UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 235

# Computational Modelling in Drug Discovery

## *Application of Structure-Based Drug Design, Conformal Prediction and Evaluation of Virtual Screening*

MARTIN LINDH

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017

ISSN 1651-6192
ISBN 978-91-513-0049-8
urn:nbn:se:uu:diva-328505

Dissertation presented at Uppsala University to be publicly examined in B/B42, Husargatan 3, Uppsala, Friday, 13 October 2017 at 09:00 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English. Faculty examiner: Professor Antti Poso.

## Abstract

Lindh, M. 2017. Computational Modelling in Drug Discovery. Application of Structure-Based Drug Design, Conformal Prediction and Evaluation of Virtual Screening. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy* 235. 47 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0049-8.

Structure-based drug design and virtual screening are areas of computational medicinal chemistry that use 3D models of target proteins. It is important to develop better methods in this field with the aim of increasing the speed and quality of early stage drug discovery.

The first part of this thesis focuses on the application of structure-based drug design in the search for inhibitors for the protein 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), one of the enzymes in the DOXP/MEP synthetic pathway. This pathway is found in many bacteria (such as *Mycobacterium tuberculosis*) and in the parasite *Plasmodium falciparum*.

In order to evaluate and improve current virtual screening methods, a benchmarking data set was constructed using publically available high-throughput screening data. The exercise highlighted a number of problems with current data sets as well as with the use of publically available high-throughput screening data. We hope this work will help guide further development of well designed benchmarking data sets for virtual screening methods.

Conformal prediction is a new method in the computer-aided drug design toolbox that gives the prediction range at a specified level of confidence for each compound. To demonstrate the versatility and applicability of this method we derived models of skin permeability using two different machine learning methods; random forest and support vector machines.

*Keywords:* drug discovery, docking, virtual screening, tuberculosis, conformal prediction

*Martin Lindh, Department of Medicinal Chemistry, Organic Pharmaceutical Chemistry, Box 574, Uppsala University, SE-75123 Uppsala, Sweden.*

ISSN 1651-6192
ISBN 978-91-513-0049-8
urn:nbn:se:uu:diva-328505 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-328505)

*This makes the rational design of new inhibitors of DXR difficult at best*[1]

Mercklé *et al*. **2005**

# List of Papers

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

I **Design, Synthesis, and X-ray Crystallographic Studies of α-Aryl Substituted Fosmidomycin Analogues as Inhibitors of *Mycobacterium tuberculosis* 1-Deoxy-D-xylulose 5-Phosphate Reductoisomerase**
Mounir Andaloussi, Lena M. Henriksson, Anna Wieckowska, Martin Lindh, Christofer Björkelid, Anna M. Larsson, Surisetti Suresh, Harini Iyer, Bachally R. Srinivasa, Terese Bergfors, Torsten Unge, Sherry L. Mowbray, Mats Larhed, T. Alwyn Jones, Anders Karlén
*Journal of Medicinal Chemistry* **2011,** *54* (14), 4964-4976

II **Substitution of the phosphonic acid and hydroxamic acid functionalities of the DXR inhibitor FR900098: An attempt to improve the activity against *Mycobacterium tuberculosis***
Mounir Andaloussi, Martin Lindh, Christofer Björkelid, Surisetti Suresh, Anna Wieckowska, Harini Iyer, Anders Karlén, Mats Larhed
*Bioorganic & Medicinal Chemistry Letters* **2011,** *21* (15), 5403-5407

III **DXR Inhibition by Potent Mono- and Disubstituted Fosmidomycin Analogues**
Anna M. Jansson, Anna Więckowska, Christofer Björkelid, Samir Yahiaoui, Sanjeewani Sooriyaarachchi, Martin Lindh, Terese Bergfors, Shyamraj Dharavath, Matthieu Desroses, Surisetti Suresh, Mounir Andaloussi, Rautela Nikhil, Sharma Sreevalli, Bachally R. Srinivasa, Mats Larhed, T. Alwyn Jones, Anders Karlén, and Sherry L. Mowbray
*Journal of Medicinal Chemistry* **2013,** *56* (15), 6190-6199

IV **Toward a Benchmarking Data Set Able to Evaluate Ligand- and Structure-based Virtual Screening Using Public HTS Data**
Martin Lindh, Fredrik Svensson, Wesley Schaal, Jin Zhang, Christian Sköld, Peter Brandt, Anders Karlén
*Journal of Chemical Information and Modeling* **2015**, *55* (2), 343–353

V **Predicting the Rate of Skin Penetration Using an Aggregated Conformal Prediction Framework**
Martin Lindh, Anders Karlén, Ulf Norinder
*Molecular Pharmaceutics* **2017**, *14* (5), 1571-1576

[1] Reprints were made with the permission of the respective publishers.

# Additional publications

I. **Trisubstituted Imidazoles as *Mycobacterium tuberculosis* Glutamine Synthetase Inhibitors**
Johan Gising, Mikael T. Nilsson, Luke R. Odell, Samir Yahiaoui, Martin Lindh, Harini Iyer, Achyut M. Sinha, Bachally R. Srinivasa, Mats Larhed, Sherry L. Mowbray, Anders Karlén
*Journal of Medicinal Chemistry,* **2012**, *55* (6), 2894–2898

# Contents

Introduction .......... 13

- Three Projects Utilising Computational Techniques Within Early Drug Development .......... 13
- Proteins as Targets for Medicinal Chemistry .......... 14
- Methods used in Structure-Based Drug Design .......... 16
- Virtual Screening .......... 17
- Ligand-Based Virtual Screening .......... 17
- Structure-Based Virtual Screening .......... 17

Computer-Aided Drug Design and Discovery (Papers I, II, III) .......... 20

- Rational Approaches to Pathogen Inhibitor Design (RAPID) .......... 20
- The DOXP/MEP Pathway and Isoprenoid Biosynthesis .......... 20
- 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (DXR) .......... 21
- Fosmidomycin – a known DXR inhibitor .......... 21
- Targeting *Mt*DXR with Fosmidomycin Analogues .......... 22

Benchmarking Virtual Screening (Paper IV) .......... 34

- Selecting the VS Protocol .......... 34
- Benchmarking VS Using HTS-Data .......... 35
- Identifying HTS-Data Useful for Benchmarking VS .......... 36

Aggregated Conformal Prediction (Paper V) .......... 38

Concluding Remarks .......... 41

Acknowledgements .......... 43

# Abbreviations

| | |
|---|---|
| AA | amino acid |
| ACP | aggregated conformal prediction |
| ASP | aspartic acid |
| CP | conformal prediction |
| DEKOIS | demanding evaluation kits for objective in silico screening |
| DMAPP | dimethylallyl diphosphate |
| DOXP | 1-deoxy-D-xylulose-5-phosphate |
| DUD | directory of useful decoys |
| DUD-E | directory of useful decoys - enhanced |
| DXP | 1-deoxy-D-xylulose-5-phosphate |
| DXR | 1-deoxy-D-xylulose-5-phosphate reductoisomerase |
| *E. coli* | *Escherichia coli* |
| GLU | glutamic acid |
| HTS | high-throughput screening |
| IPP | isopentenyl diphosphate |
| LBVS | ligand-based VS |
| MEP | 2C-methyl-D-erythritol 4-phosphate |
| *Mtb* | *Mycobacterium tuberculosis* |
| *Mt*DXR | DXR from *Mtb* |
| MD | molecular dynamics |
| MIC | minimum inhibitory concentration |
| MM | molecular mechanics |
| NADPH | nicotinamide-adenine dinucleotide phosphate |
| NMR | nuclear magnetic resonance |
| PDB | protein data bank |
| PfDXR | DXR from *Plasmodium falciparum* |
| RAPID | rapid approaches to pathogen inhibitor discovery |
| RF | random forest |
| SAR | structure activity relationship |
| SBVS | structure-based VS |
| SVM | support vector machine |

| | |
|---|---|
| TB | tuberculosis |
| TRP | tryptophan |
| VS | virtual screening |
| QM | quantum mechanics |

# Introduction

## Three Projects Utilising Computational Techniques Within Early Drug Development

In this thesis, there is first a brief introduction for people unfamiliar with the field and then a presentation of three projects related to early drug discovery for which computational methodology was used. Papers **I-III** concern an early drug discovery project that focused on the protein target 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR). Paper **IV** concerns an investigation into the use of high-throughput screening (HTS) for benchmarking virtual screening (VS) techniques. Paper **V** discusses an example of aggregated conformal prediction (ACP) when applied to predictions of skin permeability. In the description of these projects, I will focus on the aspects I find most noteworthy and when possible avoid duplication of details found in the articles.

## Proteins as Targets for Medicinal Chemistry

The lock and key model is often used in textbooks to describe drug action. In this analogy, a protein (receptor) is described as a lock with a particular shape. The small compound (ligand) acting on the target protein is seen as a key that binds to the complementary shape and unlocks the desired response. Following the unlocking of the protein, a signal is transmitted leading to a biological response that affects the disease or the symptom that the drug was intended to treat (Figure 1).

