>
<td>4.4</td>
<td>8.1</td>
<td>4.4</td>
</tr>
<tr>
<td>HS-Score</td>
<td>0</td>
<td>10</td>
<td>2.2</td>
<td>2.2</td>
<td>1.5</td>
<td>4.4</td>
<td>3.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 7. Enrichment factors at 3% subsetting for ERα for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Scoring function</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>18</td>
<td>14</td>
<td>21</td>
<td>21</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Emodel</td>
<td>12</td>
<td>20</td>
<td>19</td>
<td>14</td>
<td>11</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>FlexX</td>
<td>14</td>
<td>17</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Chem-Score</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>21</td>
<td>24</td>
<td>20</td>
<td>17</td>
<td>17</td>
<td>18</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>PMF</td>
<td>0.22</td>
<td>7.5</td>
<td>6.2</td>
<td>2.2</td>
<td>1.6</td>
<td>0.44</td>
<td>0.44</td>
<td>0.22</td>
<td>0.22</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td>GOLD</td>
<td>6.2</td>
<td>3.1</td>
<td>3.8</td>
<td>11</td>
<td>9.5</td>
<td>8</td>
<td>6.9</td>
<td>7.1</td>
<td>6.2</td>
<td>6.9</td>
<td>6.9</td>
</tr>
<tr>
<td>DOCK</td>
<td>0.67</td>
<td>0.22</td>
<td>0.22</td>
<td>6.7</td>
<td>3.1</td>
<td>2</td>
<td>1.3</td>
<td>1.8</td>
<td>0.67</td>
<td>1.1</td>
<td>0.67</td>
</tr>
<tr>
<td>HMM Score</td>
<td>13</td>
<td>7.1</td>
<td>7.3</td>
<td>9.1</td>
<td>7.5</td>
<td>20</td>
<td>19</td>
<td>17</td>
<td>17</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>HP Score</td>
<td>10</td>
<td>4.2</td>
<td>9.5</td>
<td>9.3</td>
<td>6.9</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>HS Score</td>
<td>11</td>
<td>7.8</td>
<td>14</td>
<td>6</td>
<td>4.6</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>14</td>
<td>7.5</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 8. Enrichment factors at 3% subsetting for fXa for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Scoring function</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>23</td>
<td>11</td>
<td>13</td>
<td>1.3</td>
<td>0</td>
<td>14</td>
<td>19</td>
<td>20</td>
<td>19</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>EModel</td>
<td>21</td>
<td>9.2</td>
<td>11</td>
<td>7.5</td>
<td>6</td>
<td>24</td>
<td>19</td>
<td>18</td>
<td>18</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>FlexX</td>
<td>16</td>
<td>9.9</td>
<td>15</td>
<td>4.9</td>
<td>2.8</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Chem-Score</td>
<td>14</td>
<td>4.1</td>
<td>8.4</td>
<td>0</td>
<td>0</td>
<td>4.9</td>
<td>9</td>
<td>6</td>
<td>5.4</td>
<td>6.7</td>
<td>5.6</td>
</tr>
<tr>
<td>PMF</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.2</td>
<td>8.2</td>
<td>5.4</td>
<td>2.6</td>
<td>1.7</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>GOLD</td>
<td>8.6</td>
<td>3.2</td>
<td>3.4</td>
<td>0</td>
<td>3.9</td>
<td>6.7</td>
<td>4.5</td>
<td>3.2</td>
<td>2.1</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>DOCK</td>
<td>8.6</td>
<td>1.5</td>
<td>7.1</td>
<td>0</td>
<td>3.7</td>
<td>6.2</td>
<td>3.7</td>
<td>3</td>
<td>2.6</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>HMM Score</td>
<td>9.2</td>
<td>1.5</td>
<td>3.7</td>
<td>0</td>
<td>0</td>
<td>0.86</td>
<td>1.9</td>
<td>0.86</td>
<td>3.7</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>HP Score</td>
<td>9.5</td>
<td>4.3</td>
<td>6.2</td>
<td>0</td>
<td>0</td>
<td>3.2</td>
<td>4.7</td>
<td>2.8</td>
<td>3.4</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>HS Score</td>
<td>5.7</td>
<td>3.2</td>
<td>6.7</td>
<td>0</td>
<td>0</td>
<td>0.43</td>
<td>0.64</td>
<td>0.21</td>
<td>0.64</td>
<td>0.43</td>
<td></td>
</tr>
</tbody>
</table>
Table 9. Enrichment factors at 3% subsetting for MMP3 for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Transform</th>
<th>Score Function</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>31</td>
<td>0.9</td>
<td>14</td>
<td>0</td>
<td>9</td>
<td>23</td>
<td>27</td>
<td>27</td>
<td>28</td>
<td>27</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>EMModel</td>
<td>31</td>
<td>0.45</td>
<td>24</td>
<td>5.8</td>
<td>4.9</td>
<td>24</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>FlexX</td>
<td>28</td>
<td>0.45</td>
<td>7.2</td>
<td>1.8</td>
<td>1.8</td>
<td>15</td>
<td>14</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>ChemScore</td>
<td>28</td>
<td>1.4</td>
<td>7.2</td>
<td>2.3</td>
<td>1.8</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>15</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>PMF</td>
<td>3.6</td>
<td>0.45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.9</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOLD</td>
<td>9.4</td>
<td>1.4</td>
<td>1.8</td>
<td>0</td>
<td>2.7</td>
<td>4.9</td>
<td>3.1</td>
<td>2.7</td>
<td>2.3</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOCK</td>
<td>12</td>
<td>1.8</td>
<td>0</td>
<td>0</td>
<td>3.1</td>
<td>6.7</td>
<td>3.1</td>
<td>2.3</td>
<td>1.8</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMScore</td>
<td>19</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>10</td>
<td>3.6</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>HPScore</td>
<td>22</td>
<td>0.9</td>
<td>6.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>17</td>
<td>18</td>
<td>11</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>HSScore</td>
<td>14</td>
<td>0</td>
<td>8.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.4</td>
<td>5.4</td>
<td>4</td>
<td>3.1</td>
<td>4.9</td>
<td></td>
</tr>
</tbody>
</table>

Table 10. Enrichment factors at 3% subsetting for NA for all scoring functions and ten different transforms. The best enrichment result for the non-transformed scoring function is shown in italics and the best transformed result for each scoring function is shown in bold face.

