TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

# CAVITY-BASED NEGATIVE IMAGES IN MOLECULAR DOCKING

Sami Kurkinen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1577 | MEDICA – ODONTOLOGICA | TURKU 2021

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

# CAVITY-BASED NEGATIVE IMAGES IN MOLECULAR DOCKING

Sami Kurkinen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1577 | MEDICA – ODONTOLOGICA | TURKU 2021

## University of Turku

Faculty of Medicine
Institute of Biomedicine
Pharmacology, Drug Development and Therapeutics
Drug Research Doctoral Programme

## Supervised by

Professor, Olli Pentikäinen, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

Docent, Pekka Postila, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

## Reviewed by

Professor, Daniela Schuster, Dr.
Institute of Pharmacy
Paracelsus Medical University
Salzburg, Austria

Professor, Mark Johnson, PhD
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland

## Opponent

Docent, Henri Xhaard, PhD
Centre for Drug Research
Faculty of Pharmacy
University of Helsinki
Helsinki, Finland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-8580-7 (PRINT)
ISBN 978-951-29-8581-4 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama, Turku, Finland 2021

*“Facts are meaningless. You could use facts to prove anything that’s even remotely true.”*

*Homer Simpson*

UNIVERSITY OF TURKU
Faculty of Medicine
Institute of Biomedicine
Pharmacology, Drug Development and Therapeutics
SAMI KURKINEN: Cavity-Based Negative Images in Molecular Docking
Doctoral Dissertation, 147 pp.
Drug Research Doctoral Programme (DRDP)
September 2021

## ABSTRACT

In drug development, computer-based methods are constantly evolving as a result of increasing computing power and cumulative costs of generating new pharmaceuticals. With virtual screening (VS), it is possible to screen even hundreds of millions of compounds and select the best molecule candidates for *in vitro* testing instead of investing time and resources in analysing all molecules systematically in laboratories. However, there is a constant need to generate more reliable and effective software for VS. For example, molecular docking, one of the most central methods in structure-based VS, can be a very successful approach for certain targets while failing completely with others. However, it is not necessarily the docking sampling but the scoring of the docking poses that is the bottleneck. In this thesis, a novel rescoring method, negative image-based rescoring (R-NiB), is introduced, which generates a negative image of the ligand binding cavity and compares the shape and electrostatic similarity between the generated model and the docked molecule pose. The performance of the method is tested comprehensively using several different protein targets, benchmarking sets and docking software. Additionally, it is compared to other rescoring methods. R-NiB is shown to be a fast and effective method to rescore the docking poses producing notable improvement in active molecule recognition. Furthermore, the NIB model optimization method based on a greedy algorithm is introduced that uses a set of known active and inactive molecules as a training set. This approach, brute force negative image-based optimization (BR-NiB), is shown to work remarkably well producing impressive *in silico* results even with very limited active molecule training sets. Importantly, the results suggest that the *in silico* hit rates of the optimized models in docking rescoring are on a level needed in real-world VS and drug discovery projects.

KEYWORDS: molecular docking, negative image-based rescoring (R-NiB), negative image-based (NIB) model, virtual screening (VS), computer-aided drug discovery (CADD), brute force negative image-based optimization (BR-NiB)

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Biolääketieteen laitos
Farmakologia, lääkekehitys ja lääkehoito
SAMI KURKINEN: Sitoutumistaskun negatiivikuvat molekyylien telakoinnissa
Väitöskirja, 147 s.
Lääketutkimuksen tohtoriohjelma
Syyskuu 2021

## TIIVISTELMÄ

Tietokoneiden laskentatehojen ja lääketutkimuksen tuotekehityskulujen kasvaessa tietokonepohjaiset menetelmät kehittyvät jatkuvasti lääkekehityksessä. Virtuaaliseulonnalla voidaan seuloa jopa satoja miljoonia molekyylejä ja valita vain parhaat molekyyliehdokkaat laboratoriotestaukseen sen sijaan, että tuhlattaisiin aikaa ja resursseja analysoimalla järjestelmällisesti kaikki molekyylit laboratoriossa. Tästä huolimatta on koko ajan jatkuva tarve kehittää luotettavampia ja tehokkaampia menetelmiä virtuaaliseulontaan. Esimerkiksi telakointi, yksi keskeisimmistä työkaluista rakennepohjaisessa lääkeainekehityksessä, saattaa toimia erinomaisesti yhdellä kohteella ja epäonnistua täysin toisella. Ongelma ei välttämättä ole telakoitujen molekyylien luonnissa vaan niiden pisteytyksessä. Tässä väitöskirjassa tähän ongelmaan esitellään ratkaisuksi uudenlainen pisteytysmenetelmä R-NiB, jossa verrataan ligandinsitomisalueen negatiivikuvan muodon ja sähköstaattisen potentiaalin samankaltaisuutta telakoituihin molekyyleihin. Menetelmän suorituskykyä testataan usealla eri molekyylisarjalla, lääkeainekohteella, telakointiohjelmalla ja vertaamalla tuloksia muihin pisteytysmenetelmiin. R-NiB:n näytetään olevan nopea ja tehokas menetelmä telakointiasentojen pisteytykseen tuottaen huomattavan parannuksen aktiivisten molekyylien tunnistukseen. Tämän lisäksi esitellään ns. ahneeseen algoritmiin perustuva negatiivikuvan optimointimenetelmä, joka käyttää sarjaa tunnettuja aktiivisia ja inaktiivisia molekyylejä harjoitusjoukkona. Tämän BR-NiB-menetelmän näytetään toimivan ainakin tietokonemallinnuksessa todella hyvin tuottaen vaikuttavia tuloksia jopa silloin, kun harjoitusjoukko koostuu vain muutamista aktiivisista molekyyleistä. Mikä tärkeintä, *in silico* -tulokset viittaavat optimointimenetelmän osumaprosentin telakoinnin uudelleenpisteytyksessä olevan riittävän korkea myös oikeisiin virtuaaliseulontaprojekteihin.

AVAINSANAT: molekulaarinen telakointi, negatiivikuva, NIB mallin optimointi, R-NiB, BR-NiB, uudelleenpisteytys, virtuaaliseulonta, tietokonepohjainen lääkeainekehitys

# Table of Contents

# Abbreviations

| | |
|---|---|
| 1D | one-dimensional |
| 2D | two-dimensional |
| 3D | three-dimensional |
| AUC | area under the curve |
| AR | androgen receptor |
| BCC | body-centered cubic |
| BEDROC | Boltzmann-enhanced discrimination of receiver operating characteristic |
| BR-NiB | brute force negative image-based optimization |
| CADD | computer-aided drug design |
| COX2 | cyclo-oxygenase 2 |
| CPU | central processing unit |
| CYP3A4 | cytochrome P450 3A4 |
| DUD | Directory of Useful Decoys |
| DUD-E | Directory of Useful Decoys - Enhanced |
| EF | enrichment factor |
| ER | estrogen receptor alpha |
| ESP | electrostatic potential |
| FCC | face-centered cubic |
| GR | glucocorticoid receptor |
| HTS | high-throughput screening |
| HTVS | high-throughput virtual screening |
| $IC_{50}$ | half-maximal inhibitory concentration |
| ML | machine learning |
| MR | mineralocorticoid receptor |
| NEU | neuraminidase |
| NIB | negative image-based |
| OPLS | optimized potentials for liquid simulations |
| PDB | Protein Data Bank |
| PDE5 | phosphodiesterase type 5 |
| PLP | piecewise linear potential |

| | |
|---|---|
| PPARγ | peroxisome proliferator activated receptor gamma |
| PR | progesterone receptor |
| QSAR | quantitative structure-activity relationship |
| R-NiB | negative image-based rescoring |
| RMSD | root-mean-square deviation |
| ROC | receiver operating characteristic |
| RXRα | retinoid X receptor alpha |
| VS | virtual screening |

# List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Kurkinen S.T., Niinivehmas S., Ahinko M., Lätti S., Pentikäinen O.T. & Postila P.A. Improving the Docking Performance Using Negative Image-Based Rescoring. *Frontiers in Pharmacology*, 2018; 9, 260: https://doi.org/10.3389/fphar.2018.00260.

II Ahinko M.*, Kurkinen S.T.*, Niinivehmas S.P., Pentikäinen O.T. & Postila P.A. A Practical Perspective: The Effect of Ligand Conformers on the Negative Image-Based Screening. *International Journal of Molecular Sciences*, 2019; 20, 2779: https://doi.org/10.3390/ijms20112779.

III Kurkinen S.T., Lätti S., Pentikäinen O.T. & Postila P.A. Getting Docking into Shape Using Negative Image-Based Rescoring. *Journal of Chemical Information and Modeling*, 2019; 59, 8: 3584–3599. https://doi.org/10.1021/acs.jcim.9b00383

IV Kurkinen S.T., Lehtonen J.V., Pentikäinen O.T. & Postila P.A. Optimization of cavity-based negative images to boost docking enrichment in virtual screening. Manuscript.

*authors contributed equally

The original publications have been reproduced with the permission of the copyright holders.

# 1 Introduction

Proteins are large biomolecules responsible for practically every process within a cell. These processes include metabolism, transportation of molecules and signaling, to name a few examples. In biochemistry, ligands are substances such as small molecules, peptides or even ions that interact with their target, most often a protein, and regulate its function. A simple endogenous example is testosterone that binds to and activates its target protein androgen receptor and causing, for example, muscle growth. A defect or noxious function of a protein is a typical cause of diseases that are very often treated with drugs, which are typically natural or synthetic ligands that affect their target proteins causing biological responses.

Global pharmaceutical markets increased from 390 billion US dollars in 2001 to 1.2 trillion US dollars in 2018 (Aitken et al., 2019). Similarly, the drug costs for each patient have increased as drastically during the recent decades (Morgan et al., 2020). Additionally, it has been estimated that the median cost to bring one new drug to the market has been increased in less than two decades from hundreds of millions of dollars to even billions of dollars (Adams and Van Brantner, 2006; DiMasi et al., 2003, 2016). Thus, there is an increasing interest to cut down the expenses during the drug development process.

In drug development, the initial problem is to find compounds that bind to the target protein, such as a receptor, an enzyme or an ion-channel, and determine how strong this interaction is. This can be a very laborious, expensive and time-consuming task, as typically massive amounts of molecules need to be tested in order to find the suitable one. Computer-aided drug design (CADD) has evolved as one potential solution to these problems and consists of various computational tools and software to be used in different stages of a drug development process (Macalino et al., 2015; Song et al., 2009; Veselovsky and Ivanov, 2003). Thanks to the constantly increasing amount of computing power and freely available digital data considering protein structures, chemical interactions and molecule structures, CADD has become an essential part of drug design. Computationally, it is possible to screen a huge number of molecules much faster and more economically than in laboratories without chemicals, reagents or animal experiments. Nowadays, a large collection of different tools is available to evaluate, for example, molecular interactions, stability, oral availability and toxicity of potential drug candidates.

One of the central methods in CADD is molecular docking that estimates the correct binding pose of a ligand against its target, such as a receptor, and the affinity of this interaction (Morris and Lim-Wilby, 2011; Meng et al., 2011). With docking, it is possible to virtually screen several millions of molecules and select the most promising drug candidates for further testing. As the performance of regular docking is often case-selective, there is a need for more universal and accurate methods. In this thesis, the focus was to develop rescoring methods to improve the ability of docking to recognize the potential molecules binding to the target protein from the less favourable ones.

# 2 Review of the Literature

## 2.1 High-throughput virtual screening

During the early stages of drug development, the aim is to find molecules, called hits, which show some activity in binding to the target protein of interest. Typically, a huge number of molecules need to be tested to find even a few hit molecules, and the process is often likened to searching for a needle in a haystack. In traditional drug development, high-throughput screening (HTS) is an experimental method developed for this purpose (Bleicher et al., 2003; Hertzberg and Pope, 2000; Martis et al., 2011). Nowadays, it automatically, systematically and rapidly tests even hundreds of thousands of molecules and their binding affinities to the target a day. However, large HTS set-ups demand financial resources and equipment typically possible only for large pharmaceutical companies. Accordingly, HTS has often been criticized for its lack of efficiency; although some initial hit molecules might be found, there is little chance to develop it into a lead compound, *i.e.*, a compound with likely therapeutically suitable properties.

To ease at least some of these problems, computer-aided drug design (CADD) has evolved. In CADD, an analogous process to HTS is high-throughput virtual screening (HTVS) that can easily screen several millions of molecules with only marginal costs compared to HTS (Bajorath, 2002; Klebe, 2006; Schichet, 2004). In HTVS, data from the target structure or known bioactive ligands is used to evaluate the molecule potency. By computationally selecting a smaller subset of the potential molecules to be tested *in vitro*, the hit rates can be improved remarkably in comparison to a random HTS approach. This can greatly reduce the economical investments needed for the initial or later stage testing and molecule synthesis. HTVS can be divided into structure-based and ligand-based virtual screening (VS) methods, which are discussed more closely in the next chapters. However, these methods are not necessarily mutually exclusive and it is common that a drug discovery project mixes both approaches.

### 2.1.1 Ligand-based virtual screening

In ligand-based virtual screening, chemical data from the known active ligands are used to identify new molecules binding to the target (Geppert et al., 2010; Ripphausen et al., 2011). When there are no structure data from the target available, ligand-based methods are practically the only option to proceed in a drug discovery project. These approaches are based on the structure-activity relationship, a hypothesis that presumes compounds with similar properties having similar activities, and these properties are evaluated with algorithms. Similarity properties can be one-dimensional (1D), such as affinity data or lipophilicity, two-dimensional (2D), such as common fragments or bonding information or even three-dimensional (3D) including chemical space, electrostatics, shape or 3D pharmacophore properties. These properties aim to identify common factors between the active molecules and find the best matching structures from the screened compounds.

Similarity searching is a simple and computationally inexpensive method to find molecules similar to a reference structure (Maldonado et al., 2006; Willett et al., 1998). Certain properties of a reference structure, such as 2D substructures, are determined and compared against the molecules in the database of interest. The more similar the molecule is with the reference structure, the higher score it gets. For example, one of the popular similarity methods is based on molecular fingerprints. Each property of a molecule, such as every fragment, is encoded in a binary string (the property either exists or not) called a fingerprint. The presence of similar fingerprints in both query structure and the examined structure is compared and the similarity is calculated with Jaccard/Tanimoto coefficient or with some other similarity coefficients (Haranczyk and Holliday, 2008; Willett, 2006).

Another ligand-based method is quantitative structure-activity relationship (QSAR) that aims to create correlation between a set of ligands and their pharmacological activity (Verma et al., 2010). It can be used, for instance, in the modelling of binding affinity, toxicity or absorption, distribution, metabolism, and excretion (ADME) properties, such as oral bioavailability (Hu and Aizawa, 2003; Kruhlak et al., 2007; Toropova et al., 2010). For a case example, the physicochemical properties of 232 drugs with known bioavailability were analysed by Yoshilda and Topliss to create a model for the human bioavailability (Yoshida and Topliss, 2000). Lipophilicity together with the presence of certain structural parameters, such as ketones or phenolic hydroxyl groups, were used to generate a predictive QSAR model with a Spearman rank correlation coefficient of over 0.8. The generated model can be used to predict the human bioavailability of unknown molecules by calculating the aforementioned structural parameters.

Ligand-based pharmacophore modelling aims to generate a 3D model containing the relevant features of known active ligands needed for the binding to the protein target (Yang, 2010). Algorithms consider the flexibility of a set of active ligand 3D structures and align them to recognize the relevant features to generate a pharmacophore model. These features can be, for example, possible hydrogen bond acceptors and donors, aromatic rings or hydrophobic interactions. Finally, the model is used in VS to search hits that fit best to the model (Ananthula et al., 2008; Yu et al., 2007).

### 2.1.2 Molecular docking is a key method in structure-based virtual screening

Structure-based VS uses the 3D structure of a target protein to find new drug candidates (Cheng et al., 2012; Lyne Paul D., 2002). The protein 3D structure can be obtained, for example, from X-ray crystallography, nuclear magnetic resonance spectroscopy or it can be computationally modelled. The structure-based methods can be used also for novel targets, as no ligand information is necessarily needed. Because there were over 175,000 protein structures available at RSCB Protein Data Bank (PDB; www.rcsb.org) on the 18th of March 2021, structure-based methods are very plausible approaches in drug discovery projects (Berman et al., 2002; Burley et al., 2019).

In structure-based pharmacophore modeling, the 3D information about the target protein or protein-ligand complex is used to create the model. The binding site is analysed to recognize possible interaction points for small molecules, such as locations for hydrogen bonding (Yang, 2010). In contrast to the ligand-based pharmacophore modeling, a set of known active ligands are not needed for the model generation. However, the receptor information can be included, for example, in QSAR models to improve the reliability of the approach (Cherkasov et al., 2008; Sippl et al., 2001).

Molecular docking is a fundamental part in CADD and structure-based virtual screening (Morris and Lim-Wilby, 2011; Meng et al., 2011; Shoichet et al., 2002). It aims to predict the binding pose and affinity of a molecule when it binds to the target protein (Figure 1, right). This is done by evaluating steric and electrostatic interactions of the molecule-protein complex. In the first phase, docking tries to generate the correct binding pose of a molecule with a sampling algorithm, and in the second phase, a scoring function is used to evaluate the binding affinity and to recognize the most suitable pose out of the pool of alternative docking solutions. In VS, the scoring function needs also to separate the active molecules from the inactive ones, in other words, recognize the binding affinity differences between dissimilar molecules.