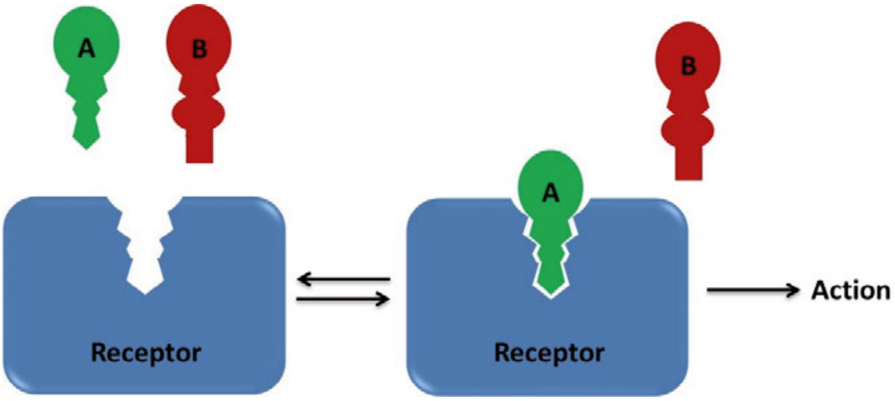


**Figure 1**. The lock and key analogy. Drug **A** binds to the receptor while molecule **B** cannot.

Three-dimensional (3D) models of proteins can be created using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. These 3D models describe the positions of the atoms that make up the protein and conform well to the lock and key analogy of drug action. In the protein data bank (PDB) there are numerous examples of proteins with and without the relevant ligand.[2]

It has long been known that shape complementarity is only one aspect of drug-protein interactions. Intermolecular forces between the ligand and the protein contribute to the strength of the interaction. These forces can be described in terms of the change in enthalpy ($\Delta H$). One example of this type of interaction is the hydrogen bond. A well positioned O-H···O hydrogen bond has been determined to bind at about 21 kJ/mol.[3] However, many hydrogen bonds do not contribute much to the binding strength as the energy value is set relative to a situation where there is no bonding. Many, if not most, ligand-protein hydrogen bonds replace hydrogen bonds that are already present between the protein and water. It is, therefore, important to incorporate the surrounding environment (usually solvent water) in models of drug-protein interaction. Another well-known oversight in the simplified lock and key model is that atoms in both the protein and the ligand are constantly moving. The importance of movement is commonly quantified in terms of entropy. In water, small compounds (ligands) move around quickly without much restriction. When binding to a macromolecule such as a protein, the compound's movement will become more restricted, the number of possible states will decrease, and the entropy of the system will be lower. On the other hand, a binding event (when a ligand is bound to a protein) often releases water molecules that were bound to the protein, which might increase their movement and thus increase the entropy of the system.

The combination of enthalpy and entropy is what drives events to occur. This was elegantly formulated by Willard Gibbs in the 1870s. Gibbs' free energy, a variable set up to evaluate whether a reaction or event occurs spontaneously, is described for constant temperatures in Equation 1. The link between Gibbs' free energy and inhibition is described by Equation 2. The inhibition constant (Ki) is often used in the experimental evaluation of the strength of ligand protein binding. The relationship between the change in Gibbs' free energy (ΔG) and the inhibition constant is expressed in Equation 2.

$$\Delta \mathrm{G} = \Delta \mathrm{H} - \mathrm{T} \times \Delta \mathrm{S}$$

**Equation 1**. Gibbs' free energy, where ΔH is the change in enthalpy, T is the temperature in Kelvin, and ΔS is the change in entropy.

$$K_i = e^{\wedge(\Delta G/(RT))}$$

**Equation 2**. The relationship between the change in Gibbs' free energy and the inhibition constant.

# Methods used in Structure-Based Drug Design

Ligand-protein binding is a complex subject. The systems in question can, however, be quite accurately described at an atomic level. It is possible to describe a protein-ligand system in terms of the atoms involved and also, with a certain degree of precision, to determine their relative positions in space. With experimental techniques such as X-ray crystallography, coordinate values can be determined for the atoms in a protein and its surroundings. The covalent bonds between the atoms can be estimated with a high degree of certainty based on the distance between the atoms. We can also describe such a system using quantum mechanical equations. I believe it is reasonable to assume that if we had the mathematical tools and the computational power to solve these equations we could perform simulations using quantum dynamics to describe the system with such accuracy that the early stages of drug discovery would be a simple task. Such calculations are, however, far out of the reach of our current tools. QM calculations are used in modern drug discovery but are limited to smaller systems.

In order to run calculations and simulations of more complex systems, atoms are instead described as moving in accordance with Newtonian laws. In molecular mechanics (MM) the energy of a molecule (or set of molecules) is described by a force field. In one analogy of MM, atoms are described as spheres connected by springs. The force constants are empirical parameters that are typically derived from physical measurements or QM calculations. In most of the calculations I have used the OPLS 2005 force field. This series of force fields has been gradually developed since 1988.[4] Interactions mediated by covalent bonds are described by equations for stretching motions, angular bending and torsional rotation. For nonbonded interactions, van der Waals interactions are often described using Lennard-Jones potentials, whereas Coulomb energy describes the electrostatic component of the interaction.[5]

In molecular dynamics (MD) simulations, the time evaluation in a system is analysed using a force field-based simulation. The force fields used in MM calculations are also used in MD. The movement of both the protein and the ligand are often important factors in a binding event.

There are also combined methods such as QM-MM, where different parts of a system are described with various levels of accuracy. In general, the computational cost of carrying out QM calculations is large, which can lead computational chemists to set limits on the investigated system that are as strict as possible.

## Virtual Screening

Traditionally VS has been used to screen, for example, a company's proprietary compound collection with the aim of selecting which compounds to test for biological and/or biochemical activity.[5] The focus has been to include as many active compounds as possible among those selected for testing.

There are in principle three categories of VS methods. Ligand-based virtual screening (LBVS), which relies on the activity data of one or more known compounds; structure-based virtual screening (SBVS), which uses structural target data; and combined methods utilising data derived from both ligands and target structure.

## Ligand-Based Virtual Screening

LBVS relies on the similarity principle; i.e. that structurally similar chemical compounds often have similar biological effects.[3] An often cited approximation is that 30 % of compounds with a daylight fingerprint similarity of above 0.85 will be active if the query molecule is active.[6] LBVS is a useful tool, but LBVS methods have limitations regarding the novelty and diversity of the hits identified.

LBVS techniques can be divided into 1D-, 2D- and 3D-based techniques. A 1D-based technique would be to use descriptors such as size (molecular weight), lipophilicity (logP) or charge. A 2D-based technique uses the connection table representation of a molecule. Fingerprint-based similarities based on the Tanimoto Index are often used.[7] 3D techniques use the three-dimensional shape of the ligand, often as it is bound to the target, with the aid, for example, of ROCS software which uses Gaussian functions centred on the atoms to describe the shape of the molecule.[8] Another widely used LBVS-technique is based on pharmacophore models. A pharmacophore model can be built from a set of active and if available inactive ligands.

## Structure-Based Virtual Screening

In SBVS different types of information about the target (often a protein) are used. Techniques such as pharmacophores or interaction fingerprints can be based on target information, but the main technique is docking. Docking can be defined as predicting the bioactive conformation of a molecule in the binding site of a target structure.[9] This could be likened to finding the minimum free energy of a protein-ligand system.

Docking involves the placement of a compound or ligand within the protein and scoring this position or pose according to the change in energy. There are many ways to construct docking algorithms. Two aspects that may not always be applicable but are nonetheless useful are pose generation and scoring.

In the pose generation step, the ligand is placed in close contact with the protein. In the first docking program to be developed (DOCK), both the ligand and protein were treated as rigid entities.[10] The pose generation part of the virtual screening process then aimed to position the ligand within the protein without overlapping with the protein atoms.

The various scoring systems are used to estimate favourable interactions between the positioned ligand and the protein. The scoring methods often use a simplified view of the problem, disregarding many levels of complexity. Scoring is ultimately an attempt to estimate the change in free energy that occurs with ligand-protein binding. There are three common types of scoring: force field-based (i.e. based on interatomic potentials), empirical and knowledge-based.[5] The docking calculations presented in this thesis have mostly been performed using Glide software.[11]

The results of docking investigations are often presented as a long list of compounds with different scores. These are usually sorted with the best score (an approximation of the largest negative change in free energy) at the top. When docking is used as a VS technique, both active and inactive compounds are scored. In most cases, compounds from both categories will be given favourable scores. An example of the distribution of scores among active and inactive compounds can be seen in Figure 2. Active compounds, in general, are more favourably scored (green curve), but there is significant overlap between the two groups.

