universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Optimizing Pharmacophore-based Virtual Screening using Greedy 3-Point Search and Enhanced Parameterization in LigandScout“

verfasst von / submitted by

Sebastian Alexander Mann, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree

Magister pharmaciae (Mag.pharm.)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet	UA 066 605
Studienrichtung lt. Studienblatt:/ degree programme as it appears on the student record sheet	Masterstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thierry Langer

Danksagungen

Ich möchte meinem Betreuer Univ.-Prof. Mag. Dr. Thierry Langer danken, der es mir ermöglicht hat, diese Masterarbeit in seiner Arbeitsgruppe zu verfassen. Fragen meinerseits wurden immer sofort beantwortet und die allgemein sehr offene und hilfsbereite Einstellung in der ganzen Gruppe hat mir im Rahmen dieser Arbeit stets sehr geholfen.

Ein großer Dank gilt ebenso Christian Permann, MSc, der mich mit seinem umfassenden Wissen über den ganzen Zeitraum dieser wissenschafltichen Arbeit unterstützt hat. Mit den gemeinsamen Überlegungen in unzähligen Meetings war er maßgeblich an dem Ergebnis dieser Arbeit beteiligt.

Auch meinen Eltern, Andrea und Alexander, möchte ich von ganzem Herzen dafür danken, dass sie mich durch mein ganzes Leben so tatkräftig unterstützt haben und mir damit erst die Möglichkeit gegeben haben, mich beruflich so wie persönlich weiterzuentwicklen und meinen Zielen nachgehen zu können.

Zuletzt möchte ich vorallem meinen Freunden einen Riesendank aussprechen, die mich durch die Jahre des Pharmaziestudiums begleitet haben und mit einer unvergleichlichen Hilfsbereitschaft und Offenheit einen Abschluss erst möglich gemacht haben. Auf die zahlreichen Stunden in Hörsälen, beim gemeinsamen Lernen oder in Laborpraktika werde ich mich ein Leben lang gerne zurückerinnern.

Contents

List of Figures

List of Tables

Part I

Introduction and Related Work

Chapter 1

Introduction

Over recent decades, the field of computer-aided drug design has become indispensable in the realm of drug discovery. Among other key applications, virtual screening of vast molecule databases has a critical role in the initial stages of drug development. Such screening procedures allow for the efficient filtration of potential compounds of interest, significantly reducing the pool of candidates for subsequent in vitro experiments. This can not only substantially reduce time but consequently save valuable financial resources [1, 2, 3].

Virtual screening techniques have exhibited significant advancements throughout their evolution. Initially, two-dimensional substructure-based similarity searches faced a challenge in commonly managing to identify molecules with identical or closely related structural configurations. The introduction of novel methods, like pharmacophore-based virtual screening, surpassed these limitations, also enabling the exploration of compounds with different structural scaffolds [4, 5].

Methods for virtual screening encompass various approaches such as quantitative structure-activity relationship (QSAR), molecular docking or pharmacophore-based virtual screening. However, among these techniques, pharmacophore-based virtual screening has emerged as an especially valuable and impactful tool [2, 4, 6, 7, 8].

Pharmacophores describe abstract representations of electrostatic and steric relationships between biologically active compounds and a specific molecular target, within defined pharmacophoric features. These constructs find versatile use across a spectrum of applications. Apart from their crucial role in virtual screening, pharmacophores prove valuable for purposes such as toxicity, pharmacodynamic and pharmacokinetic predictions, pharmacophore-based de novo drug design, molecular dynamics simulations, and their integration into machine learning protocols [4, 5, 6, 8, 9, 10, 11, 12].

The diverse array of software applications designed for pharmacophore modeling and virtual screening primarily differentiate in terms of their specification for input data in pharmacophore construction and their methods for identifying the associated pharmacophoric features [4].

The alignment strategies employed in pharmacophore-based virtual screening are predominantly centered around minimizing the root mean square deviation (RMSD) between feature pairs or maximizing the volumetric overlap using Gaussian spheres. These approaches share the limitation of prioritizing the alignment with the lowest RMSD or the highest volume overlap, which may not necessarily lead to the most optimal alignment. The primary criteria for an ideal alignment should prioritize achieving the maximum count of matching feature pairs. In response to this problem, a novel screening alignment algorithm, Greedy 3-Point Search (G3PS), was introduced by Permann et al. and implemented in an updated version

of the LigandScout software, LigandScout 5 [13, 14].

The objective of this thesis includes evaluating the potential enhancements of pharmacophore-based virtual screening in LigandScout through the implementation of the G3PS algorithm and the identification of the optimal parameter configuration.

To accomplish this, investigations were initiated by conducting preliminary tests on a selected dataset and pharmacophore model, as displayed in Chapter 4. This initial phase was dedicated to identifying optimal settings for the introduced parameters in LigandScout 5.

Subsequently, the analysis was extended by applying the determined parameter sets to a more comprehensive testing across various datasets and pharmacophore models. The key aspect of this investigation was the comparative evaluation with the established LigandScout 4.5 and LigandScout 5 version, aimed at confirming that the applicability of the identified settings holds true in a more general context.

Chapter 2

Related Work

2.1 Pharmacophore

Throughout previous decades, pharmacophores have established themselves as profoundly significant and valuable tools for a variety of purposes in the realm of computer-aided drug design [14, 15].

This section provides an overview of the pharmacophore definition, methods for their creation, illuminates some of their many applications and presents a brief list of the most frequently used software packages.

2.1.1 Pharmacophore Definition

Paul Ehrlich, who introduced the terms "phoros" and "pharmakon" in the early 1900s to signify the role of chemical or functional groups in determining biological activity, is often credited as the pioneer of the pharmacophore [16]. The actual concept of pharmacophore however was first introduced by Lemont B. Kier in the late 1960s [17]. In 1977, Günd provided additional elucidation of the concept of pharmacophores as a collection of molecules capable of recognizing receptors and delineating structural attributes contributing to molecular biological activity [2]. The formal definition established by the International Union of Pure and Applied Chemistry (IUPAC) in 1998 is articulated as follows: "A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [18]. A Pharmacophore therefore, as is often misleadingly believed, does not represent an actual molecule with its corresponding functional groups, but rather an abstracted construct that describes the electrostatic and steric relationships of active molecules and a given target [6, 9].

For three-dimensional (3D) pharmacophore models to be a valuable asset in drug development with robust predictive capabilities, they must possess the ability to represent both the nature and spatial arrangement of functional groups engaged in ligand-target interactions. Additionally, they should be able to systematically describe various non-covalent bond types and their characteristics in a manner that is easily understandable [9]. Consequently, pharmacophore models contain information about interactions or chemical features arranged in three-dimensional space. This spatial formation of chemical features serves as a representation of the crucial interactions between smaller organic ligands and a larger macromolecular receptor. Thus, a pharmacophore model is limited to include only a single mode of action, es-

sentially illustrating how its ligands bind to a desired target. The 3D pharmacophore concept is rooted in the specific types of interactions commonly observed in drug-receptor interactions, including hydrogen bond donors, hydrogen bond acceptors, charge transfer, positively and negatively charged groups, hydrophobic interactions, aromatic interactions and exclusion volumes. Uncommon interaction types that contribute to ligand binding, include metal interaction and halogen bonds. Consequently, the pharmacophore may be viewed as the most extensive common denominator shared among a group of active molecules [1, 4, 9, 14, 15].

2.1.2 Pharmacophore Creation

To create a pharmacophore, a series of three primary tasks must be performed, which involve assembling an appropriate dataset, constructing a potential pharmacophore or pharmacophores and subsequent validation of the pharmacophore or pharmacophores [3].

3D pharmacophore creation approaches can be categorized based on how they obtain pharmacophore features, organised into three main groups: feature-based, substructure pattern-based, and molecular field-based approaches. In feature-based methods, pharmacophore features are determined by filtering geometric descriptors that match the attributes of molecular interactions. Substructure pattern-based methods, which can be found in software like PHASE [19], LigandScout [14], and Catalyst [20], identify chemical feature substructures within molecules. For instance, all hydroxyl groups might be designated as both hydrogen bond donors and acceptors. While in comparison, molecular field-based methods like those used in FLAP [21] and Forge [22] examine the molecular surface of either the ligand or the macromolecular target using various chemical probes. They then calculate interaction energy maps, which can subsequently lead to pharmacophore features [4].

Another distinction can be drawn between structure-based and ligand-based pharmacophore modeling methods. Structure-based pharmacophores are based on the structural data of proteins or ligand-protein complexes. Contrary, ligand-based methods utilize a set of active ligands and structural information about their active conformations for the creation of the pharmacophores [4].

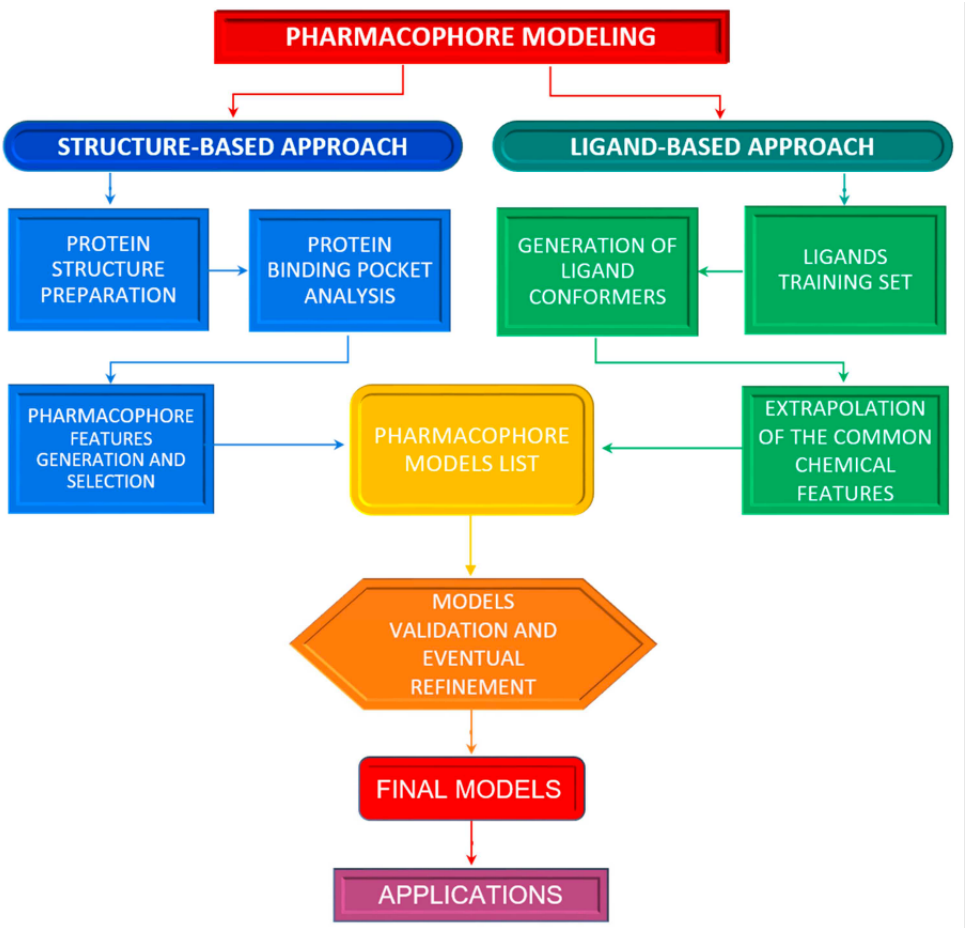


Figure 2.1: Workflow for pharmacophore modeling [1].

Figure 2.1 provides a comprehensive representation of the workflow involved in constructing pharmacophore models, highlighting the distinctions between the structure-based and ligand-based approaches. Following the initial development of pharmacophore models, which can be derived from either protein structural information or a collection of active ligands, these models typically undergo a validation and refinement phase to produce the definitive models that can subsequently be employed in further investigations.

Structure-based Pharmacophores

In order to generate a structure-based pharmacophore, it is imperative to possess relevant information regarding the target protein or the ligand-protein complex. On account of the substantial growth in the availability of 3D structures of macromolecules, particularly proteins, in recent years, the structure-based approach is gaining increasing significance in the realm of pharmacophore modeling. Ideally, a structure is available where a ligand is already present within the binding pocket of the target protein. This ligand-protein complex can then be made accessible by co-crystallisation [4].

In cases where structural data for the ligand-protein complexes or ligands for the binding pocket are unavailable, alternative methods can be employed to gather macromolecular structural insights and generate pharmacophore models. These techniques include homology modeling, binding pocket prediction, apo-site pharmacophore modeling, and molecular dock-

ing. Structure-based approaches rely on identifying potentially significant interaction sites, which can be pinpointed using energy-based or geometry-based methods and the acquired data can subsequently be utilized in the construction of a structure-based pharmacophore model [4, 23].

Essentially, the core concept of structure-based pharmacophore modeling hinges on the derivation of chemical features from macromolecular structural data. Typically, these models are constrained to a single protein-ligand complex, representing only one binding mode. To determine the optimal arrangement of chemical features, many software applications initially analyze all potential features of the ligand as well as those within the binding pocket of the target. A pharmacophore feature is subsequently only then incorporated into the model if a complementary partner is identified [24].

An additional factor to consider during the construction of both ligand- and structure-based pharmacophores is the quantity of pharmacophoric features, as it should align with practicality for subsequent experiments, such as virtual screening. A good balance is crucial, ensuring an adequate number of features for specificity while avoiding an excess that could overly constrict the model, potentially leading to false negative outcomes [4].

Taking these factors into account, the application of structure-based 3D pharmacophore modeling emerges as a promising and captivating approach for elucidating direct precise insights into interactions between a molecule and a macromolecule at the binding cite [9, 24].

Ligand-based Pharmacophores

In the absence of the structural information about the target protein, ligand-based pharmacophores are used. They include the chemical features mutually present in a group of active compounds, which are known to exhibit biological activity and a shared mechanism of action towards a desired target. The selection of ligands at this stage is highly dependent on the desired outcome. Typically, these pharmacophoric features are obtained by aligning the active molecules in their different conformations, in such a way that these features are located similarly. A pharmacophoric feature is designated to a certain location when the aligned compounds share the same chemical feature at this specific location. At this point it is relevant to mention that it is often unclear in which conformation the molecules show bioactivity. However, conformer generation algorithms are usually relatively efficient at finding bioactive conformations, but it is still never entirely clear in advance whether ligand-based methods actually predict an alignment with molecules in there bioactive conformations. In addition, it is also rather important that the molecules are as structurally similar as possible and share the same binding mechanism, otherwise with molecules getting more diverse it will become increasingly difficult to find proper alignments for the algorithms. Furthermore, it is not always certain whether diverse active molecules indeed interact with a common binding pocket, potentially complicating the pharmacophore generation process [4, 9]. The subsequent step involves a thorough examination of each conformation for every molecule, inspecting their 3D chemical attributes. The final step encompasses numerous alignment experiments focusing on the features of each conformation of every ligand, leading up to the construction of a ligand-based 3D pharmacophore model [24].

2.1.3 Pharmacophore Applications

With the growing importance of computer-aided drug discovery, pharmacophores have been effectively employed for a wide range of purposes, like gathering data on structure-selectivity relationships, ligand activity, as well as the prediction of molecular interactions including

metabolism, side-effects, toxicity and drug-drug interactions. Individually or in combination with other techniques, they have demonstrated to be most valuable in lead identification, lead optimization, and the development of new drugs [1, 3].