<table>
<thead>
<tr>
<th>Transform</th>
<th>Score Function</th>
<th>None</th>
<th>MASC</th>
<th>Mod. MASC</th>
<th>MW</th>
<th>HAC</th>
<th>SQRT (MW)</th>
<th>CRT (MW)</th>
<th>PLS1</th>
<th>PLS2</th>
<th>PLS3</th>
<th>PLS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>18</td>
<td>5.3</td>
<td>21</td>
<td>13</td>
<td>10</td>
<td>25</td>
<td>24</td>
<td>23</td>
<td>23</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>EMModel</td>
<td>27</td>
<td>27</td>
<td>31</td>
<td>28</td>
<td>28</td>
<td>30</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>FlexX</td>
<td>24</td>
<td>17</td>
<td>31</td>
<td>25</td>
<td>25</td>
<td>27</td>
<td>27</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>ChemScore</td>
<td>6.6</td>
<td>23</td>
<td>26</td>
<td>9.3</td>
<td>7.3</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>12</td>
<td>21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>PMF</td>
<td>21</td>
<td>27</td>
<td>32</td>
<td>25</td>
<td>25</td>
<td>27</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>GOLD</td>
<td>7.3</td>
<td>5.3</td>
<td>11</td>
<td>9.3</td>
<td>8.6</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>DOCK</td>
<td>11</td>
<td>14</td>
<td>19</td>
<td>12</td>
<td>10</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>HMScore</td>
<td>0</td>
<td>23</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1.3</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HPScore</td>
<td>0</td>
<td>25</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
<td>1.3</td>
<td>2</td>
<td>18</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>HSScore</td>
<td>0</td>
<td>24</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>15</td>
<td>0.67</td>
<td></td>
</tr>
</tbody>
</table>
3.3.4 Conclusions

The first conclusion from this study was that none of the evaluated methods improved the enrichments in all cases. It is therefore not clear which transform should be used in any particular case. The fact that the transforms often improved the enrichment does, however, indicate that ligand bias is an important source of false positives in SBVS. In other words, the scoring functions included in this study not only describe the ligand–target interactions of the scored poses, but also the molecular properties of the ligands themselves. This could partly explain the improved enrichments reported in Paper I.

When comparing the different scoring functions without any transforms, Glide gave the best enrichment in three cases, EModel gave the best enrichment in two cases, FlexX gave the best enrichment in two cases, and PMF gave as good enrichment as FlexX in one of those cases. Comparing transformed versus untransformed scores for all combinations of target and scoring function showed that transforming the scoring function increased the SBVS performance in 46 cases, and decreased the performance in 24 cases. Regarding the different types of transforms, MASC gave the best enrichment in 32 cases, PLS MASC in 18.5 cases and size normalisation in 19.5 cases; half cases being those where two different transform types resulted in the best enrichments at the same time. Finally, when comparing MASC to modified MASC, the latter was found to be better in 51 cases and the former in 19 cases.

Thus, if SBVS is to be performed using ranking by a scoring function as the criterion for selecting actives, the results of this study suggests that the best a priori choice of the ten scoring functions included is the Glide scoring function. In general, one should always consider transforming the scores. The best transform of those tested here, with regard to enrichment, was the modified MASC transform, defined in Equations 3 and 4.
4 Structure-based virtual screening applications

4.1 Background

As part of the interdisciplinary project RAPID, aimed at developing new drugs for the treatment of infectious diseases, SBVS was applied to two different potential drug targets. Since the assay and compound acquisition resources were limited, the goal was to predict small sets of potential actives. These were to be experimentally verified, and any confirmed hits would serve as starting points for structure-based drug design projects.

The target structures were novel, with little previously known about inhibitors or the structural features required for interaction with the targets. Below, docking studies and SBVS against a tuberculosis target, Mycobacterium tuberculosis ribose-5-phosphate isomerase (mtRpiB), and a malaria target, Plasmodium falciparum spermidine synthase (pfSpdSyn), are described. In the first study, no synthetic binders were found, but in the second study the SBVS was successful, and a set of novel binders was identified.

4.2 Mycobacterium tuberculosis ribose-5-phosphate isomerase (Paper III)

Ribose-5-phosphate isomerase (EC 5.3.1.6) is a metabolic enzyme catalysing the inter-conversion of ribulose-5-phosphate and ribose-5-phosphate (R5P)\(^{146}\) (see Figure 11). This enzyme is part of the pentose phosphate pathway, where NADPH and R5P are synthesised from glucose-6-phosphate. Ribose-5-phosphate is used in the synthesis of nucleotides and certain amino acids.
Some organisms lack the pentose phosphate pathway, instead relying on alternative pathways, but R5P isomerase seems to be present in all organisms. In spite of this, the *Escherichia coli* (*E. coli*) R5P isomerase null mutant is viable, although its growth is severely hampered.

Ribose-5-phosphate isomerase exists in two evolutionarily unrelated isoforms, RpiA and RpiB. *E. coli*, for example, has both isoforms, but MTB has only one, mtRpiB. The B isoform is almost entirely limited to bacteria. This increases the probability of a synthetic inhibitor of mtRpiB being selective for the bacterial enzyme over the human one.

In the RAPID project, mtRpiB was studied from enzymatic and structural perspectives. Docking studies were used to predict the binding pose of substrates and products, and SBVS was used to predict putative active-site binders to serve as possible starting points for the development of novel inhibitors.

### 4.2.1 Protein structure

The first RpiB structure, that of *E. coli*, was solved in 2003. The X-ray structure shows a tetramer, but the functional unit is most likely a dimer with two identical active sites. Residues from both subunits contribute to each active site.

To study the putative RpiB enzyme from MTB, it was cloned and expressed in *E. coli*. The purified enzyme was shown to have R5P isomerase activity, with a $K_M$ of 3.7 mM and a $k_{cat}$ of 120 s$^{-1}$, with R5P as substrate. The only inhibitors identified at this point were phosphate, with a $K_i$ of 130 mM, and iodoacetate, which completely abolished the isomerase activity. The structure was solved to a resolution of 1.88 Å, with phosphate from the crystallisation buffer present in the active site. The biologically active dimer (see Figure 12) had a very similar structure to the *E. coli* enzyme, and most...
active-site residues were conserved. There was, however, one interesting difference that the new structure helped to shed light on. In the \textit{E. coli} enzyme, the catalytic base was Cys66. This cysteine is a glycine in the MTB enzyme, and it was not clear how mtRpiB exercised its catalytic activity. However, in the new structure it was apparent that Glu75 occupied the same space as Cys66 in the \textit{E. coli} structure, and thus probably served as the catalytic base.

\textbf{Figure 12.} Overall fold of mtRpiB. The molecular surface of one of the two active sites, with the associated phosphate ion, is indicated.

\subsection*{4.2.2 Docking studies}

One question that had not been resolved at this point was whether RpiB catalysed the ring opening of the substrate or not. In solution, R5P exists almost entirely in the $\alpha$- or $\beta$-furanose forms (see Figure 13), only approximately 0.1\% is in the aldehyde state.\textsuperscript{150} If the enzyme were able to
bind the furanose form and catalyse its opening, this would increase the efficiency of the enzyme.