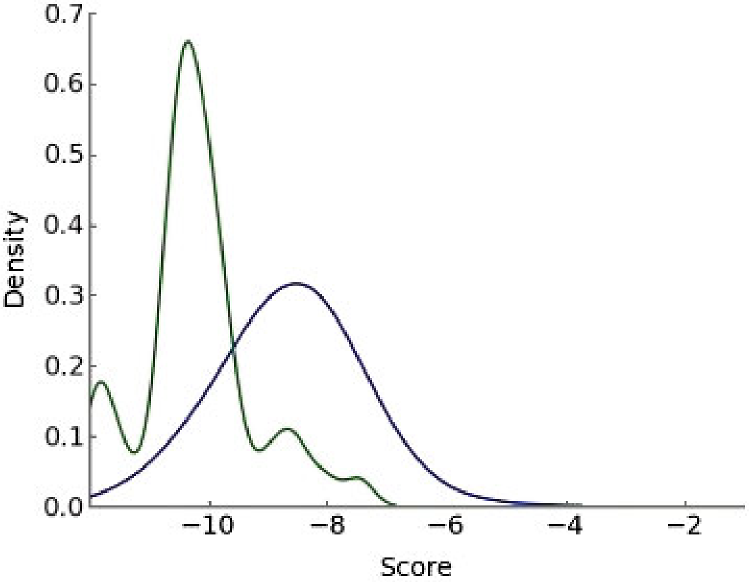


**Figure 2**. An example of docking results using Glide software. Low scores are favourable. Blue represents inactive compounds. Green represents active compounds. 90 000 compounds were screened against soluble epoxide hydrolase. Both active and inactive results have been normalised against the total number of compounds in each category.

## The Virtual Screening Funnel

In a VS funnel, computationally inexpensive methods, often ligand based approaches such as similarity search and pharmacophore screening, are used during the initial steps. Methods demanding comparatively high computational resources such as molecular docking and molecular dynamics (MD) simulation are used once the number of compounds to be screened decreases to a reasonable number. The final step in many campaigns includes the visual selection of compounds for testing.[12]

Virtual screening campaigns can be performed utilizing information about both the target and known actives. This can be done either by methods incorporating both kinds of data or by sequentially using different methods.

# Computer-Aided Drug Design and Discovery (Papers I, II, III)

## Rational Approaches to Pathogen Inhibitor Design (RAPID)

RAPID was a collaboration project between various departments at Uppsala University. The project involved scientists from research groups working with medicinal chemistry, computational chemistry, biochemistry and structural biology. The goal was to contribute to the early development of new drugs against some of the most serious diseases afflicting humanity, in particular, tuberculosis (TB).

TB is one of the major illnesses plaguing our world. WHO estimates that about 1.8 million people died from the disease in 2015.[13] It is a disease of poverty, overcrowding and unhygienic conditions and, as the incidence of these circumstances has declined, so has the incidence of TB. Existing drugs and vaccines have also helped to relieve some of the suffering and lower the mortality rate associated with the disease. The current standard treatment is a 6-month course of four antimicrobial drugs. These drugs are, however, not always effective. About half a million people are estimated to have developed multidrug-resistant TB. Consequently, there is an urgent need for new drugs which can treat drug-resistant TB.

## The DOXP/MEP Pathway and Isoprenoid Biosynthesis

Isoprenoids such as carotenoids, ubiquinone and steroid hormones are essential components of many if not all living organisms.[14] Isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are two important intermediates in the biosynthesis of these and other isoprenoids. In humans they are produced through the mevalonate pathway. In many bacteria (such as *Mycobacterium tuberculosis* [*Mtb*]) and parasites, the production of IPP and DMAPP is facilitated through the DOXP/MEP pathway, which is also called the non-mevalonate pathway. This indicates that enzymes very similar to those used in the bacterial pathway are unlikely to also exist in humans and that inhibitors of these enzymes may have a lower risk of causing selectivity problems and adverse effects.

The enzymes in this pathway are essential for bacteria such as *Mtb* and apicomplexan protozoa such as the malaria parasite *Plasmodium falciparum*.[15] It is, therefore, possible that these proteins could be useful targets for the development of new anti-infective drugs.[16]

## 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (DXR)

DXR is the second enzyme in the DOXP/MEP pathway. Alternative names for the enzyme are IspC, DXS, and Rv2682c. This enzyme catalyses the conversion of 1-deoxy-D-xylulose-5-phosphate (DXP) to 2C-methyl-D-erythritol 4-phosphate (MEP) using nicotinamide-adenine dinucleotide phosphate (NADPH) as a cofactor, see figure 3.

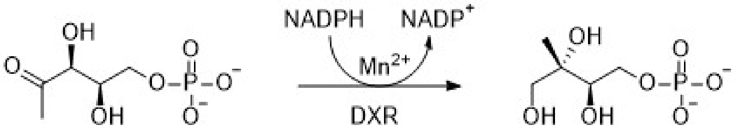
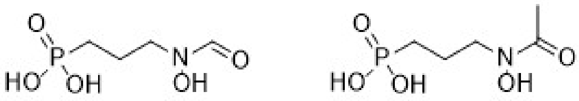


Figure 3. The reaction catalysed by DXR.

DXR is a medium-sized globular enzyme (containing about 400 amino acids (AAs), with a molecular weight of 43 kDa for each chain). The annotation is from the *Mtb* DXR (*Mt*DXR) crystal structure denoted as 4AIC in the PDB. A metal chelating site is set up by the GLU153, GLU222 and ASP151 AAs, which probably bind a magnesium (Mg), or perhaps (less likely) a manganese (Mn) ion with a 2+ charge *in vivo*. On one side of the metal is a binding site for NADPH. The binding site for DXP is on the other side. There is a 12 AA chain consisting of AAs 198-209 which forms a flexible loop, flap or lid over the binding site of DXP. TRP203 from the loop interacts with DXP (or MEP) when it is bound. This lid can exist in open or closed conformations in different crystal structures of the protein.[1]

## Fosmidomycin – a known DXR inhibitor

Fosmidomycin is a natural product isolated from *Streptomyces lavendulae* and *Streptomyces rubellomurinus*, respectively, in 1980.[17] In 1989, fosmidomycin was shown to inhibit *Escherichia coli* DXR[18] and the DOXP/MEP pathway.[19] The compound is a substrate (DXP) analogue with a different metal chelating moiety (a retro hydroxamic acid) from that in DXP (Figure 4).



**1** **2**

**Figure 4**. Fosmidomycin (**1**) and FR900098 (**2**).

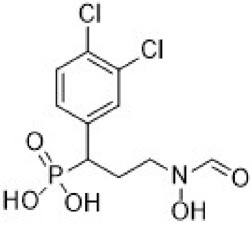
Fosmidomycin has an effect against malaria, both in *in vitro* studies against *P. falciparum* DXR (PfDXR) and against the parasite in humans.[20]

The first $IC_{50}$ value for fosmidomycin against *Mt*DXR to be reported was 310 nM (Dhiman 2005).[21] In our assay, the $IC_{50}$ value was 80 nM (Paper I). However, although fosmidomycin inhibits *Mt*DXR, it does not have antibacterial activity against *Mtb*. It has been suggested that the reason for this is that the compound does not enter across the membrane into the bacterium in sufficient concentrations, possibly because *Mtb* lacks the glycerol 3-phosphate transporter that is present in many other bacteria.[22]

## Targeting *Mt*DXR with Fosmidomycin Analogues

At the outset of our DXR project, a number of DXR inhibitors were known in the literature. A summary of this situation can be found in a paper by Silber *et al.*[23] Most data on DXR activity, as presented in that publication, were acquired from assays conducted with either *E. coli* DXR or PfDXR. It was, therefore, important to investigate whether compounds reported to be active against DXR from other species were also active against *Mt*DXR. We therefore resynthesized a few of the published active compounds and reevaluated them on *Mt*DXR. Since the work presented in this thesis was initiated a number of publications have appeared presenting many different DXR inhibitors, but unfortunately very few of those have displayed whole cell activity. Much of the work leading up to the current understanding of the SAR of these inhibitors have been thoroughly reviewed (see for example refs Hirsch, Dowd).[24,25] Today there are over 250 DXR inhibitors in the Binding Database (accessed August 2017).[26]

In short, modifications of fosmidomycin can be described as being made to the hydroxamic acid part, the phosphonate part or the carbon linker. In the present thesis we have worked along all these three strategies; a) substitution of the phosphonate for bioisosteres (paper II) and preparation of phosphate esters (paper II); b) preparation of Cα-substituted analogues (paper I) and c) preparation of hydroxamate bioisosteres (paper II) and acyl substituted analogues (paper III).



**3**

**Figure 5**. The structure of a 3,4-dichlorophenyl-substituted fosmidomycin analogue, compound **3**.

In paper **I,** a 3,4-dichlorophenyl-substituted fosmidomycin analogue, see Figure 5, which had been shown by Haemers *et al.* to have *E. coli* DXR activity, was synthesised and cocrystallized with *Mt*DXR.[27] A set of similar compounds was synthesised to evaluate whether a nearby hydrated pocket could be used to improve binding. For synthetic reasons, we chose to base our investigation on compound **14** (Table 1), the unsubstituted phenyl derivative of compound **3**.