This section will highlight some of the most prominent application fields.

Virtual Screening

Most frequently, pharmacophores are used as filters to search vast databases of molecules on their particular features, allowing researchers to identify molecules having different chemotypes and scaffolds that align with features of interest. Pharmacophore-based virtual screening is a prominent in-silico tool for discovering and developing new drugs, drastically reducing the pool of potential candidates, enabling further in-vitro experiments to evaluate their efficacy [2, 5, 11].

Toxicity, pharmacodynamic and pharmakokinetic predictions

Pharmacophore models are often employed to guide absorption, distribution, metabolism, excretion and toxicity (ADMET) predictions. Suboptimal pharmacodynamic or pharmacokinetic characteristics frequently constitute the primary factors contributing to the ineffectiveness of prospective pharmaceutical agents. Therefore, it is a well advised strategy to employ pharmacophore models for predictive purposes during the initial phases of drug discovery. These models can be instrumental in recognizing potential interactions between novel compounds and predominant metabolizing enzymes [8, 12].

Pharmacophore-based de novo design

The design of novel compounds that are of interest to specific targets offers another potential application for pharmacophores. De novo design enables molecules with new structures desired features to be developed in order to synthesise potential novel drugs. In most cases, however, the structures obtained only provide prototypes for new active substances, as they often lack in activity. Subsequently, with the aid of pharmacophore models completely novel chemical compounds are designed, which subsequently often have to be modified manually. That is why it is particularly important to design them in such a way, that the molecules aquired are also accessible for chemical synthesis [5].

Molecular dynamics simulation

As previously discussed, pharmacophores primarily emerge from superimposing atomic models of ligand-protein complexes, with the disadvantage that they represent a rigid image of the structure. Nevertheless, there are alternative approaches for forecasting and employing pharmacophoric information. Molecular dynamics (MD) simulations, for instance, are founded on the premise that both macromolecules and ligands are flexible, consequently implying that their complexes should also demonstrate dynamic behaviors. MD simulations are employed to forecast the temporal positions of atoms within molecular systems, relying on Newton's laws of motion. It involves estimating the forces between interacting atoms through the utilization of an appropriate force field, ultimately leading to the calculation of the system's total energy. Grounded in this underlying principle, numerous methods and software applications have been devised, to generate pharmacophores or refine pre-existing models [4, 5, 10].

Machine learning

Given the recent surge in interest and enthusiasm surrounding artificial intelligence and machine learning, approaches have emerged that combine pharmacophoric methods with machine learning techniques. For instance, the software HS-pharm [25], employs machine learning techniques to condense the number of pharmacophoric features. Pharm-IF [26] utilizes diverse machine learning algorithms to evaluate the docking poses of ligands. DeepSite programs like KDeep [27] or LigVoxel [28] harness the combined power of pharmacophores and machine learning to train neural networks, subsequently facilitating the prediction of binding affinities or the design of novel compounds [4].

2.1.4 Pharmacophore Modeling Software

This section is concluded by providing an inventory of some of the most recent commercial software applications and online tools that employ pharmacophore modeling, accompanied by brief information about their services.

Software	Input data	Method of Identification
FLAP [21]	Ligand, complex, apo	Molecular field
Pharmer [29]	Ligand, complex	Feature, substructure pattern
LigandScout [14]	Ligand, complex, apo	Feature, substructure pattern, molecular field
Catalyst [20, 30]	Ligand, complex, apo	Feature, substructure pattern, molecular field
MOE [31]	Ligand, complex, apo	Feature, substructure pattern, molecular field
PHASE [19]	Ligand, complex, apo	Feature, substructure pattern, molecular field
Pharao [32]	Ligand	Substructure pattern
UNITY [4]	Ligand, complex	Feature, substructure pattern
Forge [22]	Ligand	Molecular field

Table 2.1: Software for pharmacophore modeling [4].

Freely accessable online tools

PharmaGist [33] serves as an web server for ligand-based pharmacophore modeling. It offers the capability to upload a maximum of 32 ligand molecules for the generation of a pharmacophore model from the provided input utilizing the common feature method [1, 10].

Pocketv.2 [34] encompasses both a web server and a stand-alone version designed to construct structure-based pharmacophore models by using the pocket module within LigBuilder. Necessary input data for generating the model includes a Protein Data Bank (PDB) [35] structure of a macromolecule, either with or without an associated ligand [1, 10].

2.2 Virtual Screening

Given that virtual screening ranks among the most prevalent applications of pharmacophores, and the research in this thesis predominantly focuses on virtual screening, the subsequent section will provide a more detailed elucidation of this subject matter.

2.2.1 Virtual Screening Definition

Virtual screening describes the process of filtering molecular databases with the primary aim of predicting activity of specific compounds at a desired biological target. Due to the fact that virtual screening is a computer-based method and therefore does not require the in vitro synthesis of molecules in advance, it is an excellent supportive tool for the early stages in the development of new drugs. The great ability of virtual screening lies in helping to selectively filter bioactive compounds out of large virtual databases while at the same time excluding the ones that are not relevant [3].

2.2.2 Virtual Screening Approaches

Over the last decades, the evolution of virtual screening methods has witnessed substantial advancements, leading to increased levels of user-friendliness, utility, and overall performance enhancements [3]. Computational systems and their software applications have undergone rapid progress, significantly contributing to the reduction of both temporal and financial resources required for the development of novel drugs [2].

Initial two-dimensional (2D) substructure-based similarity searches have primarily been constrained in their capacity to identify molecules of identical structural types or those sharing closely resembling molecular frameworks. In contrast, novel approaches, especially 3D pharmacophore-based screening, expands the scope to include the identification of compounds with different scaffolds still fitting searched for features without being constrained by any particular molecular structure [4, 5].

The following sections highlight some of the most prominent approaches for virtual screening.

Quantitative structure-activity relationship

Quantitative structure-activity relationship (QSAR) investigates the interaction dynamics between small molecules and larger macromolecules. However, QSAR is specifically concerned with establishing correlations between computed molecular properties and the biological activities of molecules. QSAR proves notably advantageous in the context of drug development, in cases where a receptor's structure remains undisclosed. This method can be exceptionally valuable in identifying compounds with inhibitory attributes and minimal potential for toxicity for a desired target. During the 1980s and 1990s, this approach underwent substantial refinement with the incorporation of a 3D component. Within 3D-QSAR, the investigation extends to the 3D characteristics of both ligands and target structures. It encompasses an analysis of the molecules' 3D configurations, energy alterations, and specific interaction patterns between the active compounds and their targets. By combining physicochemical attributes, 3D structural information, and quantitative relationships, this method has proven effective in the anticipation and enhancement of novel drug candidates. Conse-

quently, 3D-QSAR has evolved into a pivotal tool for advancing drug development since the 1990s [2].

Molecular docking

Much like QSAR, molecular docking is a fundamental approach to predict the interactions between small molecules and proteins, or even between two proteins themselves. This process relies on the analysis of both the energetic and spatial configurations to discern the accurate binding mechanisms or conformations. Frequently, molecular docking simulations are employed for the screening of compound datasets against a specific target, with subsequent ranking based on their predicted binding affinity [2, 8].

In general, docking methods encompass the utilization of the available information from the macromolecules environment to assess various potential interaction modes through different alignments. Initially, docking involves the flexible alignment of the ligand molecule within the macromolecular surroundings and subsequently evaluates the strength of the interaction using diverse scoring functions [8, 14].

Molecular docking can be systematically classified into three distinct categories: rigid docking, semi-flexible docking, and flexible docking. Within rigid docking, the molecular structures remain unaltered. In semi-flexible docking, the small molecules have the ability for adjustment in their conformations, while the proteins or macromolecules maintain their rigidity. Lastly, flexible docking permits both molecules and proteins to move freely within their conformations. The selection among these methods is dependant upon the specific research objectives. A more flexible approach, while offering higher accuracy in predicting interactions, necessitates increased time and resources, making it important to weigh these factors in relation with the goals of the experiment [2, 8].

It is worth noting that employing docking for the screening of extensive compound databases can be computationally demanding. Consequently, some approaches have emerged that combine docking-based virtual screening with pharmacophore-based virtual screening to address these computational challenges [8, 14].
One possibility is to employ pharmacophores as preliminary filters for ligand databases, followed by the application of docking simulations. Alternatively, pharmacophores can be employed post-docking to filter and exclude ligands that have achieved favorable docking scores but do not align with the pharmacophore features. The utilization of pharmacophores during the actual docking simulation is also possible. In this context, the pharmacophore model can assist in guiding the ligand's placement during the docking process [8].

Pharmacophore-based virtual screening

Pharmacophore-based virtual screening utilizes 3D pharmacophores, usually derived from a set of active ligands or a ligand-protein complex facilitating the exploration of extensive virtual molecular databases to identify compounds that satisfy the criteria and align with the pharmacophoric features. Through this filtering process, potential candidates for future active substances can be identified at an early stage of drug development [4].

Pharmacophore-based virtual screening finds application in various domains, including drug discovery, lead identification, selectivity and toxicity profiling, scaffold hopping, structure-activity relationships, and in synergy with complementary methods such as docking or molecular dynamics simulations [6].

Given that the primary focus of this thesis regards this subject matter, the following section will delve into it with greater detail.

2.3 Pharmacophore-based Virtual Screening

In the actual process of pharmacophore-based virtual screening, the molecules in the libraries are scanned for the desired pharmacophoric features. The associated methods can be broken down into two different concepts, fingerprint-based and 3D alignment-based methods [4].

Fingerprint-based techniques predominantly capture data regarding the presence of features and interfeature geometries in fingerprint descriptors. This allows for efficient similarity comparisons between the query pharmacophore and a library of conformers. In contrast, alignment-based methods involve the 3D alignment of the pharmacophore feature set. A match is recorded when the pharmacophoric feature set of a specific conformation of a molecule can be aligned with the feature set of the pharmacophore [4].

3D alignments often involve preliminary prefiltering procedures utilizing rapid distance assessments, which significantly reduce computational efforts. The Catalyst software, for example, uses an algorithm employing a "pruned exhaustive search" technique to gradually construct shared 3D pharmacophores from two-feature pharmacophores identified in molecule conformers. In order to assert the presence of a shared 3D pharmacophore, at each step, a precomputed list encompassing all interfeature distances within the molecule is initially consulted. Following this prefiltering stage, alignment occurs through a least-squares fit of the features. LigandScout, for example, identifies optimal alignments by evaluating the best pairings between two sets of pharmacophore features based on interfeature distances before proceeding with the actual alignment [4].
The most commonly employed software used include those listed in 2.1.4: FLAP [21], Pharmer [29], LigandScout [14], Catalyst [20], MOE [31], PHASE [19], Pharao [32], UNITY [4], Forge [22]. The aforementioned software programs are all suitable for pharmacophore-based virtual screening, as they incorporate functions for both pharmacophore modeling and virtual screening [4].

2.3.1 Workflow for Pharmacophore-based Virtual Screening

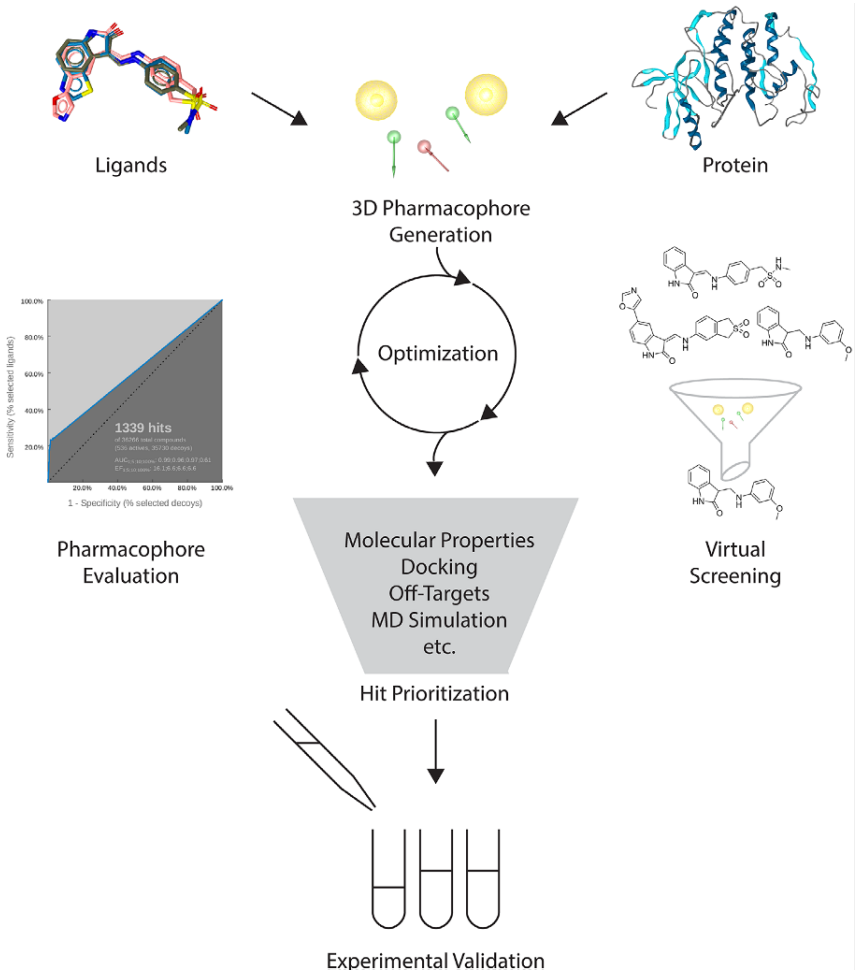


Figure 2.2: Workflow of pharmacophore-based virtual screening [4].

Figure 2.2 provides a representation of the workflow involved in constructing pharmacophore models, highlighting the distinctions between the structure-based and ligand-based approaches. Following the initial development of pharmacophore models, which can be derived from either protein structural information or a collection of active ligands, these models typically undergo a validation and refinement phase to produce the definitive models that can subsequently be employed in further experiments.

Up next, the individual steps of this procedure are examined in greater detail.

Creation of the pharmacophore

The initial stage in pharmacophore-based virtual screening involves the construction of a pharmacophore model, serving as a representation of the specific features that must be satisfied by the screened compounds [9].

Precise methods for developing pharmacophores have already been discussed in Section 2.1.2.

Database creation

As previously mentioned, the flexibility of molecule conformations represents an important issue in the screening process. Addressing this matter can be pursued through two distinct approaches. Databases can be pregenerated, containing precomputed conformations of molecules, offering the substantial benefit of a one-time computational effort. Alternatively, conformations can be calculated in real-time during the actual screening process. Presently, the pre-calculation and storage of conformations is the favored method, primarily due to its cut down of computer resources (as conformations can be reused for multiple screens), time efficiency, and cost-effectiveness [5].

Database screening

As for actual screening of databases, it is highly relevant to implement a series of pre-filtering measures. These measures encompass various techniques, including the application of filtering criteria such as the Lipinski Rule of Five or PAINS (pan-assay interference compounds). However, these rules do not necessarily always have to be applied, as there may also be targets for which ligands do not conform to these rules but still fit. Additionally, feature types and feature counts are taken into consideration to assess the potential exclusion of molecules as non-matching. Subsequently, 3D matching algorithms are deployed, typically wtih higher computational costs, though in favor of the increased precision they offer [1, 5].