To approach this question, and to further study the interactions in the active site, the substrates/products shown in Figure 13 were docked to the two available RpiB structures (PDB ID 1NN4 being the *E. coli* structure and 1USL being the MTB structure). The docking program used was Glide\textsuperscript{31} with the XP scoring function.\textsuperscript{32}

![Image of molecules](image)

**Figure 13.** The molecules docked against the *E. coli* and MTB RpiB structures.

The docked poses of the ring-closed and open forms of R5P in mtRpiB are shown in Figure 14. The \(\alpha\)- and \(\beta\)-anomers docked with roughly the same poses. In these dockings, it could be seen that (a) the phosphate group of the sugars invariably docked in the same position as the phosphate ion in the mtRpiB structure, (b) O2 of the substrate docked near the catalytic base (Cys66 or Glu75 in the *E. coli* and MTB enzymes, respectively) and (c) the furanose ring oxygen docked close to His102 (His99 in *E. coli*), a conserved histidine. Based on the third observation, His102 was hypothesised to be important in the presumed catalysis of the ring opening.
Figure 14. The docked poses of β-furanose R5P (above) and open R5P (below), in mtRpiB. All residues from both subunits that form direct hydrogen bonds to the docked molecules are shown. Note that His102 from the A-chain forms a hydrogen bond to the ring oxygen of the ring-closed sugar, and that the presumed catalytic base, Glu75 from the B-chain, forms a hydrogen bond to O2 in the open sugar.
4.2.3 Structure-based virtual screening

Having solved the first mtRpiB structure, SBVS was used to try to identify novel synthetic active-site binders. The protein structure was prepared by firstly removing all the water molecules, adding hydrogen atoms, and assigning tautomers to histidines and side-chain amide orientations to glutamines. This was done using the convertObject* macro in ICM. Secondly, the structure was relaxed in the OPLS-AA force-field, using the pprep and impref programs (Impact 2.5, Schrödinger LLC.). The compound test library used was the ExpressPick May 2003 library from ChemBridge. This contained 315 102 structures, which were filtered using Lipinski’s Rule of 5 and a set of undesirable chemical fragments, leaving 236 191 structures. These remaining structures were ionised using the same SPL scripts as in the study described in Paper I, and the 3D coordinates were generated using the premin script (Schrödinger LLC).

Docking was performed using Glide, storing one pose per docked structure. Compounds were required to have an ‘hbond score’ of less than -0.1. The ‘hbond score’ is the hydrogen-bond-specific component of the Glide scoring function. Since the known binders showed extensive hydrogen bonding (see Figure 14), it was reasoned that true positives should obtain some of their predicted affinity from hydrogen bonding. After docking, scoring and application of this filter, 234 297 compounds remained. The top half of these, according to the Glide SP score, were investigated further.

This was done prior to the studies described in Paper II, and at this point an alternative method was explored. The idea was to create a scoring matrix from several scoring functions, as in Paper I, and to perform PCA on the data. The first principal component was then used to rank the compounds and predict actives. The principal components are linear combinations of all scoring functions, and in that sense are composite scoring functions. This approach had been tested on a set of random compounds and known actives for fXa, with good results. A scoring matrix was therefore constructed for the 117 000 docked compounds, using seven scoring functions: Glide, EModel, and the five scoring functions implemented in CScore. The 1 109 highest-ranking compounds were selected, sorted according to the first principal component of the PCA. These were re-docked using Glide XP, and the resulting poses were visually inspected in Maestro (Schrödinger LLC). In the visual inspection, hierarchical clustering based on Tanimoto distances was used to cluster and view similar compounds together. This was done to assist in selecting a diverse set of molecules to test. As a result of this

* http://www.molsoft.com/icm/icm-macros.html#convertObject
Figure 15. The first set of thirty-two compounds acquired and tested with regard to mtRpiB inhibition and binding.
analysis, thirty-two compounds were finally chosen and acquired (see Figure 15). As is further described below, these compounds were tested regarding both inhibition of mtRpiB and binding to the active site. Unfortunately, only one potential weak binder was identified, compound 1. This binder could not be confirmed in the enzymatic assay, and it was not possible to solve an X-ray structure representing a complex with this compound.

At this point, new RpiB transition-state analogues were synthesised, and two new complex structures were obtained. The two new inhibitors were 4-phospho-D-erythronohydroxamic acid (4PEH, in the complex structure seen in PDB ID 2BES) and 4-phospho-D-erythronocarboxylic acid (4PE, PDB ID 2BET), see Figure 16. Structure-based virtual screening was repeated with this new structural information.

![Chemical structures of the mtRpiB transition-state analogues 4PE and 4PEH.](image)

Figure 16. Chemical structures of the mtRpiB transition-state analogues 4PE and 4PEH.

This time a combination of 3D pharmacophore searching and docking and scoring was employed, using UNITY (Tripos Inc.) and Glide. To identify a suitable protein target structure for the SBVS, the substrates and the two new inhibitors were docked to one of the active sites of the three available mtRpiB structures (1USL, 2BET and 2BES). The protein structure that resulted in the highest number of correct poses was chosen as the target structure. This was the complex with 4PEH ($K_i = 57 \mu M$, PDB ID 2BES). It was refined using the ProtPrep script from Schrödinger, with 4PEH in place during refinement. The compound database was a composite of all commercially available screening compounds from eleven different suppliers, with 1.6 million unique structures. A subset comprised of all compounds with a ClogP between 0 and 3.5 and a molecular weight below 400 Da was created (158,917 unique structures). These compounds were transformed into neutral 3D structures using LigPrep (Schrödinger LLC), and used to build a UNITY database.
The first pharmacophore was an eight-point (including site points) partial match query. A 2.0 Å negatively ionisable feature was placed at the phosphate of 4PEH. A 0.6 Å hydrogen bond donor-acceptor pair was placed on the hydrogen bond between His102 and 4PEH, another on the hydrogen bond between Gly70 and 4PEH, and finally a donor with two alternative acceptor sites, the partial match part, was placed on the hydrogen bond between Ser71 and the carbonyls of 4PE and 4PEH, respectively. This rather elaborate pharmacophore yielded 911 hits when used to flexibly search the UNITY database. The hits were expanded to all ionisation states, tautomers and stereo-isomers, using LigPrep. These were subsequently docked using Glide, storing nine poses per structure. In total, 9 943 poses were generated. These were then matched rigidly to the original pharmacophore. However, none of the docked poses matched the pharmacophore. Therefore, the negatively ionisable feature was removed from the pharmacophore and a second round of flexible 3D pharmacophore searching was performed, followed by docking and rigid pharmacophore filtering of docked poses. This yielded eighteen hits, from 70 339 matches to the initial flexible search. An additional fourteen compounds were selected from those matching the flexible, but not the rigid, pharmacophore filter. These thirty-two compounds were docked with Glide XP, and a final set of sixteen compounds was chosen by visual inspection (see Figure 17).