It was suggested that we use a Suzuki reaction with boronic acids to attach substituents in the ortho position of the phenyl ring of compound **14**. This kind of chemistry would allow relatively late differentiation and could be an effective strategy for synthesis. The structures of the boronic acids were obtained from the chemical supplier Sigma-Aldrich. Using Legion software,[28] the boronic acids were connected virtually to compound **14** in the ortho position of the phenyl ring. The resulting 441 virtual compounds were docked (using Glide software) to the crystal structure of *Mt*DXR, which had been cocrystallized with $Mn^{2+}$, NADPH and compound **3.**

Redocking of compound **3** was successful, with a root-mean-square deviation of the atomic positions of 0.33 Å, see Figure 6. The docking results for the 441 compounds indicated that many of the compounds were too large. A set of aliphatic substituents was also rejected, for synthetic reasons. After this process, only two compounds remained; they had a pyridine substituent (**7)** and a thiophene substituent (**9)**. In the enzymatic assay of *Mt*DXR, these two compounds resulted in 20 % and 30 % inhibition at 100 µM. Two possible reasons for their weak activity were that it was energetically unfavourable to disturb the network of water molecules in the cavity and that the rather rigid nature of the compounds resulted in steric clashes with the protein.

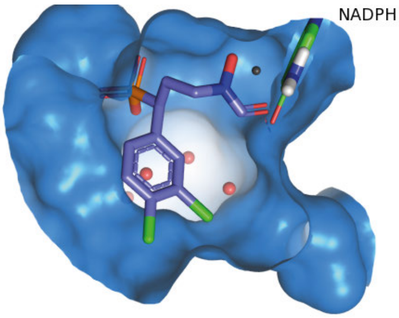
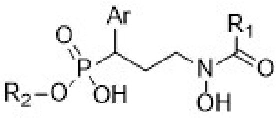


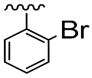
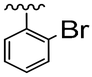
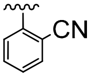
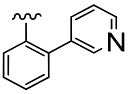
Figure 6. Crystal structure of *Mt*DXR with compound 3 and NADPH, displaying the hydrated cavity. The protein surface is shown in blue, compound 3 is purple, the four water molecules in the hydrated cavity are red, $Mn^{2+}$ is black, and the nicotinamide moiety of the NADPH cofactor is green.

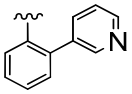
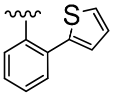
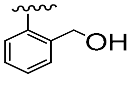
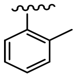
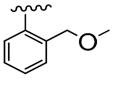
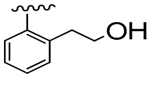
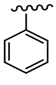
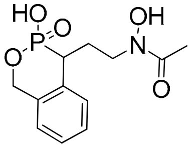
A set of analogues with smaller, hydrophilic substituents was then docked. Of these, compounds **6**, **10** and **13** were synthesised. However, none of these compounds displayed activity above 100 µM, illustrating the difficulties of identifying the correct substituent for reaching into the hydrated cavity.

Four additional compounds (**5**, **11**, **12**, **15**) were synthesised as they could be accessed from the later stages of the synthetic routes used; however, all had lower potency than compound **13.**

**Table 1**. Inhibition of *Mt*DXR by α-aryl-substituted fosmidomycin analogues



| | Ar | $R_1$ | $R_2$ | % inhibition at 100 µM | $IC_{50}$ (µM) |
|---|---|---|---|---|---|
| **4** |  | H | H | 93 ± 0.2 | 5.6 ± 5.9 |
| **5** |  | $CH_3$ | H | 38 ± 8 | 210 ± 48 |
| **6** |  | $CH_3$ | H | 30 ± 11 | Nt |
| **7** |  | $CH_3$ | H | 20 ± 10 | Nt |

| | | | | | |
|---|---|---|---|---|---|
| **8** |  | $CH_3$ | Et | 20 ± 5 | Nt |
| **9** |  | $CH_3$ | H | 30 ± 10 | Nt |
| **10** |  | $CH_3$ | H | 36 ± 5 | 465 ± 156 |
| **11** |  | $CH_3$ | H | 55 ± 6 | 205 ± 27 |
| **12** |  | $CH_3$ | H | 30 ± 11 | Nt |
| **13** |  | $CH_3$ | H | 35 ± 5 | 150 ± 47 |
| **14** |  | $CH_3$ | H | 92 ± 7 | 7.4 ± 2.6 |
| **15** |  | | | 12 ± 6 | Nt |

Nt = not tested.

Monoethyl and diethyl phosphonate esters of some of the compounds were prepared. Their activity was much lower in the enzyme assay. The monoethyl ester of compound **3** had an $IC_{50}$ of 38 μM, compared to 0.15 μM for compound **3**. Fosmidomycin analogues with larger phosphate esters have antimicrobial activity in whole cell assays of *Mtb*,[29–31] which indicates that these modifications might improve their activity against the bacteria. None of the compounds described in paper I had a minimum inhibitory concentration (MIC) above 32 μg/ml. Previous reports had indicated that the phenylethyl ester **22** (Table 2) was active against *Mtb*,[29] but the resynthesised compound was tested and did not display activity against the enzyme or whole cell *Mtb*.

We believe the charged phosphonate group found in most fosmidomycin analogues could be responsible for the lack of uptake and bacterial activity.

This idea was the initial motivation for the project described in paper II. Replacing the phosphate with different bioisosteres had been tried earlier with compounds **16** and **17.**[32] A small series of novel fosmidomycin analogues incorporating phosphonate bioisosteres was also prepared **18, 19, 20,** and **21**. Compound **21** had an $IC_{50}$ of 151 μM. The docking pose of the compound can be seen in Figure 7. Although the activity is modest, it is interesting because the compound could be a useful starting point for further investigation. One could imagine a set of close analogues with increased enzyme binding and *Mtb* whole-cell activity. Increased activity could be obtained by extending other parts of the molecule. Other interesting groups that could be used to replace the phosphonate is found in investigations of PTP1b inhibitors.[33]

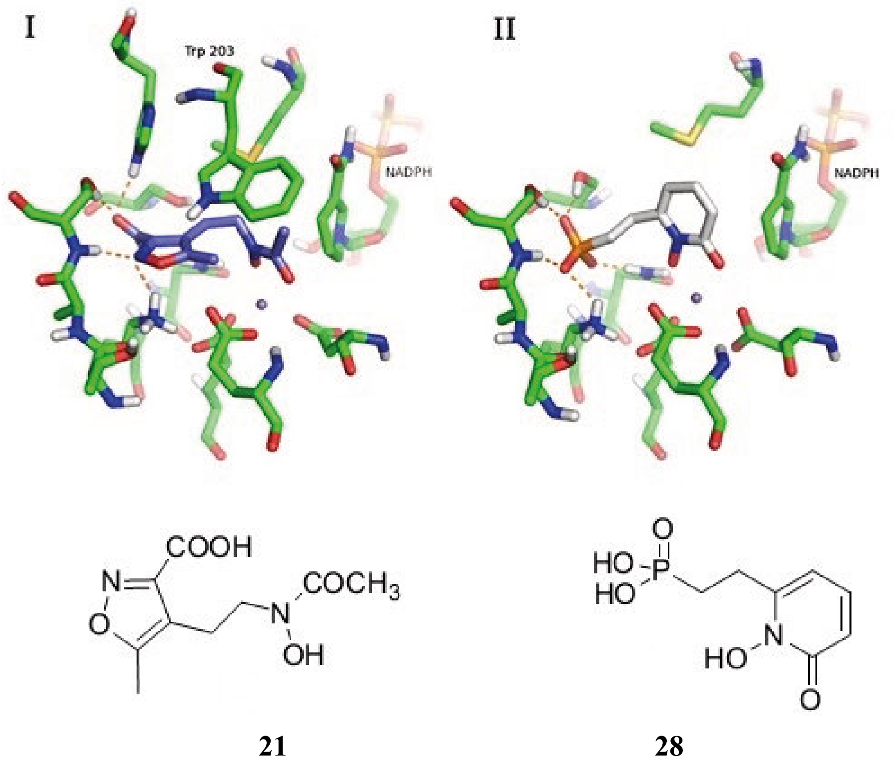
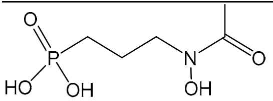
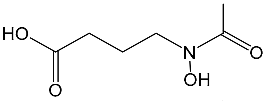
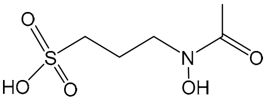
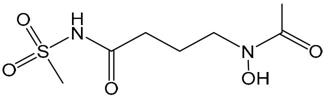
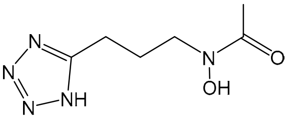
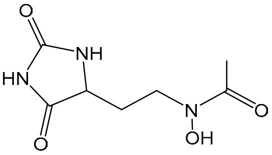
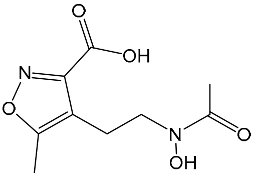
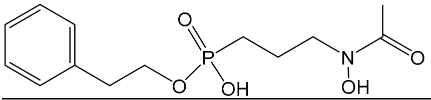
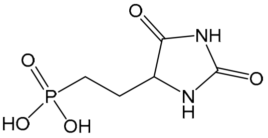
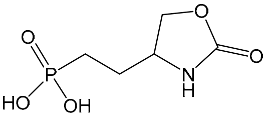
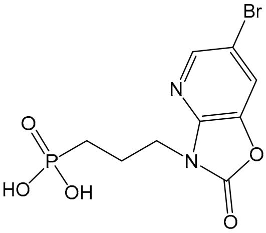
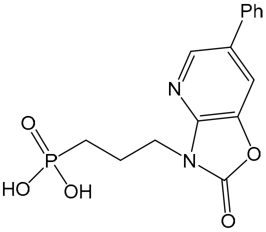
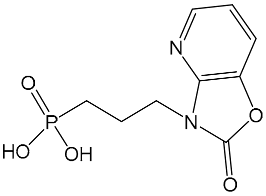
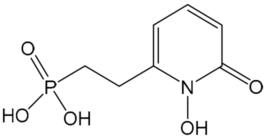
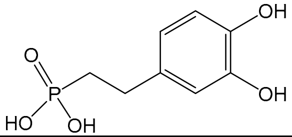


**Figure 7**. Compound **21** (in blue) and **28** (in white) docked to *Mt*DXR. *Mt*DXR and NADPH are shown in green, the metal ion in purple, and hydrogen bonds in orange.