It is essential to consider that certain methods may also filter out molecules that could potentially conform to the fitting criteria later on. While some approaches tolerate this occurrence, acknowledging that it leads to more efficient screening, others, such as LigandScout, employ lossless filter methods that exclusively eliminate molecules lacking geometric compatibility [5].

During the actual matching phase, the primary objective is to find out whether the molecules that have progressed to this stage align with the pharmacophores and get included in the hitlist. This step bears significant importance and exerts a direct influence on the quality of the results. The utilization of greedy algorithms and two-point pharmacophores, which involve pure feature pair comparisons were introduced at an early stage. However, they lacked the capability to distinguish between the pharmacophore and its enantiomer, necessitating the implementation of a 3D overlay pharmacophore for accurate match prediction. Noteworthy software packages for pharmacophore modeling, renowned for their state-of-the-art screening functions, include Catalyst [20], Phase [19], MOE [31], and LigandScout [14]. These programs all employ some form of geometric alignment during the 3D pharmacophore matching stage, with a focus on minimizing the RMSD (root mean square deviation) between feature pairs. While all programs pursue n-point distance-based searches to achieve matches, LigandScout employs a pattern-matching technique, which imposes fewer constraints on the number of features within the pharmacophore model, contributing to its distinct approach [5].

Analysis of the hitlist

When validating derived 3D pharmacophore models, it is immanent to have experimental data for searched molecules. Typically, a validation set for 3D pharmacophores comprises reported active, inactive, or decoy molecules. Two critical considerations come into play when assembling the validation set: Firstly, 3D pharmacophores delineate a specific binding pose. Consequently, the active set should encompass ligands that share the same binding

mechanism within the target protein. Secondly, caution must be exercised when incorporating reported inactives, as observed inactivity may arise from factors unrelated to the binding mode, such as insolubility or an inability to reach the target in cell-based assays. Therefore, it can be advisable to favor selected decoys over inactive molecules. Decoys are compounds presumed to be inactive and exhibit a high degree of physicochemical similarity to the active compounds. However an issue with decoy compounds lies in their lack of empirical validation against the target, potentially leading to designing active compounds unintentionally. The Directory of Useful Decoys (DUD-E) [36] for example offers a convenient web-based tool for generating decoys. With the advantages and disadvantages in minde, the choice between decoys or inactives for model validation has to be adjusted within the particular cases at hand. Subsequent screening against the selected validation set serves to assess the quality of the developed 3D pharmacophore and provides opportunities for further optimization [4]. To ensure success, a comprehensive evaluation of the outcomes becomes essential. An initial step in this evaluation process involves the validation of the hitlist. The following criteria are most frequently used for the purpose of assessment [5].

Enrichment parameters categorize the compounds within the data set into subsequent four distinct classifications [4]:

- Actives (true positives, TP)
- Inactive compounds that are falsely identified as actives (false positives, FP)
- Genuinely inactives (true negatives, TN)
- Active compounds that are misclassified as inactive (false negatives, FN)

The **yield of actives (Ya)** quantifies the ratio of true positives within the comprehensive list of hits (n) obtained through the pharmacophore model [4, 5].

$$Ya = \frac{TP}{n}$$

The **enrichment factor (EF)** measures the yield of actives relative to the ratio of active compounds within the database. It is expressed by the Ya divided by N, which denotes the total number of database molecules, excluding their conformations [5].

$$EF = \frac{Ya}{N}$$

Sensitivity (Se) describes the obtained true positives (TP) compared to the sum of TP and false negatives (FN) [5].

$$Se = \frac{TP}{TP + FN}$$

Sensitivity values vary between 0 and 1. A sensitivity value of Se = 0 signifies that the search failed to identify any of the active compounds within the database, while Se = 1 denotes that the search successfully retrieved all active compounds [9].

Specificity (Sp) characterizes the correctly identified true negatives (TN) relative to the combined total of TN and the detected false positives (FP) [5].

$$Sp = \frac{TN}{TN + FP}$$

Specificity can be measured on a scale from 0 to 1. A specificity value of 0 implies that none of the inactive compounds were correctly identified, while a value of 1 indicates that all inactive compounds were accurately dismissed during the procedure of screening [5, 9].

The **Goodness of hitlist (GH)** displays a combination of yield of actives (Ya), Sensitivity (Se), and Specificity (Sp). It is calculated as the sum of Ya and Se, multiplied with Sp. The possibility of weighting Ya and Se allows for the prioritization of certain factors; for instance, assigning greater weight to Ya would require a higher number of correctly identified actives with minimal false negatives to achieve a high GH value [5].

Exceptionally valuable and arguably the most illustrative manner for the evaluation of screening outcomes are **receiver operating characteristic (ROC)** curves. ROC-curves describe the true positive rate relative to the false positive rate, thereby combining the sensitivity and specificity. These curves provide insight into the model's ability to identify active hits while also revealing the extent to which it correctly classifies inactive compounds as inactive or decoys as decoys [4, 5, 9, 37].

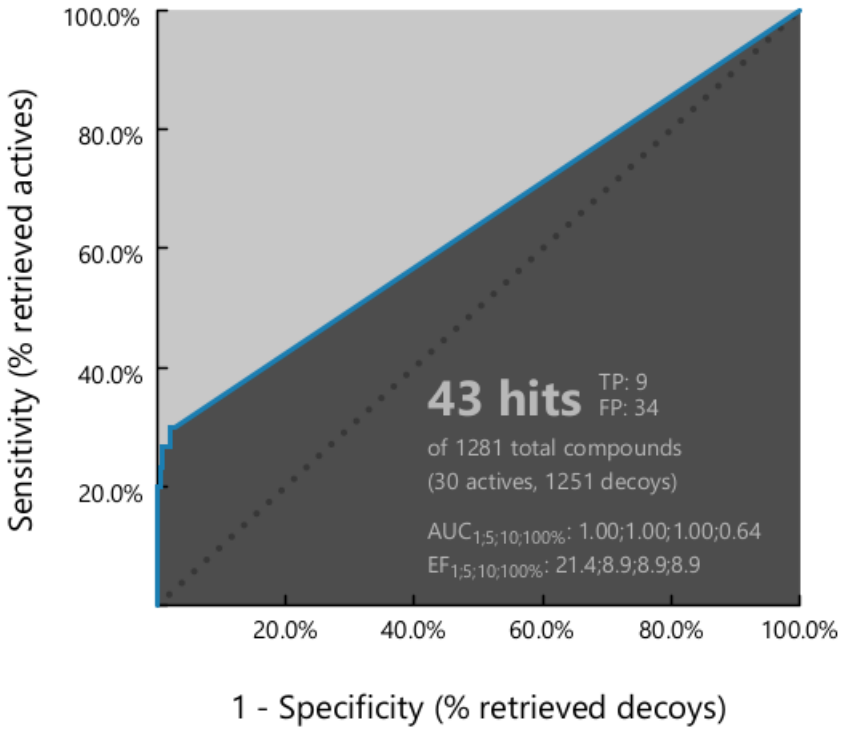


Figure 2.3: ROC-curve generated within virtual screening in LigandScout.

The Y-coordinate of the ROC-curve in Figure 2.3 displays the true-positive rate, while the X-coordinate represents the corresponding false-positive rate. Effective models exhibit a large area under the curve (AUC). It's crucial for the curve to rise rapidly along the Y-axis, with a maximum true positive rate of 1, indicating that the hit-list successfully identified all potential active compounds. In contrast, ineffective models display flatter curves with a significantly smaller area under the curve. In such scenarios, the performance of the pharmacophore model may be suboptimal, often comparable to random assignment, as it retrieves an excessive number of inactive compounds in relation to the genuine active hits, showing an undesirable outcome in this context. The ROC-curve for a random database search is typically represented by the median [9, 37].

In conclusion, a parameter is introduced, that predominantly served as a point of comparison in the experiments commencing with Chapter 4.

The **F1-Score** incorporates Precision and Recall, serving as a valuable comparative metric for evaluating the balance between true positives (TP), false positives (FP) and true negatives (TN). Precision quantifies the proportion of correctly identified positive hits, while Recall aligns with Sensitivity, indicating the fraction of actual positives captured. The F1-Score ranges from 0 to 1, with 0 indicating the poorest performance and 1 signifying optimal performance [38].

They are expressed as follows:

$$\begin{aligned} F1_{score} &= 2 \cdot \frac{precision \cdot recall}{precision + recall} \\ &= 2 \cdot \frac{TP}{TP + FP + FN} \end{aligned}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall (= Sensitivity) = \frac{TP}{TP + FN}$$

Refinement of the pharmacophore

With the data obtained from the analysis of the hitlist, it is now possible to deliberate on potential refinements aimed at enhancing the method. This may involve a detailed examination of the pharmacophore model, wherein foremost adjustments of the features are considered. Alternatively, a reevaluation of the database can be undertaken with the objective of optimizing the representation of the pre-calculated conformers [5].

Generally, it is important to acknowledge that both the model development and screening process are characterized by a high degree of complexity, depending on the factors mentioned above. Consequently, it is advisable to systematically assess the model's validity in comparison to the input data, often necessitating multiple iterations of the defined steps to achieve the desired outcomes [1, 5].

With the successful development and thorough validation of the pharmacophore model, it can subsequently be applied to the extensive screening of databases in search of specific compounds. Researchers have the opportunity to screen commercially accessible databases like the compound libraries offered by Enamine [39], MDL Drug Data Report, open-access databases like the Protein Data Bank (PDB) [35], PubChem [40], ChEMBL [41], Zinc [42], Drugbank [43] or individually created and owned databases [1, 4].

Furthermore, it can be of advantage, to additionally integrate techniques such as molecular docking or molecular dynamics simulations to enhance the knowledge of structural insights, ultimately clearing the way for the identification of the most promising molecules for subsequent in vitro experiments [4].

Part II

LigandScout

Chapter 3

LigandScout

3.1 LigandScout Introduction

LigandScout is a software application that has undergone ongoing development since 2005, aimed at molecular modeling and design. In its initial version, LigandScout primarily focused on the creation of pharmacophore models derived from protein-ligand complexes. However, its evolution over the years has been marked by significant expansion, encompassing a diverse array of functions. Presently, it additionally offers the capability to extract pharmacophores from ligands, investigate binding pockets, perform docking, create extensive databases, apply advanced filtering techniques, and execute comprehensive virtual screening procedures making it possible to search vast compound libraries. Equipped with a user-friendly interface, LigandScout empowers scientists worldwide to engage in molecule design, filtration, and retrieval. It has demonstrated its efficacy in early hit and lead discovery, as well as in evaluating ligand and macromolecule activity. Notably, LigandScout's pharmacophore models and virtual screening functionalities have demonstrated superior performance compared to alternative methods [11, 15, 44, 45]

3.2 LigandScout Tools

LigandScout stands out due to its distinct algorithms, which enhance the definition of pharmacophore features, the generation of ligand conformations, and the alignment of molecules through pattern matching. Furthermore, it offers the advantage of not restricting the number of features that can be incorporated into a model [7].

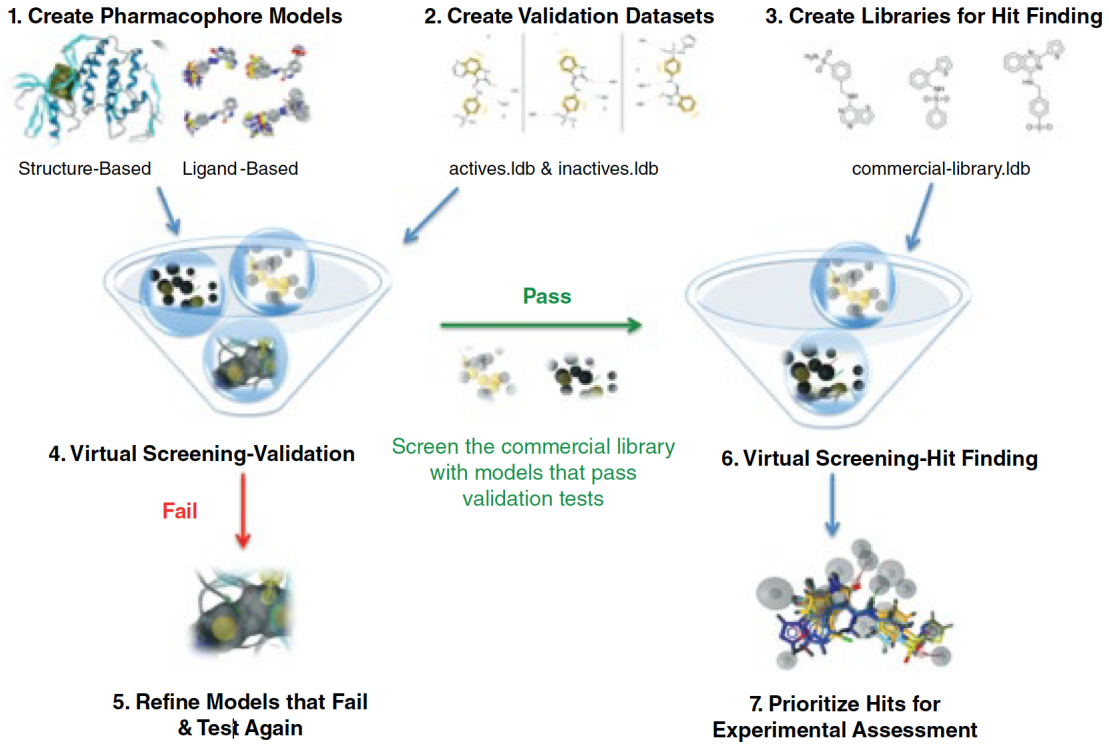


Figure 3.1: Pharmacophore-based virtual screening workflow in LigandScout [15].

Next to Figure 3.1 providing a comprehensive rundown of the typical workflow steps when applying pharmacophore-based virtual screening, an overview of the functionalities offered by LigandScout 4.5 in the context of pharmacophore modeling and virtual screening will be presented [15].