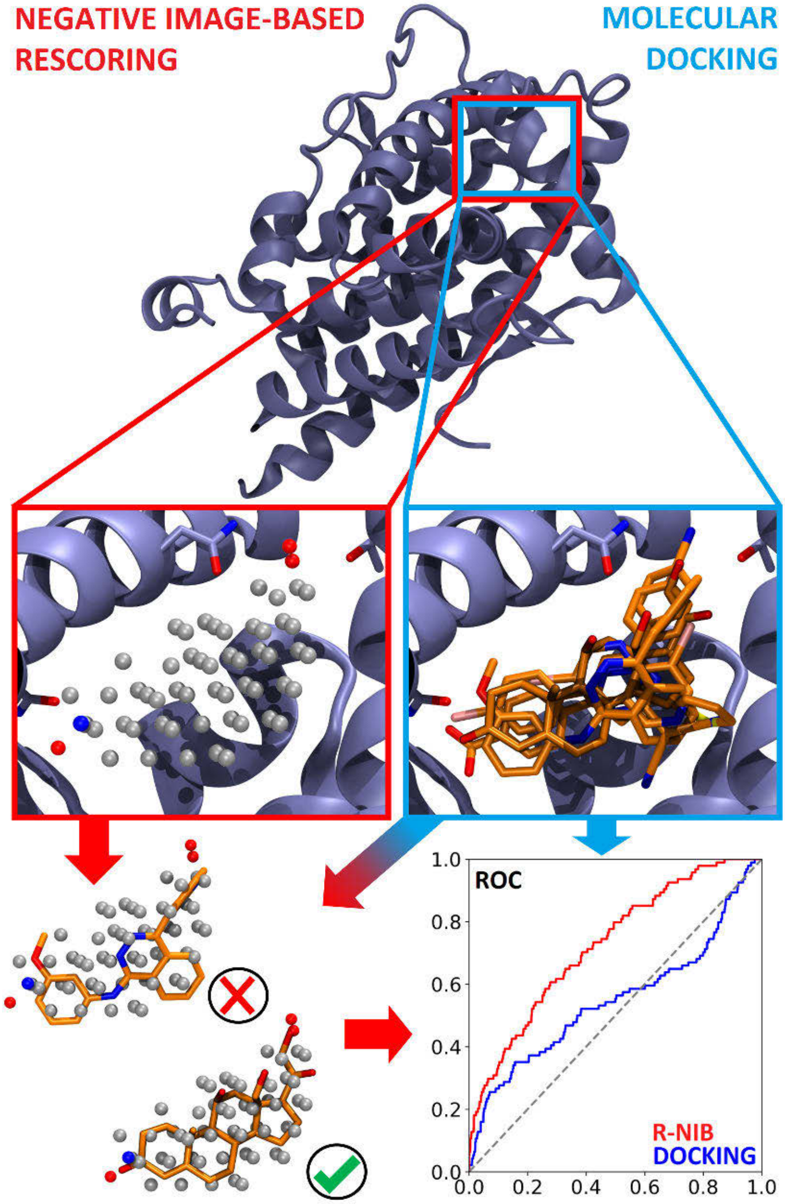

**Figure 1.** Principles of negative image-based rescoring and molecular docking. First, in both methods, the centroid of the target protein ligand binding cavity is determined to generate a negative image-based model and dock the molecules (top, center). The generated model considers both shape and electrostatics of the cavity (center, left). The gray cavity atoms are nonpolar whereas the red and blue atoms correspond negative and positive charges of the pocket, respectively, caused by, for example, polar amino acids side chains. The shape/electrostatic similarity of the docking solutions (center, right) is compared against the generated NIB model, and the new ranking is calculated based on this similarity score (bottom, left). The rescoring approach (red curve) outperforms regular docking approaches (blue curve) and greatly improves the enrichment metrics (bottom, right). Modified from Studies II and III.

Thanks to the increasing computing power, the original rigid docking with a simple lock-and-key approach, in which the molecule and target are treated as rigid bodies, has evolved to flexible docking that allows the flexibility of a molecule during the docking process (Kuntz et al., 1982; Rosenfeld, 1995; Sousa et al., 2006). This has still led to induced-fit docking, a method that treats the desired parts of the target protein as flexible. Nowadays, plenty of docking software is available from commercial to freeware that use different sampling approaches and scoring functions (Pagadala et al., 2017). A selection of them is introduced in Table 1.

**Table 1.** A collection of docking programs used in virtual screening.

| Software | Sampling method | Scoring function | License* | References |
|---|---|---|---|---|
| AutoDock | Stochastic (genetic algorithm) | Empirical | Free | Morris et al., 2010 |
| AutoDock Vina | Stochastic (quasi-Newton method) | Empirical+knowledge-based | Free | Trott and Olson, 2010 |
| DOCK | Systematic (incremental construction) | Force field-based | Academic | Allen et al., 2015 |
| FlexX | Systematic (incremental construction) | Empirical | Commercial | Rarey et al., 1996 |
| Glide | Systematic+stochastic (incremental construction with Monte Carlo sampling) | Empirical+force field-based | Commercial | Friesner et al., 2014; Halgren et al., 2004 |
| GOLD | Stochastic (genetic algorithm) | Originally force field-based, also empirical possible | Commercial | Gareth et al., 1997 |
| MCDOCK | Stochastic (Monte Carlo sampling) | Force field-based | Academic | Liu and Wang, 1999 |
| LigandFit | Shape complementarity with stochastic (Monte Carlo sampling) | Empirical+shape comparison+similarity clustering | Academic | Venkatachalam et al., 2003 |
| PLANTS | Stochastic (ant colony optimization) | Empirical | Academic | Korb et al., 2006; 2009 |
| Surflex | Systematic (Incremental construction algorithm) with surface-based molecular similarity method | Empirical | Academic | Jain, 2003 |

* Software licensing information: Free is freely downloadable for everyone, academic is only for non-profit institutions, and commercial usage needs a purchasable license.

Computationally even more demanding molecular dynamics (MD) and quantum mechanical approaches treat the entire molecular system, such as protein and its ligand, flexibly and simulates its behavior over time (Ganesan et al., 2017). MD simulations give information about the relevant conformational changes and other time-dependent variations that can have a significant role in the protein-ligand binding properties. Although not suitable for HTVS or simulating long time periods, MD gives one solution to the fundamental docking problem regarding protein flexibility.

#### 2.1.2.1 Sampling methods

In the sampling phase, docking considers the ligand flexibility and aims to predict its plausible binding orientation in the protein binding site (Brooijmans and Kuntz, 2003; Kitchen et al., 2004; Sousa et al., 2006). The search methods can be roughly divided into two categories: systematic and random/stochastic search methods.

Six degrees of rotational and translational freedom describe the movement of a rigid molecule. In the case of a flexible molecule, the conformational degrees of freedom must also be considered during the sampling process. While the number of rotatable bonds in the molecule increases, also the conformational degrees of freedom rapidly increase. The principal idea in a systematic search is to cover all the degrees of freedom during the sampling (Friesner et al., 2004; Halgren et al., 2004; Sousa et al., 2006). Because this is typically computationally impossible, as the number of combinations easily gets far too high, several methods have been developed to facilitate the calculations (Table 1). For example, only a limited number of fixed rotations is performed per rotatable bond to cover the whole 360° scene. In the fragmentation method that utilizes the incremental construction algorithm, the docked molecule is first broken into fragments, and the main fragment is anchored in the cavity and fixed. The remaining fragments are then docked separately before rejoining them back to the anchor fragment again (Allen et al., 2015; Jain, 2003).

In random search methods, the conformational space of the molecule is screened by making random changes to the rotatable bonds. Thus, the results might vary a little when repeating the process. To name a few examples, in the Monte Carlo method, the starting point is a randomly generated initial conformation (Hart and Read, 1992). This conformation is then treated in a stepwise process generating small random changes to the molecule. If the new conformation fulfils a certain energy threshold, it is selected for the next round. This process is continued until a new conformation cannot be created. Genetic algorithms are based on the idea of biological evolution and utilize it in docking (Gareth et al., 1995, 1997; Guan et al., 2017). It starts from the initial population of randomly generated molecule conformations. With operations such as crossovers, mutations and recombinations, a new population is made from the parent population. The fitness of each individual is calculated with a certain energy function and the best conformation is selected for the next cycle. These steps are repeated until the optimization is finished.

In the last part of docking, a scoring function is used to evaluate the binding affinity of each conformation to recognize the correct pose (Kitchen et al., 2004; Sousa et al., 2006). These typically aim to evaluate the differences in binding free energies. For HTVS, computationally expensive free-energy simulation methods, such as Molecular Mechanics/Generalized Born Surface Area or Poisson-Bolzmann Surface Area (MM-GBSA and MM-PBSA), are usually too time-consuming. Thus, several simplifications need to be done to enable the efficient use of the scoring functions, and particularly the effects of entropy, protein flexibility and solvent are difficult to evaluate.

#### 2.1.2.2 Scoring functions

Docking scoring functions can be divided into different subtypes (Huang et al., 2010; Liu and Wang, 2015). Force field-based scoring functions evaluate the sum of the ligand-target interaction energy and the internal energy of the ligand (Gareth et al., 1997; Morris et al., 2010; Sousa et al., 2006). There are several force fields available, and they are used to calculate electrostatic, van der Waals and steric interactions between the docking pose and the binding site to rank the molecules and their conformations. Empirical scoring functions utilize experimentally studied binding energies and 3D data to calculate coefficients for several terms with regression analysis (Eldridge et al., 1997; Guedes et al., 2018). These coefficients are used to approximate the binding energies between the compound and the target protein by calculating a diverse set of terms, such as hydrophobic contacts or a number of hydrogen bonds, between them. Knowledge-based scoring functions utilize PDB or other large 3D databases to collect information about the intermolecular interactions of the functional groups between the ligand and protein (Gohlke et al., 2000; Muegge, 2000). These are used to evaluate the corresponding atomic interaction potential in the docking solutions. Thus, the focus is in reproducing the experimentally confirmed poses rather than energetically the most favorable ones. Furthermore, it is common that the docking software use combinations of several scoring functions (Table 1) (Friesner et al., 2004; Halgren et al., 2004; Korb et al., 2009; Trott and Olson, 2010).

During the last decade, the machine learning (ML) approaches have become commonplace also in CADD (Lo et al., 2018; Vamathevan et al., 2019). In the field of scoring function development, ML and neural networks have been shown to be promising (Ain et al., 2015). The construction of ML scoring functions is based on diverse training sets that contain, for example, the structures of protein-ligand complexes and their affinity data. The training sets are used to train and optimize the scoring function, whereas a smaller test set is used for validation. As the freely available online structural databases are increasingly comprehensive, the construction of diverse training and test sets is becoming effortless (Table 2). For example, a freely available RF-Score-VS scoring function is trained with over 15,000 active and 890,000 inactive molecules docked into 102 targets, and it is shown to easily outperform the docking program Autodock Vina and its scoring function (Wójcikowski et al., 2017).

**Table 2.** A selection of molecular databases available online for virtual screening.

| Database | No of molecules | Website | Other |
|---|---|---|---|
| Asinex | 530,000 | www.asinex.com | Commercial |
| ChEMBL | 2,000,000 | www.ebi.ac.uk/chembl | Non-commercial |
| Chembridge | 1,300,000 | www.chembridge.com | Commercial |
| DrugBank | 14,000 | www.drugbank.ca | Non-commercial, contains data of drugs and their targets |
| Enamine | 2,700,000 | www.enamine.net | Commercial |
| Molport | 7,000,000 | www.molport.com | Commercial, compiles data from other suppliers |
| NCI | 260,000 | www.cactus.nci.nih.gov | Non-commercial, for cancer and AIDS research |
| Specs | 350,000 | www.specs.net | Commercial, drug-like molecules |
| SuperNatural II | 330,000 | http://bioinf-applied.charite.de/supernatural_new | Non-commercial natural compound collection |
| Vitas-M | 1,400,000 | www.vitasmlab.biz | Commercial |
| Zinc | 230,000,000 | https://zinc.docking.org | Non-commercial |

## 2.2 Measuring the docking performance

Although docking is a popular and effective method in HTVS, there is an ongoing debate as to which of the docking algorithms and scoring functions, if any, is better than the other in reproducing and selecting the correct binding pose among the different conformations and inactive molecules (Bursulaya et al., 2003; Cross et al., 2009; Elokely and Doerksen, 2013; Ferrara et al., 2004; Mohan et al., 2005; Pagadala et al., 2017; Wang et al., 2003, 2016). Rather, the docking success is typically case-specific, and the results vary depending on the target and docking program. However, the problem has mainly focused on the scoring functions rather than the sampling algorithms, *i.e.*, docking is often able to create the correct binding pose but it has difficulties in recognizing it (Dariusz et al., 2010; Pagadala et al., 2017; Warren et al., 2006). Several studies show that the consensus scoring strategy, which combines the results of several scoring methods or docking software, typically generates the best results (Charifson et al., 1999; Cheng et al., 2009; Houston and Walkinshaw, 2013; Oda et al., 2006). However, the selection of a suitable type of consensus scoring might be challenging, and docking with several programs is very time-consuming. Thus, the development of more accurate and universal scoring functions and other post-processing techniques is still a topical issue.

Not only the unbiased comparison of different docking programs is difficult, but also the evaluation of the docking performance can be complicated. The most obvious way for that is to examine how well docking is able to reproduce the experimentally verified binding conformation, such as the co-crystallized ligand in an X-ray structure. This can be done, for example, by calculating the root-mean-square deviation (RMSD) between the atoms in the superimposed co-crystallized ligand structure and the docked molecule (Kramer et al., 1999; Li et al., 2010; Pagadala et al., 2017). The smaller the value, the better the docked molecule superimposes with the crystallized binding conformation. The RMSD method is not an ideal approach, as the difficulty of evaluating small molecules or determination of a suitable threshold is difficult. Moreover, ligands may have multiple binding orientations, and, for example, the crystallographic binding pose is not necessarily the only correct one (Mobley and Dill, 2009). Furthermore, docking success is highly dependent on the target structure, and it is possible that the "verified binding conformation" is not the correct one, as the crystal structure represents only a certain snapshot of a dynamic protein-ligand complex formation (Jain, 2009). Although containing several problems, the RMSD method is probably the most used approach when evaluating the docking poses against the experimentally determined ones (Kirchmair et al., 2008).

Typically, a significant number of active ligands lack any crystal structure making previously described comparisons impossible. Another approach is to examine how well a docking program is able to separate active molecules from the inactive ones and ignore the RMSD comparison to the actual structures. For this approach, a reliable benchmarking set is needed that contains a set of known active molecules and a sufficiently large number of decoy molecules that are supposed to be inactive (Lagarde et al., 2015; Réau et al., 2018). Although originally used in medicinal imaging, the receiver operating characteristic (ROC) curve is a graphical presentation that can be used to plot the active molecule rate against the decoy rate illustratively (Swets, 1979). In other words, the plot describes the sensitivity vs. specificity relation by plotting, for example, the ranking list from docking so that the x-axis contains the percentage of the found decoys whereas the y-axis contains the percentage of the found actives. Area under the curve (AUC) is a metric that is calculated from the ROC curve and roughly describes the probability of a single molecule in a test set to be recognized as active or decoy (Figure 1, bottom right) (Hanley and McNeil, 1982; Truchon and Bayly, 2007). It takes values from 0 to 1 in which 0 indicates the perfectly inaccurate ability to separate actives from decoys: all decoys get a higher score than the active molecules. On the contrary, AUC of 1 indicates that all active molecules are ranked higher than decoys. In practice, AUC varies between 0.5 (random picking or the 50/50 probability to separate molecules) to 1.

In HTVS, in which even millions of compounds are screened against one target, the ability to find those few active molecules out of the inactive ones is essential. Because only a tiny part of the best-ranked molecules from HTVS can be screened *in vitro*, the active molecules need to be ranked high enough. Thus, in addition to AUC, another interesting metric is the early enrichment. Although there are several approaches to define the enrichment factors (EFs), in this thesis, EF describes the percentage of actives ranked higher than the certain percentage of decoys (Lätti et al., 2016). For example, EF 5 % = 10.5 implies that 10.5 % of the active molecules are ranked higher than 5 % of the top-ranked decoy molecules in the set. Typically, the main interest is in maximizing the very early enrichment, such as EF 1 % or EF 0.1 %: if there are 1,000,000 molecules in a database and only 1,000 of them can be screened *in vitro* according to the docking score, the active molecules should be ranked higher than 0.1 % of the best inactive ones.

A high AUC value does not tell anything about the early enrichment. Rather, it describes the overall performance of the method. Ironically, the AUC value can be high, but the method is still too inefficient for practical HTVS. Similarly, a high EF factor does not guarantee the overall success. A high EF with the low AUC value describes that the method recognizes only a part of the active molecules well: it can be specific only to a certain ligand subgroup. Thus, although typically more complex and less intuitive, metrics that consider better both the early enrichment and the overall performance have been developed, such as the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) or the robust initial enhancement (RIE) (Truchon and Bayly, 2007; Zhao et al., 2009). BEDROC includes a user adjustable parameter so that the importance of the early part of the ROC curve can be adjusted. However, the ratio of actives and inactives as well as the test set size need also be considered to minimize the error.

## 2.3 Molecule preparation for virtual screening

As well as selection and preparation of the target structure, the molecule preparation is a critical part in HTVS (Figure 3) (Madhavi et al., 2013). In molecular databases, the compounds are typically downloadable in 2D structure-data file (SDF) format, which is capable of including associated data, such as rotatable bond number, lipophilicity or molecular weight (Dalby et al., 1992). Other formats are also possible, such as a simplified molecular-input line-entry system (SMILES), which describes the structure as a 1D line notation saving a significant amount of storage capacity (Weininger, 1988). However, the molecules need to be converted to the 3D format before docking. Additionally, stereochemistry of the structures needs to be retained during the conversion, or created extensively if not determined, and the relevant protonation state or states should be considered (Brink and Exner, 2009; Castaño and Maurer, 2015). Instead of covering all possible ionization states, in HTVS, covering the most relevant ones at physiological pH is more desirable. This minimizes the number of the screened compounds and prevents the occurrence of false positives caused by infeasible protonation states.

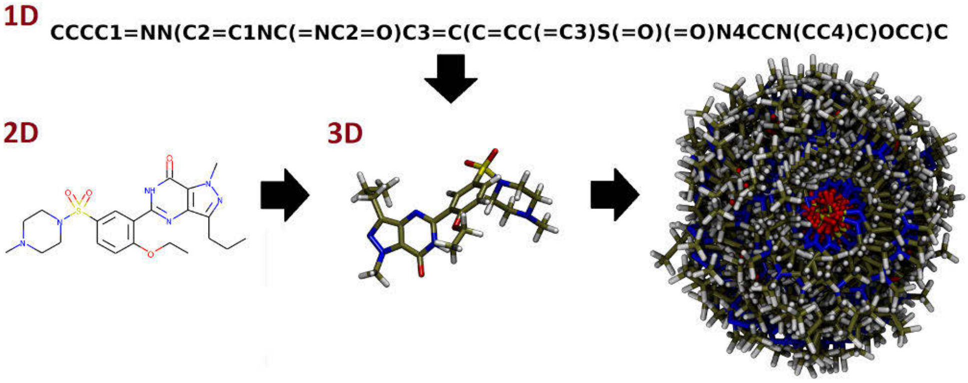
1D
CCCC1=NN(C2=C1NC(=NC2=O)C3=C(C=CC(=C3)S(=O)(=O)N4CCN(CC4)C)OCC)C
2D
3D

**Figure 2.** Sildenafil conversion from 1D or 2D to 3D format. One- or two-dimensional structures are converted to 3D with plausible protonation states and partial charges. If needed, molecule conformations are created to cover the energetically plausible poses. In the case of molecules with a high number of rotatable bonds, such as sildenafil with seven of them, the conformation number easily gets very large.