**Table 2**. Biological evaluation of fosmidomycin analogues

Modification of the phosphonic acid part

| Chemical Structure | # | % inhibition | $IC_{50}$ (μM) |
|---|---|---|---|
|  | **2** | - | 0.16±0.03 |
|  | **16** | 0 | - |
|  | **17** | 36±14 | ›100 |
|  | **18** | 29±12 | |
|  | **19** | 0 | |
|  | **20** | 13±4 | |
|  | **21** | 41±1 | 151±22 |
|  | **22** | 23±7 | |
| Modification of the hydroxamic acid part | | | |
|  | **23** | 8±6 | |
|  | **24** | 0 | |

| | | | |
|---|---|---|---|
|  | **25** | 7±4 | |
|  | **26** | 11±6 | |
|  | **27** | 11±7 | |
|  | **28** | 61±5 | 53±13 |
|  | **29** | 78±3 | 41±10 |

In Paper **II,** we also produced a series of fosmidomycin analogues in which the hydroxamic acid part of the molecule was replaced (compounds **23-29**). Compound **26** was synthesised to follow up on a suggestion made by Mincheva et al.[34] They showed that the compound was active in a whole-cell assay of *Catharanthus roseus* and suggested that this could be due to DXR inhibition. Compounds **25-27** were therefore synthesised and evaluated for *Mt*DXR inhibition and *Mtb* whole-cell activity. Compound **26** did not show activity against *Mt*DXR or on the bacteria (MIC of 256 µg/mL).

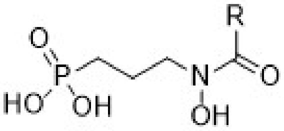
Compounds **23**, **24**, **28,** and **29** had different hydroxamic acid replacements. Compound **29**, earlier described by Deng et al[35] to have an *E-coli* DXR $IC_{50}$ of 4.5 µM, was resynthesized and had an *Mt*DXR $IC_{50}$ of 41 µM. Docking studies of this compound did not indicate how it would bind; it would, therefore, be interesting to see the compound co-crystallized with *Mt*DXR.

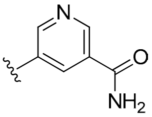
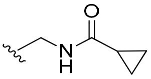
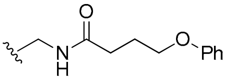
The novel compound **28** had an $IC_{50}$ of 53 µM. The proposed binding mode is shown in Figure 7.

In paper **III,** a set of fosmidomycin analogues (**30-39**) with substituents that extended from the hydroxamic acid group was prepared to investigate their activity (Table 3). The activity decreased with increased size for the smaller substituents: H > methyl > ethyl > cyclopropyl. This was probably due to interactions with the indole ring of TRP203. Pushing the indole ring away would induce the lid/flap of the protein to be in a more open state. Larger substituents (phenyl, 3-pyridyl, 2-naphthyl, 2-furyl, benzyl) would be more likely to interact with the protein with the flap/lid in the open position. For those compounds, $IC_{50}$ values were in the 1-100 µM range. Compounds **38** and **39** were resynthesized and evaluated on *Mt*DXR because they had previously been shown to show reasonable potency on *E-coli* DXR and PfDXR.[36] Compound **30** was later shown to be a strong inhibitor of PfDXR.[37]

The proposed binding modes of the compounds were predicted using Glide docking software using the crystal structure of *Mt*DXR co-crystallized with compound **3** (PDB entry 2Y1D) and the protein flap/lid was in an open position. When superimposing the docking pose of **33** with the co-crystallised ligand **3,** it was clear there would be room for a disubstituted compound with both rings in close contact. Building on this hypothesis compound **40** was synthesised, tested and co-crystallised (3ZHX, 3ZHY).

**Table 3.** Inhibition constants for fosmidomycin analogues. Literature values are included for comparison.



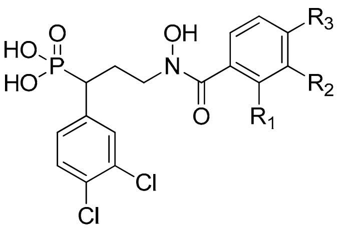
| | R | *Mt*DXR<br>$IC_{50}$ (μM) | *E. coli* DXR<br>$IC_{50}$ (μM) | *Pf*DXR<br>$IC_{50}$ (μM) |
|---|---|---|---|---|
| **1** | H | 0.08 ± 0.02[a] | 0.050[a] | 0.032[a] |
| **2** | Methyl | 0.16 ± 0.03[a] | 0.051[a] | 0.018[a] |
| **30** | Ethyl | 27.2 ± 7.7 | | |
| **31** | 2-furyl | 48.7 ± 6.2 | | |
| **32** | Cyclopropyl | 156 ± 68 | 10-4.44[a] | |
| **33** | Phenyl | 2.0 ± 0.6 | 0.13[a] | 0.061[a] |
| **34** | 3-pyridyl | 1.6 ± 0.1 | | |
| **35** |  | 3.6 ± 0.5 | | |
| **36** | 2-naphthyl | 2.0 ± 0.3 | | |
| **37** | Benzyl | 67.7 ± 23.3 | 13 | |
| **38** |  | >200 | 7.1[a] | 3.3[a] |
| **39** |  | >200 | 1.0[a] | 0.4[a] |

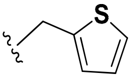
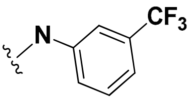
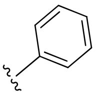
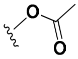
[a] Test results from the literature, see paper **III** for details.

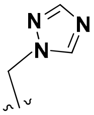
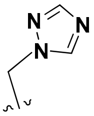
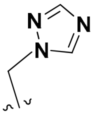
The crystal structure of compound **40** was used to prepare a set of compounds (compounds **41-50** in Table 4). The docking poses of these compounds suggested they would reach towards and perhaps into the binding site of the cofactor NADPH. High concentrations of NADPH were used in the assay for measuring $IC_{50}$ values, which would affect the $IC_{50}$ measurements of any compound competing with NADPH for binding. A secondary direct fluorescence assay was used to investigate the strength of the binding constant $K_d$ for these compounds. The $K_d$ of for example compound **48** (0.04 μM) indicated that the compound binds more strongly than suggested by the $IC_{50}$ value (13 μM) and is thus more likely to interfere with the binding of NADPH.

The region of DXR that is of interest here has shown substantial flexibility when binding to different ligands. For example two bisphosphonate ligands bound to *E. coli* DXR structures showed induced-fit binding.[38] Docking studies were performed with some compounds**,** but their predictive value was estimated to be low. The crystal structures of compounds **45** and **48** showed induced-fit binding and movement of several AAs in the vicinity. In the crystal structures, the two compounds reached towards and into the NADPH binding site. However, the crystal structures of compounds **45** and **48** did not show the compounds chelating to the metal. The induced fit seen in the crystal structures does seem to echo the claim by Mercklé *et al.* that the rational design of new inhibitors of DXR is, at best, difficult.[1]

**Table 4.** Inhibition of *Mt*DXR activity and *Plasmodium* growth by disubstituted fosmidomycin analogues.



| | $R_1$ | $R_2$ | $R_3$ | *Mt*DXR $IC_{50}$ (µM) | $K_d$ with $MnCl_2$ (µM) | *P. falciparum* growth *in vitro* $IC_{50}$ (µM) |
|---|---|---|---|---|---|---|
| **40** | H | H | H | 0.32 ± 0.05 | 0.21 | 0.04 |
| **41** | $CH_3$ | H | H | 0.83± 0.08 | | >10 |
| **42** | $CF_3$ | H | H | 9 ± 8 | | No fit |
| **43** | $OCF_3$ | H | H | 8 ± 2 | 0.09 | >10 |
| **44** |  | H | H | 7± 3 | | >10 |
| **45** |  | H | H | 19 ± 3 | 0.09 | >10 |
| **46** |  | H | H | 21 ± 4.1 | | >10 |
| **47** | H |  | H | > 100 | | 3.9 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **48** |  | H | H | 13 ± 3 | 0.04 | >10 |
| **49** | H |  | H | 0.14± 0.04 | 0.17 | 0.39 |
| **50** | H | H |  | 1.2 ± 0.2 | 0.20 | 0.19 |

# Benchmarking Virtual Screening (Paper IV)

## Selecting the VS Protocol

Scientists planning to perform VS are confronted with a broad array of possible methods to use. Choosing the methods to use, how to combine different methods and how to set the myriad of possible parameters is not a trivial task. There are plenty of examples in the literature of successful VS procedures, which could help in the choice. However, picking a method based on previous achievements does not guarantee success. VS is target-dependent and fundamentally based on chance and likelihoods. It should also be borne in mind that publication bias may play a role. Scientists do not often publish unsuccessful VS procedures. In addition to looking at previous prospective screenings, retrospective analysis is also used.