3.2.1 Pharmacophore Creation in LigandScout

Depiction of pharmacophores in LigandScout

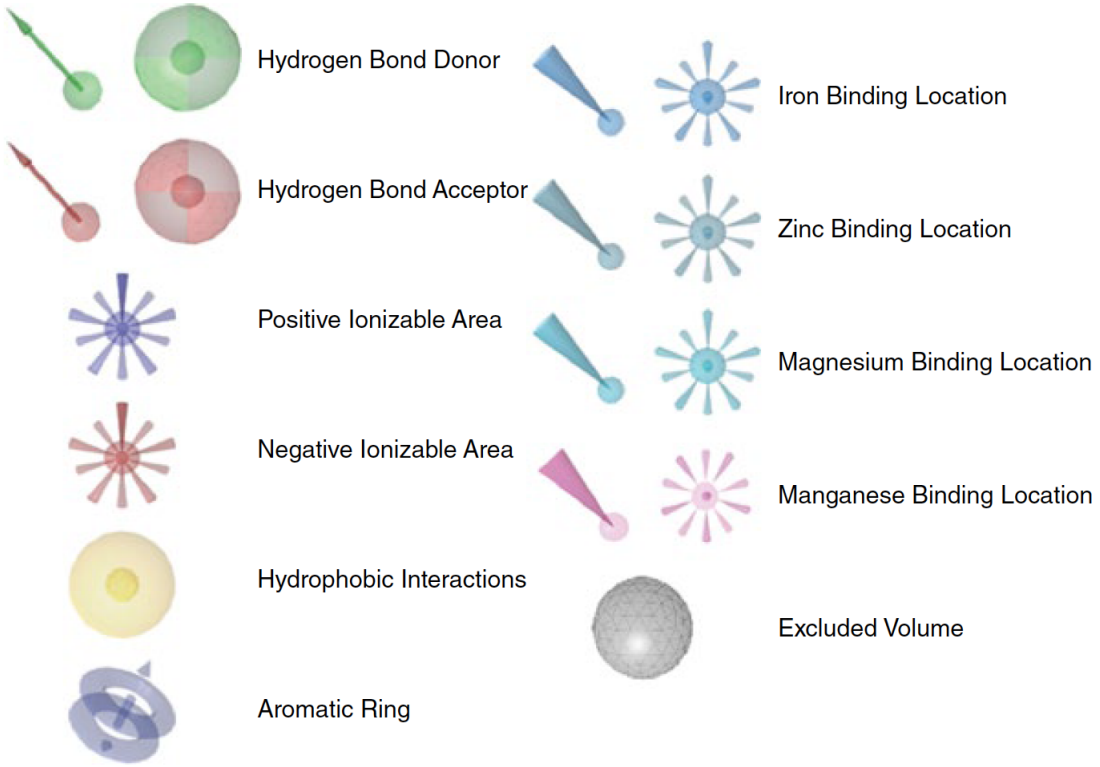


Figure 3.2: Representation of pharmacophoric features in LigandScout [15].

Figure 3.2 depicts the available pharmacophoric features in LigandScout, along with their illustrative representations. The display of hydrogen-bond donors or acceptors may vary, with options including spheres or vectors depending on the specific model [15].

Manual pharmacophore construction in LigandScout

LigandScout offers the capability for manual pharmacophore model creation. Nevertheless, with the complexity of feature selection, coupled with the valuable insights provided by other approaches, the relevance of this function has diminished, thus this subject will not be delved into deeper at this point [15].

Structure-based pharmacophores in LigandScout

LigandScout also provides the capability to construct structure-based pharmacophore models by leveraging data extracted from PDB (Protein Data Bank) files. This process involves utilizing the 4-letter PDB code (part 1 in Figure 3.3) to retrieve the associated protein-ligand complex within the software, which is subsequently visualized in 3D (parts 2-3 in Figure 3.3). Within this view, the bound ligand is distinctly highlighted by a yellow box, which can be examined in a 3D representation within the macromolecule or in a 2D format presented in a separate field. At this stage, a review of the ligands bonds is advisable to ensure their

accurate representation, and adjustments can be made as needed (parts 4-5 in Figure 3.3). Once a satisfactory depiction has been verified, one can proceed to instruct LigandScout to generate a pharmacophore model.

With the protein-ligand complex uploaded, LigandScout conducts an analysis to fix the hybridization states of unsaturated bonds and aromatic rings. Subsequently, both the ligand and the amino acids within the binding pocket are inspected to identify atoms and groups capable of participating in interactions such as hydrogen bonding, hydrophobic, aromatic, ionic, and metal binding. In cases where complementary interaction partners between the ligand and binding site functionalities are present, LigandScout automatically incorporates a corresponding feature into the pharmacophore model. The detection of pharmacophoric features can be tailored to specific interactions by adjusting geometric characteristics like allowable distances and angle ranges. Inclusion of a feature in the final pharmacophore model depends on its spatial relationship relative to a complementary feature within the binding site. For instance, a hydrogen-bond acceptor feature located on a ligand's acceptor atom is included only if there exists a corresponding hydrogen-donor feature on the receptor side, within specified distance and angle parameters. Following the comprehensive analysis of all complementary feature pairs within the complex and the integration of corresponding ligand features into the derived pharmacophore model, exclusion volume spheres are introduced to emulate the shape of the binding pocket [15, 44].

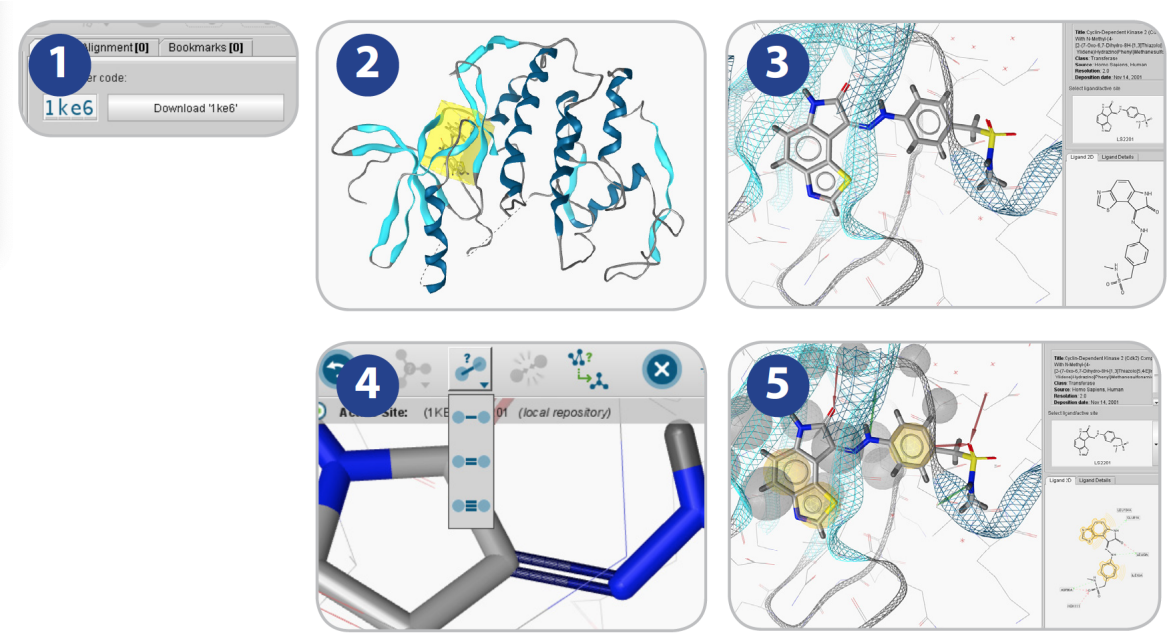


Figure 3.3: Screenshots highlighting the steps of creating a structure-based pharmacophore model in LigandScout [44].

LigandScout also provides the option to generate pharmacophores from multiple PDB structures. The process involves initially creating any number of distinct pharmacophores, as previously described, and subsequently superimposing them to produce a shared feature pharmacophore (Figure 3.4), aligning the associated ligands [5, 44].

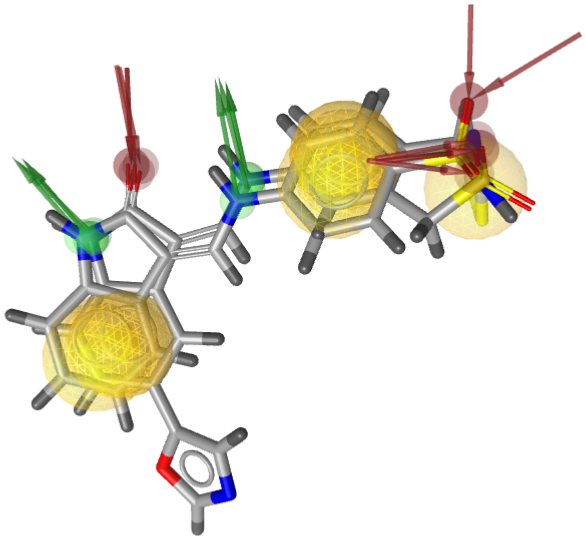


Figure 3.4: Screenshot of a shared pharmacophore and the associated ligands in LigandScout.

Another capability to generate pharmacophores supplied is using proteins that do not yet have associated ligands bound to them. The process begins by uploading a protein into the software using the 4-letter PDB code. Subsequently, the LigandScout Pocket Finder tool is employed to compute a druggable pocket and binding site within the macromolecule. This allows for the calculation of all potential pharmacophoric interactions and provides the flexibility to adjust the specific features. After configuring the features to satisfaction and determining the desired feature count, the software can be utilized to create a pharmacophoric model [44].

Ligand-based pharmacophores in LigandScout

An essential and crucial step in developing a ligand-based pharmacophore involves the availability of relevant ligands. The significance of establishing a well-suited database and the conformers of the corresponding molecules is also encompassed within LigandScout's tools. Like in most cases, conformers are estimated before the actual screening step in order to avoid exceedingly long runtimes during virtual screening later on. Utilizing the IDBGEN molecular database generator, users can input a multitude of ligands in the smi, sdf or mol file format, before the iCon tool is subsequently able to generate a set encapsulating the created low energy 3D conformations of the molecules. Following this, physicochemical parameters and molecular descriptors can be computed, as well as diverse filters can be applied to attain a comprehensive overview of the results. Apart from the option to save the results as ldb files, users can also export the outcomes as 2D sdf files or as a list of molecules in a Microsoft Excel file. Once a desired set of ligands is uploaded to the software the Espresso algorithm can be applied to perform clustering of the molecules based on their 3D pharmacophore features. The resulting clusters are organized according to their cluster ID and cluster size. After this process, ligands from the clusters can be chosen, for instance with the largest cluster size and same cluster ID for the development of a ligand-based pharmacophore model. In the next step, the actual core of the process, the software tries to find a pattern of chemical features that matches all of the designated ligands in at least one conformation. As often more than

one pattern emerges, LigandScout creates several pharmacophore models and compares them sorted by fitness function, making it possible for the user to choose a preferred model. Each model not only represents the identified features but also links them to the associated ligands [11, 15, 44, 46].

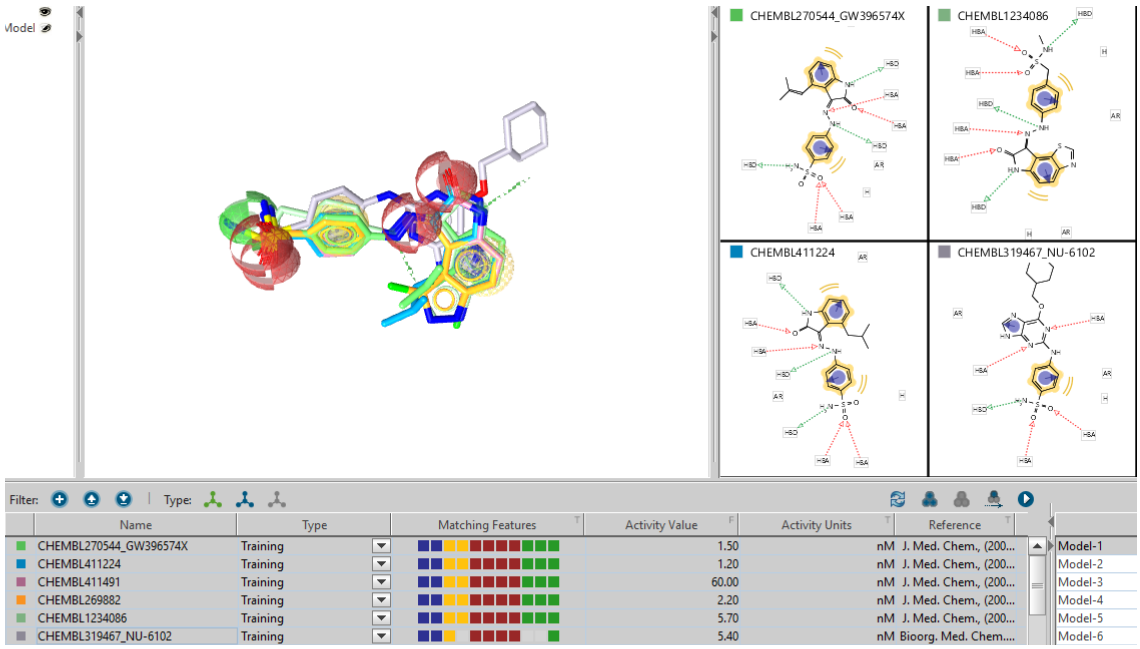


Figure 3.5: Ligand-based pharmacophore modeling in LigandScout.

Figure 3.5 illustrates the user interface following the generation of a ligand-based pharmacophore in LigandScout.

3.2.2 Virtual Screening in LigandScout

Virtual screening in LigandScout utilizes a pharmacophore-based method, examining molecules in inserted databases to determine if and how well they fit into a selected pharmacophore [11].

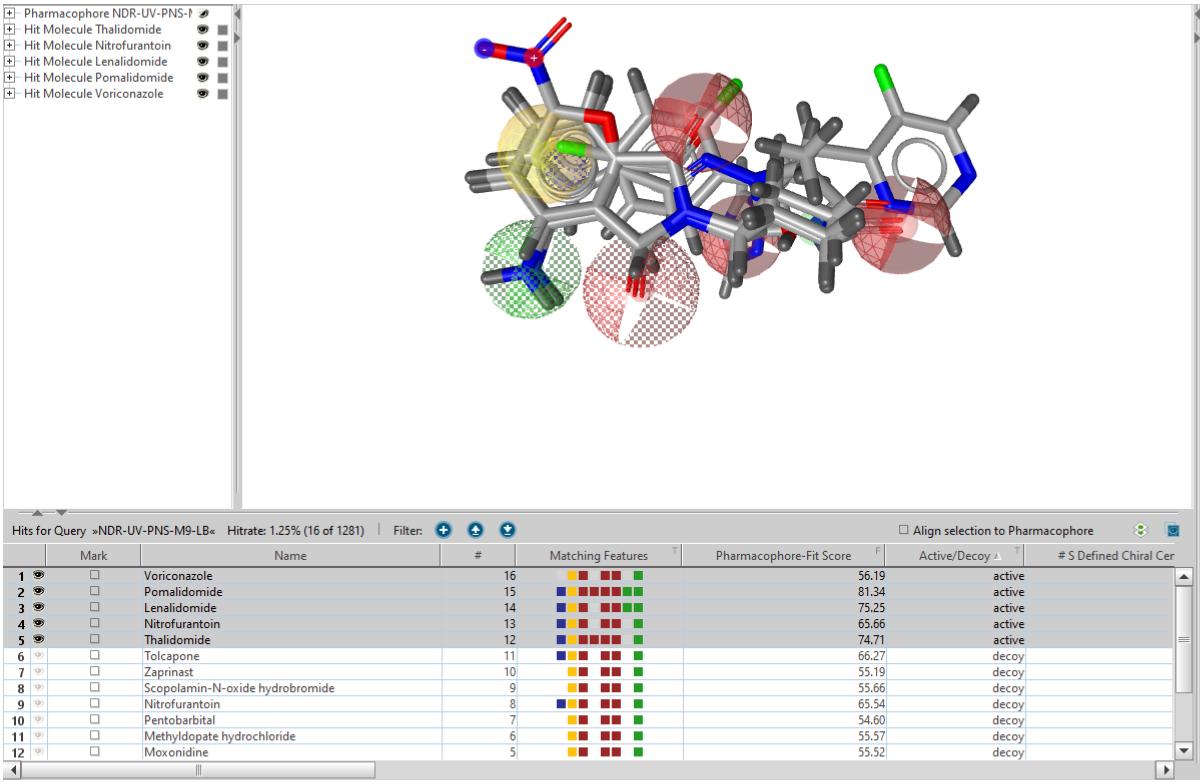
For the initiation of virtual screening, two types of data are required: a pharmacophore (which can be conveniently be provided in pml or pmz file format) and the ldb files containing the databases designated for screening. Once these essentials have been successfully imported into LigandScout, the virtual screening protocol can be put in place. Usually, virtual screening in LigandScout runs through several filter phases. LigandScout distinguishes itself by providing a lossless prefiltering process that ensures geometric precision, resulting in an accurate screening algorithm. During the actual alignment step, LigandScout carefully assesses the conformations of the remaining molecules, verifying the potential to align with the spatial arrangement of the inquired features. In order to accurately identify a molecule's compatibility with the specified feature tolerances within the query pharmacophore, a 3D spatial overlay is essential. This is also crucial for assessing and scoring additional constraints imposed by vector features like hydrogen-bond acceptors or donors, planar features such as aromatic rings, and exclusion or inclusion volume spheres. Latest commercial software for pharmacophore modeling, equipped with advanced screening capabilities, including

Ligandscout all employ a form of geometric alignment during the 3D pharmacophore matching process. Typically, this involves minimizing the root mean square deviation (RMSD) between corresponding feature pairs [4, 15].