Furthermore, although usually underestimated, tautomerism is another issue that need to be considered, and particularly the most common tautomers should be identified and considered (Martin, 2009; Pospisil et al., 2003). Tautomerism does not only affect the shape, hydrophobicity or fragmentation of the molecule, but it may also have a critical effect on the hydrogen bonding.

Determining the partial atomic charges is another crucial factor in the molecule preparation (Kukic and Nielsen, 2010; Pissurlenkar et al., 2009). Today, there are several force fields, such as Gasteiger-Marsili, Merck molecular force field (MMFF) and optimized potentials for liquid simulations (OPLS), to assign partial charges for the molecules and the development is ongoing (Gasteiger and Marsili, 1980; Halgren, 1996; Jorgensen and Tirado-Rives, 1988; Roos et al., 2019). In order to efficiently approximate the partial charges of even millions of molecules in a database, several simplifications for the calculations need to be made, the greatest of which is treating the atomic charges as fixed. To overcome this issue, computationally more demanding polarizable force fields have also been developed, mainly for MD simulations (Halgren and Damm, 2001; Lin and MacKerell, 2019).

After all these steps have been considered, the database is typically ready for docking. However, if the database is used in similarity comparisons or the negative image-based (NIB) screening (described in the next chapter), also the molecule conformations need to be generated (Figure 3, right) (Hawkins, 2017). As described previously in Chapter 2.1.2.1, the number of rotatable bonds is a key factor in determining the number of conformations to be generated, and the conformation generation is based on the same principles as the docking sampling. To cover all energetically plausible poses of the molecule without the combinatorial explosion, a balance between the number of conformations and rotatable bonds needs to be found (Figure 3).

For molecule preparation, there are several computer programs from commercial to freely distributed ones that are able to handle most, if not all, of the aforementioned issues, such as ChemAxon (www.chemaxon.com) or Maestro tools (www.schrodinger.com), RDKit (www.rdkit.org), OpenBabel, SPORES and Balloon (Brink and Exner, 2009; O'Boyle et al., 2011; Vainio and Johnson, 2007). Accordingly, different workflows are available for molecule preparation and docking, as well as for other cheminformatics approaches to ease the process (Gally et al., 2017; Kooistra et al., 2018; Pearce et al., 2009).

## 2.4 Negative image-based approaches

Although the computing power is constantly increasing and parallel computing can be utilized with many central processing units (CPU) even with regular computers, docking is a relatively slow method. It is typical that several docking runs are needed for numerous protein conformations to obtain satisfactory results. Furthermore, it might be necessary to screen several molecular databases, the biggest of which can easily contain millions of molecules (Table 2). Additionally, several conformations for each compound might be needed to generate to find the correct one. Thus, balancing between speed and accuracy is the essence in docking, as with any other HTVS method.

In NIB screening, the idea is to determine the properties of the ligand binding area by utilizing protein structural data to create a model that can be used in HTVS (Bauer and Mackey, 2019; Fukunishi et al., 2006; Niinivehmas et al., 2015; Tran-Nguyen et al., 2019; Virtanen and Pentikäinen, 2010). This thesis focuses on the use of software called Panther that generates a negative image from the ligand binding area and is specially developed for HTVS purposes (Niinivehmas et al., 2015). The ligand binding site is filled with positively and negatively charged atoms together with neutral atoms to create a ligand-like model that preserves the complement shape and electrostatic properties of the area. This model is then used in similarity searches to screen molecular databases using the program ShaEP (Vainio et al., 2009). Originally developed for the similarity evaluation of ligand-sized molecules, it superimposes and compares the shape and electrostatic similarity between the NIB model and the molecule of interest and calculates the similarity score based on these two properties. As no tedious sampling of molecules against the target protein is needed, the NIB screening is much faster in comparison to docking but still considers the properties of the binding pocket. Accordingly, it gives a high weight for the importance of shape in protein-ligand interaction while still allowing some overlap with the protein. A more detailed description of ShaEP is given in Chapter 4.5.2.

The idea of determining the binding cavity originates from the binding site prediction problem for which a high number of computational methods have been developed, such as VOIDOO, GRID, SiteMap and VolSite (Desaphy et al., 2012; Goodford, 1985; Halgren, 2007, 2009; Kleywegt and Alwyn Jones, 1994). The importance of shape in HTVS has been recognized for a long time and there are several shape comparison programs available including ROCS, MSC and USR, for example (Ballester and Richards, 2007; Masek et al., 1993; Rush et al., 2005). A docking program DOCK that determines the binding area with spheres was among the first programs where the shape complementarity was considered (Allen et al., 2015; Kuntz et al., 1982). Some newer docking programs, such as LigandFit, MS-DOCK and QSDock, are even based on the shape complementarity of the ligand binding cavity and this information is used for the compound selection (Goldman and Wipke, 2000; Sauton et al., 2008; Venkatachalam et al., 2003). Nevertheless, although a very important aspect in the ligand binding, the importance of shape complementarity is typically underestimated in docking software, and the scoring functions focus more on electrostatics (Hawkins et al., 2007; Kahraman et al., 2007; Kirchmair et al., 2009; Virtanen and Pentikäinen, 2010; Warren et al., 2006).

This thesis introduces two methods that are based on the NIB screening. The first method, called negative image-based rescoring (R-NiB), ranks the molecule poses generated by a docking program (Table 1) based on their similarity to the negative image of the binding cavity. It compares the shape and electrostatic similarity between the NIB model and the docking poses without superimposing them, *i.e.*, the docked molecule is kept in place in the binding pocket during the similarity comparison (Figure 1, left). In contrast to the NIB screening, R-NiB requires docking and only rescores the docking solutions, whereas the NIB screening scores the molecule conformations generated *ab initio*. The second method, labelled as brute force negative image-based optimization (BR-NiB), optimizes the NIB model using known active and decoy molecules as a training set (Figure 2). Based on the observation that Panther typically generates too bold NIB models, the NIB models are optimized by removing excess cavity atoms one by one.

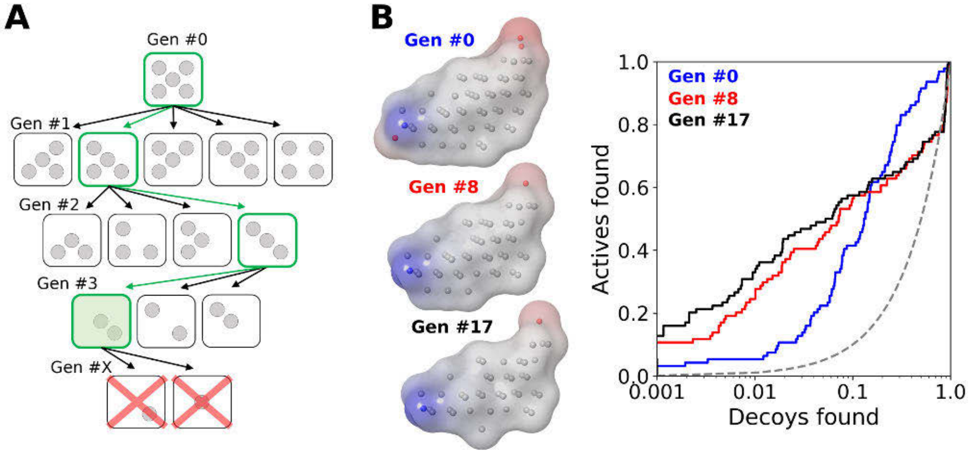

**Figure 3.** A simplified presentation of the brute-force negative image-based optimization. A) A diverse set of active and inactive molecules are used as a training set for the negative image-based (NIB) model optimization. Generation #0 (Gen #0, in this example only five cavity atoms) corresponds to the original NIB model. Cavity atoms (gray spheres) are removed one by one and the enrichment metrics are calculated for each new model similarly to the NIB rescoring (Figure 1, left). The model, now containing *n*-1 cavity atoms (Gen #1), that produces the best enrichment, is selected for the next round (green arrow). This cycle is continued until the attainment of Gen #X in which the enrichment improvement is no longer achieved. B) Model shrinks and the enrichment improves during the optimization process. The semi-logarithmic ROC plot shows the improvement of the early enrichment from Gen #0 to Gen #17. Modified from Study IV.

BR-NiB is based on the greedy algorithm principle, which has already been utilized in bioinformatics, such as in DNA alignment and phylogenetics (Florea et al., 1998; Steel, 2005). Furthermore, greedy approaches have been used in docking sampling in incremental construction algorithms, for example (Allen et al., 2015; Rarey et al., 1996). Greedy approaches avoid the combinatorial explosion of the exhaustive search by always selecting the locally optimal solution, such as a certain number of the energetically best conformations, to the next step. Despite being ideal for solving only problems with the optimal substructure, such as the local energy minimum of a molecule conformation instead of the global minimum, greedy algorithms are often able to find a solution close enough to the optimal one.

# 3 Aims

The aim of this thesis was to develop methods for using the NIB models in improving molecular docking-based VS performance. Preliminary testing by Prof. Olli Pentikäinen and M.Sc. Sakari Lätti suggested that, in addition to the shape-based NIB screening (Niinivehmas et al., 2015), the NIB models could also be utilized in rescoring of the docking solutions. Inspired by this idea, this thesis introduces two novel methods that use the NIB models in re-ranking docked compounds: 1) The R-NiB method ranks the docking poses based on their shape and electrostatic similarity to the cavity-based NIB model. 2) The BR-NiB method optimizes the NIB model using known active and inactive molecules included in a training set. In Study I, the aim was to show that the NIB models could effectively be used to rescore the docking solutions of PLANTS docking program with different target proteins. In Study II, it was shown that the R-NiB method works effectively also with alternative docking software. In Studies I and II, the performance of the R-NiB method was compared to a few other scoring functions and docking approaches. In Study III, practical instructions are provided for performing the NIB screening and R-NiB, and the effect of the computational molecule 3D preparation and conformation generation to their performance was analysed. In Study IV, the aim was to show that by optimizing the NIB model with a greedy search method utilizing benchmarking sets, the ability of the model to separate the active molecules from the inactive ones could be improved remarkably for docking rescoring.

# 4 Materials and Methods

The most relevant methods used in the thesis are described here, and the used programs are listed in Table 3. A more detailed description of the methods can be found from the original publications.

**Table 3.** The most central methods and programs used in Studies I-IV.

| | **Method** | **Version** | **References** | **Original publication** |
|---|---|---|---|---|
| **Molecular databases** | DUD<br>DUD-E | | (Huang et al., 2006)<br>(Mysinger et al., 2012) | I, II<br>I-IV |
| **Negative image generation** | Panther | 0.8.15 | (Niinivehmas et al., 2015) | I-IV |
| **Protein preparation** | Reduce<br>Bodil | 3.24<br>0.9 | (Word et al., 1999)<br>(Lehtonen et al., 2004) | I-IV<br>I-IV |
| **Docking** | PLANTS | 1.2 | (Korb et al., 2006, 2009) | I-IV |
| | Glide | 2018-1 | (Friesner et al., 2004; Halgren et al., 2004) | III |
| | GOLD | 5.6.3 | (Gareth et al., 1997) | III |
| | DOCK | 6.8 | (Allen et al., 2015) | III |
| | AutoDock | 4.2.6 | (Morris et al., 2010) | III |
| | AutoDock Vina | 1.1.2 | (Trott and Olson, 2010) | III |
| **Alternative rescoring methods** | X-Score<br>Smina | 1.2.1<br>11.9.2017 | (Wang et al., 2002)<br>(Koes et al., 2013) | I<br>III |
| **Visualization** | Raster3D<br>MolScript<br>VMD | 3.0.2<br>2.1.2<br>1.9.2 | (Merritt and Murphy, 1994)<br>(Kraulis, 1991)<br>(Humphrey et al., 1996) | I,II<br>I,II<br>I-IV |
| **Data analysis** | ShaEP | 1.0.7.915<br>1.1.2.1036<br>1.1.3<br>1.3.1 | (Vainio et al., 2009) | I<br>II<br>III, IV<br>IV |
| | Rocker | 0.1.4 | (Lätti et al., 2016) | I-IV |

## 4.1 Benchmarking sets

To evaluate the performance of different scoring functions, docking software and other VS methods, reliable benchmarking sets are essential. In this thesis, the Directory of Useful Decoys (DUD) and the Database of Useful Decoys: Enhanced (DUD-E) were used for the validation of the methods (Huang et al., 2006; Mysinger et al., 2012).

DUD is a benchmarking set that contains 40 protein targets with 2950 known ligands (Huang et al., 2006). For each ligand, there are 36 physicochemically similar but topologically different decoy molecules selected from the ZINC database (Irwin and Shoichet, 2005). DUD-E is an upgraded version of the DUD set and it should be a more comprehensive, less biased and more challenging benchmarking set (Mysinger et al., 2012). It contains 102 protein targets with 22,886 active molecules in total. For each active molecule, the number of decoy molecules is increased to 50 obtained from the ZINC database.

Although criticized for containing biases or a too low ratio of active and decoy molecules, the DUD and DUD-E databases are widely used and still reasonable as benchmarking sets for VS protocols (Chaput et al., 2016; Chen et al., 2019; Good and Oprea, 2008; Sieg et al., 2019). They were also used in the original NIB screening study with Panther (Niinivehmas et al., 2015). Thus, these benchmarking sets were a logical selection also for the thesis work to easily compare the performance of the methods.

ChEMBL is a freely available online database of bioactive molecules (Mendez et al., 2019). It was used in Study IV to select active molecules for mineralocorticoid receptor (MR) and neuraminidase (NR) validation sets. Because the ratio of active molecules in the DUD-E set is as "high" as at least 1.5 %, the validation sets were generated to better correspond a more realistic situation in which the ratio between actives and decoys was lowered to 0.014 %. Active molecules with different affinity values (the half maximal inhibitory concentration ($IC_{50}$) of < 1 µM, < 50 µM and 1-50 µM) and not present in the DUD-E set were randomly selected for the validation sets using compounds from the commercial Specs database (Table 2) as decoy molecules. Naturally, some molecules in the Specs set might be false negatives by binding to the protein targets. However, this unlikely occurrence should increase the reported hit rates only slightly.

## 4.2 Target protein preparation

All protein structures used in the studies were acquired from the PDB (Berman et al., 2002; Burley et al., 2019). For a certain target, mainly the same PDB entry was used as listed in the DUD and DUD-E databases. Because of the usage of several docking programs, it was essential to use different tools for the protein preparation.

However, the aim was to keep the preparation process as similar as possible between the docking programs. For PLANTS (Studies I-IV) and GOLD (Study III) docking, the necessary structure editing was done with Bodil Molecular Modeling Environment and the protonation was performed with Reduce (Lehtonen et al., 2004; Word et al., 1999). In the case of Glide docking (Study III), the target preprocessing was performed with Protein Preparation Wizard in Maestro (Schrödinger Release 2018-1, Epik, Schrödinger, LLC, New York, NY, USA). In AutoDock and Vina docking (Study III), AutoDockTools provided with AutoDock was used in protein editing. When using DOCK for docking (Study III), Dock Prep tool was used in molecule visualization software UCSF Chimera 1.12 (Goddard et al., 2004).

## 4.3 Small molecule preparation

Molecular database preparation and 3D conversion were done with Maestro tools (Studies I, III and IV) using OPLS3 force field. In Study II, the preparation and molecule conformation generation were performed for comparison also with OpenBabel, RDKit and Marvin tools. Commercial Maestro tools are widely used in CADD and considered as one of the state-of-the-art tools in the field. Maestro tools are fast and easy to use for large molecular database conversions including all the essential steps in molecule preparation: the 3D conversion, protonation at physiological pH, tautomerization, partial charge calculation and conformer generation. Maestro tools were also used in the original publication of Panther (Niinivehmas et al., 2015).

However, alternative tools were also used for comparison in Study II. The aim was to select software that is freely available, reliable enough, widely used and cover tools as comprehensively as possible for complete molecule preparation. OpenBabel is a freely available tool for molecule file conversion and is also able to, for example, 3D conversion, conformer generation and partial charge calculation. RDKit, also freely available software, is based on Python programming language and contains tools and scripts for molecule 3D conversion and conformer generation. However, at least during the preparation of Study II, it lacked some functionalities such as tautomer generation. Marvin tools are freely available for academic institutions and contain several software for thorough molecule preparation and file conversion.

In Study III, ligand-based screening with ShaEP was performed. The *ab initio* generated conformers were prepared with ConfGenX in Maestro (Schrödinger Release 2018-1, Epik, Schrödinger, LLC, New York, NY, USA) using OPLS3 force field. The conformer number was limited to 64.

## 4.4 Molecular docking

In this thesis, PLANTS was used for docking in all of the studies. PLANTS is docking software based on a stochastic ant colony optimization sampling method and includes two empirical scoring functions: PLANTS$_{\mathrm{CHEMPLP}}$ as a default and PLANTS$_{\mathrm{PLP}}$ (Table 1). PLP function was only used for docking rescoring in Study I. Two piecewise linear potential (PLP) functions are used in both functions, one for repulsive or attractive interactions and the other for only repulsive interactions, to evaluate the steric complementarity between the ligand and the protein target. In the case of ChemPLP, the empirical scoring function ChemScore implemented in the docking program GOLD is also used to determine angle-dependent terms in hydrogen and metal bonds (Eldridge et al., 1997; Verdonk et al., 2003). ChemScore evaluates the total free energy change during the ligand binding and is based on the affinity data from 82 protein-ligand complexes. A more detailed description of the PLANTS scoring functions or sampling method is provided in the original PLANTS publications (Korb et al., 2006, 2009).