One way of validating docking algorithms has been to dock co-crystallized ligands and evaluate how closely the docking pose resembles the pose of the co-crystallized ligand. Another way has been to assess the score for active compounds and make correlations between the biological activity and the docking score. Both these methods focus on docking and scoring of active compounds, ignoring the behaviour of inactive compounds in the VS protocol. In any random set of compounds screened for activity, the inactive compounds will vastly outnumber the active compounds. It is therefore more important to differentiate active from inactive compounds than to correctly score active compounds (i.e. accurately estimate the ΔG of binding).

Benchmarking sets have been used extensively to validate, optimise and compare the performances of different VS techniques.[39] In most cases, a set of known active compounds is combined with a set of random compounds presumed to be (mostly) inactive. Two main problems have emerged from these benchmarking sets: artificial enrichment and analogue bias.

Artificial enrichment is the false enrichment of the data that can occur if the active and inactive compounds differ in physicochemical properties. A simple example: if all the active compounds are negatively charged, and none of the inactive compounds are negatively charged, a perfect enrichment of the active compounds can then be achieved using a simple descriptor based on the charge of the molecules.

Analogue bias is false enrichment that occurs when the active compounds are too similar to each other. In many benchmarking sets, the active compounds are fetched from earlier lead generation projects in which sets of analogues have been synthesised. This leads to groups of similar compounds among the active compounds.

Recent benchmarking data sets have compensated for these problems and artificial enrichment is rarely a problem, although analogue bias may still occur. The most cited benchmarking data set today is the directory of useful decoys (DUD) and the follow-up DUD-E.[40,41] This data set contains a broad array of targets and the data are of reasonably high quality. In my opinion, however, they still have problems associated with picking active and presumed inactive compounds from different sources. It is therefore unclear how well the results acquired using the DUD or DUD-E data set can be translated into guidelines for prospective VS projects.

There are three main problems I have associated with benchmarking sets which use active compounds from the literature and presumed inactive compounds.

Firstly, due to analogue bias, enrichment by ligand-based methods will be overestimated (or artificially underestimated if methods are used to remove the bias). Ligand-based or combined methods are therefore difficult to evaluate. The problem of a remaining analogue bias in DUD-E has recently been reported by Chaput et al.[42]

Secondly, the ratio of active to inactive compounds is set by the researcher constructing the data set, and these ratios will often not represent real case VS scenarios. It is likely that the results will be affected by the ratio of active to inactive compounds and how similar and/or dissimilar they are. I believe this problem will be magnified when evaluating VS funnels with sequentially reduced sets of molecules being transferred to later stages.

Thirdly, the VS method being evaluated may not have to deal with the kind of inactive compound found in a prospective VS method. There might be certain types of compounds which are especially problematic for some VS methods. For example, in my experience some docking software gives highly flexible molecules high scoring values. If these compounds are not included in the benchmarking data set, this problem will go unnoticed.

These and other problems provide reason for caution when trusting the results given by benchmarking data sets based on active and presumed inactive compounds derived from the literature.

## Benchmarking VS Using HTS-Data

It may seem easy to set up a reliable VS-benchmarking data set. One simply needs to collect a (large) set of compounds, test them in a biochemical assay to determine which are active and which are inactive. The target protein needs to be crystallised, preferably with a ligand that can act as a query for LBVS. If a sufficient number of compounds is reliably screened against an adequate number of targets one would end up with a good benchmarking data set.

However, looking more closely at what would be required in practice, one would have to (arbitrarily) set an activity level. For example, a Ki of 10 μM, to define when a compound could be regarded as active. To make the benchmarking set useful, one would need to attain a sufficient number of active compounds. If we assume that 1 in 2500 compounds is active, and that about 300 active compounds are required for each target, about 750 000 compounds would need to be screened for each target. Further, in order to make the benchmarking set attractive, many different targets would need to be investigated. The benchmarking set DUD-E includes 102 targets and DEKOIS 2.0 includes 81 targets. An HTS-based benchmarking set would probably add value even if it is much smaller, but a large number of targets is clearly beneficial to get around some of the problems associated with benchmarking.

Performing this many HTS procedures takes significant effort, and is perhaps outside the scope of a single scientist. However, such a project is not particularly large when compared to scientific projects such as the Human Protein Atlas which has been funded by the Swedish research funding community.[43,44]

Many HTS procedures are performed in industry and academia. It is, therefore, possible that data from these screens could be collated and used for benchmarking. Results from academic HTS procedures are deposited in the publicly available PubChem bioassay database.[45] We decided to see whether this database could be used to extract the necessary data to create a benchmarking data set. Paper **IV** deals with this investigation.

HTS-data from PubChem have been used previously to create benchmarking data sets. The Maximum Unbiased Validation data set was one such attempt; it was, however, aimed at evaluating LBVS techniques.[46]

## Identifying HTS-Data Useful for Benchmarking VS

In paper **IV**, we attempted to assemble the best HTS-data we could find and connect it with crystal structures binding a drug-like ligand. HTS-data were derived from the PubChem bioassay database, and related crystal structures were accessed from the PDB.[45,47] Targets that had been co-crystallized with a drug-like ligand were used. This ligand could act as a query for LBVS and could be used to locate the binding site in SBVS.

We mined the thousands of possible entries at PubChem and PDB using python scripts. To avoid ambiguity about the resulting targets, we initially only chose targets with one binding site. Different quality criteria were used to weed out undesirable targets. The aim of implementing the criteria was to obtain targets with a reasonable number of true positives and true negatives while reducing the risk of false positives and false negatives. The exact limits set for these criteria could be modified to obtain different results. It would be easy to argue for the use of stricter criteria than those used; however, using very strict criteria could result in no targets passing the filters. The exercise first led to five targets. We later added two protein kinases, which had been sorted out since they bound cofactors. The data set then included seven targets with usable HTS and PDB data.

The diversity of the active compounds in each of the seven data sets was assessed using Bemis-Murcko scaffolds.[48] The ratio between the  number of scaffolds and number of actives was high indicating that analogue bias is not a big problem in these data sets. Analogue bias was also investigated using a 2D similarity search, with the X-ray ligand from each data set as the template. The average enrichment factor (at 1 %) for 2D similarity searches using the seven data sets was 2.3. Active compounds were also subjected to filtration using the Pan Assay Interference Compounds (PAINS) filter. This exercise indicated that a few of the compounds that were reported as active were probably false positives.

We also evaluated the performance of some representative VS techniques using the data sets: 1D using similarity/distance in simple descriptor space, 2D similarity using fingerprints, 3D similarity using ROCS, and docking using Glide software. These results indicated some differences from the evaluations reported in the most cited current benchmark data set, DUD. The overall performance was much lower in our data sets. I would speculate that this was a consequence of analogue bias in the DUD data set. In our data sets, we also noted that docking outperformed ligand-based methods.

This preliminary study indicates that using HTS-data for benchmarking will result in substantially different results from current benchmarking data sets. The quality of the HTS-data employed in this study may not be good enough, and the number of data sets may be too few to replace the current benchmarking data sets. The data sets can, however, be used to complement the results from DUD-E, DEKIOS 2.0 and other modern benchmarking data sets.

There would be great value in creating a new benchmarking data set using novel high quality HTS-data. This data set could be used for comparing different docking methods so that the results could be reliably transferred to prospective VS projects. In addition, and perhaps more importantly, such a data set could also help the scientific community to focus on the parts of the VS funnel that are in most need of improvement. Additionally, such a data set could also be valuable for developing machine learning or applying deep learning approaches to VS.

# Aggregated Conformal Prediction (Paper V)

Conformal Prediction (CP) is a machine-learning framework that uses past experience to relate precise levels of confidence to new predictions. Many commonly used machine-learning algorithms produce point predictions, while conformal predictors produce prediction ranges.[49] The method was presented in 2005 by Vladimir Vovk, Alex Gammerman and Glenn Shafer. They presented CP as a framework that makes it possible to hedge predictions while simultaneously allowing the algorithms to learn and predict. CP can be inductive or transductive. The transductive inference refers to methods that use specific (training) cases to predict specific (test) cases. This type of method was used in the original CP approach but is computationally demanding since all the calculations must be made for each new prediction. Inductive CP methods use training cases to formulate general rules, which are then applied to the test cases. Inductive CP is more computationally efficient and was used in paper V.