LigandScout presents users with the flexibility to personalize some screening configurations according to their specific requirements. This adaptability encompasses the capacity to set the number of allowed or omitted features, thereby modulating model restrictiveness. Furthermore, users are empowered to directly change or cut features at the pharmacophore and adjust the dimensions of feature tolerance spheres, either expanding or contracting them as needed [15, 44]. LigandScout also provides the option to conduct screenings across multiple databases or pharmacophores concurrently. Various databases can be uploaded in ldb format, with the ability to designate their active or inactive/decoy status, which is crucial for pharmacophore validation. When it comes to pharmacophores, multiple ones can be selected and there is an option to define boolean expressions, providing the user with even more flexibility in the process. In principle, there are no restrictions to the number of databases or pharmacophores included at this stage [15, 44].

Analyzing screening Results

Following the completion of the screening step, users will obtain a hitlist, featuring the molecules identified during the process. Additionally, the software allows the loading and further analysis of pre-existing files at this stage. The list can be sorted based on various parameters, for instance including the Pharmacophore-Fit Score, or users can perform Gaussian Shape Similarity Score calculations within the program and subsequently sort the list accordingly. For visualization purposes, LigandScout allows for individual or overlapping 3D viewing of the molecules, with the additional option to apply color adjustments. When conducting simultaneous screenings involving active and inactive/decoy databases, a direct feature for generating ROC-curves from the results is also provided [15, 44].



	Mark	Name	#	Matching Features	Pharmacophore-Fit Score	Active/Decoy	# S Defined Chiral Cer
1		Voriconazole	16		56.19	active	
2		Pomalidomide	15		81.34	active	
3		Lenalidomide	14		75.25	active	
4		Nitrofurantoin	13		65.66	active	
5		Thalidomide	12		74.71	active	
6		Tolcapone	11		66.27	decoy	
7		Zaprinast	10		55.19	decoy	
8		Scopolamin-N-oxide hydrobromide	9		55.66	decoy	
9		Nitrofurantoin	8		65.54	decoy	
10		Pentobarbital	7		54.60	decoy	
11		Methyldopate hydrochloride	6		55.57	decoy	
12		Moxonidine	5		55.52	decoy	

Figure 3.6: Screenshot of the obtained hitlist in LigandScout.

Figure 3.6 exhibits the hitlist attained following virtual screening in LigandScout. Organized after the identified active compounds, which are also visually illustrated aligned to the pharmacophore model.

Refinement of pharmacophores

The examination of the screening results outlined above provides an initial foundation for refining a pharmacophore model. Nevertheless, it is typically necessary to iterate through the entire process multiple times to achieve an optimal pharmacophore model capable of effectively discriminating between true actives and inactives. After that, utilizing the model for screening extensive libraries of untested molecules becomes meaningful and practical [15].

Chapter 4

Improving Virtual Screening in Ligandscout

In addition to the currently commercially available LigandScout 4.5 version, a more recent iteration, LigandScout 5, is under development. This chapter aims to elucidate the distinctions in virtual screening methods and the resultant outcomes between LigandScout 4.5 and LigandScout 5, ultimately exploring strategies for enhancing virtual screening within the software. With LigandScout 5 still being relatively new, it has not yet been researched thoroughly which settings appear most effective and how this can impact research outcomes. In this regard a number of experiments have been performed commencing with Section 4.2.

Prior to advancing to the tests, a brief look at the disparities in alignment methods employed by the two versions is provided.

Ligandscout 4.5 uses the algorithm of Wolber et al. [47] for virtual screening and pharmacophore alignment. This approach encompasses the generation of cost matrices and further the utilization of the Hungarian Algorithm [48] to identify the optimal strategy for minimizing the total cost. Subsequently, matching feature pairs are derived, which are then employed in the actual 3D pharmacophore alignments using Kabsch´s Method [13, 49, 50].

LigandScout 5 implements a novel alignment algorithm known as Greedy 3-Point Search (G3PS). This method operates by encoding pharmacophore features and assessing the likelihood of feature matches. Given that a minimum of three feature pairs are required for a distinct 3D transformation, initial experiments are generated for precisely these three feature pairs. Subsequently, these pairs undergo refinement in the following step, wherein additional pairs are gathered without disrupting the alignment of previously collected pairs. This iterative process serves to optimize the alignment with the new found pairs. Following this procedure, there is an option to incorporate exclusion volumes. Last assessments involve additional checks to exclude alignments that do not fulfill more specific feature demands. Previous alignment techniques primarily focused on minimizing the root mean squared deviation (RMSD) between feature pairs or maximizing volume overlap using Gaussian Spheres. Nonetheless, these approaches face a challenge as their objectives do not sympathize with the essence of a pharmacophore model. Features exhibit specific tolerances for matching, meaning that the optimal alignment may not necessarily entail the lowest RMSD or the highest volume overlap. Instead, the ideal alignment should encompass the maximum number of

matching feature pairs within a specified position and orientation. Given that G3PS strives to identify the set of maximum matching feature pairs for optimal alignment, it stands as an exceptional new approach for conducting virtual screening with pharmacophore models [13].

This section is concluded by presenting three practical implementations of LigandScout in the context of pharmacophore-based virtual screening and providing a brief overview of potential results.

Nazarshodeh et al. made use of LigandScout version 3.12 to construct a pharmacophore model for carbonic anhydrase isoform XII (CA XII) with the aim of identifying potential inhibitors. In the validation phase, they evaluated their model against a dataset consisting of an active Set including 13 molecules and a decoy Set of 642 decoy molecules [51]. Virtual Screening findings are outlined below in Table 4.1.

Hits	Actives	Decoys
13	10	3

Table 4.1: Virtual screening results from validating the CA XII pharmacophore.

Moussa et al. implemented LigandScout version 4.4.1 for the validation of their Cyclooxygenase-2 (COX-2) pharmacophore. The dataset comprised five active ligands, while the decoy database contained 703 molecules [52]. Virtual screening yielded the subsequent results, displayed in Table 4.2.

Hits	Actives	Decoys
3	3	0

Table 4.2: Virtual screening results from validating the COX-2 pharmacophore.

Karaboga et al. employed LigandScout for the identification of CXR4 antagonists. Through the utilization of the virtual screening tool, they guided a validation of their five feature CXR4 pharmacophore. This validation was carried out by screening the pharmacophore against two distinct datasets: an active set incorporating 211 well-established, highly potent CXR4 antagonists, and a decoy database encompassing 4695 molecules [45]. Table 4.3 displays the resultant findings.

Hits	Actives	Decoys
298	184	114

Table 4.3: Virtual screening results from validating the CXR4 pharmacophore.

The outcomes presented in the tables above demonstrate the powerful utility of pharmacophores in the identification of novel active substances.

4.1 Virtual Screening Methods in LigandScout

This section will examine the functionalities for virtual screening offered by the different LigandScout versions, alongside the approach of conducting the experiments.

LigandScout 4.5

In general, virtual screening in LigandScout 4.5 operates in accordance with the description provided in Section 3.2.2. LigandScout 4.5 features a fully developed graphical user interface (GUI), thereby offering the capacity to directly visualize and edit pharmacophores and virtual screening results within the software environment. Nevertheless, for the purpose of facilitating a more comprehensive comparative analysis with the new LigandScout version, screenings have been carried out by utilizing the iScreen tool within the command-line interface (CMD).

LigandScout 5

The updated implementation of the software employed for experimental investigations, a LigandScout 5 pre-alpha version, lacks a finalized GUI, necessitating the utilization of the iScreen tool through command-line interface. The basic workflow of the virtual screening is corresponds to the method described in Section 3.2.2. An important alteration from the previous LigandScout version is the transition from ldb to ldb2 file format for the databases applied for screening. For that LigandScout 5 introduces an integrated tool called ldbupgrader, facilitating the direct conversion of existing ldb files into the ldb2 format.

The most significant adjustment, as discussed at the beginning of this chapter, centers around the G3PS algorithm. Within the introduction of the novel alignment algorithm, LigandScout 5 incorporates several additional options for potential fine-tuning of the virtual screening process, which will be explored in greater depth in section 4.2. It is crucial to note at this point, that LigandScout 5 is not a final version but rather a preview for research purposes.

Hardware used for the experiments

All experiments detailed henceforth were executed using the Microsoft Surface Book 3 with the following hardware and software:

- Processor: Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz, 64-bit operating system , x64-based
- Installed RAM: 32.0 GB
- Software: Windows 10 Enterprise

4.2 Parameterization of LigandScout 5

The aim of the experiments is to find out, to what extent establishing the new virtual screening method and finding the optimal settings within, can lead to improved outcomes that, ideally, enhance the detection of active compounds while minimizing the number of total hits, including decoys. It is important to note, that during the phase of parameter adjustments, it is plausible that the model may initially get softer, resulting in the novel method obtaining a relatively higher quantity of total hits and decoys. This, in turn, can potentially lead to less favorable active-decoy ratios and F1-Scores, prior to finding the optimal parameter configurations.

For illustration purposes, a specific pharmacophore model and dataset was selected for the initial experimentation to determine the optimal parameter sets for virtual screening within the updated LigandScout version. To facilitate a comparative analysis, both the outcomes of the screenings obtained by LigandScout 4.5 and LigandScout 5 were examined. Primary points of comparison encompassed the total hits, actives, decoys, F1-Scores and runtimes.

The following dataset was applied for the preliminary testing phase:

- Active set: PNS-30-Neurotoxic-compds.ldb (30 active compounds)
- Decoy set: PNS-Decoys-PCL.ldb (1251 decoys)
- Pharmacophore Model: NDR-UV-PNS-M9-LB.pml

4.2.1 Findings in LigandScout 4.5

Before engaging in the LigandScout 5 experiments, an initial virtual screening of the NDR-UV-PNS-M9-LB model using LigandScout 4.5 under default settings was conducted. This preliminary step served to measure the performance of the current method and establish a baseline for subsequent comparisons with LigandScout 5.

LS Version	Hits	Actives	Decoys	F1-Score
LS 4.5	24	6	18	0.222

Table 4.4: NDR-UV-PNS-M9-LB virtual screening results within LigandScout 4.5 under default settings.

Observed in Table 4.4, virtual screening of NDR-UV-PNS-M9-LB in LigandScout 4.5 under default settings, yields the discovery of 6 active compounds and 18 decoys, representing an F1-Score of 0.222.

Subsequently, an examination of the adjustable parameters and their potential effects on the outcomes in LigandScout 5 was performed.

4.2.2 Number of Alignments

The Number of alignments (N) parameter determines the quantity of alignment attempts performed. A value of N=50 signifies the utilization of 50 three-point pairs in the search for an optimal alignment. Number of alignments predominantly relate to optimizing the accuracy of the method, rather than refining the pharmacophore model itself. When lowering the N value, the screening process is expected to perform with a relatively reduced accuracy,

while operating faster compared to setting N at a higher value, where the screening algorithm is anticipated to identify matching hits for the associated pharmacophore model more accurately, however, with longer runtimes. The objective is to discover the best settings to identify a greater number of hits, predominantly active hits within a fast parameter set and an accurate setting, leading to an even more precise outcome, while still retaining relatively fast processing times. To investigate this, initial screenings with the NDR-UV-PNS-M9-LB pharmacophore in its default settings for feature tolerances, without applying RMSD thresholds, at various settings for the number of alignments were conducted. Subsequently, these results were compared with the reference values obtained from the LigandScout 4.5 screening, as it does not implement this setting. The ensuing graphs represent the outcomes of these assessments.

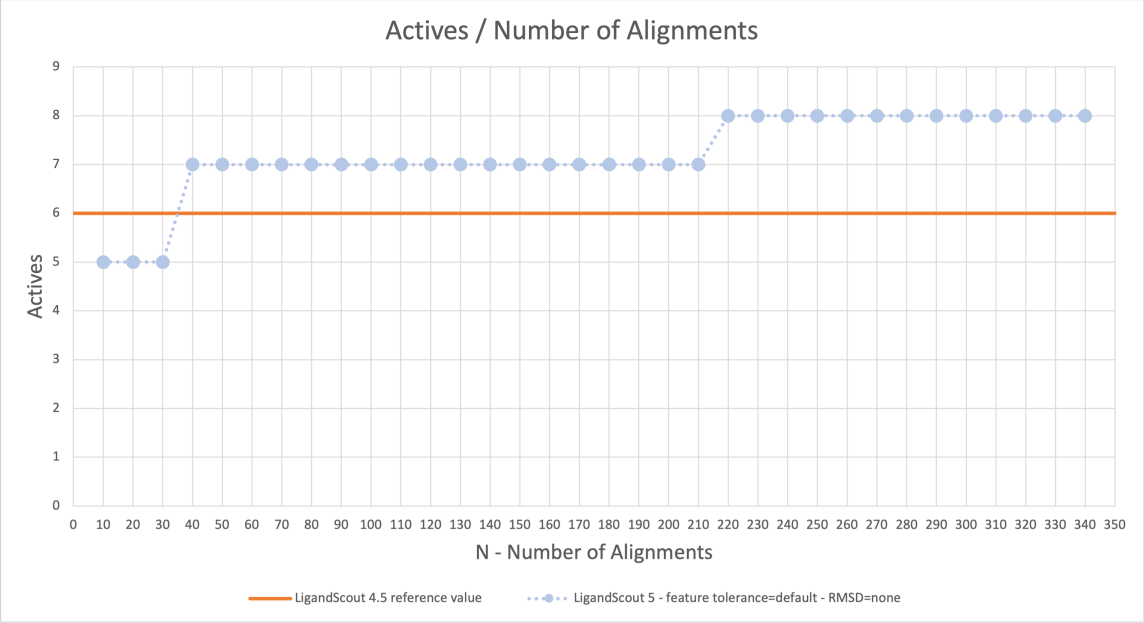


Figure 4.1: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified actives under default settings with varying N values (10-350 in increments of 10).

As illustrated in Figure 4.1, an initial increase in the count of identified active compounds (7) becomes noticeable at N=50. Starting at this point, a greater number of active compounds were found compared to LigandScout 4.5 (6) at all higher N values. A secondary notable increase is observed at N=230, where 8 active compounds were discovered. Raising the number of alignments any further, did not yield more actives.