PLANTS was selected as the primary docking program, because it is straightforward to use and considered reliable in our experience. It always generates the desired amount of docking poses. Moreover, PLANTS practically always docks all the molecules in the dataset and does not skip them if they, for instance, lack some physicochemical properties. PLANTS is also reviewed to be very convincing docking software in reproducing the binding pose of a ligand co-crystallized with the protein (Ren et al., 2018).

In Study III, four other docking software were used to study the performance of the R-NiB method with other popular docking programs. GOLD uses genetic algorithm for sampling (Table 1) but the same default scoring function ChemPLP as PLANTS in versions newer than 5.0. However, ChemScore, the original GoldScore and some other functions are also available.

Glide (Study III) is a part of Maestro tools and uses a combination of systematic and stochastic sampling with a combination of empirical and force field-based scoring functions. The empirical scoring function GlideScore, based on ChemScore, is used to rank the different molecules and evaluate the ligand binding affinity. Then, the force field-based Emodel scoring function is used to select the best pose of a single docked ligand. It is based mainly on the Coulombic and van der Waals energies between the receptor and ligand but some contribution from GlideScore is also received. There are three modes available in Glide docking based on the balance between speed and accuracy, HTVS, SP and XP. Only Glide HTVS and SP were used in the Study II, as XP is too time-consuming for VS approaches. Glide SP is described being the recommended choice for a standard VS study and performs a more exhaustive sampling than Glide HTVS that is designed for docking very large databases.

DOCK (Study III) uses an incremental construction algorithm in sampling. Particularly in the newer versions of DOCK, several scoring functions, or even their combinations, can be used. In Study II, the grid-based scoring was selected as it was also used in the rigid and flexible ligand docking tutorials (Lang, 2018). In DOCK, a high number of parameters are user adjustable. However, the principle was to use as default settings as possible to keep the results somewhat comparable between the other docking programs and protein targets.

Since version 4, AutoDock has been utilized a genetic algorithm for docking sampling and a semiempirical free energy force field for scoring (Table 1). The scoring function evaluates the free energy change during ligand binding with pair-wise atomic terms including hydrogen bonding, desolvation, electrostatics, dispersion and repulsion (Huey et al., 2007). The empirical approach is used to evaluate the contribution of dissolved water by utilizing the data from 188 protein-ligand complexes.

AutoDock Vina is a newer docking program than AutoDock both developed by the same Molecular Graphics Lab. It uses a quasi-Newton sampling method with a combination of empirical and knowledge-based scoring functions by utilizing data from both known protein-ligand complexes and affinity measurements (Table 1). The scoring function calculates the protein-ligand binding affinity by evaluating steric, hydrophobic and hydrogen bonding interactions.

## 4.5 Negative image-based rescoring

The NIB screening is based on the usage of two software tools. First, Panther is used to generate a NIB model from the binding cavity (Figure 1). Secondly, the actual rescoring of the docked molecules is done with ShaEP by comparing the shape and electrostatics similarity between the NIB model and the docking solutions using the *noOptimization* option, which keeps both the orientation of the docking solutions and the NIB model fixed. Finally, Rocker is used to calculate the enrichment metrics (Lätti et al., 2016). The EF's were calculated as true positive rates when 1 or 5 % of the decoys have been discovered.

### 4.5.1 Cavity detection and negative image-based model generation using Panther

Panther is software developed to predict small molecule binding into proteins with NIB screening. It generates a NIB model based on the shape and electrostatic complementarity of the binding cavity properties considering protein environment, such as possible explicit water molecules, ions or cofactors. The outputted model is a negative image of the binding cavity and it can be used in VS approaches. Although there are plenty of parameters the user can adjust, in this thesis, the principle was to keep the setting as simple as possible. The settings and NIB model input files are described and available in the original studies (I-IV). The main options adjusted in the studies are listed below:

- *Center* determines the centroid of the generated NIB model. Typically, the center coordinates of the co-crystallized ligand in the protein structure file were used.
- *Box radius* defines the radius how far the filler atoms are generated from the centroid in ångströms. This value varied depending on the target and the size of its binding cavity.
- *Ligand distance limit* determines the dimensions of the NIB model in ångströms. Generated atoms are not farther than this limit from the specified ligand. The *box radius* option still affects, *i.e.,* the atoms that are farther from the center than defined with *box radius*, are removed anyway.
- *Packing method* determines the lattice how the atoms are packed in the NIB model. Either the body-centered cubic (BCC) or the denser face-centered cubic (FCC) method was used.

### 4.5.2 Similarity comparison with ShaEP

ShaEP is freely downloadable software for similarity comparison and is used in the original NIB screening protocol (Niinivehmas et al., 2015). Thus, it is an obvious choice also for R-NiB and BR-NiB to evaluate the similarity between the NIB model and the docked molecule without geometry optimization. Accordingly, in comparison to the other similarity comparison software, the advantage of ShaEP is that it compares both the shape and electrostatic similarity in a relatively straightforward manner: both scores get values from 0 to 1, and the user can adjust the weight of both similarities to the total score if the default equal 50/50 weight distribution is not suitable (Vainio et al., 2009). The 3D conformation of the molecule is used to generate vertexes, connected with graphs, around the molecule, and at these points, the electrostatic potential (ESP) is calculated with a Coulombic function and computed in volts:

$$\varphi_E = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\sum_i \frac{q_i}{d_i}$$

in which $1/4\pi\varepsilon_0$ is the Coulomb constant, $\varepsilon_r$ is the relative static permittivity of the medium, $q_i$ is the partial charge of atom $i$ in Coulombs, and $d_i$ is the Euclidean distance between the atom and the vertex in meters.

Shape is described as a histogram vector. The shape-density at the vertex coordinates r is the sum of all individual atomic densities and expressed as a spherical Gaussian surface:

$$\rho_i(r) = p_i^{-\alpha_i(r-R_i)^2}$$

in which $R_i$ is the atomic coordinate, $p_i$ is the amplitude (set to $2\sqrt{2}$), and $\alpha_i$ is the decay factor. In the next phase, graph matching is calculated with a backtracking search algorithm to find maximal subgraph isomorphism between a graph of the NIB model and docked molecule pose.

Both R-NiB and BR-NiB are based on the similarity comparison without overlay optimization, *i.e.*, ShaEP only scores the similarity between the existing docking poses and the NIB model rather than generating the optimal alignment between the template and target. The scoring of the template and target alignment is based on the overlap of their shape-densities and field-graphs. A more detailed description of ShaEP algorithm is reported in the original publication (Vainio et al., 2009).

### 4.5.3 Brute force negative image-based optimization

BR-NiB (Study II) is the next step for R-NiB, where the NIB model is optimized using a training set of molecules (Figure 2). The size of the benchmarking set used for the model optimization was varied to evaluate this effect on the results: typically, large benchmarking sets are not necessarily available in the early-stage drug discovery projects. The model optimization starts by calculating the shape and ESP score between the original NIB model and every conformation of the docked molecules in the benchmarking set with ShaEP. Then, Rocker is used to calculate the enrichment metrics AUC, EF and BEDROC based on the ranking order provided by the similarity comparison. In the next step, the NIB model cavity atoms are removed one by one, and the same enrichment metrics are calculated for every new NIB model containing $n$-1 atoms. If the selected enrichment is improved, the model with the best enrichment is picked for the next editing round, and the cycle starts again.

The NIB model optimization was tested with alternative target metrics AUC, EF 1 % and BEDROC with $\alpha = 20$ (in Study IV referred as BR20) to see which of the metrics produce the best results. The $\alpha$ value gives a weight for the early part of the ROC curve in the BEDROC calculation, *i.e.*, it determines the importance of the early recognition (Truchon and Bayly, 2007). If $\alpha$ is high, more weight is given for the early part of the accumulation curve. When $\alpha$ is 20, approximately 85 % from the BEDROC results comes from the first 10 % of the ranked molecules in the benchmarking set.

## 4.6 Other rescoring methods

To compare the R-NiB performance to other rescoring methods, X-Score and SMINA were used in Studies I and II, respectively. X-Score has three empirical scoring functions calibrated with 200 protein-ligand complexes, and they are combined to form a single consensus scoring function (Wang et al., 2002). All the scoring functions evaluate the binding free energy by calculating the sum of van der Waals, hydrogen bonding, deformation and hydrophobic effects. In X-Score, three different algorithms are used to calculate hydrophobic effects, resulting in three different scoring functions.

Smina is a fork of AutoDock Vina and developed for scoring and minimization (Koes et al., 2013). It enables the usage of custom scoring functions but also contains its own empirical scoring function trained with 293 protein-ligand structures. Based on Pearson's correlation and manual selection, van der Waals, solvation, hydrogen bond and torsion terms were used in the default scoring function.

## 4.7 Figure preparation and data analysis

Figures used in Studies I-IV as well as Figures 1-4 used in this thesis presenting protein structures, NIB models or ligands were generated using Bodil, VMD, MolScript, Raster3D and Maestro. The enrichment metrics were calculated and the ROC plots were generated with Rocker that uses the Wilcoxon statistics in the error estimation (Hanley and McNeil, 1982).

# 5 Results

## 5.1 The effect of ligand preparation (II)

Ligand-based virtual screening methods, such as similarity comparison or pharmacophore modelling, require a careful generation of molecule 3D structures and conformations. Unlike R-NiB that uses the molecule conformations generated by external docking sampling, the NIB screening requires low-energy 3D molecule conformations for the similarity comparison with rigid superimposition. In Study II, molecule preparation for cyclo-oxygenase 2 (COX2) benchmarking set obtained from DUD and DUD-E databases was performed using four different conformer generators: Maestro tools, Obabel, RDKit and Marvin tools. The effect of these molecule preparation tools for the NIB screening results was evaluated using different NIB model compositions. In the case of Maestro tools, ConfGen (Watts et al., 2010) was used to generate the conformers, and it was shown to generate a remarkably smaller number of conformations in comparison with the other software (Study II; Table 1).

When only a single molecule conformation was used, RDKit was shown to be the best program for COX2 molecule preparation with the DUD set while Maestro tools came in a close second (Study II; Table 2). In the case of the DUD-E set, molecules generated with Marvin tools produced the best results. When working with multiple conformations, RDKit was shown to be the best program for the DUD set whereas Marvin tools were the best choice for DUD-E (Study II; Table 3). Surprisingly, the computationally expensive generation of multiple conformations was not very beneficial for COX2, and the enrichment metrics, particularly the EF, was increased only when generating the conformers with RDKit, or occasionally with Marvin tools, for the DUD set (Study II; Table 2 vs. 3).

## 5.2 Negative image-based rescoring improves docking results (I, III)

In Study I, eleven benchmarking sets from both DUD and DUD-E were docked and rescored with R-NiB. The model generation was based either on the co-crystallized ligand dimensions (*ligand distance limit)* in the protein structure or a certain radius (*box radius*) from the co-crystallized ligand centroid but keeping the other settings as default as possible. For R-NiB, the results are summarized in Table 4.

**Table 4.** Summary table of the R-NiB results using DUD and DUD-E benchmarking sets (Study I). If the AUC or the EF at 1 % of the R-NiB method was higher and outside the error range than that of PLANTS docking, it is marked with x. If the EF 1 % was over five percentage points higher than in the case of original docking, the x is underlined. The DUD set did not contain NEU and CYP3A4 sets, whereas DUD-E lacked $ER_{ag}$ and $ER_{antag}$ sets. Modified from Study I.

| **Target**[1] | **DUD** | | | | **DUD-E** | | | |
|---|---|---|---|---|---|---|---|---|
| | **AUC** | | **EF 1%** | | **AUC** | | **EF 1%** | |
| | LIG[2] | BOX[3] | LIG[2] | BOX[3] | LIG[2] | BOX[3] | LIG[2] | BOX[3] |
| **ER** | x | x | x̲ | x | | | | |
| **$ER_{ag}$** | | | x̲ | x̲ | - | - | - | - |
| **$ER_{antag}$** | x | | x̲ | x̲ | - | - | - | - |
| **AR** | x | x | x̲ | | x | x | x̲ | x |
| **GR** | x | x | x̲ | x̲ | x | x | x | x |
| **MR** | x | x | x̲ | | x | x | x̲ | x |
| **PPARγ** | | | x̲ | | | | | |
| **PR** | | | x̲ | x̲ | x | x | x | x |
| **RXRα** | x | x | x̲ | x̲ | x | x | | |
| **COX2** | x | x | x̲ | x̲ | x | | | |
| **PDE5** | | | x̲ | x̲ | | | | |
| **NEU** | - | - | - | - | x | x | x̲ | x |
| **CYP3A4** | - | - | - | - | | | x | x̲ |

[1] ER, estrogen receptor alpha (mixed set of agonists and antagonists); $ER_{ag}$, estrogen receptor alpha agonists; $ER_{antag}$, estrogen receptor alpha antagonists; AR, androgen receptor; GR, glucocorticoid receptor; MR, mineralocorticoid receptor; PPARγ, peroxisome proliferator activated receptor gamma; PR, progesterone receptor; RXRα, retinoid X receptor alpha; COX2, cyclo-oxygenase 2; PDE5, phosphodiesterase type 5; NEU, neuraminidase; CYP3A4, cytochrome P450 3A4

[2] NIB model generated with the ligand distance limit option

[3] NIB model generated with the box radius option

According to AUC, PLANTS docking performed moderately well with the DUD and DUD-E benchmarking sets, although the results varied significantly depending on the target (Study I; Tables 2 and 3). AUC varied from 0.60 to 0.95 with DUD sets and 0.54 to 0.85 with the more demanding DUD-E sets. At its best, the R-NiB approach improved AUC even 0.20 units, such as in the case of glucocorticoid receptor (GR) in the DUD dataset (0.60 to 0.80) or androgen receptor (AR) in the DUD-E dataset (0.54 to 0.76), but typically the improvement was about 0.10 units or fewer. Some targets, such as peroxisome proliferator activated receptor gamma (PPARγ) and phosphodiesterase type 5 (PDE5), were particularly demanding for the R-NiB approach.

The performance of docking in the early enrichment (EF 1% and EF 5 %) varied remarkably depending on the target (Study I; Tables 4 and 5). Particularly some nuclear receptors, such as estrogen receptor alpha (ER) in DUD (EF 1% = 0.0) or GR in DUD-E (EF 1% = 1.2), were problematic cases. However, EF 1% was even 69.1 with PPARγ (DUD) and 21.7 with ER (DUD-E). Still, the R-NiB methodology performed convincingly and was able to improve the early enrichment greatly for every DUD set tested. In the case of DUD-E, the success was not as systematic. However, in most cases, EF was improved also for DUD-E. At its best, EF 1% was improved even 10 percentage points (pp) (1.5 to 13.0 and 4.1 to 13.3 in the cases of AR and neuraminidase (NEU), respectively).

With cytochrome P450 3A4 (CYP3A4), the advantage of R-NiB can be only seen in the late enrichment (EF 5% and later); otherwise, the performance was on the same level with docking (Study I; Tables 3 and 5 and Figure 3). PPARγ, with a large binding cavity, was clearly a more difficult case for R-NiB. Similarly, ER and PDE5 with mixed ligand sets were demanding to rescore with a single NIB model. Separating the ER agonists and antagonists from the DUD-E set and rescoring them individually showed that the NIB model separated the antagonists much better than the agonists: for agonists, AUC and EF 1% were only 0.73 and 16.1 whereas for antagonists, the values were 0.88 and 61.1, respectively. In Study III, the R-NiB results for PDE5 were improved when using two different NIB models. The first one was made by limiting the cavity dimensions based on the crystal structure of sildenafil and the other on tadalafil (Study III; Figure S12). The best results were obtained when two models were fused together (Study III; Tables 3 and S7).

Overall, R-NiB was particularly able to improve the early enrichment of the datasets (Study I; Figures 2 and 3). This result can also be seen from the summarization table (Table 4). Furthermore, the *ligand distance limit* option in the NIB model generation was shown to be more efficient than using only the *box radius* option. For both docking and R-NiB, the DUD-E set turned out to be more demanding. Especially, PDE5, PPARγ and ER were demanding targets for R-NiB.

A practical guide for performing R-NiB or the regular NIB screening was described in detail in Study II.

### 5.2.1 The performance with alternative docking software (III)

After testing the R-NiB performance with PLANTS, a logical continuation was to study the rescoring performance also with other docking software. Thus, R-NiB was tested together with five other popular docking programs, Glide (HTVS and SP modes), GOLD, DOCK, AutoDock and AutoDock Vina (referred as Vina), using five diverse targets from the more demanding DUD-E test set: retinoid X receptor alpha (RXRα), COX2, PDE5, mineralocorticoid receptor (MR), and NEU. Again, the philosophy was to use each docking software settings as default as possible. However, the settings were fine-tuned during the NIB model generation in the case of some targets. Thus, the models and results are not fully comparable with the previous Study I.

Especially Glide and DOCK skipped a relatively large number of molecules, some of them even the active ones, during the docking process. In order to keep the results comparable, the skipped molecules were added to the bottom of the ranking list in the order that corresponds random picking. However, no docking software was clearly better than the other, and AUC results varied from a complete failure (*e.g.*, AUC 0.48 in the case of MR with Glide) to great (*e.g.*, AUC 0.88 in the case of RXRα with AutoDock) between the targets and the used software (Study III; Table 2; Figure 3). From the early enrichment point of view, Glide and Vina produced typically slightly better enrichment than the others did. Although DOCK seemed to perform worse than the other software, it is probably because it cannot be used as productively as the other programs with "default settings", and fine-tuning the settings could likely have improved the performance.

R-NiB was shown to work well with different docking software and targets and it improved the results in most cases, although some docking programs are more suitable for R-NiB than others (Study III; Table 3, Figure 3). The typical AUC improvement was as high as 0.10-0.20 units, but even higher improvement could occasionally be seen (*e.g.*, AUC improvement of 0.29 in the case of MR with GOLD or 0.23 in the case of RXRα with DOCK docking). However, also moderate AUC improvement between 0.04 and 0.09 was seen in several cases (*e.g.*, COX2 and MR with DOCK or RXRα and NEU with PLANTS docking). Comparably, the EF improvement was usually over 10 pp although many cases performed even better (*e.g.*, EF 1% improvement of 40.5 pp in the case of RXRα with DOCK or EF 5% improvement of 66.3 pp in the case of NEU with GOLD docking). R-NiB failed in the early enrichment enhancement mainly in the case of RXRα when docking with Vina or AutoDock, that performed exceptionally well already without rescoring (*e.g.*, AUC=0.88 and EF 1%= 54.2 in the case of AutoDock). Particularly, Glide was the difficult case for the R-NiB method, and neither AUC nor EF values were improved with the only exception being NEU (*e.g.*, AUC improvement of 0.21 pp and EF 1% improvement of 21.4 pp with HTVS docking).