These compounds were acquired and tested regarding inhibition, in the same manner as the first thirty-two compounds. Unfortunately, again none showed any significant inhibition. They were also tested regarding binding to the active site by ESI-MS experiments, but no specific active-site binders were found.

4.2.4 Experimental testing of predicted inhibitors

Enzymatic assays
To screen the compounds for inhibition of the mtRpiB catalytic activity, an assay based on the measurement of the ribulose-5-phosphate formed as a function of time was first used. Ribulose-5-phosphate was detected using UV absorption at a wavelength of 290 nm. This turned out not to be ideal, since many of the predicted actives showed significant background absorption at this wavelength.

Due to the problem of background absorption, a second colorimetric assay was developed. In this assay the reaction was stopped at different points in time, using concentrated sulphuric acid. The amount of ribulose-5-phosphate formed was determined by adding a carbazole-cysteine solution to the reaction mixture, upon which a purple colour developed. The colouring,
which is proportional to the concentration of ribulose-5-phosphate, was quantified by measuring the absorption at 540 nm. At this wavelength, fewer of the predicted actives absorbed.\textsuperscript{155,156}

![Chemical structures of compounds](image)

**Figure 17.** The second set of sixteen compounds acquired and tested for mtRpiB inhibition and binding.

Unfortunately, none of the compounds showed any significant inhibition. Nine of the compounds could not be tested due to precipitation at the screening concentration.

**NMR binding studies**

The first thirty-two compounds, shown in Figure 15, were tested regarding binding to the active site using NMR. NMR binding experiments are not hampered by compound interference, as is the UV absorption-based assay described above. These types of experiments are also able to detect weaker binders than most biochemical assays, with no risk of false positives.\textsuperscript{157-159}
Saturation transfer difference NMR (STD-NMR) was used to detect binding. 4PEH was used as the reference compound. This method is described further in Section 4.3.2, and the experimental procedures are described in detail in the Appendix.

Compound 2 was not soluble in DMSO, and compounds 9, 12, 17, 30 and 32 were not sufficiently soluble in buffer to be visible in the NMR spectra, and could thus not be further investigated. Compounds 1 and 11 bound to the protein, according to the STD spectra. To assess whether these compounds bind in the active site, attempts were made to displace 4PEH with the test compounds. Both compounds showed partial displacement of 4PEH. When trying to displace the compounds with 4PEH instead, only compound 1 was partially displaced. No competition could be observed for compound 11. Additionally, no competition was observed when trying to displace compound 1 with compound 11, or vice versa. In summary, compound 1 was identified as a binder that bound both unspecifically and (weakly) in the active site. Compound 11 was probably a purely unspecific binder.

In spite of these results, compound 1 could not be confirmed as an inhibitor in the enzymatic assays. Neither was it possible to obtain crystals of compound 1 bound to the protein.

**ESI-MS binding studies**

In the second SBVS campaign, where compounds 33–48 were identified and acquired, NMR binding experiments were not used to test for active-site binding. Instead electrospray ionisation mass spectrometry (ESI-MS) was used. Although the physical principle of ESI-MS is different from that of NMR, the methods are similar in the type of information obtained, when used to detect binding. ESI-MS binding experiments are also useful for detecting weak binders, and the binding event is measured directly, with no risk of false positives. Basically, binding is measured by mixing the compound and the protein, and detecting the mass corresponding to the complex. ESI-MS also gives information on the stoichiometry of the binding.

Again 4PEH was used as the reference compound. Figure 18 shows ESI-MS spectra demonstrating the binding of 4PEH. The fact that there are almost no peaks corresponding to the unbound protein in the lower spectrum means that 4PEH bound fairly strong in this experiment. Note that the interactions between 4PEH and mtRpiB probably have an entirely ionic/polar character. Such interactions are favoured in the gas phase, and this is what is observed in mass spectrometry.
Figure 18. ESI-MS spectra of mtRpiB (above) and mtRpiB mixed with 4PEH (below). The theoretical mass of the protein dimer is 36 632 Da, and the mass of 4PEH is 229 Da. The broadness and number of peaks in each spectrum is due a large number of adducts, presumably ions from the buffer.

Compounds 33–48 were tested for their binding to mtRpiB. Eight compounds did show weak binding, but upon inspecting the spectra more closely, it was apparent that there were peaks corresponding to 2:1, 2:2, 2:3 and in some cases even 2:4 complexes (see Figure 19 for one example). Hence, the observed binding was probably unspecific, and there is thus no reason to believe that any of the sixteen additional compounds acquired in the second round of SBVS were mtRpiB active-site binders.
Figure 19. An example from the ESI-MS testing of the potential active-site binders. Compound 36 (80μM) and mtRpiB (5 μM) above, and mtRpiB only below. The molecular weight of the compound is 315 Da. There were discernable peaks at 36 694, 37 026, 37 354, 37 691 and 37 958 Da in the upper spectrum, probably corresponding to free protein dimer, and 2:1, 2:2, 2:3 and 2:4 complexes, respectively.

4.2.5 Discussion

Why were these attempts to identify new mtRpiB inhibitors or active-site binders through SBVS not successful? One explanation may be that although there were such compounds in the starting test library, the SBVS protocols
used were unsuccessful in identifying them. This could be due to a number of reasons, ranging from flaws in the algorithms, force fields or their implementations to errors in the preparation of the protein structures. The reason could of course also be erroneous decisions in the visual inspection of docked poses. This is common to all SBVS campaigns.

However, one can ask how many true positives there actually were in the starting library, for this particular target. The active site of this protein is rather small, it is exposed to solvent, and evolved by nature to bind phospho-sugars. The protein-ligand interactions evident in the available structures are all ionic or hydrogen-bond interactions (see Figure 14). Also, the active-site has no large hydrophobic patches. For a compound to fit in this binding pocket, probably all or most of the hydrogen bonds made by solvent molecules prior to being displaced by the synthetic binder, need to be made again by that binder. Since there are no hydrophobic interactions that add to the net free energy of binding, any loss of hydrogen bonds due to a displaced water molecule will have a considerable negative impact on the binding energy.

In an interesting paper, Hajduk et al.\textsuperscript{163} used the result from twenty-three fragment-based NMR screens to discuss the ‘drugability’ of different binding pockets. They found that the ‘ability’ of a pocket to identify small-molecule binders in an experimental screen, was negatively correlated to the polar surface area of the pocket and the number of charged residues, and that it was positively correlated to the total surface area and the apolar surface area, if this was not too large. Their ‘drugability’ model would probably predict the active site of mtRpiB to be poorly ‘drugable’. Whatever the reason, no non-sugar like active-site binders of RpiB have been reported to date.