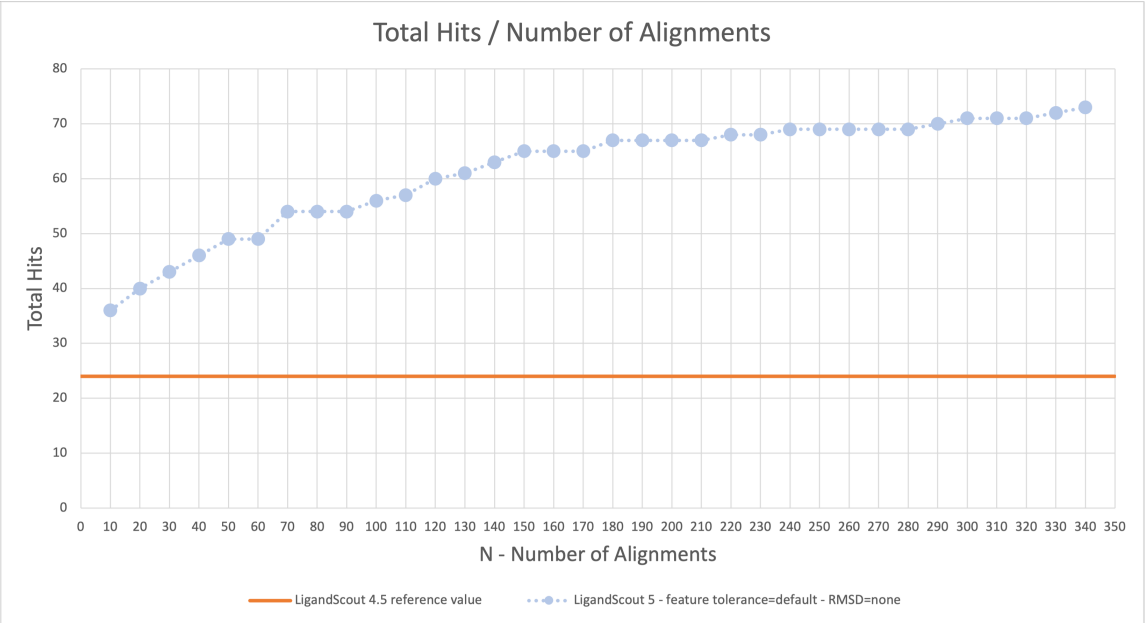


Figure 4.2: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified total hits under default settings with varying N values (10-350 in increments of 10).

Figure 4.2 demonstrates that increasing the number of alignments results in an upswing in the count of total hits and subsequently decoy compounds, while at every stage being above the LigandScout 4.5 reference value.

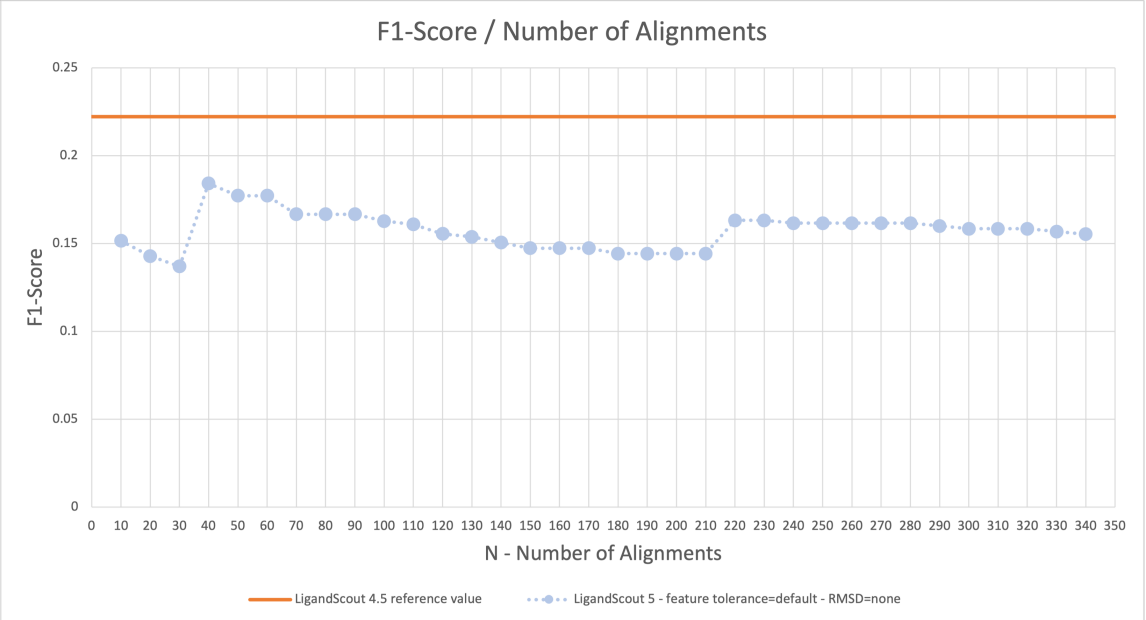


Figure 4.3: NDR-UV-PNS-M9-LB virtual screening results illustrating the F1-Scores under default settings varying N values (10-350 in increments of 10).

Figure 4.3 depicts an increase in the F1-Score at N=50 and N=230 after dropping at higher values of N, while consistently remaining below the LigandScout 4.5 value.

4.2.3 RMSD-Thresholds

The root mean square deviation (RMSD) assesses the mean distance between two aligned molecules and compared to the feature tolerance parameter it offers a relatively simple option for users to filter and enhance outcomes within virtual screening in LigandScout 5 [53].

Fundamentally, the underlying principle of the novel alignment technique is to identify all relevant hits without considering RMSD. For that matter, the default setting in LigandScout 5 is not to apply RMSD thresholds. The current query involves finding out whether establishing an RMSD threshold while screening might help eliminate specific compounds, particularly those falling below a designated threshold value, can ultimately lead to a better active-decoy ratio and better F1-Scores. To acquire an initial understanding of the alteration of the outcomes with varying RMSD thresholds, N was set to a relatively high value of 10,000 and carried out screenings with NDR-UV-PNS-M9-LB under the default feature tolerances, with varying RMSD thresholds. It's worth noting that LigandScout 4.5 lacks this tool, hence, the subsequent graphs will depict comparisons with the respective default reference values of LigandScout 4.5.

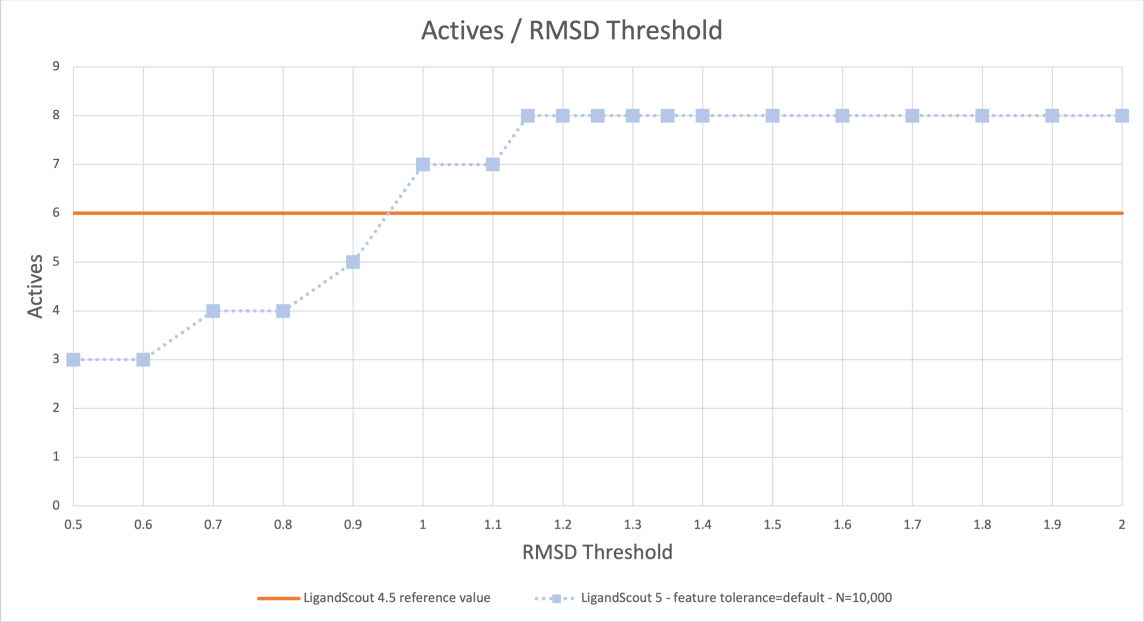


Figure 4.4: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified actives with N=10,000, feature tolerances at default and varying RMSD thresholds (0.10-2.00 in different increments).

In Figure 4.4, early consequences of implementing RMSD thresholds can be observed. Commencing with a threshold value of 1.00, LigandScout 5 discovered more active compounds (7) than LigandScout 4.5 (6). Setting the threshold to 1.15, the new method identified 8 active compounds, with no further increases beyond this point.

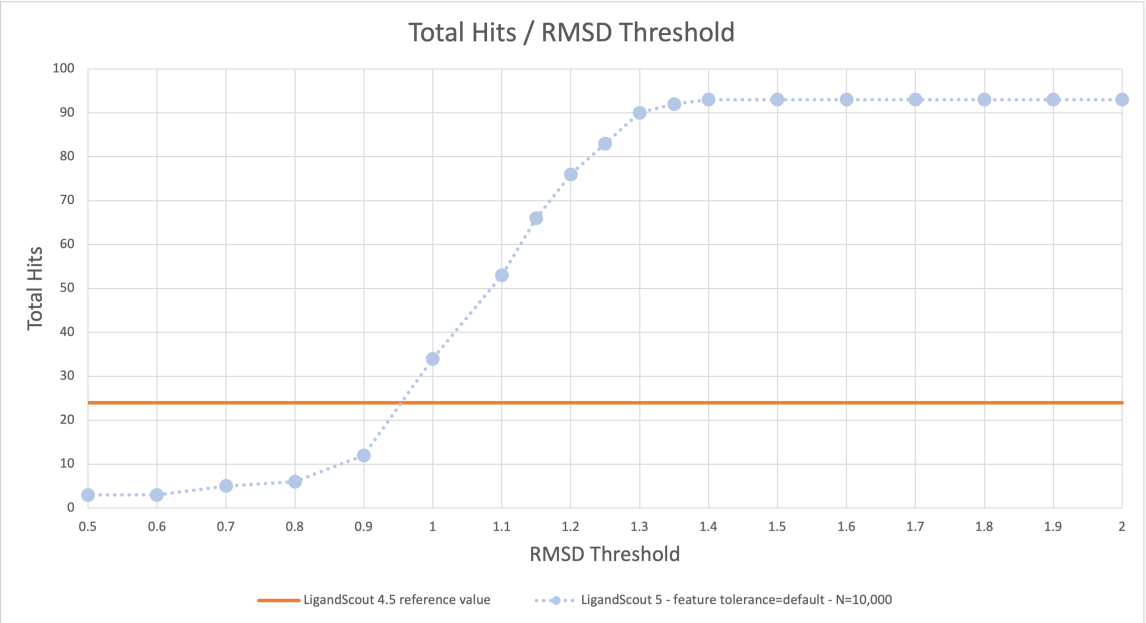


Figure 4.5: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified total hits with N=10,000, feature tolerances at default and varying RMSD thresholds (0.10-2.00 in different increments).

Figure 4.5 illustrates that, in LigandScout 5, when employing RMSD thresholds ranging from 0.10 to 1.00, there was a notable reduction in the quantity of total hits and decoys, compared to LigandScout 4.5. Nonetheless, the curve displays a steep climb at the thresholds 1.10 to 1.40, subsequently flattening again, when the potential maximum of hits is reached.

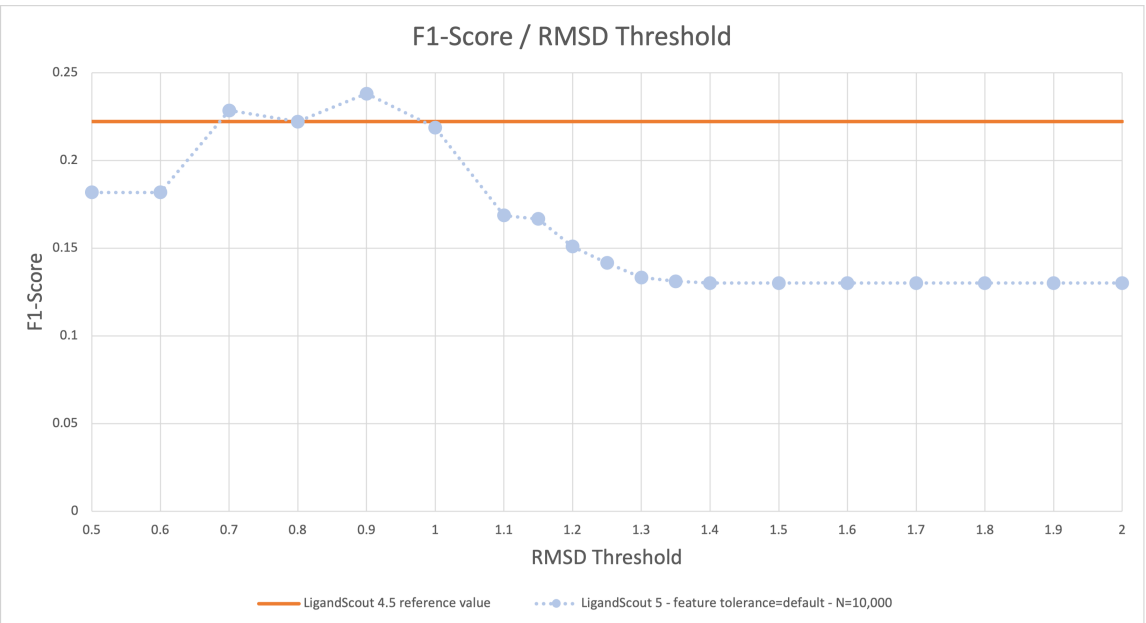


Figure 4.6: NDR-UV-PNS-M9-LB virtual screening results illustrating the $F1$-Scores with N=10,000, feature tolerances at default and varying RMSD thresholds (0.10-2.00 in different increments).

Figure 4.6 reveals that RMSD thresholds of 0.70 and 0.90, resulted in higher F1-Scores than the reference value of LigandScout 4.5. This finding is followed by a subsequent decrease, as a result of the increase in the number of total hits and decoys observed beyond this threshold, as demonstrated in Figure 4.8. The improved F1-Scores can be attributed to the fact, that despite fewer active hits were observed at lower RMSD values, a substantial reduction of decoys lead to better active-decoy ratios in relation.