The stringent mathematical framework that allows CP to present valid predictions is based on the assumption that the model is built using data that follow the exchangeability assumption. This assumption is standard in machine learning and means that new observations (compounds) behave like earlier observations, i.e. have the same distribution. CP is commonly applied to classification problems, but can also be used for regression problems, as was done in paper V. The framework can in principle be used with any machine-learning algorithm and the predictions are presented as easily interpretable prediction ranges. In paper V we created models using both random forest (RF) and support vector machine (SVM) software. In order to reduce the variance in the predicted ranges, we opted to perform 100 selections of (proper) training and calibration sets (see Figure 8). This method has been named Aggregated Conformal Prediction (ACP).[50]

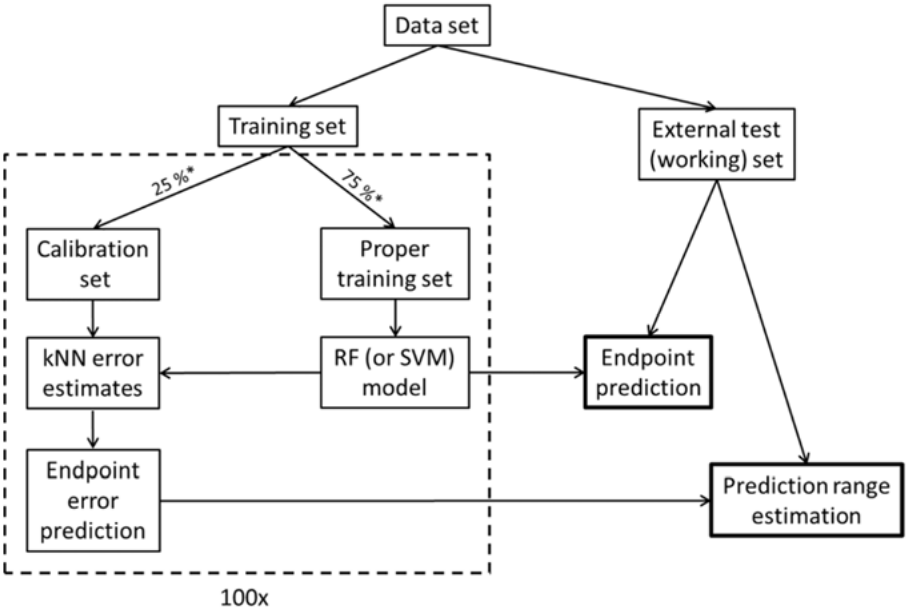


**Figure 8**. Conformal Prediction scheme. The scheme becomes an aggregated conformal prediction scheme when the split (*) into calibration and proper training sets is performed multiple times. kNN = k-nearest neighbours algorithm; RF = random forest software; SVM = support vector machine software.

Our calculation using the CP framework is schematically presented in Figure 8. Data were randomly split into a training set for building the model and a test set for evaluating the model. In addition to the training and test sets, the CP framework used an internal test set, called the calibration set, to evaluate how similar new compounds were to the compounds used to build the CP model. For this, the training set was randomly split into a proper training set (75% of the compounds) and a calibration set (25% of the compounds). A model was developed using the proper training set. This model was applied to predict the endpoint (skin permeability in our study) of the compounds in the calibration set. This random split can be performed many times, thus generating many models. The absolute errors of these predictions (compared to the experimental values) are used to calculate the prediction ranges of new compounds. In the k-nearest neighbours model, the attributes of the k compounds from the calibration set that is closest to the test compound in normalised descriptor space are used to predict the prediction range of the test compound. This model was used to construct the prediction ranges for the new compounds at the specified significance level by averaging the absolute errors from the neighbours of the calibration set. In our study, the CP significance level was set to 0.2. This means that the derived models at most showed 20% prediction errors, if the exchangeability criterion for the data set was fulfilled.

In paper V we exemplified the use of CP with models of skin permeability. It is easy to imagine that similar models could be built to predict many other parameters important in early drug discovery. CP appears to be a useful tool in many areas of computer-aided drug design, and can also be used to support VS results.

Machine-learning algorithms can be used in combination with docking to increase the number of active compounds one might find. If this procedure is performed within a CP framework, the researcher may be able to obtain guidance on the number of compounds they ought to screen.[51]

# Concluding Remarks

The aim of the DXR project, presented in papers I-III, was to find *Mt*DXR inhibitors that had the potential to be developed into drugs against *Mtb*. In this project, several crystal structures of *Mt*DXR were produced, and a number of novel compounds were synthesised and evaluated for enzyme activity and whole cell activity. Of these compounds, several inhibited the enzyme and a few had also encouraging activity. Some compounds were also active in an assay of whole-cell *P. falciparum*. Overall, however, the project failed from both a modelling perspective and a drug development perspective when the initial goals were considered: no new DXR inhibitors with activity against *Mtb* were prepared.

DXR and the other enzymes in the non-mevalonate pathway continue to be reasonably attractive as drug targets. The results presented in this thesis and published by other groups have in no way exhausted the possibilities. Most DXR projects found in the literature have described fosmidomycin analogues. By the use of a prodrug strategy some phosphate esters have been prepared that exhibited activity on *Mtb*.[29–31] The flexible nature of DXR does put some restraint on the use of computational methods in this work. Data about the compounds synthesised and tested in the project could be used to validate any future modeling efforts.

New VS methods and variants are frequently presented in the literature. Validating and quantifying their performance is central to continued progress. A number of VS benchmarking data sets are currently available for this purpose. However, there are a number of problems associated with these data sets, and the outcome of investigations using them may be in doubt. Their use seems particularly problematic when comparing different types of VS methods and VS funnels. The problems are thought to stem from the use of known active compounds and unknown decoys when the benchmarking data sets were created.

HTS studies results in both active and inactive compounds obtained from a common pool. Data from HTS can, therefore, be used for VS benchmarking. HTS-data is however often associated with uncertainty. Paper IV presents our search for high quality HTS-data in the public domain. Unsurprisingly, we were not able to find data of sufficient quality to be able to create a reliable VS benchmarking data set. This was probably not because HTS cannot be used to acquire high quality data; it was more likely associated with the aims of the HTS studies. New data are needed, and a project aimed at obtaining high quality HTS-data to be used for VS benchmarking would be of great value.

CP is a useful, new framework that could be applied to many problems in computational drug discovery. In paper V we have shown how CP could work for models of skin permeability.

# Acknowledgements

I would like to thank Saga and Viktor, who make any struggle worth the pain.

My sincere gratitude to Linnea, Ingela, Yngve, Katarina, Björn and Lotta for the love and support you've offered during these years.

I would like to thank Fredrik Svensson and Ulf Norinder for choosing to work with me on the projects leading to papers **III** and **IV**. Without you, these articles would not have been written. I would also like to thank the people who participated in writing all the papers: Mounir Andaloussi, Lena Henriksson, Anna Więckowska, Christofer Björkelid, Anna Larsson, Surisetti Suresh, Harini Iyer, Bachally Srinivasa, Terese Bergfors, Torsten Unge, Sherry Mowbray, Alwyn Jones, Anna Jansson, Samir Yahiaoui, Sanjeewani Sooriyaarachchi, Shyamraj Dharavath, Matthieu Desroses, Rautela Nikhil, Sharma Sreevalli, Harini Iyer, Wesley Schaal, Mats Larhed, Peter Brandt and Jin Zhang. Additionally I would like to thank Sorin Srbu and Daniel Muthas.

I would also like to thank the people who have directed this work, my supervisor Anders Karlén, and assistant supervisor Christian Sköld.

# References

(1) Mercklé, L.; de Andrés-Gómez, A.; Dick, B.; Cox, R. J.; Godfrey, CR: Fragment-Based Approach to Understanding Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase. *Chembiochem* **2005**, *6* (10), 1866–1874.

(2) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112* (3), 535–542.

(3) Schneider, G.; Baringhaus, K.; Kubinyi, H.: Molecular Design: Concepts and Applications; Wiley-VCH, **2008**.

(4) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988,** 110 (6): 1657–1666.

(5) Sotriffer C., Mannhold R., Kubinyi H., Folkers G.; Virtual Screening: Principles, Challenges, and Practical Guidelines*;* Wiley-VCH, **2011**.

(6) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45* (19), 4350–4358.

(7) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 20.

(8) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82.

(9) Blaney, J. M.; Dixon, J. S. A Good Ligand Is Hard to Find: Automated Docking Methods. *Perspect. Drug Discov. Des.* **1993**, *1* (2), 301–319.

(10) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.

(11) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11* (5), 425–445.

(12) Kumar, A.; Zhang, K. Y. J. Hierarchical Virtual Screening Approaches in Small Molecule Drug Discovery. *Methods* **2015**, *71*, 26–37.