Although the results varied depending on the target, the best results were usually obtained when the docking results of GOLD were rescored with R-NiB: AUC varied between 0.70-0.93 and EF 1% between 10.3 and 62.6 with every target (Study III; Table 3). Particularly, DOCK, GOLD and PLANTS were well suited for R-NiB: the AUC and EF values for PLANTS or DOCK rescoring were almost as high as with GOLD if excluding some docking failures, such as MR docked with DOCK. Rescoring of AutoDock and Vina docking poses improved the results in most cases, but technically the rescoring process itself was more laborious. Accordingly, Vina and AutoDock docking solutions contain only polar protons, so the R-NiB methodology was also tested when adding all protons to the docking poses. Again, the results varied and it is difficult to say whether R-NiB works better or not when including all protons.

### 5.2.2 Comparing the performance to other methods (I, III, IV)

Rescoring of the datasets with the alternative program X-Score did not outperform R-NiB (Study I; Tables 2-5). In fact, X-Score showed to be more case-specific performing exceptionally well, for example, with RXRα (AUC 0.97 and EF 1% = 70 with DUD set). However, some nuclear receptors, such as MR and AR, were particularly difficult for X-Score. Furthermore, the sets difficult for R-NiB were problematic also for X-Score: PDE5, PPARγ and ER, to some extent. X-Score outperformed R-NiB only with the RXRα and GR sets of DUD.

The docking solutions were also scored with PLANTS$_{\mathrm{PLP}}$ scoring function included in PLANTS. However, the success of this scoring function was not very convincing, and only COX2 set seemed to benefit from this scoring method (Study I; Tables 2-5). Although occasionally, PLANTS$_{\mathrm{PLP}}$ produced slightly better AUC or EF than the original docking scoring function PLANTS$_{\mathrm{CHEMPLP}}$, only the EF 1% and EF 5% values of the COX2 set from DUD-E were improved in comparison to the R-NiB method.

In Study III, the R-NiB performance was compared with the default empirical scoring function of Smina. Although clearly more case-specific than R-NiB, Smina seemed to be a more efficient competitor than the previously mentioned approaches (Study III; Table 4, Figure 3). As well as R-NiB, also Smina worked well with PLANTS, GOLD and DOCK. Particularly, NEU and MR were difficult targets for it, as well as the docking program Vina and Glide. However, Smina outperformed R-NiB with PDE5 and in some other occasional cases. Similar to R-NiB, also Smina improved particularly the early enrichment rather than AUC.

In HTVS methods, balancing between speed and accuracy is a key issue. Thus, the time consumption of the R-NiB similarity comparison was compared with X-Score and PLANTS$_{PLP}$ rescoring in Study I. If excluding the time used for the NIB model generation, R-NiB is an ultrafast method being at least 10 times faster than X-Score and took approximately 2 to 4 ms/comp depending on the size of the NIB model and docked ligands. Scoring of the docked molecules with a single scoring function PLANTS$_{PLP}$ was still two times slower than R-NiB.

The intended usage of ShaEP is ligand-based similarity comparison: the structure of a query molecule is used to identify similar compounds from molecular databases. As the NIB model can be made by restricting the cavity dimensions with the co-crystallized ligand pose in the target structure (*ligand distance limit* option), it was probed in Study III if the performance of a simple ligand-based screening is analogous to R-NiB (Study III; Table S6). A ligand pose included in the crystal structure was used as a query, and the molecule conformations of the set were generated *ab initio*. The results show that although the method occasionally produced better AUC and EF values than docking, such as in the case of COX2 (AUC = 0.71 and EF 1% = 19.3) that contains structurally mainly similar ligands as celecoxib that was used as the reference structure, R-NiB clearly outperformed the traditional ligand-based approach.

In Study IV, similarity comparison was performed between the crystal structure ligand and docking poses generated by a docking program without any geometry optimization similarly to the R-NiB method (Study IV; Table S6). Again, the results show that particularly the COX2 but also the MR set suited well for ligand-based similarity comparison outperforming the original docking. In the case of COX2, the EF 1 % enrichment was even better in comparison to R-NiB, but this can be counted as an exception and otherwise, R-NiB performed clearly better.

To exclude the possibility that R-NiB only finds a certain subgroup of ligands, *i.e.,* the ligands similar to the one used in the NIB model generation, the active molecules of the benchmarking sets were clustered based on Daylight's Fingerprint and Tanimoto similarity (Study III; Figure S11). The clustering shows that the ability of R-NiB to find different molecule subgroups was not any weaker than that of Smina or the original docking.

### 5.2.3 Shape is the determining factor in the scoring process (III)

In Study III, the shape and ESP scores were separated from the R-NiB results to evaluate the importance of both factors in the rescoring process (Study III; Table S5). The absolute numbers of the shape score were practically always at least two times higher than the ESP score making the shape similarity the major element in the total score.

When calculating the AUC and EF values using only the shape or ESP score, in most cases, the best enrichment metrics were obtained when using the shape score (Study III; Table S5). In some cases, such as MR and PDE5, the EF values were even better with many docking programs if using only the shape score rather than the equal 50/50 weight distribution for both scorings. However, typically the shape score alone did not produce the best enrichment, and the ESP score needed to be considered when calculating the total score. Particularly, in the case of RXRα, the ESP score alone produced better enrichment metrics with all the other docking programs but PLANTS. Nevertheless, it did not outperform the default 50/50 scoring.

The better the molecule aligns against the NIB model, the better similarity score it acquires. Thus, it could be presumed that the R-NiB performance is just based on its ability to give the molecules docked further from the cavity, and NIB model, center a lower score instead of actually considering the shape and electrostatic similarity. As the same center coordinates were used for both docking and NIB model generation, this presumption was evaluated by comparing the average distance of 10 % of the top-ranked docking and R-NiB poses from the cavity center (Study III; Table S4). The results show that in the case of RXRα, COX2 and MR, the average distance of the R-NiB-selected poses is not remarkably shorter than that of the docking poses. However, in the case of PDE5 with a spacious cavity and NEU containing a surface pocket, the average distance is over 0.5 Å shorter for the R-NiB poses with several docking programs.

### 5.2.4 Consensus scoring approach has potential (I, II)

Several studies indicate that a combination of several scoring functions or docking methods generally produce better enrichment than any of the methods alone (Charifson et al., 1999; Cheng et al., 2009; Houston and Walkinshaw, 2013; Oda et al., 2006). As previously described, this can be also seen in ShaEP scoring in which the combination of shape and ESP scores produced better enrichment than either of the scoring functions alone. Inspired by these observations, this approach was tested by normalizing and combining the original PLANTS docking score and ShaEP similarity score and adjusting their relative weight (from 0 to 1).

In Study I, the results show that it is difficult to determine an optimal weight between docking and similarity score, and the results vary depending on the target (Study I; Tables 6 and 7). For example, in the case of DUD sets, the best EF 1% enrichment was obtained when using the weight of 1.0 for ShaEP scoring with MR and RXRα sets, *i.e.*, the total score came entirely from the ShaEP rescoring. In the case of AR, only the weight of 0.25 for ShaEP rescoring was needed for the best enrichment. When using the optimal weight, all DUD and DUD-E datasets produced better EF 1% and EF 5 % enrichments than the original docking. However, the AUC was not necessarily improved. Similar results were obtained also from Study II where the consensus scoring was tested with DUD and DUD-E sets for COX2 using two different target structures and three alternative NIB model compositions. The optimal weight for ShaEP was shown to vary between 0.60-0.95 depending on the used target structure and NIB model (Study II; Table 4).

However, determining the optimal weight requires testing and is not possible with new targets or targets that lack a benchmarking set. Thus, equal 50/50 weight for both docking and ShaEP score was used to test if it is possible to find a more universal approach for consensus scoring (Study I; Tables 6 and 7; Figures 2-3). In Study I, the results show that although the equal weight always produced better early enrichment than the original docking scoring (PPARγ in DUD-E set being the only exception), the regular R-NiB approach without considering the docking score produced better enrichment in several cases. In Study II, the equal weight did not work as well with the COX2 set of DUD as with the DUD-E (Study II; Table 4). Although the early enrichment was practically improved in every case when using the DUD-E set, the equal weight approach worked only when the NIB model was generated using the *box radius* option in the case of the DUD set.

### 5.2.5 Can the correct ligand pose be found? (I, II, III)

In Study III, the binding poses selected by R-NiB, SMINA rescoring or docking software were compared with the co-crystal ligand poses, if available. Out of the five benchmarking sets, 31 active molecules with a protein-ligand structure available in the PDB database were found. Although there was a relatively limited amount of structures to study, the similarity between the crystal structure and the docked pose was compared using the root mean square deviation (RMSD) calculations. By finding 18 out of 31 poses (61 %), Vina was shown to be the best software in recognizing the correct binding orientation with the RMSD value less than 2.0 (Study III; Table 5). If considering only the poses with the RSMD similarity less than 1.0, GOLD and Glide (in SP mode) were the most successful programs (39 % recognition). For every docking software, MR was the easiest case in reproducing the correct binding orientation whereas PDE5 was the most difficult.

From the rescoring point of view, it is more relevant to recognize if docking software is able to sample the correct binding pose despite the docking scoring function does not recognize it as the best solution. Therefore, all outputted poses generated by the docking programs were evaluated to study if any of the docking solutions is the correct pose. With this approach, PLANTS and Vina were shown to be the best programs by reproducing 87 % and 84 % of the crystallized ligand poses, respectively, with the RMSD less than 2.0 (Study III, Table 5). When studying the R-NiB ability to select the pose closest to the crystal structure as the best pose, it was shown that R-NiB was slightly more successful than the original docking software or Smina rescoring. 27 out of 31 ligands with a crystal structure, PLANTS generated a conformation with less than 2.0 Å of RMSD in comparison to the crystal structure pose. 16 of them were recognized as the best pose by R-NiB (original docking recognized 15, Smina 8). GOLD was able to generate corresponding conformations for 23 ligands, and R-NiB recognized 21 of them as the best pose (original docking and Smina recognized 17).

In Studies I and II, the best-ranked docking pose and the R-NiB-selected pose were compared with the experimentally determined crystal structures. The best-ranked docking pose of and COX2 inhibitor, that closely reminds celecoxib, was ranked to be the 8585th best molecule by PLANTS (Study II; Figure 6). R-NiB scored another conformation of that molecule to be the best pose and it resembles more closely the crystal structure of celecoxib. More importantly, R-NiB ranked this molecule significantly higher, 3rd best, than the original docking. In the case of hydrocortisone, which is an agonist of the MR receptor and highly resembles aldosterone, docking with PLANTS resulted as a binding pose that is the most probably docked a wrong way around (Study I; Figure 5). The R-NiB methodology selected the conformation that resembles more the crystal structure of aldosterone to be the best pose. However, the improvement in the ranking was only minor: this pose was ranked to be as the 13th best pose by R-NiB, whereas PLANTS ranked it to be the 17th best.

### 5.2.6 Model generation is a critical step (I, II, III)

There are two different approaches to generate a NIB model with Panther after determining the cavity location. The model dimensions can be simply determined with a certain radius, and cavity atoms farther from the determined centroid are removed (*box radius* option). This approach considers the entire shape of the binding pocket. Another option is to restrict the NIB model dimensions with the help of the crystallized ligand in the target protein structure (*ligand distance limit* option). This approach removes the cavity atoms farther than a certain radius from the ligand and highlights the shape occupancy of the existing ligand. It is also important to consider the used lattice when filling the model with cavity atoms (*packing method* option), which affects directly the density of the model.

If possible, the model is typically better to be constrained using an existing ligand present in the target structure (*ligand distance limit* option). This approach was shown to produce the best results in most cases in Study I (Table 4) and was also used when preparing the NIB models in Study III. Similar results were obtained also in Study II. Here, the R-NiB performance was evaluated using different approaches in the NIB model generation (Study II; Figure 4). Although only one target, COX2, was used, NIB models generated with the *ligand distance limit* option produced clearly better improvement when rescoring both DUD and DUD-E benchmarking sets (Study II, Table 4). However, the results were not as straightforward when considering the effect of the used lattice. Although denser FCC packing performed better in most of the cases, BCC worked better in DUD-E when using PDB structure complexed with celecoxib (PDB code 3ln1) as a target structure.

In fact, the similar trend to the *ligand distance limit* option can be seen also with the original NIB screening method based on the rigid superimposition (Niinivehmas et al., 2015). Typically, the usage of this setting produced higher AUC and enrichment values than using the *box radius* option regardless of the program used in the molecule preparation (Study II; Tables 2 and 3). The usage of a single or multiple ligand conformers did not affect these results. However, a regular NIB screen was shown to work better for COX2 when using a less dense BCC lattice in the model generation.

In Study I, the NIB models used in the original NIB screening publication (Niinivehmas et al., 2015) were shown to give quite different enrichment results in R-NiB in comparison to the models generated intentionally for the R-NiB study (Study I; Tables 2-5). Similarly, adjusting the NIB models for Study III, such as just removing a single critical polar atom, improved particularly the early enrichment even remarkably (Study III; Table S2). For instance, the EF 1% of RXRα set was increased from 6.9 to 21.4 in Study III in comparison to Study I.

## 5.3 Model optimization pushes the performance to the next level (IV)

Optimization of the NIB model was performed based on a greedy optimization method, named brute force negative image-based rescoring (Study IV). The cavity atoms are removed one by one and after every removal, enrichment metrics for the training set of molecules are calculated (Figure 2). The new model producing the best enrichment, being short by one cavity atom, is selected for the next round: another cavity atom is removed, and the enrichments are calculated again. This cycle is continued until the improvement is no longer acquired. To learn which of the enrichment metrics, AUC, EF or BEDROC is the most suitable one for BR-NiB, the optimization was done by using all the three metrics separately (Study IV, Table S4). Although there were target-specific differences, the BEDROC was generally shown to produce the best enrichment metrics. Optimizing the model using EF caused BR-NiB to stop relatively fast (e.g., only two and nine generations for MR and COX2, respectively) and it was clearly shown to be the worst metrics to use in the optimization. With BEDROC, the optimization proceeded typically the largest number of generations (e.g., 17 and 24 generations in the case of MR and COX2, respectively). In the case of MR, the optimization with BEDROC produced good EF metrics (EF 1% = 33.0) but relatively low AUC (0.76). In contrast, the optimization with AUC produced better AUC (0.86) but lower EF metrics (EF 1% = 18.1).

### 5.3.1 Enrichment metrics are boosted comprehensively (IV)

The performance of the BR-NiB optimization was tested with seven benchmarking sets from the DUD-E database (COX2, RXRα, MR, NEU, PDE5, ER and PPARγ) divided randomly into training and test sets of different sizes. The full 100:100 ratio was used as a starting point where the whole benchmarking set was used for the model optimization (Study IV; Table S1). In typical ML approaches, the training set size is larger than 50 %, commonly 70 % or 80 %, of the whole data set (Rácz et al., 2021). From this basis, 70:30 ratio between training and test set was selected also for this study to represent a situation in which the user has a large set of active compounds available. The second ratio, 10:90, represents a more realistic case when the user has only a limited number of known active molecules available (*e.g.*, NEU with nine active molecules in the training set). The enrichment metrics were calculated using the default 50/50 weight distribution for shape and ESP score as well as using only the shape score (Study IV; Table 1).

Although it was difficult to determine the best scoring method (50/50 of shape and ESP or only the shape score) as it depended on the target, it was clear that the BR-NiB optimization improved both AUC and EF metrics considerably in comparison to the original docking or R-NiB (Study IV; Tables 1 and S2, Figures S2-S9). Only the optimization of the PPARγ model with smaller training sets did not outperform the original docking but still generated relatively satisfying enrichments. On average, the AUC values improved from 0.74 to 0.83 with 100:100 and 70:30 sets. Even with 10:90 sets, AUC improved from 0.74 to 0.81. BR-NiB worked particularly well, for instance, in the case of NEU and RXRα, and at its best, the EF 1% improvement was even 20-fold. However, the EF 1% improvement varied from 1.3 to 25.3-fold depending on the success of docking. With smaller 70:30 and 10:90 ratios, the results were typically less dramatic, but, importantly, the BR-NiB optimization worked well also with these small training sets. Optimization of COX2, PDE5 and PPARγ worked better when using only the shape score than the 50/50 weight distribution in the enrichment calculations.

Similar to the Study III, the NIB model for PDE5 set was generated using two models, sildenafil- and tadalafil-based, and this merged model was shown to work better than either of the models alone (Study IV; Table S3). After the optimization, at its best, the combined model produced the AUC value of 0.87, EF 1% of 27.6 and BEDROC of 0.46 when using 100:100 set. For example, the sildenafil-based model alone produced only AUC of 0.72, EF 1 % of 15.1 and BEDROC of 0.29.

A similar RMSD comparison was performed also for the molecules selected by BR-NiB as described in Section 5.2.5 for R-NiB. When analysing the molecules with experimentally determined binding poses using 1-3 Å RMSD range, it was shown that BR-NiB was not remarkably better or worse than R-NiB or docking scoring selecting the binding poses closely related to the experimentally determined ones (Study IV; Table S13).

### 5.3.2 The optimization process with different docking algorithms (IV)

The BR-NiB method was tested with three alternative docking programs (DOCK, GOLD and Glide in SP mode) using four targets (COX2, RXRα, NEU and MR). The BR-NiB method was shown to work well also with different docking software. In some cases, such as NEU with Glide or COX2 with GOLD, it worked even better than with PLANTS docking (Study IV; Table S9). In the case of Glide, BR-NiB improved the enrichment metrics less than with the other software, as the results were already relatively high, and the AUC was not improved with COX2 and RXRα. However, the best enrichment metrics were often obtained with Glide and/or by optimizing the NIB model with Glide docking solutions (*e.g.*, COX2 and NEU sets). Nevertheless, it should be noted that the tendency of Glide to skip molecules during the docking process made it too biased to even calculate the enrichment metrics for the MR set.