4.2.4 Feature Tolerances

Feature Tolerance Spheres define the radius within which a pharmacophoric feature must be located in order to be suitable as matching to the pharmacophore model. Tolerance values can be adjusted to larger or smaller sizes, either through the LigandScout 4.5 graphical user interface (GUI) or by directly modifying the parameters within the file. In essence, the concept behind altering feature tolerances is to establish the best sphere sizes for refining the pharmacophore model to a point where it exclusively identifies fitting molecules at the most possible quantity. Setting these spheres either excessively small or overly large may lead to the omission of relevant active compounds or result in an unrestricted number of hits that do not properly fit the model. Up to this point, examinations have incorporated the default feature tolerance values of NDR-UV-PNS-M9-LB, which were predominantly configured at 1.50 for the majority of features, as this represents the standard setting within LigandScout. However, specific feature tolerances have been adjusted during prior refinement procedures. The practicability to create a stricter model by generally adjusting all feature tolerance spheres using the new alignment method shall be investigated. Ideally, this approach may yield an increased number of identified active compounds while finding less decoys. For the purposes of illustration, a considerably high value of N at 10,000 was selected, enabling us to extract the maximum information for each feature tolerance setting.

As it is possible to configure the feature tolerances both within LigandScout 4.5 and directly through the files, it was possible to conduct testing with varying feature tolerances in both LigandScout versions. This allowed for a direct comparison of this parameter, the results of which will be presented in the upcoming graphs showcasing the screening outcomes.

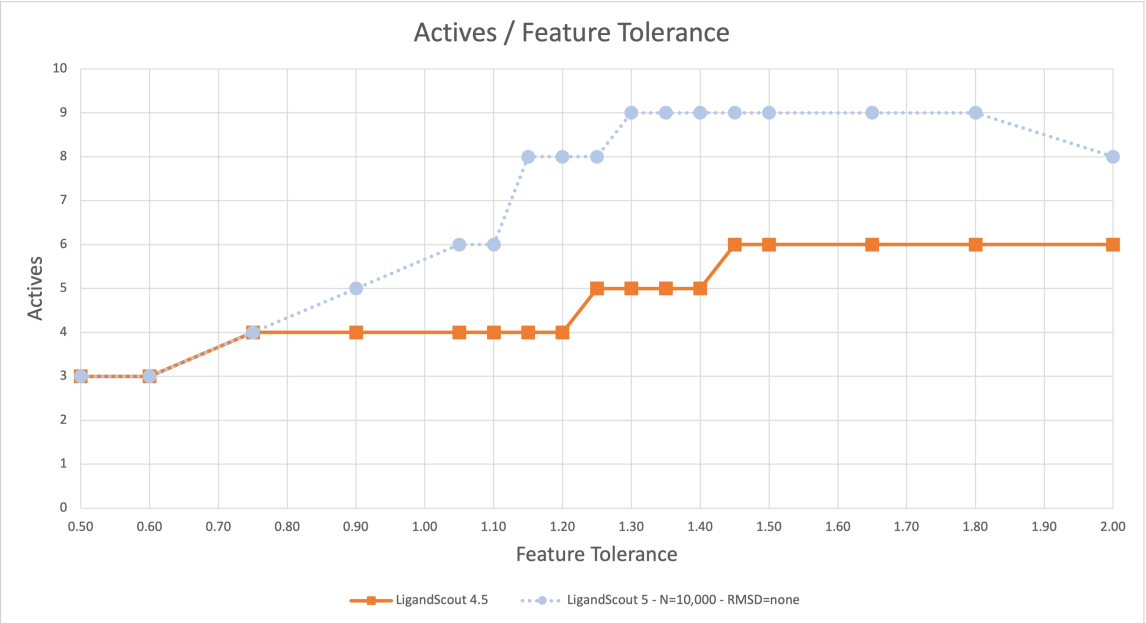


Figure 4.7: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified actives with N=10,000 and varying feature tolerances (0.50-2.00 in different increments).

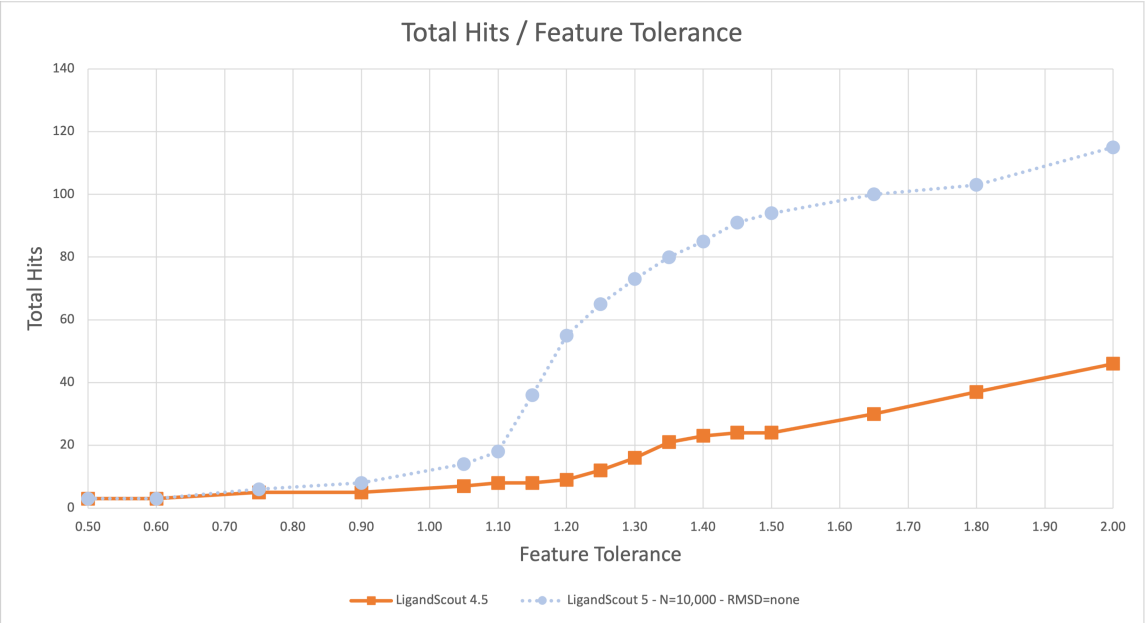


Figure 4.8: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified total hits with N=10,000 and varying feature tolerances (0.50-2.00 in different increments).

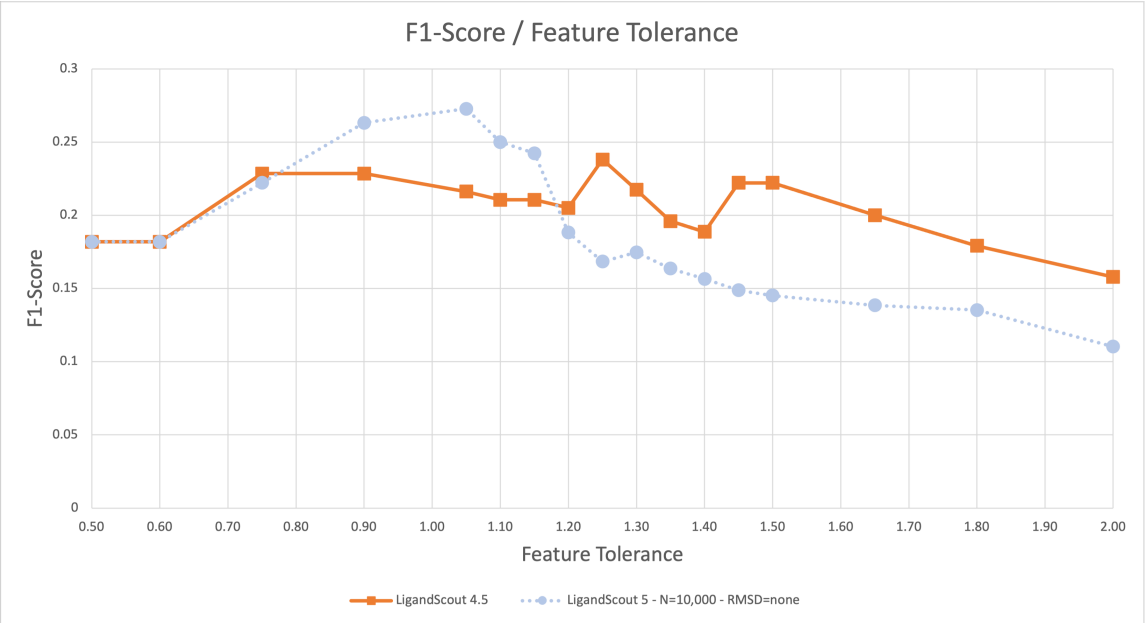


Figure 4.9: NDR-UV-PNS-M9-LB virtual screening results illustrating the F1-Scores with N=10,000 and varying feature tolerances (0.50-2.00 in different increments).

Figure 4.7 highlights that, with number of alignments set rather high at 10,000 in LigandScout 5, feature tolerances 0.90 and above, consistently identified a greater number of active compounds compared to LigandScout 4.5. Notably, at feature tolerance=1.15, 8 active compounds were discovered, further increasing to 9 at feature tolerance=1.30. The trend continues at higher feature tolerances, where the constant maximum of 9 active hits were detected, With a minor deviation observed at 2.00, which can most probably be attributed to the algorithm´s pursuit of RMSD optimization, among other factors, after discovering matches. Consequently, it is possible, that certain hits are discarded due to misalignment of directional vectors. Furthermore, with the use of larger tolerance spheres, the number of available options for features to fit increases, making it progressively more challenging to identify all of the optimal matches. However, Figures 4.8 and 4.9 reveal that, as the N setting increases, the total number of hits and decoys rises, resulting in the F1-Scores falling down. A notable observation is highlighted in 4.9 with feature tolerances 0.90 to 1.15 revealing higher F1-Scores than LigandScout 4.5.

Wir suchen einen Match, der nach dem match schon noch auf RMSD optimieren und mit mehr Threshold kanns sein, dass man mit den Richtungsvektoren rausfliegt

4.2.5 Brief Conclusion of First Tests

For an enhanced overview, tables summarizing the encouraging outcomes of the observed parameters have been compiled herein.

LS Version	N	F	R	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	x	24	6	18	0.222
LS 5	50	-	-	46	7	39	0.184
LS 5	230	-	-	68	8	60	0.163
LS 5	250	-	-	69	8	61	0.161
LS 5	300	-	-	70	8	62	0.160

Table 4.5: NDR-UV-PNS-M9-LB best virtual screening results for number of alignments parameter. N=number of alignments; F=feature tolerance; R=RMSD threshold; -=default setting; x=setting not available.

As can be taken from the graphical representations in Section 4.2.2 and Table 4.5, particular settings stood out regarding the number of alignments parameter. When N was set to 50, 7 active compounds alongside 39 decoys were identified, which represents an increase of 1 active compared to LigandScout 4.5 (6). Furthermore, with N at 230, 250 and 300 8 active compounds along 60, 61, and 62 decoys were discovered, marking an improvement of 2 actives compared to the previous version. Besides the faster N setting (50), given the minimal differences among the higher N settings (230, 250, 300) concerning runtimes, the quantity of total hits and decoys identified, along with the resulting F1-Scores, all these configurations will be included into subsequent experiments. This approach will enable to investigate the potential benefits of raising the number of alignments to achieve a more precise parameter set, with the aim of consistently improving the outcomes.

LS Version	N	F	R	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	x	24	6	18	0.222
LS 5	10,000	-	-	93	8	85	0.130
LS 5	10,000	1.15	-	36	8	28	0.242
LS 5	10,000	1.30	-	73	9	64	0.175
LS 5	10,000	1.50	-	94	9	85	0.145

Table 4.6: NDR-UV-PNS-M9-LB best virtual screening results for feature tolerance parameter. N=number of alignments; F=feature tolerance; R=RMSD threshold; -=default setting; x=setting not available.

Regarding the feature tolerance settings, seen in Table 4.6, the analysis revealed that, at a relatively high N value of 10,000, a maximum of 9 active hits were detected with setting all feature tolerances to 1.50, marking an improvement of 3 to LigandScout 4.5 (6 actives), and 1 to the pharmacophore in its default state (8 actives), however this setting also leads to the observation of a relatively substantial number of decoys (85). Remarkably, with reduced feature tolerances, more active compounds have been identified compared to the LigandScout 4.5 version as well, while simultaneously yielding fewer decoys than with the larger tolerance values. Specifically, feature tolerances 1.15 lead to the discovery of 8 active compounds alongside only 28 decoys, and tolerances 1.30 revealed 9 active compounds with 64 decoys.

LS Version	N	F	R	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	x	24	6	18	0.222
LS 5	10,000	-	1.15	66	8	58	0.167

Table 4.7: NDR-UV-PNS-M9-LB best screening results for RMSD Threshold parameter. N=number of alignments; F=feature tolerance; R=RMSD threshold; -=default setting; x=setting not available.

Given the functionality of the RMSD thresholds, it is rational to observe that, with the default feature tolerance settings, no more active compounds can be identified than the specified tolerance spheres permit, as demonstrated by the detection of a maximum of 8 active compounds in 4.4. In consideration of the parameter itself, an RMSD threshold of 1.15 came forth as the most promising configuration, identifying 8 active compounds among 58 decoys (F1-Score=0.167), as seen in 4.7. Reducing the thresholds constrained the count of active hits, while raising the values exclusively resulting in the recognition of additional decoys, consequently causing a reduction in the F1-Score. The significance of the RMSD threshold parameter is anticipated to become more apparent once the optimal settings for the number of alignments and feature tolerances have been determined.

4.3 Optimization of the Parameters

The previous section has provided a preliminary insight into which settings for each of the three parameters appeared most promising when considered individually. This section will focus on merging these optimal settings to determine the most effective parameter combinations for the new virtual screening method in LigandScout 5. To achieve this, the best number of alignment settings will be weighed up with varying feature tolerances. Subsequently, attempting to enhance the outcomes by applying RMSD thresholds to the discovered parameter sets.

4.3.1 Combining Number of Alignments and Feature Tolerances

Section 4.2 revealed, that the number of alignments parameter values of 50, 230, 250, and 300 were particularly intersting. As a result, these settings were defined as constants in the following experiments and reexamined different feature tolerances in combination with them.

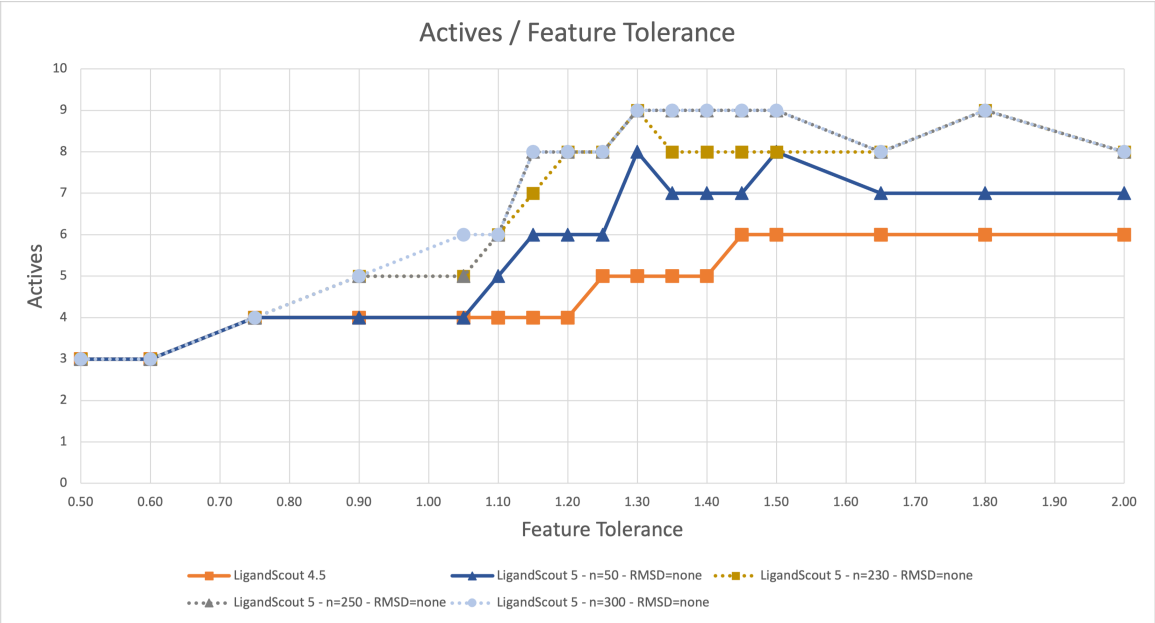


Figure 4.10: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified actives with the best N settings and varying feature tolerances (0.50-2.00 in different increments).