(13) WHO | Global Tuberculosis Report 2016. *WHO* **2017**.

(14) Sacchettini, J. C.; Poulter, C. D. Creating Isoprenoid Diversity. *Science* **1997**, *277* (5333), 1788–1789.

(15) Rohmer, M. The Discovery of a Mevalonate-Independent Pathway for Isoprenoid Biosynthesis in Bacteria, Algae and Higher Plants. *Nat. Prod. Rep.* **1999**, *16* (5), 565–574.

(16) Hirsch, A. K. H.; Diederich, F. The Non-Mevalonate Pathway to Isoprenoid Biosynthesis: A Potential Source of New Drug Targets. *Chim. Int. J. Chem.* **2008**, *62* (4), 226–230.

(17) Okuhara, M.; Kuroda, Y.; Goto, T.; Okamoto, M.; Terano, H.; Kohsaka, M.; Aoki, H.; Imanaka, H. Studies on New Phosphonic Acid Antibiotics. III. Isolation and Characterization of FR-31564, FR-32863 and FR-33289. *J. Antibiot. (Tokyo).* **1980**, *33* (1), 24–28.

(18) Kuzuyama, T.; Shimizu, T.; Takahashi, S.; Seto, H. Fosmidomycin, a Specific Inhibitor of 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase in the Nonmevalonate Pathway for Terpenoid Biosynthesis. *Tetrahedron Lett.* **1998**, *39* (43), 7913–7916.

(19) Shigi, Y. Inhibition of Bacterial Isoprenoid Synthesis by Fosmidomycin, a Phosphonic Acid-Containing Antibiotic. *J. Antimicrob. Chemother.* **1989**, *24* (2), 131–145.

(20) Missinou, M. A.; Borrmann, S.; Schindler, A.; Issifou, S.; Adegnika, A. A.; Matsiegui, P.-B.; Binder, R.; Lell, B.; Wiesner, J.; Baranek, T.; Jomaa, H.; Kremsner, P. G. Fosmidomycin for Malaria. *Lancet (London, England)* **2002**, *360* (9349), 1941–1942.

(21) Dhiman, R. K.; Schaeffer, M. L.; Bailey, A. M.; Testa, C. a; Scherman, H.; Crick, D. C. 1-Deoxy- D -Xylulose 5-Phosphate Reductoisomerase ( IspC ) from Mycobacterium Tuberculosis: Towards Understanding Mycobacterial Resistance to Fosmidomycin. *J. Bacteriol* **2005**, *187* (24), 8395–8402.

(22) Brown, A. C.; Parish, T. Dxr Is Essential in Mycobacterium Tuberculosis and Fosmidomycin Resistance Is due to a Lack of Uptake. *BMC Microbiol.* **2008**, *8* (1), 78.

(23) Silber, K.; Heidler, P.; Kurz, T.; Klebe, G. AFMoC Enhances Predictivity of 3D QSAR: A Case Study with DOXP-Reductoisomerase. *J. Med. Chem.* **2005**, *48* (10), 3547–3563.

(24) Masini, T.; Hirsch, A. K. H. Development of Inhibitors of the 2 *C* -Methyl- D -Erythritol 4-Phosphate (MEP) Pathway Enzymes as Potential Anti-Infective Agents. *J. Med. Chem.* **2014**, *57* (23), 9740–9763.

(25) R. Jackson, E.; S. Dowd, C. Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (Dxr): A Review of the Synthesis and Biological Evaluation of Recent Inhibitors. *Curr. Top. Med. Chem.* **2012**, *12* (7), 706–728.

(26) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35,* 198–201.

(27) Haemers, T.; Wiesner, J.; Poecke, S. Van; Goeman, J.; Henschker, D.; Beck, E.; Jomaa, H.; Calenbergh, S. Van. Synthesis of α-Substituted Fosmidomycin Analogues as Highly Potent Plasmodium Falciparum Growth Inhibitors. *Bioorg. Med. Chem. Lett* **2006**, 16(7):1888-91.

(28) Tripos. Legion. St. Louis, MO **1998**.

(29) Perruchon, J.; Ortmann, R.; Altenkämper, M.; Silber, K.; Wiesner, J.; Jomaa, H.; Klebe, G.; Schlitzer, M. Studies Addressing the Importance of Charge in the Binding of Fosmidomycin-like Molecules to Deoxyxylulosephosphate Reductoisomerase. *ChemMedChem* **2008**, *3* (8), 1232–1241.

(30) San Jose, G.; Jackson, E. R.; Haymond, A.; Johny, C.; Edwards, R. L.; Wang, X.; Brothers, R. C.; Edelstein, E. K.; Odom, A. R.; Boshoff, H. I.; Couch, R. D.; Dowd, C. S. Structure–Activity Relationships of the MEPicides: *N* -Acyl and *O* -Linked Analogs of FR900098 as Inhibitors of Dxr from *Mycobacterium Tuberculosis* and *Yersinia Pestis*. *ACS Infect. Dis.* **2016**, *2* (12), 923–935.

(31) Jackson, E. R.; San Jose, G.; Brothers, R. C.; Edelstein, E. K.; Sheldon, Z.; Haymond, A.; Johny, C.; Boshoff, H. I.; Couch, R. D.; Dowd, C. S. The Effect of Chain Length and Unsaturation on Mtb Dxr Inhibition and Antitubercular Killing Activity of FR900098 Analogs. *Bioorg. Med. Chem. Lett.* **2014**, *24* (2), 649–653.

(32) Woo, Y. H.; Fernandes, R. P. M.; Proteau, P. J. Evaluation of Fosmidomycin Analogs as Inhibitors of the Synechocystis Sp. PCC6803 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase. *Bioorganic Med. Chem.* **2006**, *14* (7), 2375–2385.

(33) Wan, Z. K.; Follows, B.; Kirincich, S.; Wilson, D.; Binnun, E.; Xu, W.; Joseph-McCarthy, D.; Wu, J.; Smith, M.; Zhang, Y. L.; Tam, M.; Erbe, D.; Tam, S.; Saiah, E.; Lee, J. Probing Acid Replacements of Thiophene PTP1B Inhibitors. *Bioorganic Med. Chem. Lett.* **2007**, *17* (10), 2913–2920.

(34) Mincheva, Z.; Courtois, M.; Andreu, F.; Rideau, M.; Viaud-Massuard, M.-C. Fosmidomycin Analogues as Inhibitors of Monoterpenoid Indole Alkaloid Production in Catharanthus Roseus Cells. *Phytochemistry* **2005**, *66* (15), 1797–1803.

(35) Deng, L.; Sundriyal, S.; Rubio, V.; Shi, Z.; Song, Y. Coordination Chemistry Based Approach to Lipophilic Inhibitors of 1-Deoxy- D -Xylulose-5-Phosphate Reductoisomerase. *J. Med. Chem.* **2009**, *52* (21), 6539–6542.

(36) Gießmann, D.; Heidler, P.; Haemers, T.; Van Calenbergh, S.; Reichenberg, A.; Jomaa, H.; Weidemeyer, C.; Sanderbrand, S.; Wiesner, J.; Link, A. Towards New Antimalarial Drugs: Synthesis of Non-Hydrolyzable Phosphate Mimics as Feed for a Predictive QSAR Study on 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase Inhibitors. *Chem. Biodivers.* **2008**, *5* (4), 643–656.

(37) Cobb, R. E.; Bae, B.; Li, Z.; DeSieno, M. A.; Nair, S. K.; Zhao, H. Structure-Guided Design and Biosynthesis of a Novel FR-900098 Analogue as a Potent Plasmodium Falciparum 1-Deoxy- D -Xylulose-5-Phosphate Reductoisomerase (Dxr) Inhibitor. *Chem. Commun.* **2015**, *51* (13), 2526–2528.

(38) Yajima, S.; Hara, K.; Sanders, J. M.; Yin, F.; Ohsawa, K.; Wiesner, J.; Jomaa, H.; Oldfield, E. Crystallographic Structures of Two bisphosphonate:1-Deoxyxylulose-5-Phosphate Reductoisomerase Complexes. *J. Am. Chem. Soc.* **2004**, *126* (35), 10824–10825.

(39) Lagarde, N.; Zagury, J.-F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, 55 (7), 1297–1307

(40) von Korff, M.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **2009**, *49* (2), 209–231.

(41) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

(42) Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminform.* **2016**, *8* (1), 56.

(43) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Björling, L.; Ponten, F. Towards a Knowledge-Based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248–1250.

(44) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; Feilitzen, K. Von; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; Heijne, G. Von; Nielsen, J.; Pontén, F. Tissue-Based Map of the Human Proteome. *Science* **2015**, 347(6220).

(45) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, 400–412.

(46) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184.

(47) RCSB Protein Data Bank - RCSB PDB - 3F81 Structure Summary http://www.rcsb.org/pdb/explore/explore.do?structureId=3F81 (accessed Apr 11, **2013**).

(48) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.

(49) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer-Verlag: New York, **2005**.

(50) Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction; Springer Berlin Heidelberg, **2014**; 231–240.

(51) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57* (3), 439–444.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-328505

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017