4.2.6 Subsequent results

Since the publication of Paper III, a number of studies have been carried out. As mentioned above, the complex with phosphate was followed by two complexes with the two different transition state analogues shown in Figure 16. The observed hydrogen bonding pattern was not identical to that predicted in Paper III (see Figure 14), but the interactions and hence the predicted roles of His102 and Glu75, were confirmed.\textsuperscript{153}

The mtRpiB structure in complex with R5P and ribulose-5-phosphate was recently solved.\textsuperscript{164} In that study, an \textit{E. coli} RpiB His99Asn mutant was prepared. Histidine 99 in \textit{E. coli} RpiB corresponds to His102 in mtRpiB. This mutant showed the same $K_M$ for R5P as the wild-type protein, but the
The enzyme spermidine synthase (SpdSyn) is part of the polyamine pathway, which ultimately generates the polyamines putrescine, spermidine and spermine (see Figure 20). These molecules are present in most eukaryotic cells, including all parasitic protozoans.  

The polyamines are essential for eukaryotic cell growth and differentiation. In this, their binding to and stabilisation of the negatively charged DNA is believed to be important. Additionally, spermidine participates in the activation of the eukaryotic initiation factor eIF5 through hypusinylation.

The polyamine pathway has been suggested as a target in treating both cancer and protozoan infections, including malaria. The cancer hypothesis has not been successful so far, but the protozoan infection hypothesis has shown more promise. The ornithine decarboxylase (ODC) inhibitor α-difluoromethylornithine is used clinically to treat sleeping sickness caused by Trypanosoma brucei gambiense. This drug has also
been shown to have effect on *Plasmodium falciparum*. Wider use of the drug against different protozoans is, however, limited by the very high turnover of the target enzyme, ODC. Since the drug is an irreversible mechanism-based inhibitor, rapid turnover of the target requires a constant supply of the drug to the patient. The turnover of ODC in *Trypanosoma brucei gambiense* is much lower than in other protozoans.

Other inhibitors of AdoMetDC and ODC have been shown to inhibit Pf growth *in vitro*. In 2005, Haider et al. published a study in which some simple amines were tested with regard to their *in vitro* inhibition of SpdSyn as well as Pf growth in cell cultures. The most potent SpdSyn inhibitor was 4-methyl-cyclohexylamine (4MCHA), with an enzyme inhibition IC$_{50}$ of 1.4 μM and a growth inhibition EC$_{50}$ of 34 μM.

In the current work SBVS was used to try to identify novel active-site binders of Pf SpdSyn. The goal was to find starting points for the development of new inhibitors of Pf SpdSyn, and ultimately new drug candidates to treat malaria.

### 4.3.1 Structure-based virtual screening

The first structure of Pf SpdSyn, a complex with one of the products, MTA, was solved in 2007 (PDB ID 2HTE). Later the same year, the complex with the transition state analogue S-adenosyl-1,8-diamino-3-thio-octane (AdoDATO) was solved (PDB ID 2I7C). These two structures were the starting point for SBVS against this target. After the SBVS study was completed, the apo structure (PDB ID 2PSS), the complex with dcAdoMet alone (PDB ID 2PT6) and together with 4MCHA (PDB ID 2PT9), the complex with spermidine (PDB ID 2PWP), and the complex with spermine (PDB ID 3B7P) were released to the PDB.

Through docking of known binders and inhibitors, and evaluation of the resulting poses, the complex with AdoDATO was chosen as the target structure in the virtual screen. This time, the compound test library was a composite of commercially available screening compounds from thirteen different suppliers. It contained a total of 3 million compounds and 2.8 million unique structures. Those with molecular weight less than 550 Da, and a number of rotatable bonds less than eleven were filtered out. The limit on rotatable bonds was applied to exclude too flexible compounds, which would cause problems in the generation of conformers in the pharmacophore filtering step (see below). More elaborate filters were not applied since only a limited number of compounds would be acquired and tested. This meant...
that all compounds that would be acquired would first be subjected to visual inspection and evaluation.

After filtering, 2.7 million structures remained. In the translation to 3D, an additional 100,000 structures were removed. The remaining 2.6 million structures were used to create a Phase (Schrödinger LLC) database, which was the starting point for the SBVS.

---

**Figure 21.** The virtual screening protocol described in Paper IV.

The SBVS approach was similar to the second mtRpiB virtual screen, and similar to some previously published SBVS studies\(^{186-188}\) (see Figure 21). In the first step, Phase\(^{189-191}\) was employed. This is a 3D pharmacophore modelling software package, used to generate and search databases of pre-computed conformational ensembles. The searches can include excluded volumes. These are defined as spheres where atoms from the ligands are forbidden when matching the ligand to the other pharmacophore features.
Two separate pharmacophores were used, one based on the adenosine part of AdoDATO, and one based on the putrescine part (see Figure 22). These were constructed manually from the AdoDATO complex. Excluded volumes representing the protein atoms lining the active site were used.

![Figure 22. The two Phase pharmacophores used in the SpdSyn SBVS. Light blue spheres are hydrogen-bond donor, red spheres are hydrogen-bond acceptor, green spheres are hydrophobic, dark blue spheres are positively ionisable and orange toruses are aromatic features. AdoDATO is shown in ball and stick. The adenosine-based pharmacophore (left) was applied requiring at least four features to match, three of which had to be the features associated with the adenine ring. The putrescine-based pharmacophore (right) was applied requiring both features to match.](image)

The pharmacophores were used to search the Phase database. The adenosine-based pharmacophore resulted in 5,489 hits, and the putrescine-based pharmacophore resulted in 1,866 hits. As for mtRpiB, these compounds were expanded to all feasible tautomers, stereo-isomers and ionisation states, using LigPrep and Epik. The resulting structures were docked to the target protein structure, using Glide. Nine poses were stored for each structure. These poses were matched against the original pharmacophores, keeping the compounds rigid. In the version of Phase used here, 2.0, it was not possible to turn off translation and rotation of the compounds. Therefore, the pharmacophore matching was somewhat less stringent than originally intended in the SBVS protocol.
Figure 23. The 28 compounds predicted to bind in the SpdSyn active site.