The optimized models generated using different docking sampling methods remarkably resemble each other, but they still have some different cavity atoms (Study IV, Figure 5). To study how well an optimized NIB model based on the docking solutions of a certain docking program works in rescoring of poses from other docking software, the BR-NiB models were cross-used (Study IV; Table S11). The optimized models were shown to work relatively well in the cross-usage and occasionally produced even better enrichments than the original BR-NiB approach with PLANTS. This was the case when rescoring the NEU docking solutions generated with Glide using any of the NIB models optimized for PLANTS, DOCK or GOLD, for example.

### 5.3.3 Which cavity atoms are removed? (IV)

During the optimization process, the NIB model shrinks remarkably. On average, 40 % of the cavity atoms were removed during the process. However, in the case of large models (PDE5 and PPARγ with 129 and 144 cavity atoms, respectively) over 50 % of the cavity atoms can be removed (Study IV; Table S7). The percentage of polar atoms increased 2-5 % during the optimization regardless of the scoring method used (50/50 weight distribution of shape and ESP score or only the shape score). This is not surprising as the polar cavity atoms are typically relevant showing the locations of hydrogen bonds or ionic interactions, while the nonpolar cavity atoms fill the pocket and define the model dimensions, which is a much more elusive concept to determine.

It is easy to understand that the nonpolar cavity atoms not overlapping with the docked active molecules can be considered irrelevant in separating the actives from the inactives. This is typically the case with the cavity atoms located on the NIB model surface and these atoms are removed first (Study IV; Figure 4). Particularly, during the last generations, cavity atoms are removed from inside the NIB model generating even holes and shafts. Although these removals might be harder to understand intuitively, even a tiny improvement in the BEDROC value is enough to remove a cavity atom during the optimization. If there is a cavity atom that practically overlaps with every molecule in the training set, its effect on the overall enrichment is minor. Thus, it can be either removed or retained depending on the changes in the fifth decimal of the BEDROC value.

### 5.3.4 The model enhancement takes time (IV)

Although the original R-NiB rescoring is very fast, the BR-NiB approach is more time-consuming despite the fact it only follows a certain path by selecting the most advantageous choice at each stage. In the case of a NIB model composed of 50 cavity atoms, the greedy search needs to perform the rescoring calculation for 50 different NIB models until it can proceed to the next generation (Gen #1) with the model that has the best enrichment (Figure 2A). In the next generation (Gen #2), the rescoring needs to be done for 49 different NIB models, and so on. Thus, the duration of the optimization depends not only on the size of the training set but also heavily on the size of the NIB model and, obviously, the number of generations. In a regular BR-NiB run, the rescoring calculation needs to be done several hundred times.

The computational demands of BR-NiB were tested using two small benchmarking sets, MR (n = 39,090, includes alternative docking poses) and NEU (n = 48,860), and one large set COX2 (n = 176,830) with 70/30 training and test set ratio and 15 CPUs. The first generation was shown to take 8 min for MR, which had 57 atoms in the NIB model. In the case of NEU with 79 cavity atoms, the first generation took 12 min. COX2 had the smallest NIB model with 44 cavity atoms, but a large number of molecules increased the calculation time of the first generation to 21 minutes. In total, COX2 took about 5 h 50 min (13 generations), MR 2 h (15 generations) and NEU 4 h 40 min. A much more time-consuming case was PPARγ, with 144 cavity atoms and about 180,000 docking poses. However, using parallel computing with 40 CPUs instead of 20, the calculation time of the first generation was shortened from 1.5 h to 1 h.

The simplest ways to speed up the optimization are to ensure that the input NIB model is not any larger than necessary and limiting the number of outputted docking poses as low as possible. The more poses are generated, the higher is the probability reproducing the biologically relevant binding orientation. On the other hand, this easily leads to the generation of energetically unfavourable docking poses that only distract the NIB screening or rescoring, which do not consider the internal energies of the molecule poses at any level. However, it is also possible to stop the optimization after a certain number of generations to save time and potentially avoid overfitting of the model with the training set data. Accordingly, with several benchmarking sets, it was shown that the highest boost to the enrichment metrics was achieved during the first generations of the BR-NiB optimization (Study IV, Figures S2 and S3). In most cases, BR-NiB produced good enrichment metrics already at the halfway of the process. However, in some cases, such as with EF 1% in PDE5, the improvement increased relatively constantly over the generations. If the optimization was stopped after 35 generations, the EF 1 % would be < 15. However, it was even 27.6 after the last generation (#72). In the case of ER, the improvement in AUC was achieved as late as during the generations 20-35 (AUC improvement from 0.67 to 0.82, 41 generations in total).

It was also tested if several cavity atoms could be removed already during the first generation based on their effect on the enrichment metrics instead of removing them one by one (Study IV, Table S15 and Figure S10). The results show that it could be possible to recognize the important and detrimental cavity atoms in the beginning of the optimization process. This could greatly speed up the BR-NiB run without affecting significantly the enrichment metrics.

### 5.3.5 Are the optimized models suitable for practical usage? (IV)

In the DUD-E set, the ratio of active and decoy molecules is relatively high, at least 1.5 % (Study IV, Table S1). However, in real HTS studies, the actual hit rates are much lower (Zhu et al., 2013). To mimic the conditions of actual drug discovery projects, in Study IV, validation sets with the active/decoy ratio of 0.014 were generated for MR and NEU using a real commercial HTVS database as a decoy set. 20 active molecules, which were verified to not have been included in the original DUD-E sets, were picked randomly from the ChEMBL database based on their affinity ($IC_{50}$) to generate validation sets with different activity ranges ($IC_{50}$ < 1 μM, < 50 μM and 1–50 μM).

It was clearly shown that BR-NiB outperformed the original docking with these validation sets (Study IV, Table 2). More importantly, when examining the very early enrichments EF 0.1 and 0.5 %, corresponding the top 140 and 700 compounds, respectively, the results show that some active molecules were ranked above these thresholds. In the case of MR, the original docking scoring was not able to rank hardly any active molecules among the best 140 or 700 compounds. BR-NiB instead ranked 1–3 active molecules among these groups depending on the activity thresholds given to the validation sets. The NEU sets were easier for the PLANTS docking scoring: at its best, it was able to rank one molecule among the top 140 compounds and three molecules at the top 700. However, BR-NiB was much more efficient ranking even 7 active molecules at the top 140 compounds and 11 at the top 700. In general, validation sets with high-level potency ligands performed better than the low-level ones at least with BR-NiB, but this trend, although logical, was not entirely consistent.

Overall, the results show that the performance of the BR-NiB method has potential to work satisfactorily in actual drug discovery projects in which at least a couple hundred of molecules are typically selected for *in vitro* screening depending on the available resources.

# 6 Discussion

## 6.1 Defining the cavity dimensions

R-NiB was shown to be a very fast and efficient docking rescoring method, and its performance is not based solely on refocusing of docking or simple ligand-based similarity. However, it was also shown that careful generation of the NIB model is critical, and the results can be improved notably if paying attention to the model composition rather than just using the "default" settings. The best results were obtained when limiting the NIB model dimensions based on the bound ligand present in the target structure. This indicates that certain parts of the cavity are more important than others in ligand binding. When generating a NIB model and limiting its volume using the co-crystallized ligand, the most critical parts of the cavity are covered, as the binding pocket is defined more tightly than just applying a simple cavity center. Using a cavity center radius in the model generation makes the NIB model bulkier and enables it to cover subcavities more or less irrelevant for the ligand binding. However, excluding certain parts of the cavity from the NIB model should always be performed carefully: the benchmarking sets only contain known active molecules, which typically occupy similar parts of the cavity. The results are likely worsened if the model is increased including a subcavity that none of the active molecules in the benchmarking set occupies, but it does not mean that a novel drug could not bind there.

Selecting a suitable lattice (FCC or BCC) is a demanding part as the results in Studies I-III showed this parameter to be case-specific. Thus, it might be necessary to test both lattices in the model generation. In addition, the polar cavity atoms in the model need to be inspected carefully as they might have a big impact on the results. For example, if the generated NIB model lacks an important polar cavity atom that should represent a hydrogen bonding partner in a central cavity position, the settings should be optimized so that this atom is included in the model. On the contrary, some useless polar atoms should be deleted even manually.

The performance of the R-NiB or BR-NiB methods in docking rescoring is mostly based on the shape match of the ligand binding cavity, and the electrostatics have a smaller, yet important, role in the similarity comparison that should not be underestimated. For instance, even the electrostatic complementarity alone has been used to predict the binding affinity of small molecules in the binding site (Bauer and Mackey, 2019). In fact, shape complementarity has been considered carefully particularly in some newer docking software. For example, QSDock focuses only on the shape complementarity, and DOCK determines, although in a relatively coarse way, the shape of the cavity prior to the docking progress (Allen et al., 2015; Goldman and Wipke, 2000; Kuntz et al., 1982). A commercial docking program LigandFit considers both the shape and electrostatic complementarity of the binding pocket during the docking reminding the NIB principle, at least to some extent (Venkatachalam et al., 2003). However, this software has not been shown to be superior to other docking approaches in the comparison studies (Wang et al., 2016). The idea behind the R-NiB approach is different: as docking software is already successful at generating the relevant binding conformations and often recognizes well the electrostatics, such as possible hydrogen bonding, the docking poses need only to be scored better by highlighting the importance of shape similarity. ShaEP unites shape and electrostatics straightforwardly and picks the best parts from both approaches.

However, determining the cavity dimensions can be a difficult task. The presence of water and possible side chain rotations during the ligand binding affect the cavity shape and, thus, the final composition of the NIB model. These differences can be notable even when handling relatively similar ligands (Boström et al., 2006; Wang et al., 2008). In these cases, it is difficult to generate a universal NIB model. Furthermore, for NEU and other surface pockets, defining the binding cavity limits is a much more arbitrary operation than for the buried cavities, such as MR. In these cases, the NIB performance comes also from its ability to refocus the docking solutions: the NIB model highlights the site where the docking solutions should be located. This is valuable information providing that the NIB model dimensions are determined correctly.

## 6.2 Some targets are difficult for negative image-based rescoring

Although R-NiB was shown to be an efficient method, it is clear that some targets are more demanding than others, and alternative methods can work better than R-NiB with certain targets. For R-NiB, particularly ER, PDE5 and PPARγ sets were problematic. The original docking scoring performed already well with these targets achieving convincing AUC and early enrichment results. In the case of ER, at its best, R-NiB only slightly improved the early enrichment while AUC decreased, which suggests that only a certain ligand subgroup, most likely the antagonists, can be found when using the generated NIB model. This was not a surprising result as the ER set contains both agonists and antagonists, the former of which lacks the long tail typical for antagonists, and the NIB model was generated by limiting its dimensions based on the antagonist binding pose (Figure 4A). Separation of the agonist and antagonists from the benchmarking set showed, also unsurprisingly, that the NIB model worked excellent for the antagonists set than for the mixed set (agonists and antagonists together). Thus, it is clear that improved results would be acquired by searching for agonists and antagonists separately, which requires the generation of a distinct NIB model focusing only on the agonist space occupancy.

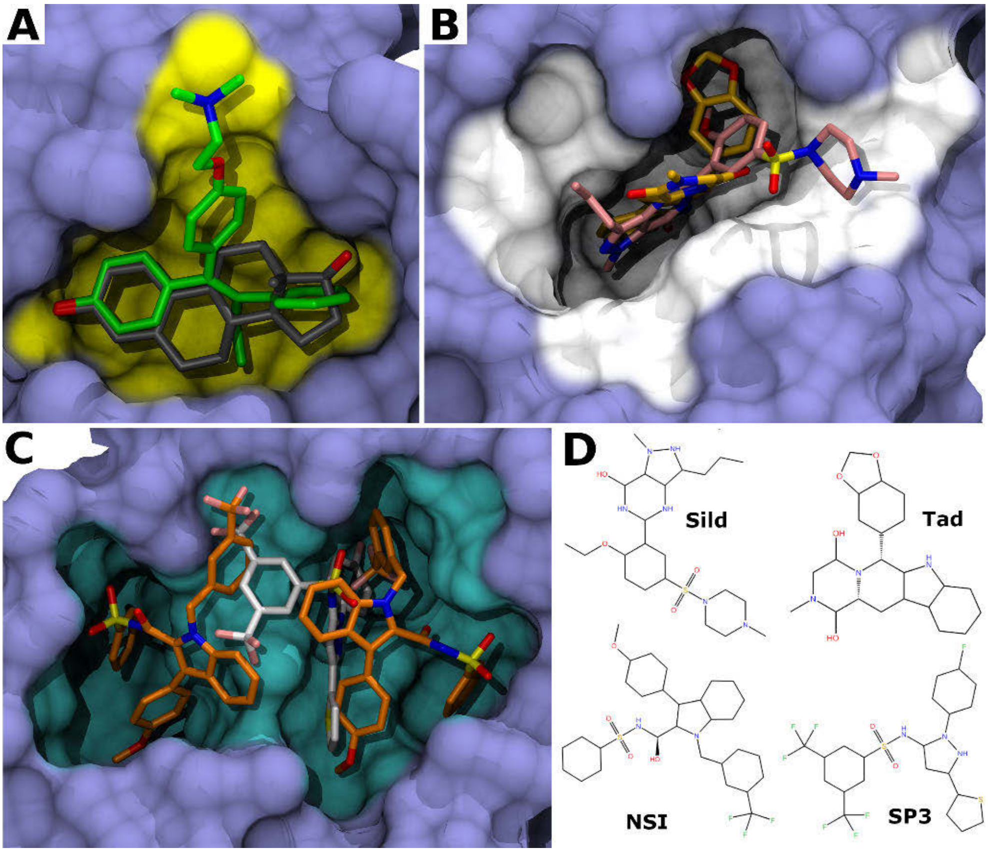

**Figure 4.** The cavity dimensions of ER, PDE5 and PPARγ with certain ligands. Figure A represents a binding cavity of ER, highlighted in yellow, with agonist (estrogen, gray) and antagonist (hydroxytamoxifen, green). The NIB model of ER covers approximately the same area. Although agonists and antagonists otherwise cover the same ligand binding site, the tail part of antagonists inhabits a distinct subcavity (PDB codes 1A52 and 3ERT). Figure B highlights the large binding cavity of PDE5 (white). Although creating the NIB model using two different ligands, sildenafil (Sild, pink) and tadalafil (Tad, yellow), even together they only occupy a small part of the cavity. Furthermore, the space occupancy of these two molecules is different (PDB codes 1UDT and 1XOZ). Figure C shows the binding cavity of PPARγ. Two molecules of a bigger ligand (NSI, orange, PDB code 2HFP) is shown to bind in the binding cavity. Another ligand (SP3, white, PDB code 2G0H) shows very different binding orientation occupying the middle part of the cavity. For clarification, 2D structures of PDE5 and PPARγ ligands are shown in Figure D.

PDE5 set enrichment metrics were not improved by the R-NiB approach. The PDE5 binding cavity is relatively large, and the bound ligands, sildenafil and tadalafil, used for the NIB model generation each occupy only a small sub-cavity (Figure 4B; Study III, Figure S12). In addition, these two molecules are structurally very different, and their space occupancy differ considerably. When two separate NIB models, one limited by the dimensions of sildenafil and the other by tadalafil, were combined, the enrichment metrics were improved only slightly in comparison to using either of the models alone. Still, R-NiB outperformed the original docking only occasionally. Although sildenafil- and tadalafil-like molecules might be discovered in the actual rescoring, if docked properly, the hybrid-NIB model still occupies a relatively small part of the cavity. Alternatively, generating a NIB model that fills the entire cavity (*box radius* option) was shown to perform even worse. This is likely because the actual cavity is relatively spherical (in Figure 4B towards the viewer) allowing also the NIB model to be globular. Thus, it is difficult to consider the shape parameter in the "shapeless" pocket of PDE5. Additionally, there are water molecules and even ions in the binding pocket that affect the shape of the cavity during the binding of a certain ligand (Wang et al., 2008).

With PPARγ, the situation is quite similar, and R-NiB did not improve any of the enrichment metrics. Molecules that bind to PPARγ are structurally very diverse (Sauer, 2015). Moreover, the size of the ligands is relatively large: the active ligands of PPARγ in DUD-E have an average rotatable bond number as high as 9.5 and the average molecular weight of 464 g/mol (for comparison, 5.3 and 408 g/mol for RXRα actives). In fact, it is impressive that the docking scoring performed as well as it did with the molecules of this size. Additionally, The PPARγ cavity is really spacious. Perplexingly, there is an X-ray crystal structure in which the same small molecule also included in the DUD-E set, has acquired two utterly different binding poses simultaneously within the cavity (Figure 4C, orange) (Hopkins et al., 2006). To make the matter even more complex, the verified binding pose of yet another DUD-E compound is highly different from either of the alternative poses (Figure 4C, white). Considering the sizes of the ligands and their cavity occupancies, it is not surprising that the NIB model generated with the *box radius* option performed the best, as it is impossible to cover all ligand binding orientations using a single NIB model limited by the dimensions of a single bound ligand (*ligand distance limit* option). However, the rescoring performance was not very satisfying: as in the case of the PDE5 set, in general, a large NIB model generated with the *box radius* option has a limited specificity to separate active molecules, which only occupy certain parts of the cavity volume, from the inactive ones. This should be considered during the NIB model generation, and it could be advantageous to generate several models each focusing on a certain molecule subgroup.

Similarly, resolution of R-NiB and BR-NiB for recognizing small differences in the ligand structures might not be high enough if using a single model. This could be the case when aiming to generate a model that should recognize ligands selective only for certain receptor subtype. If the differences between the binding cavities and selective ligands are minor, such as in the case of estrogen receptor alpha and beta (Manas et al., 2004; Shanle and Xu, 2010), it is likely that a single NIB model is unable to separate molecules that bind only one of these targets.