The rigid pharmacophore filter resulted in 1866 remaining compounds, compared with zero in the mtRpiB virtual screen. These were re-docked using Glide XP, and the poses were reviewed manually. Both Glide XP and modified MASC scores, as defined in Paper II, were used, to guide in the final selection of hits. However, since the total number of compounds was manageable all compounds were reviewed manually, regardless of their scores. Twenty-eight compounds were finally chosen and acquired for testing (see Figure 23).
4.3.2 Experimental testing of predicted binders

NMR binding experiments were used to confirm active-site binding as we did not have access to an enzymatic assay to measure SpdSyn inhibition. The known binder MTA (IC$_{50} = 159 \pm 27$ μM)\textsuperscript{166} was used as both positive control and competitor. For this purpose, Pf SpdSyn was expressed in *E. coli* and purified according to existing protocols.\textsuperscript{185,193}

The NMR binding experiments started with the collection of conventional $^1$H 1D reference spectra for all compounds. Compounds 50 and 64 were not soluble in buffer, and could not be further investigated. The other twenty-six compounds were tested regarding binding to SpdSyn, using STD-NMR.\textsuperscript{160}

The principle of STD-NMR is to saturate the protons of the protein with repeated radio frequency pulses at a single protein resonance. Saturation will rapidly spread throughout the entire protein by spin diffusion. If a molecule interacts with the protein, inter-molecular spin diffusion will transfer the saturation to that molecule. By collecting two NMR spectra, with and without saturation, and comparing them, such inter-molecular spin diffusion and thus binding, can be measured. STD-NMR can be used to detect weak binders, but too potent or covalent binders will not be detected. This is because the molecules to which saturation has been transferred will not dissociate from the protein and be measurable in solution during the time of the experiment. The site of binding can not be determined from a single STD-NMR experiment. Therefore, a known binder with a known binding site is needed to perform competition experiments. Methylthioadenosine was used for this purpose.

The STD-NMR experiments identified nine binders. Of these, compounds 49 and 74 clearly competed for binding with MTA, compounds 58 and 62 showed partial competition, compounds 53, 65 and 73 showed some competition, compound 52 showed stronger binding upon addition of MTA and compound 59 showed no competition. The STD-NMR results for compound 49 are shown in Figure 24.
Figure 24. STD-NMR results for compound 49. The top spectrum is the aromatic region of the $^1$H 1D spectrum of MTA and 49. The middle spectrum is the same region of the STD-NMR spectrum of SpdSyn and 49, showing that the compound clearly binds to the protein. The bottom spectrum shows the same region of the STD-NMR spectrum for SpdSyn, 49 and MTA. The decrease in the peak resulting from compound 49 shows that MTA and 49 have a common binding site, namely the active site.

To confirm and clarify these results, a second NMR binding technique was applied, $T_{1,\rho}$ relaxation filter NMR. The principle of this technique is to
mix the compound and the protein in approximately equimolar amounts, and
detect binding as the decrease in the signal from the binding molecule. This
decrease is caused by the increase in relaxation rate in slowly tumbling
molecules. A small molecule binding to a macromolecule tumbles more
slowly than a small molecule free in solution. The greater the decrease in the
signal, the more compound is bound to the protein, and hence the stronger
the binding. This method can detect strong as well as weak binders, but still
requires competition experiments with a known binder to confirm the site of
binding.

Figure 25. T₁,ρ relaxation filter NMR results for compound 49. The top spectrum is
from compound 49 only, the middle spectrum is from compound 49 and protein, and
the bottom spectrum is from compound 49, protein and MTA. The height of the
relevant peak is given in arbitrary units.

In the T₁,ρ relaxation filter experiments, the protein and compound concen-
trations were both 30 μM. In the competition experiments, the MTA
concentration was 1880 μM. Hence, an almost 63-fold molar excess of MTA
was used. At these concentrations, with MTA present at a concentration of more than ten times its IC$_{50}$ value, the competition in the active site should be almost complete.

![Chemical Structure](image)

**Figure 26.** T$_{1\rho}$ relaxation filter NMR results for compound 52. The top spectrum is from the compound only, the middle spectrum is from the compound and the protein, showing some binding, and the bottom spectrum is from the compound, the protein and MTA, showing strong binding.

The T$_{1\rho}$ relaxation filter experiments confirmed compounds 49 and 74 as active-site binders, as they were completely competed out by MTA. The results for compound 49 are shown in Figure 25. The middle spectrum shows that the compound binds to the protein, and the bottom spectrum shows that this binding is completely abolished by the addition of 1880 μM MTA. Compounds 58, 62 and 65 were confirmed as binders, partially
competing with MTA. This meant that they bound in both the active site and at other sites on the protein. Compound 53 could not be confirmed to bind in the active site. Weak binding was evident, but the measured competition, if any, was too close to the noise to be significant. Compound 59 was confirmed as a completely non-competitive, most probably unspecific binder. Finally, compounds 73 and 52 were both shown to bind more strongly in the presence of MTA (see Figure 26). This was much more pronounced for compound 52 than for compound 73.

4.3.3 Results

In total, seven of the twenty-eight compounds predicted to bind in the active site of Pf SpdSyn were confirmed as active-site binders. The five compounds that showed competition with MTA were all predicted by the adenosine-based pharmacophore, and the binding poses predicted by docking would result in competition with MTA. The two compounds that bound more strongly when MTA was added were predicted by the putrescine-based pharmacophore. Their predicted binding poses would not necessarily require them to compete for binding with MTA. The predicted binding poses of compounds 49 and 52 are shown in Figure 27.

![Figure 27. Predicted binding poses of compound 49 (left) and compound 52 (right).](image)

Regarding the stronger binding, Tamu Dufe et al. saw something similar in their structural studies of PfSpdSyn. They were not able to obtain crystals with 4MCHA, despite the fact that the compound concentration was 2.5
mM. Only when the substrate dcAdoMet was added first, were they able to see 4MCHA in the active site. Their interpretation was that dcAdoMet affects the stability and binding characteristics of the rest of the pocket upon binding. This is in agreement with the current observations regarding the binding of compounds 52 and 73 together with MTA.

4.4 Discussion

The approach taken in this SBVS was quite similar to the SBVS against mtRpiB. The most important differences were probably the type of binding pockets involved, as discussed above, and improvements in the algorithms employed. Also, a larger starting test library of compounds was used. Finally, the starting target structure was a complex with a potent inhibitor covering the entire active site, as opposed to an apo structure or a complex with a low-potency inhibitor, as was the case for mtRpiB.

I believe that the approach taken here, with rather severe filtering of the test library using 3D pharmacophores, is well suited for identifying a few true positives from a large starting test library. When trying to identify some starting points for a target with no previously known inhibitors, while testing only a handful of compounds experimentally, it is not important to find all the actives in the test library. What is most important is finding some true actives. In other words, it is more important to avoid false positives than to avoid false negatives. Hence, the work described here focused on identifying compounds able to form the same interactions as an experimentally verified inhibitor with a known binding pose. Many other compounds in the test library were probably able to interact with the target through other protein conformations, and/or through other sets of specific interactions. These compounds would be false negatives in the current approach.