## 6.3 Model optimization causes overfitting?

When using different settings in the model generation, comparison of the R-NiB results already suggests that slimmer models typically produce higher enrichment than the bulkier ones. From this basis, the success of removing the excess cavity atoms, albeit with a greedy algorithm, was presumable. However, it was surprising that the enrichment improvement was so consistent: when the training set was comprehensive enough, the BR-NiB approach always produced better results than docking, or at least very good enrichment metrics (AUC > 0.80 and EF 1% > 25). Interestingly, also the targets difficult for R-NiB, such as PDE5 and ER, were shown to work remarkably better with the BR-NiB approach. BR-NiB succeeded also well when using a small training set for the model optimization (*e.g.*, only nine active molecules for MR and NEU). Although some important interactions might be omitted when using a small number of actives, it is typically a realistic situation in actual drug discovery projects. Thus, the BR-NiB optimization could truly be a useful tool for drug development.

Most importantly, the BR-NiB was shown to give great enrichment for validation sets with the active molecule ratio of only 0.014 % that corresponds to a typical situation in an actual VS project. Nevertheless, the results do not suggest that highly different or novel compounds can be found using the BR-NiB-optimized models in rescoring: as novel ligands are often just variations of the known drugs, the active molecules in the validation sets are likely similar to the ones in the training and test sets (Chen et al., 2019; Eckert and Bajorath, 2007; Müller, 2003). Thus, the results might vary depending on the composition of the active molecules in the validation sets (only 20 compounds).

Because the model optimization was selected to be based on BEDROC, it is possible that the model could focus only a certain subgroup of molecules explaining why AUC may even lower while the early enrichment increases (*e.g.* ER and PDE5 sets). This could happen particularly if the active molecule subgroup is large enough: the early enrichment could be remarkably improved over several generations although some other molecule subgroup is separated worse from the decoy molecules. Similarly, if the active molecule group in a training set is small, it is possible that some chemotypes included in the test set are completely missing. In these cases, it is possible that the NIB models are overfitted with respect to the certain training set. This might be the case, for example, with PDE5 that produced a lot better enrichment results with the training set than with the test set. However, it should be noted that the BR-NiB method itself did not perform any worse in recognizing different active molecule chemotypes in comparison to docking or other rescoring methods.

If there is not clearly a major molecule subgroup, it is also possible that the model gets scattered when trying to overlap with all active molecules in the training set. This is done by removing the cavity atoms that enable the decoy molecules to be ranked high in the scoring list. In fact, the composition of the decoy set is a surprisingly relevant issue for the optimization (Réau et al., 2018). As the BR-NiB optimization only tries to separate active molecule poses from the decoy ones, the size and content of the decoy set can remarkably affect the final NIB model composition, and typically, the quality is at least as important as the quantity. For example, it is easy to get great enrichment results if using a small number of decoys, but it is also easy to recognize the active molecules when using a large set of decoys structurally very distinct to the active ones. In other words, the decoy set should be complex and comprehensive enough to enable BR-NiB to train the model sufficiently diversely.

However, in the BR-NiB approach, the overfitting problem (Hawkins, 2004) might be quite persistent. Typically, the further the BR-NiB run is continued, the smaller the improvement effect is when removing a cavity atom and the more difficult it is to rationalize the reason for the removal. In fact, one could argue that the model automatically gets overfitted if the BR-NiB optimization is allowed run to the very end. Dividing the DUD-E sets into training and test sets does not necessarily solve the problem, as the benchmarking sets are still relatively homogenic (Chen et al., 2019; Lagarde et al., 2015). The BR-NiB-optimized model only finds the molecules that are the most similar to the active ones in the training set. In many cases, the successful separation of actives and decoys is actually a relatively easy task, as the ligands of a certain target are often structurally relatively similar, at least in the literature, and there are not many really different molecule chemotypes that would bind to the same location (Müller, 2003). By getting a high similarity score for one active molecule, the model likely gives a good score for many other active molecules as well, and the targets with a structurally more diverse set of active molecules are intelligibly more demanding cases for the NIB methodologies. However, other methods, such as QSAR, struggle with these similar problems as well (Gramatica, 2013; Yang, 2010; Zhao et al., 2017). In actual VS usage, stopping the BR-NiB run somewhere at mid-point could be beneficial to limit the overfitting, but determination of the best end point is difficult. It could be also used models from several generations to calculate the total score for the molecule poses.

The best results were typically obtained when combining the shape and ESP score instead of using them separately in R-NiB. The only exceptions were the very nonpolar cavities, such as MR, in which the shape score alone produced the best results. However, with BR-NiB, the NIB model optimization based only on the shape score was relatively often the most successful approach. In some cases, it is possible that the shape score just recognizes better a certain subgroup of molecules than the combination score. However, it is difficult to evaluate why BR-NiB produced constantly better results for PDE5 and PPARγ when optimizing the cavity using only the shape score. Although neither of the binding cavities is exceptionally nonpolar, both are very large. It is possible that in spacious cavities, the shape is a more universal and, thus, more useful feature than electrostatics in separating the active molecules from the inactive ones, as the polar interactions are already well optimized in the docking sampling. It is also probable that polar cavity atoms are too specific for certain types of ligands. As none of the ligands contains a matching part for every polar atom in the NIB model, the ESP gives too conflicting results.

## 6.4 The practical usability of the methods

Although the results are very promising, the further confirmation of the usefulness of the BR-NiB or R-NiB method would require large *in vitro* testing beyond the scope of this thesis. In fact, testing only a handful of the top-ranked molecules for each target and finding one or two hits can always be considered just good luck, as well as not finding anything can be diagnosed as bad luck. Moreover, the performance of the R-NiB and BR-NiB methods, as well as any other rescoring technique, is always dependent on the performance of the initial docking sampling. If the docking fails in reproducing at least somewhat relevant binding poses, the rescoring process has little chances.

However, the requirements for performing the NIB methods do not otherwise differ from other structure-based approaches. Both R-NiB and BR-NiB require at least a target protein structure. To perform BR-NiB, at least some active (and decoy) ligand data is also required, and even R-NiB would benefit from having an active ligand pose available to limit the NIB model dimensions. Thus, at least the BR-NiB method is not suitable for targets without any known compounds. The main weakness of these NIB methodologies is that they are typically based on only one target protein structure. Although it is possible to use several target structures for docking, or create NIB models based on different binding cavity orientations or cavities bound by diverse ligands, the successful usage of these approaches in the actual screening remains unclear. However, side chain rotations, more or less conserved water molecules and very spacious cavities are a challenge also for other structure-based methodologies (Orgován et al., 2019; Teague, 2003). When correctly applied, the NIB model optimization could be a solution for at least some of these problems, but it always needs a reliable benchmarking set.

The quality of the benchmarking sets is a relevant issue, as they always include some erroneous data (Lagarde et al., 2015; Réau et al., 2018). Active molecules in the COX2 set are mostly very similar to celecoxib (Study III; Figure S11), which is the reason why ligand-based similarity comparison worked well with this particular set. Although several features are said to be unbiased in the DUD-E sets, such as molecular weight, net charge and rotatable bond number, the benchmarking sets also include, for example, analogue biases that cause problems for ML approaches (Chen et al., 2019; Mysinger et al., 2012; Sieg et al., 2019). However, similar problems occur also with other benchmarking sets, and a completely bias-free molecule set is difficult to generate (Wallach and Heifets, 2018). It should be also noted that in DUD and DUD-E, as with several other benchmarking sets, the decoys are generated computationally or randomly picked from a large compound library, and there is no guarantee that all of these picks are actually inactive and could not be false decoys (Niinivehmas et al., 2016; Réau et al., 2018)

Effective optimization of the NIB model with exhaustive search would be an optimal problem only for quantum computers (Grover, 1998). Although based on a greedy search method, in practice, the most limiting part in the usage of BR-NiB is still the time consumption. However, thanks to the possibilities of parallel computing, the BR-NiB can be performed within a day or two even for the demanding cases. Nevertheless, large input models and benchmarking sets take time to process. In the case of any target, several NIB models need typically to be generated with different settings and optimized with the BR-NiB method to select the best one. Thus, speeding up the optimization process would be highly beneficial. Although it seemed to be favourable to remove (or retain) several cavity atoms clearly worsening the results already after the first generation, defining the universal thresholds for the automatic search might turn out to be difficult. However, it should be noted that the proof of concept of the BR-NiB method, rather than optimizing the time consumption, was the main focus in Study IV.

# 7 Summary and conclusions

Because the scoring functions of the docking programs have very case-selective ability to separate the active molecules from the inactive ones, there is a need for more efficient methods. In this thesis, two novel docking rescoring methods, which rely on the negative image of the protein binding cavity, are introduced. Studies I-III show that the first method, R-NiB, is fast and effective to separate active molecules from the inactive ones. Based on the *in silico* benchmarking results, R-NiB works well with alternative docking software and various protein targets. Because the calculation times are rapid, the R-NiB method suites well for HTVS. Instructions for performing the NIB techniques and model generations are described in Study II. These studies show that by carefully generating the applied NIB models, R-NiB can greatly improve the results of VS studies. However, with targets containing diverse ligand sets, it is advisable to generate several NIB models selectively, for example, focusing on agonists and antagonists binding volumes.

When reliable benchmarking sets are available, it is possible to optimize the NIB model with the second method, termed BR-NiB, before using it for the actual virtual screening. It was shown in Study IV that BR-NiB improved the scoring performance to the next level, outperformed the original docking scoring function with all tested targets and worked also with alternative docking programs. In particular, the early enrichment, a key factor when working with libraries containing even millions of molecules, was remarkably improved. Additionally, it was shown *in silico* that the hit rates of the BR-NiB method could be satisfying for real HTVS experiments. Although the optimization process of a single NIB model can be time-consuming, the investment is clearly profitable. In the next step, the usability of the R-NiB and BR-NiB methods should be confirmed in practical use with extensive *in vitro* studies.

# Acknowledgements

This work was started at the Nanoscience Center at the University of Jyväskylä and carried out at the Integrative Physiology and Pharmacology research unit, Institute of Biomedicine at the University of Turku. The thesis work got funding from the Finnish Cultural Foundation and the Turku University Foundation. The computational resources were provided by CSC – IT Center for Science.

I would really like to thank my supervisor Professor Olli Pentikäinen for all the guidance, opinions and ideas during these years, and for showing me both the academic and commercial fields. It was indeed a favourable decision when I went skiing to Laajavuori on that winter day in 2017, at least for me. I would also like to thank my second supervisor, Docent Pekka Postila, with whom I performed most of the practical work. We had great and not so great moments in the amazing world of negative images.

I acknowledge my follow-up group member Outi Salo-Ahen for advice during the PhD process. Thank you Docent Henri Xhaard for accepting the invitation to be my opponent. I am also thankful to Professor Mark Johnson and Daniela Schuster for the preliminary examination of this thesis.

I am grateful to all my collaborators, co-authors and co-workers. Dr. Jukka Lehtonen, without your help in scripting Pekka and I would probably still be removing cavity points manually. I also thank my workmates and friends Sakari Lätti, Sanna Niinivehmas and Elmeri Jokinen for the fluent collaboration in different projects during these years. We have had fun moments at both work and free time. Thank you Mira Ahinko for the efforts in our shared publication and Kseniia Petrova-Szczasiuk for your work in our mutual projects.

I thank the Drug Research Doctoral Program, its other PhD students and the entire Farmis group. Especially, I thank Outi Irjala for ensuring that the manuscript of this thesis made it to the preliminary examination before the summer break.

Finally, I would like to thank my wife and family for all the support during my 23 years at school. I guess the dumbest ones of us just need some extra years.

Kuopio, August 2021
*Sami Kurkinen*

# References

Adams, C.P., and Van Brantner, V. (2006). Market watch : Estimating the cost of new drug development: Is it really $802 million? Health Aff. *25*, 420–428.

Ain, Q.U., Aleksandrova, A., Roessler, F.D., and Ballester, P.J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip. Rev. Comput. Mol. Sci. *5*, 405–424.

Aitken, M., Kleinrock, M., Simorellis, A., and Nass, D. (2019). The Global Use of Medicine in 2019 and Outlook to 2023. IQVIA Institute. p. 49. <https://www.iqvia.com/insights/the-iqvia-institute/reports/the-global-use-of-medicine-in-2019-and-outlook-to-2023>

Allen, W.J., Balius, T.E., Mukherjee, S., Brozell, S.R., Moustakas, D.T., Lang, P.T., Case, D.A., Kuntz, I.D., and Rizzo, R.C. (2015). DOCK 6: Impact of new features and current docking performance. J. Comput. Chem. *36*, 1132–1156.

Ananthula, R.S., Ravikumar, M., Pramod, A.B., Madala, K.K., and Mahmood, S.K. (2008). Strategies for generating less toxic P-selectin inhibitors: Pharmacophore modeling, virtual screening and counter pharmacophore screening to remove toxic hits. J. Mol. Graph. Model. *27*, 546–557.

Bajorath, J. (2002). Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. *1*, 882–894.

Ballester, P.J., and Richards, W.G. (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. J. Comput. Chem. *28*, 1711–1723.

Bauer, M.R., and Mackey, M.D. (2019). Electrostatic Complementarity as a Fast and Effective Tool to Optimize Binding and Selectivity of Protein-Ligand Complexes. J. Med. Chem. *62*, 3036–3050.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. (2002). The protein data bank. Acta Crystallogr. Sect. D Biol. Crystallogr. *58*, 899–907.

Bleicher, K.H., Böhm, H.J., Müller, K., and Alanine, A.I. (2003). Hit and lead generation: Beyond high-throughput screening. Nat. Rev. Drug Discov. *2*, 369–378.

Boström, J., Hogner, A., and Schmitt, S. (2006). Do structurally similar ligands bind in a similar fashion? J. Med. Chem. *49*, 6716–6725.

Brink, T. Ten, and Exner, T.E. (2009). Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. J. Chem. Inf. Model. *49*, 1535–1546.

Brooijmans, N., and Kuntz, I.D. (2003). Molecular recognition and docking algorithms. Annu. Rev. Biophys. Biomol. Struct. *32*, 335–373.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., et al. (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. *47*, D464–D474.

Bursulaya, B.D., Totrov, M., Abagyan, R., and Brooks, C.L. (2003). Comparative study of several algorithms for flexible ligand docking. J. Comput. Aided. Mol. Des. *17*, 755–763.

Castaño, A., and Maurer, M.S. (2015). Protonation and pK changes in protein-ligand binding. Q. Rev. Biophys. *20*, 163–178.

Chaput, L., Martinez-Sanz, J., Quiniou, E., Rigolet, P., Saettel, N., and Mouawad, L. (2016). VSDC: A method to improve early recognition in virtual screening when limited experimental resources are available. J. Cheminform. *8*.

Charifson, P.S., Corkery, J.J., Murcko, M.A., and Walters, W.P. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J. Med. Chem. *42*, 5100–5109.

Chen, L., Cruz, A., Ramsey, S., Dickson, C.J., Duca, J.S., Hornak, V., Koes, D.R., and Kurtzman, T. (2019). Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS One *14*.

Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. J. Chem. Inf. Model. *49*, 1079–1093.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S.H. (2012). Structure-based virtual screening for drug discovery: A problem-centric review. AAPS J. *14*, 133–141.

Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M., and Hammond, G.L. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. J. Med. Chem. *51*, 2047–2056.

Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y., and Humblet, C. (2009). Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. J. Chem. Inf. Model. *49*, 1455–1474.

Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., and Laufer, J. (1992). Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci. *32*, 244–255.

Dariusz, P., Michal, L., Rafal, A., and Krzysztof, G. (2010). Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. J. Comput. Chem. *32*, 741–755.

Desaphy, J., Azdimousa, K., Kellenberger, E., and Rognan, D. (2012). Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. J. Chem. Inf. Model. *52*, 2287–2299.

DiMasi, J.A., Hansen, R.W., and Grabowski, H.G. (2003). The price of innovation: New estimates of drug development costs. J. Health Econ. *22*, 151–185.

DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. *47*, 20–33.

Eckert, H., and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov. Today *12*, 225–233.

Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G. V., and Mee, R.P. (1997). 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. *11*, 425–445.

Elokely, K.M., and Doerksen, R.J. (2013). Docking challenge: Protein sampling and molecular docking performance. J. Chem. Inf. Model. *53*, 1934–1945.

Ferrara, P., Gohlke, H., Price, D.J., Klebe, G., and Brooks, C.L. (2004). Assessing scoring functions for protein-ligand interactions. J. Med. Chem. *47*, 3032–3047.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. *8*, 967–974.

Friesner, R.A., Banks, J.L., Murphy, R.B., T, Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., et al. (2004). Glide: A new approach for rapid, accurate docking and scoring 1. Method and assessment of docking accuracy. J. Med. Chem. *47*, 1739–1749.

Fukunishi, Y., Kubota, S., Kanai, C., and Nakamura, H. (2006). A virtual active compound produced from the negative image of a ligand-binding pocket, and its application to in-silico drug screening. J. Comput. Aided. Mol. Des. *20*, 237–248.

Gally, J.M., Bourg, S., Do, Q.T., Aci-Sèche, S., and Bonnet, P. (2017). VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. Mol. Inform. *36*.

Ganesan, A., Coote, M.L., and Barakat, K. (2017). Molecular dynamics-driven drug discovery: leaping forward with confidence. Drug Discov. Today *22*, 249–269.

Gareth, J., Willett, P., and Glen, R.C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. *245*, 43–53.

Gareth, J., Willett, P., Glen, R.C., Leach, A.R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. *267*, 727–748.

Gasteiger, J., and Marsili, M. (1980). Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. Tetrahedron *36*, 3219–3228.

Geppert, H., Vogt, M., and Bajorath, J. (2010). Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J. Chem. Inf. Model. *50*, 205–216.

Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera-A visualization system for exploratory research and analysis. J. Comput. Chem. *25*.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. J. Mol. Biol. *295*, 337–356.

Goldman, B.B., and Wipke, W.T. (2000). QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). Proteins Struct. Funct. Genet. *38*, 79.

Good, A.C., and Oprea, T.I. (2008). Optimization of CAMD techniques 3. Virtual screening enrichment studies: A help or hindrance in tool selection? J. Comput. Aided. Mol. Des. *22*, 169–178.

Goodford, P.J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. J. Med. Chem. *28*, 849–857.

Gramatica, P. (2013). On the development and validation of QSAR models. Methods Mol. Biol. *930*, 499–526.

Grover, L.K. (1998). Quantum computers can search rapidly by using almost any transformation. Phys. Rev. Lett. *80*, 4329–4332.

Guan, B., Zhang, C., and Ning, J. (2017). Genetic algorithm with a crossover elitist preservation mechanism for protein–ligand docking. AMB Express *7*.