On the other hand, if the purpose of SBVS is to predict a larger set of molecules (10 000 to 100 000 compounds), suitable for HTS, it would be more important to avoid false negatives. The reason is that the goal of an experimental screen of an entire compound collection is to find all or most actives in that collection. Approaches such as ensemble docking, to model target structure flexibility and induced fit, would be more important in this case.
5 Concluding remarks

The present study has contributed to the field of structure-based virtual screening. It has also resulted in the identification of novel binders in the active site of *Plasmodium falciparum* spermidine synthase, a potential target to treat malaria. Furthermore, molecular modelling has been applied in the characterisation of *Mycobacterium tuberculosis* ribose-5-phosphate isomerase. The main conclusions can be summarised as follows.

- A method for structure-based virtual screening was developed and evaluated. Scores of known actives were used to train multivariate classifiers, which were then used to predict new binders. The resulting enrichment of actives was significantly increased compared with consensus scoring or single scoring functions.

- So-called multiple active-site correction or size normalisation was evaluated. Transformation of scores from a single scoring function often increased the enrichment compared with the untransformed scoring function.

- Scores from structure-based virtual screening often show a significant correlation to the size of the molecules being scored. This is a source of false positives.

- According to docking studies, *Mycobacterium tuberculosis* ribose-5-phosphate isomerase probably binds the ring-closed sugar substrate and catalyses ring opening. Histidine 102 is predicted to be important in this catalysis.

- Through structure-based virtual screening, seven novel binders, binding in the active site of *Plasmodium falciparum* spermidine synthase were identified. Two of these compounds bind cooperatively with the product methylthioadenosine, while five compete for binding with this product.
Acknowledgements

First of all, I would like to thank my supervisor, Prof. Anders Karlén, for all his ideas and advice and for sharing his knowledge in computational and medicinal chemistry, scientific writing and science in general. He is one of the best managers I have met: indulgent when necessary, firm when necessary, but always positive and supportive. Thank you also for enabling me to combine my PhD studies with my day job.

I would also like to thank Prof. Anders Hallberg for making it possible for me to become a member of his excellent department, and for sharing his vast knowledge in medicinal chemistry through courses, projects and in other ways. Thanks to Prof. Mats Larhed for sharing his great energy and knowledge in organic and medicinal chemistry, and for his calm guidance through the intricacies of the academic world. I would also like to thank the other staff members, especially Gunilla Eriksson for helping me with everything, from finding a bike to ensuring I was paid my salary, and for handling all the impossible invoices from strange Russian chemical companies.

I would also like to express my thanks to all my co-authors, especially: Dr. Ulf Norinder, Per Lidén and Dr. Henrik Boström, for perfect collaboration on the multivariate classifier paper. Special thanks to Eva Stjernschantz, first my MSc student and later colleague, for the LIE paper, for all her great scientific work, and extremely interesting discussions on every possible subject. Thanks to Dr Annette Roos and Prof. Sherry Mowbray for all their work on RpiB, for testing all my compounds again and again, and for their expertise in structural biology. Thanks to Magnus Gäredal for the great work on his Master’s dissertation that led to the spermidine synthase paper. Thanks also to Dr Johan Schultz, my colleague for the past eight years, both for the work with RpiB and spermidine synthase. He has taught me much about biophysics and NMR, beer, good food, travelling and everything else that makes life worth living.

Thanks to all members of the computational chemistry group, it has been a pleasure working with you: Dr Anna Ax, Dr Shane Petersen, Dr Christian Sköld, Daniel Muthas, Anneli Nordqvist, Dr Yogesh Sabnis, Martin Lindh
and Dr Aparna Vema. Thanks also to Christian, Daniel, Aparna and Martin for help with proof-reading this thesis.

I would also like to thank all my other past and present fellow PhD students at the Department of Organic Pharmaceutical Chemistry. It was great to be part of such an energetic scientific environment. Thanks also to MSc students Helena Hammar and Ludvig Holmberg.

I would also like to express my appreciation to my colleagues at Pharmacia, Biovitrum and iNovacia. I learned everything I know about drug discovery, wine tasting and exuberant parties while working together with you for the past eight years.

Thanks to my managers, who have supported me and allowed me to combine work with my PhD studies: Mats Kihlén, also my assistant supervisor and the most enthusiastic and productive person I know, fixing everything from multi-million contract negotiations to tea cookers, without breaking a sweat; Dr. Martin Norin, heading the best drug discovery departments for the past ten years; Dr Anna-Lena Gustavsson, for her support and encouragement, and for everything she taught me about wine, work and life in general.

Special thanks to all my other friends and colleagues at the former Structural Chemistry department (R.I.P.) at Pharmacia/Biovitrum. It was a wonderful scientific and social environment.

Thanks to my colleagues in the computational chemistry group: Dr Maria Ek, a great room-mate who shares my poor sense of direction; Dr Evert Homan, the nicest Dutchman I know; Dr Carmen Medina, always engaged and enthusiastic; Dr Jerk Vallgård, Tripos expert and relentless project manager; Dr Jonas Nilsson, expert diver and chemometrician; Dr Erik Nordling, a cool guy and fellow ICM nerd; and Dr Peter Brandt, an avid chemist and really smart guy.

Thanks also to all the former members of the macromolecule group who opened my eyes to the tantalizing world of biophysics: Dr. Agneta Tjernberg, a real expert in mass spectrometry and human relations, for helping me with mass-spectrometry experiments on RpiB, for everything she has taught me about MS and biophysics, and for her encouragements; Dr. Natalja Markova, the best Russian biophysicist and double doctor I know, for helping me with calorimetric and other biophysical experiments on RpiB and spermidine synthase, as well as for her encouragement and guiding to St Petersburg; Carina Norström, lab wizard and a really nice person, for her help with spermidine synthase experiments, for rescuing the protein from the bin, and help with other experiments over the years; Dr. Johan Weigelt, for
suggesting spermidine synthase as a virtual screening target and for allowing us to produce the protein at SGC, and for all his energy and ideas over the years; Dr. Mats Wikström, for the collagen paper, for being the most far-out person I know, for all the good parties and all the deep truths; Dr. Dan Hallén, for the AUC measurements on RpiB and for always being willing to help with calorimetric experiments; Dr. Stefan Svensson, for the LXR paper and for being such fun to work with; Dr. Thomas Lundbäck, an extremely dedicated drug discovery project leader and a great scientist; Dr. Toshiaki Nishida, for helping with NMR experiments for spermidine synthase, for all our discussions about drug discovery and the world in general; Fredrik Lindqvist, protein expert and cultural elite; Kristina Zachrisson, expert protein analyst and fellow theatre frequenter; Dr. Jonas Uppenberg, a fount of X-ray expertise and life wisdoms. Thanks also to Dr. Thomas Åkerud, Åsa Enhörning, Dr. Björn Elleby, Dr. Maria van Dongen and Dr. Derek Ogg.