Guedes, I.A., Pereira, F.S.S., and Dardenne, L.E. (2018). Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges. Front. Pharmacol. *9*.

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. Chem. Biol. Drug Des. *69*, 146–148.

Halgren, T. (2009). Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. *49*, 377–389.

Halgren, T.A. (1996). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J. Comput. Chem. *17*, 520–552.

Halgren, T.A., and Damm, W. (2001). Polarizable force fields. Curr. Opin. Struct. Biol. *11*, 236–242.

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

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

Haranczyk, M., and Holliday, J. (2008). Comparison of similarity coefficients for clustering and compound selection. J. Chem. Inf. Model. *48*, 498–508.

Hart, T.N., and Read, R.J. (1992). A multiple-start Monte Carlo docking method. J. Mol. Graph. *10*, 58.

Hawkins, D.M. (2004). The Problem of Overfitting. ChemInform *35*.

Hawkins, P.C.D. (2017). Conformation Generation: The State of the Art. J. Chem. Inf. Model. *57*, 1747–1756.

Hawkins, P.C.D., Skillman, A.G., and Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. J. Med. Chem. *50*, 74–82.

Hertzberg, R.P., and Pope, A.J. (2000). High-throughput screening: New technology for the 21st century. Curr. Opin. Chem. Biol. *4*, 445–451.

Hopkins, C.R., O'Neil, S. V., Laufersweiler, M.C., Wang, Y., Pokross, M., Mekel, M., Evdokimov, A., Walter, R., Kontoyianni, M., Petrey, M.E., et al. (2006). Design and synthesis of novel N-sulfonyl-2-indole carboxamides as potent PPAR-γ binding agents with potential application to the treatment of osteoporosis. Bioorganic Med. Chem. Lett. *16*, 5659–5663.

Houston, D.R., and Walkinshaw, M.D. (2013). Consensus docking: Improving the reliability of docking in a virtual screening context. J. Chem. Inf. Model. *53*, 384–390.

Hu, J.Y., and Aizawa, T. (2003). Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. Water Res. *37*, 1213–1222.

Huang, N., Shoichet, B.K., and John J. Irwin (2006). Benchmarking sets for molecular docking. J. Med. Chem. Med *49*, 6789–6801.

Huang, S.Y., Grinter, S.Z., and Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions. Phys. Chem. Chem. Phys. *12*, 12899–12908.

Huey, R., M. Morris, G., Olson, A.J., and Goodsell, D.S. (2007). A Semiempirical Free Energy Force Field with Charge-Based Desolvation. J. Comput. Chem. *28*, 1145–1152.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

Irwin, J.J., and Shoichet, B.K. (2005). ZINC - A free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. *45*, 177–182.

Jain, A.N. (2003). Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. J. Med. Chem. *46*, 499–511.

Jain, A.N. (2009). Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. J. Comput. Aided. Mol. Des. *23*, 355–374.

Jorgensen, W., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc *110*, 1657–1666.

Kahraman, A., Morris, R.J., Laskowski, R.A., and Thornton, J.M. (2007). Shape Variation in Protein Binding Pockets and their Ligands. J. Mol. Biol. *368*, 283–301.

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., and Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes? J. Comput. Aided. Mol. Des. *22*, 213–228.

Kirchmair, J., Distinto, S., Markt, P., Schuster, D., Spitzer, G.M., Liedl, K.R., and Wolber, G. (2009). How to optimize shape-based virtual screening: Choosing the right query and including chemical information. J. Chem. Inf. Model. *49*, 678–692.

Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat. Rev. Drug Discov. *3*, 935–949.

Klebe, G. (2006). Virtual ligand screening: strategies, perspectives and limitations. Drug Discov. Today *11*, 580–594.

Kleywegt, G.J., and Alwyn Jones, T. (1994). Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr. Sect. D Biol. Crystallogr. *50*, 178–185.

Koes, D.R., Baumgartner, M.P., and Camacho, C.J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J. Chem. Inf. Model. *53*, 1893–1904.

Kooistra, A.J., Vass, M., McGuire, R., Leurs, R., de Esch, I.J.P., Vriend, G., Verhoeven, S., and de Graaf, C. (2018). 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery. ChemMedChem *13*, 614–626.

Korb, O., Stützle, T., and Exner, T.E. (2006). PLANTS: Application of ant colony optimization to structure-based drug design. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) *4150 LNCS*, 247–258.

Korb, O., Stützle, T., and Exner, T.E. (2009). Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. J. Chem. Inf. Model. *49*, 84–96.

Kramer, B., Rarey, M., and Lengauer, T. (1999). Evaluation of the FlexX incremental construction algorithm for protein- ligand docking. Proteins Struct. Funct. Genet. *37*, 228–241.

Kraulis, P.J. (1991). MOLSCRIPT. A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 947–950.

Kruhlak, N.L., Contrera, J.F., Benz, R.D., and Matthews, E.J. (2007). Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. Adv. Drug Deliv. Rev. *59*, 43–55.

Kukic, P., and Nielsen, J. (2010). Electrostatics in proteins and protein–ligand complexes. Future Med. Chem. *2*, 647–666.

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

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

Lang, P.T. (2018). Rigid and Flexible Ligand Docking. accessed 5.8.2018 <http://dock.compbio.ucsf.edu/Dock_6/tutorials/ligand_sampling_dock/ligand_sampling_dock.html>

Lätti, S., Niinivehmas, S., and Pentikäinen, O.T. (2016). Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. J. Cheminform. *8*.

Lehtonen, J. V., Still, D.J., Rantanen, V. V., Ekholm, J., Björklund, D., Iftikhar, Z., Huhtala, M., Repo, S., Jussila, A., Jaakkola, J., et al. (2004). BODIL: A molecular modeling environment for structure-function analysis and drug design. J. Comput. Aided. Mol. Des. *18*, 401–419.

Li, X., Li, Y., Cheng, T., Liu, Z., and Wang, R. (2010). Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. J. Comput. Chem. *31*, 2109–2125.

Lin, F.Y., and MacKerell, A.D. (2019). Force Fields for Small Molecules. Methods Mol. Biol. *2022*, 21–54.

Liu, J., and Wang, R. (2015). Classification of current scoring functions. J. Chem. Inf. Model. *55*, 475–482.

Liu, M., and Wang, S. (1999). MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. J. Comput. Aided. Mol. Des. *13*, 435–451.

Lo, Y.C., Rensi, S.E., Torng, W., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discov. Today *23*, 1538–1546.

Lyne Paul D. (2002). Structure-based virtual screening: an overview. Drug Discov. Today *7*, 1047–1055.

Macalino, S.J.Y., Gosu, V., Hong, S., and Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. Arch. Pharm. Res. *38*, 1686–1701.

Madhavi, S.G., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aided. Mol. Des. *27*, 221–234.

Maldonado, A.G., Doucet, J.P., Petitjean, M., and Fan, B.T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. Mol. Divers. *10*, 39–79.

Manas, E.S., Xu, Z.B., Unwalla, R.J., and Somers, W.S. (2004). Understanding the selectivity of genistein for human estrogen receptor-β using X-ray crystallography and computational methods. Structure *12*, 2197–2207.

Martin, Y.C. (2009). Let's not forget tautomers. J. Comput. Aided. Mol. Des. *23*, 693–704.

Martis, E.A., Radhakrishnan, R., and Badve, R.R. (2011). High-throughput screening: The hits and leads of drug discovery-An overview. J. Appl. Pharm. Sci. *1*, 2–10.

Masek, B.B., Merchant, A., and Matthew, J.B. (1993). Molecular Shape Comparison of Angiotensin II Receptor Antagonists. J. Med. Chem. *36*, 1230–1238.

Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. (2019). ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. *47*, D930–D940.

Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. Curr. Comput. Aided. Drug Des. *7*, 146–157.

Merritt, E.A., and Murphy, M.E.P. (1994). Raster3D version 2.0 A program for photorealistic molecular graphics. Acta Crystallogr. Sect. D Biol. Crystallogr. *50*, 869–873.

Mobley, D.L., and Dill, K.A. (2009). Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." Structure *17*, 489–498.

Mohan, V., Gibbs, A., Cummings, M., Jaeger, E., and DesJarlais, R. (2005). Docking: Successes and Challenges. Curr. Pharm. Des. *11*, 323–333.

Morgan, S.G., Bathula, H.S., and Moon, S. (2020). Pricing of pharmaceuticals is becoming a major challenge for health systems. BMJ *368*.

Morris, G., Huey, R., Linkstrom, W., Sanner, M., Belew, R., Goodsell, D., and Olson (2010). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J. Comput. Chem. *31*, 2967–2970.

Morris, G. M., and Lim-Wilby, M. (2008) Molecular Docking. In: Kukol A. (eds) Molecular Modeling of Proteins. Methods Molecular Biology™, vol 443. Humana Press.

Muegge, I. (2000). A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. Perspect. Drug Discov. Des. *20*, 99–114.

Müller, G. (2003). Medicinal chemistry of target family-directed masterkeys. Drug Discov. Today *8*, 681–691.

Mysinger, M.M., Carchia, M., Irwin, J.J., and Shoichet, B.K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. J. Med. Chem. *55*, 6582–6594.

Niinivehmas, S.P., Salokas, K., Lätti, S., Raunio, H., and Pentikäinen, O.T. (2015). Ultrafast protein structure-based virtual screening with Panther. J. Comput. Aided. Mol. Des. *29*, 989–1006.

Niinivehmas, S.P., Manivannan, E., Rauhamäki, S., Huuskonen, J., and Pentikäinen, O.T. (2016). Identification of estrogen receptor α ligands with virtual screening techniques. J. Mol. Graph. Model. *64*, 30–39.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An Open chemical toolbox. J. Cheminform. *3*.

Oda, A., Tsuchida, K., Takakura, T., Yamaotsu, N., and Hirono, S. (2006). Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. J. Chem. Inf. Model. *46*, 380–391.

Orgován, Z., Ferenczy, G.G., and Keserű, G.M. (2019). The role of water and protein flexibility in the structure-based virtual screening of allosteric GPCR modulators: an mGlu5 receptor case study. J. Comput. Aided. Mol. Des. *33*, 787–797.

Pagadala, N.S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. Biophys. Rev. *9*, 91–102.

Pearce, B.C., Langley, D.R., Kang, J., Huang, H., and Kulkarni, A. (2009). E-Novo: An automated workflow for efficient structure-based lead optimization. J. Chem. Inf. Model. *49*, 1797–1809.

Pissurlenkar, R.R.S., Shaikh, M.S., Iyer, R.P., and Coutinho, E.C. (2009). Molecular mechanics force fields and their applications in drug design. Antiinfect. Agents Med. Chem. *8*, 128–150.

Pospisil, P., Ballmer, P., Scapozza, L., and Folkers, G. (2003). Tautomerism in Computer-Aided Drug Design. J. Recept. Signal Transduct. *23*, 361–371.

Rácz, A., Bajusz, D., and Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. Molecules *26*.

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. *261*, 470–489.

Réau, M., Langenfeld, F., Zagury, J.F., Lagarde, N., and Montes, M. (2018). Decoys selection in benchmarking datasets: Overview and perspectives. Front. Pharmacol. *9*.

Ren, X., Shi, Y.S., Zhang, Y., Liu, B., Zhang, L.H., Peng, Y.B., and Zeng, R. (2018). Novel Consensus Docking Strategy to Improve Ligand Pose Prediction. J. Chem. Inf. Model. *58*, 1662–1668.

Ripphausen, P., Nisius, B., and Bajorath, J. (2011). State-of-the-art in ligand-based virtual screening. Drug Discov. Today *16*, 372–376.

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J.M., Lu, C., Dahlgren, M.K., Mondal, S., Chen, W., Wang, L., et al. (2019). OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J. Chem. Theory Comput. *15*, 1863–1874.

Rosenfeld, R. (1995). Flexible Docking and Design. Annu. Rev. Biophys. Biomol. Struct. *24*, 677–700.

Rush, T.S., Grant, J.A., Mosyak, L., and Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J. Med. Chem. *48*, 1489–1495.

Sauer, S. (2015). Ligands for the Nuclear Peroxisome Proliferator-Activated Receptor Gamma. Trends Pharmacol. Sci. *36*, 688–704.

Sauton, N., Lagorce, D., Villoutreix, B.O., and Miteva, M.A. (2008). MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. BMC Bioinformatics *9*.

Schichet, B.K. (2004). Virtual screening of chemical libraries. Nature *432*, 862–865.

Shanle, E.K., and Xu, W. (2010). Selectively targeting estrogen receptors for cancer treatment. Adv. Drug Deliv. Rev. *62*, 1265–1276.

Shoichet, B.K., McGovern, S.L., Wei, B., and Irwin, J.J. (2002). Lead discovery using molecular docking. Curr. Opin. Chem. Biol. *6*, 439–446.

Sieg, J., Flachsenberg, F., and Rarey, M. (2019). In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. J. Chem. Inf. Model. *59*, 947–961.

Sippl, W., Contreras, J.M., Parrot, I., Rival, Y.M., and Wermuth, C.G. (2001). Structure-based 3D QSAR and design of novel acetylcholinesterase. J. Comput. Aided. Mol. Des. *15*, 395–410.

Song, C.M., Lim, S.J., and Tong, J.C. (2009). Recent advances in computer-aided drug design. Brief. Bioinform. *10*, 579–591.

Sousa, S.F., Fernandes, P.A., and Ramos, M.J. (2006). Protein-ligand docking: Current status and future challenges. Proteins Struct. Funct. Genet. *65*, 15–26.

Steel, M. (2005). Phylogenetic diversity and the greedy algorithm. Syst. Biol. *54*, 527–529.

Swets, J.A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. Invest. Radiol. *14*, 109–121.

Teague, S.J. (2003). Implications of protein flexibility for drug discovery. Nat. Rev. Drug Discov. *2*, 527–541.

Toropova, A.P., Toropov, A.A., Benfenati, E., Leszczynska, D., and Leszczynski, J. (2010). QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL. J. Math. Chem. *48*, 959–987.

Tran-Nguyen, V.K., Da Silva, F., Bret, G., and Rognan, D. (2019). All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening. J. Chem. Inf. Model. *59*, 573–585.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. J. Comput. Chem. *31*, 455–461.

Truchon, J.F., and Bayly, C.I. (2007). Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. J. Chem. Inf. Model. *47*, 488–508.

Vainio, M.J., and Johnson, M.S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. J. Chem. Inf. Model. *47*, 2462–2474.

Vainio, M.J., Puranen, J.S., and Johnson, M.S. (2009). ShaEP: Molecular overlay based on shape and electrostatic potential. J. Chem. Inf. Model. *49*, 492–502.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

Venkatachalam, C.M., Jiang, X., Oldfield, T., and Waldman, M. (2003). LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model. *21*, 289–307.

Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., and Taylor, R.D. (2003). Improved protein-ligand docking using GOLD. Proteins Struct. Funct. Genet. *52*, 609–623.

Verma, J., Khedkar, V., and Coutinho, E. (2010). 3D-QSAR in Drug Design - A Review. Curr. Top. Med. Chem. *10*, 95–115.

Veselovsky, A. V., and Ivanov, A.S. (2003). Strategy of computer-aided drug design. Curr. Drug Targets - Infect. Disord. *3*, 33–40.

Virtanen, S.I., and Pentikäinen, O.T. (2010). Efficient virtual screening using multiple protein conformations described as negaVirtanen, S.I., and Pentikäinen, O.T. (2010). Efficient virtual screening using multiple protein conformations described as negative images of the ligand-binding site. J. Ch. J. Chem. Inf. Model. *50*, 1005–1011.

Wallach, I., and Heifets, A. (2018). Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. J. Chem. Inf. Model. *58*, 916–932.

Wang, H., Ye, M., Robinson, H., Francis, S.H., and Ke, H. (2008). Conformational variations of both phosphodiesterase-5 and inhibitors provide the structural basis for the physiological effects of vardenafil and sildenafil. Mol. Pharmacol. *73*, 104–110.

Wang, R., Lai, L., and Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comput. Aided. Mol. Des. *16*, 11–26.

Wang, R., Lu, Y., and Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. J. Med. Chem. *46*, 2287–2303.

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., and Hou, T. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: The prediction accuracy of sampling power and scoring power. Phys. Chem. Chem. Phys. *18*, 12964–12975.

Warren, G.L., Andrews, C.W., Capelli, A.M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., et al. (2006). A critical assessment of docking programs and scoring functions. J. Med. Chem. *49*, 5912–5931.

Watts, K.S., Dalal, P., Murphy, R.B., Sherman, W., Friesner, R.A., and Shelley, J.C. (2010). ConfGen: A conformational search method for efficient generation of bioactive conformers. J. Chem. Inf. Model. *50*, 534–546.

Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. *28*, 31–36.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. Drug Discov. Today *11*, 1046–1053.

Willett, P., Barnard, J.M., and Downs, G.M. (1998). Chemical similarity searching. J. Chem. Inf. Comput. Sci. *38*, 983–996.

Wójcikowski, M., Ballester, P.J., and Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. Sci. Rep. *7*.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. *285*, 1735–1747.

Yang, S.Y. (2010). Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. Drug Discov. Today *15*, 444–450.

Yoshida, F., and Topliss, J.G. (2000). QSAR model for drug human oral bioavailability. J. Med. Chem. *43*, 2575–2585.

Yu, H., Wang, Z., Zhang, L., Zhang, J., and Huang, Q. (2007). The discovery of novel vascular endothelial growth factor receptor tyrosine kinases inhibitors: Pharmacophore modeling, virtual screening and docking studies. Chem. Biol. Drug Des. *69*, 204–211.

Zhao, L., Wang, W., Sedykh, A., and Zhu, H. (2017). Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. ACS Omega *2*, 2805–2812.

Zhao, W., Hevener, K.E., White, S.W., Lee, R.E., and Boyett, J.M. (2009). A statistical framework to evaluate virtual screening. BMC Bioinformatics *10*.

Zhu, T., Cao, S., Su, P.C., Patel, R., Shah, D., Chokshi, H.B., Szukala, R., Johnson, M.E., and Hevener, K.E. (2013). Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis. J. Med. Chem. *56*, 6560–6572.

TURUN YLIOPISTO
UNIVERSITY OF TURKU

Painosalama, Turku, Finland 2021

ISBN 978-951-29-8580-7 (PRINT)
ISBN 978-951-29-8581-4 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)