Thanks to my fellow computer nerds in the informatics group: Mats Dahlberg, with probably the highest IQ I have encountered. Thanks for all the well-spent hours in various pubs on Kungsholmen, the trip to Rome with the nice musical evening, wine-tasting in Budapest, solving all my Linux problems and many other things. Thanks to Dr. John Marelius, room-mate, expert programmer and wildlife fanatic; Mikael Mireiborn, all things databases, Star Wars and WoW; Sarah Hunter, a really nice Englishwoman with a knack for databases; Dr. Torsten Sejlitz, Swedish Latino, for introducing me to aged rum.

Thanks also to the current and former members of the CAMP group: Per Persson, the nicest communist I know; Dr. Bengt Norén, the humorous grand old man of mass spectrometry; Dr. Thomas Wehler, a.k.a. Herr CAMP; Elisabeth Olsson, for helping with all the tricky small-molecule NMR interpretations; Åsa Brunström, social engine and party genius; Joakim Staffas, analytical chemist and secret agent, for answering all my spectrometry questions; Lars Gustavsson, globetrotter and chronicler.

Thanks also to all the protein experts in the former TEP group, Dr. Kristina Köhler van Alstine, Robert Svensson, Dr. Erik Holmberg, Dr. Joakim Nilsson, Dr. Susanne Gräslund and others.

I would like to extend my thanks to Dr Lena Jendeberg, an excellent project manager, and Dr Tove Östberg, for their collaboration in the dERR triple mutant study and other projects.

Thanks also to all my other colleagues at iNovacia, probably the best CRO in the world: Dr. Thomas Olin, hard-working CEO and Gotland jet-setter;
Dr. Isabel Climent-Johansson, hot-blooded Española and fellow project leader; Henrik Alexandersson, pro poker player, for all our late-night political and philosophical discursions, and for keeping me in shape; Elisée Wiita, an assay development wizard with the right attitude towards dancing; Peter Halkjaer-Knudsen, travelling salesman, brilliant networker and business development Nestor; Carlo Zaniol, Italian robotician, cook and automation expert; Ingrid Almlöf, an excellent cell biologist and assay developer; Andreas Svahn, problem solver with the best % CV I have seen; Jan Holmberg, drug expert and storyteller; Åsa Slevin, a cynic with a great sense of humour; Hanna Axelsson, a.k.a. Ms. High-throughput screening; Mari Siggesjö, administrative genius; Dr Åsa Edman, dedicated biochemist and project leader; Dr. Jan Vågberg, a living medicinal chemistry encyclopaedia; Dr. Katarina Färnegårdh, an excellent medicinal chemist and fellow screen-to-hitter; Mattias Jönsson, who can make more molecules than anyone else; Dr. Charlotta Löfberg, a.k.a. Ms. ADME; Helena Almqvist, determined ion-channel expert and runner; Nasrin Najafi, who knows her GPCRs; Dr. Philippe Verwaerde, Gallic soccer aficionado, who knows everything worth knowing about assays and screening; Charlotta Häggström, analytical chemist and Espanophile; Dr. Karl-Johan Fasth, resident mechanic.

I also want thank my family for always being there when I need them. First of all my mother, Ann, who has supported my studies and scientific ambitions ever since I was a four-year old boy wanting to become a “naturforskare”. Thank you also to my stepfather, Conny, and to my wonderful siblings, nieces and nephews: Matilda and Stefan and their children Amanda, Linnéa and William; Gert-Ove; Linda; Anna-Lill and little Anton. Thanks also to all my part-time siblings. Thanks to my father, Roger, and Maj, and to my half-siblings Linnéa and Martin.

Most of all, thank you, beautiful Maria, my love, my wonderful wife and lifetime partner. You are my best friend, greatest support, wisest advisor and sharpest discussion partner. Without you, this thesis would not have been written.
References


84. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "Receiver operating characteristic" Curve


148. Sorensen, K. I.; Hove-Jensen, B. Ribose catabolism of *Escherichia coli*: Characterization of the RpiB gene encoding ribose phosphate isomerase B and


Appendix

Experimental procedures for biophysical evaluation of predicted binders to mtRpiB

NMR binding experiments

All NMR measurements were performed using a Varian Unity Inova 600 MHz NMR spectrometer equipped with a cryogenically cooled probe and a flow cell insert. The samples were loaded into a 96-well plate placed on a Peltier cooling device in order to keep samples waiting to be measured at 6 °C. The NMR experiments were all performed at 20 °C.

A stock solution of the reference compound, 4PEH, was prepared in buffer. Compounds 1-32 were dissolved in DMSO-d6, to a concentration of 25 mM. The buffer used in all experiments was 20 mM Tris-d11, pH=7.5, 150 mM NaCl, in D2O.

Firstly binding of 4PEH was established. The binding signal was somewhat weaker than would be expected from the Kᵢ value of 57 μM, and it was necessary to use 20 μM protein and 1 mM compound to obtain a strong binding signal. This signal was, however, stable over more than 16 h, and no degradation of the protein signals was observed in this timeframe.

To screen compounds 1-32 for binding, 1H 1D spectra were first collected for all compounds at 250 μM. The compounds that were sufficiently soluble were tested for binding to the protein, by mixing 10 μM protein and 250 μM compound. To test for competition against 4PEH by the test compounds, a mixture of 20 μM protein, 500 μM 4PEH and 250 μM test compound was used. To test for competition against the test compounds by 4PEH, a mixture of 20 μM protein, 10 mM 4PEH and 250 μM test compound was used. Finally, when competing compounds 1 and 11 against each other, 10 μM protein, together with 250 μM compound 1 and 250 μM or 10 mM compound 11 was used.
ESI-MS binding experiments

The ESI-MS measurements were performed using a quadrupole time-of-flight Ultima API instrument (Waters, Milford, MA) operating in positive ion mode. The samples were pumped into the electrospray ion source at a flow rate of 5 μl/min. The mobile phase was 10 mM ammonium acetate, pH 7. For the ESI interface, the capillary voltage was set to 3000 V, cone voltage 100 V, source block temperature 45 °C, desolvation temperature 60 °C, desolvation gas flow 200 l/h, and cone gas flow 50 l/h. An elevated instrument pressure in the source region, as well as collisional cooling, was applied in order to preserve the noncovalent interactions in the gas phase. The data were acquired by scanning over the m/z range 1000-5000, with data accumulation of 2 s per spectrum and an inter-scan time delay of 0.1 s. Mass spectra were averaged over 80 scans. The spectra were deconvoluted over the mass range 36000-38500 Da.

All experiments were performed at room temperature. Compounds 33–48 were dissolved in a 1:1 mixture of acetonitrile and ammonium acetate, to a stock concentration between 400 μM and 2.5 mM, depending on the apparent solubility of the compound. The buffer used in all experiments was 20 mM ammonium acetate. In the binding experiments, the final compound concentration was 50 μM and the final protein concentration was 5 μM.
A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy”.)