In Figure 4.10, when considering an N value of 50, two specific feature tolerances stand out. Firstly, feature tolerance 1.15, displaying an initial rise in the count of active compounds, reaching 7. Subsequently, the highest peak occurs at feature tolerance 1.30, resulting in the discovery of 8 active compounds. N at 230 shows an initial upswing in active compounds at feature tolerance 1.15, where 8 active compounds were detected, continuing to reach a maximum of 9 actives identified at feature tolerance 1.30. Interestingly, the curves for the number of alignments setting at 250, and 300 are overlapping, indicating identical outcomes across these settings. For these configurations peaks are perceptible at the same values for the feature tolerances at 1.15, where 8 active compounds and 1.30, obtaining 9 actives.

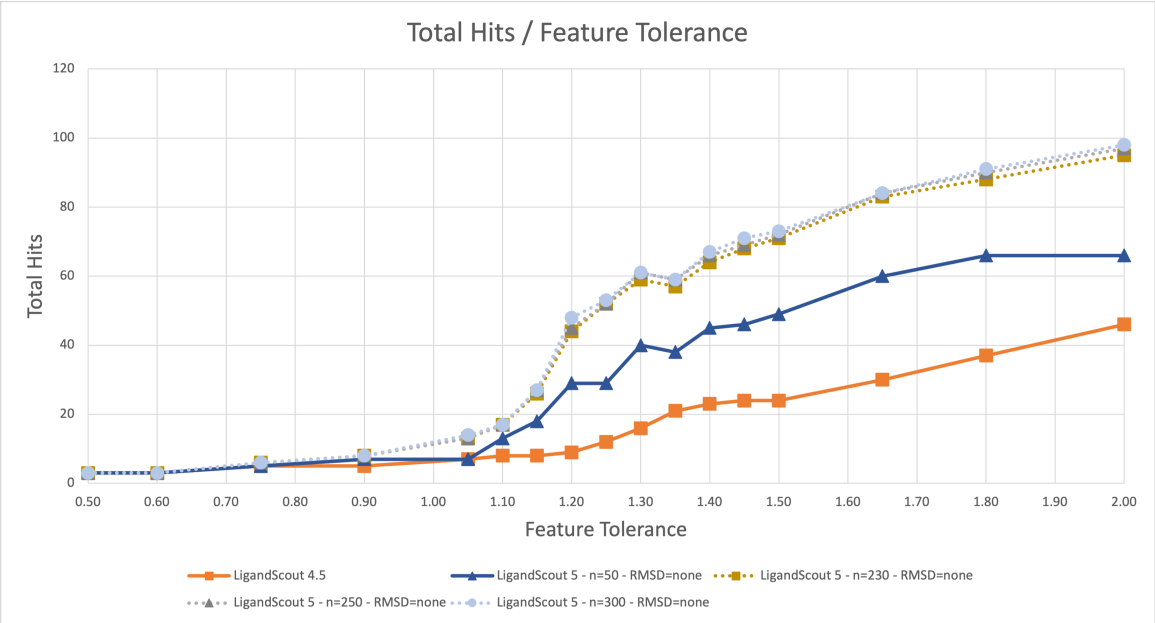


Figure 4.11: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified total hits with the best N settings and varying feature tolerances (0.50-2.00 in different increments).

The curves in Figure 4.11 reveal a moderate rise in total hits and decoy compounds, within feature tolerances ranging from 0.50 to 1.00. Eventually, a more rapid ascent indicates that larger feature tolerances correspond to the detection of relatively higher quantities of total hits and decoys. The results for N=230, 250, 300 are almost overlapping, while N=50 identified less total hits. However all of the LigandScout 5 settings displayed finding more total hits than LigandScout 4.5 with feature tolerances 1.05 and higher.

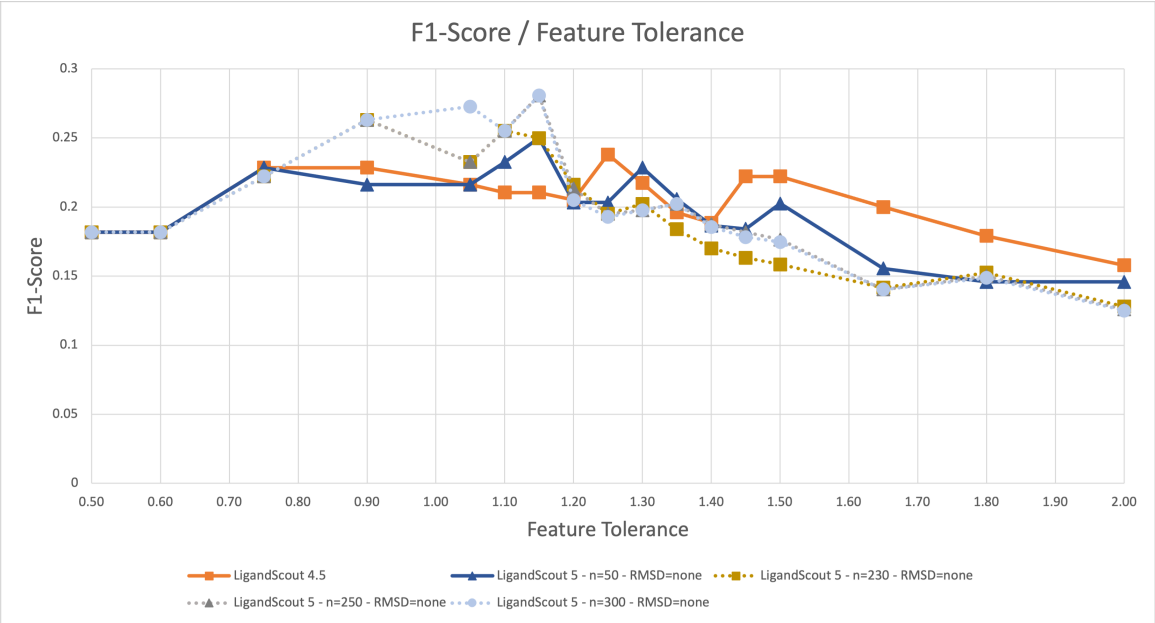


Figure 4.12: NDR-UV-PNS-M9-LB virtual screening results illustrating the F1-Scores with the best N settings and varying feature tolerances (0.50-2.00 in different increments).

Upon initial observation, Figure 4.12 may appear a bit disorienting. The frequent fluctuations in the number of active compounds and decoys retrieved, specifically within the feature tolerances of 0.90 to 1.30, contribute to numerous alterations in the F1-Scores, being partly better than the LigandScout 4.5 values. At higher values they begin worsening again, primarily due to the fact that, as depicted in Figures 4.10 and 4.11, only an increase in the number of decoy compounds is observed alongside the already identified active compounds within higher feature tolerances. A more thorough analysis of the F1-Scores is deferred to a later Section 4.3.2, as it is more suitable to address this matter when both the number of alignment and feature tolerance parameters are set and held constant.

Prospects after the conducted tests

LS Version	N	F	R	Hits	Actives	Decoys	F1-Score
LS 4.5	x	default	x	24	6	18	0,222
LS 5	50	1.30	-	40	8	32	0.229
LS 5	230	1.30	-	59	9	50	0.202
LS 5	250	1.30	-	61	9	52	0.198
LS 5	300	1.30	-	61	9	52	0.198

Table 4.8: NDR-UV-PNS-M9-LB best virtual screening results for number of alignments paired with the best feature tolerance setting. N=number of alignments; F=feature tolerance; R=RMSD threshold; -=default setting; x=setting not available.

As evident from the experimental data presented in Section 4.3.1 and indicated in the associated Table 4.8, a feature tolerance of 1.30 emerges as the most promising setting for this specific parameter. Regarding the number of alignments, with N=50 the lowest value

identifying more active hits (8) than LigandScout 4.5 (6 actives) was discovered. With N increased to 230, 250 and 300 the number of active hits further expanded to 9, while in the cases of N=250 and N=300, a slightly higher quantity of total hits were identified, with only minimal alterations observed in the F1-Score. Given the primary aim is determining a fast and a more accurate parameter set, with the focus on maximizing the possible outcomes, subsequent investigations in this thesis will concentrate on N values of 50 and 300.

4.3.2 Combining Promising Feature Tolerances and Number of Alignments with RMSD-Thresholds

Now that the most promising parameters for the number of alignments (50 and 300) alongside the most advantageous feature tolerance value (1.30) have been established, upcoming investigations aim to determine the extent to which the outcomes, specifically the active-decoy ratio and the associated F1-Score, can be further enhanced by applying RMSD thresholds.

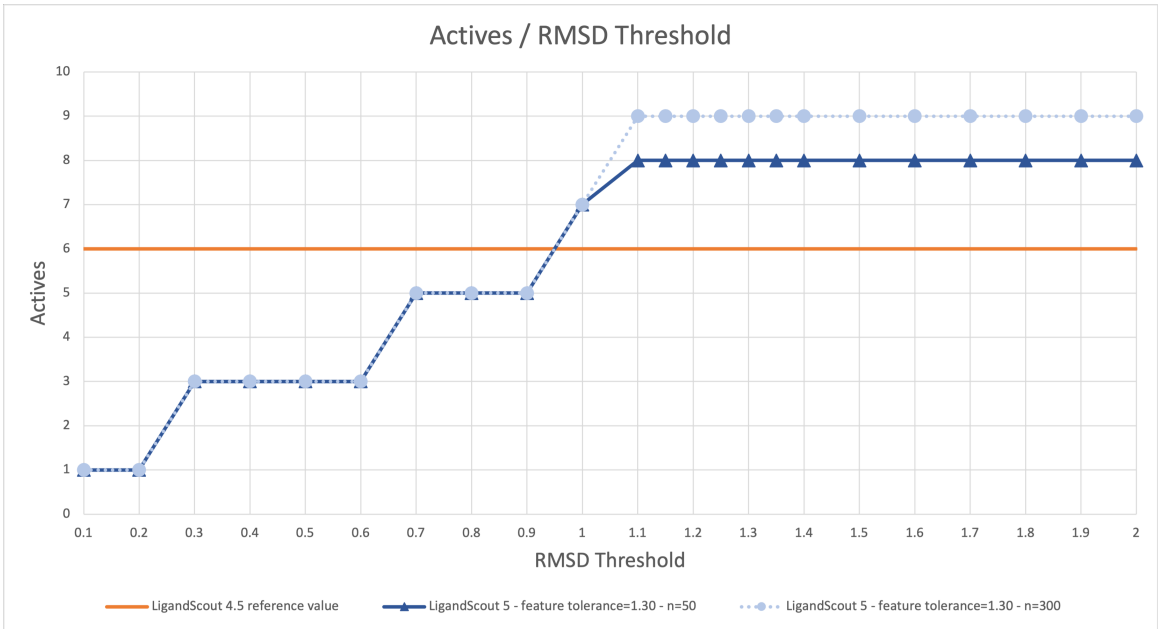


Figure 4.13: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified actives with the best N settings (50, 300), feature tolerances at 1.30 and varying RMSD thresholds (0.10-2.00 in different increments).

In Figure 4.13, two notable shifts are visible in the number of active compounds identified. When N was set to 50, 8 active compounds were detected with an RMSD threshold of 1.10, with N set to 300, 9 actives were retrieved at this stage. Beyond this point, the quantity of discovered active hits was constant.

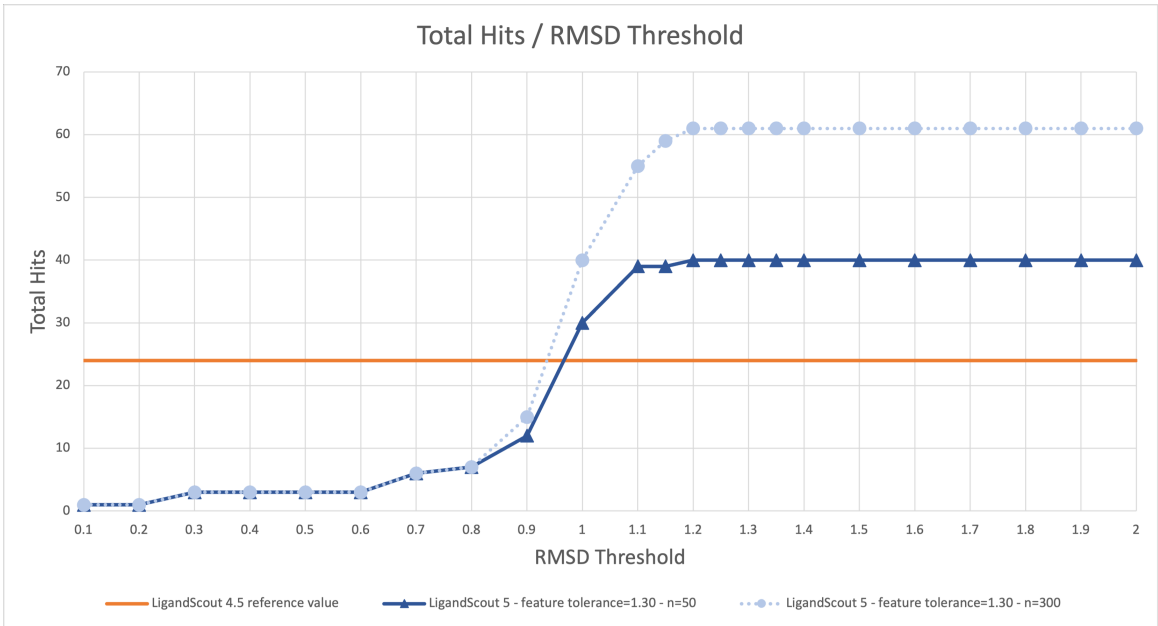


Figure 4.14: NDR-UV-PNS-M9-LB virtual screening results illustrating the identified total hits with the best N settings (50, 300), feature tolerances at 1.30 and varying RMSD thresholds (0.10-2.00 in different increments).

Figure 4.14 indicates a steep progression in the number of total hits starting with an RMSD threshold of 0.90 reaching a maximum at 1.20 and displaying constant outcomes at even higher RMSD thresholds. Comparing the higher number of alignment setting, it is evident that N=300 reveals a curve slightly above the lower N towards thresholds 0.90 and higher indicating a larger quantity of total hits identified. Starting with RMSD threshold 1.00 LigandScout 5 constantly retrieved more total hits than LigandScout 4.5, with N=50 permanently identifying fewer hits than the more accurate setting, N=300.

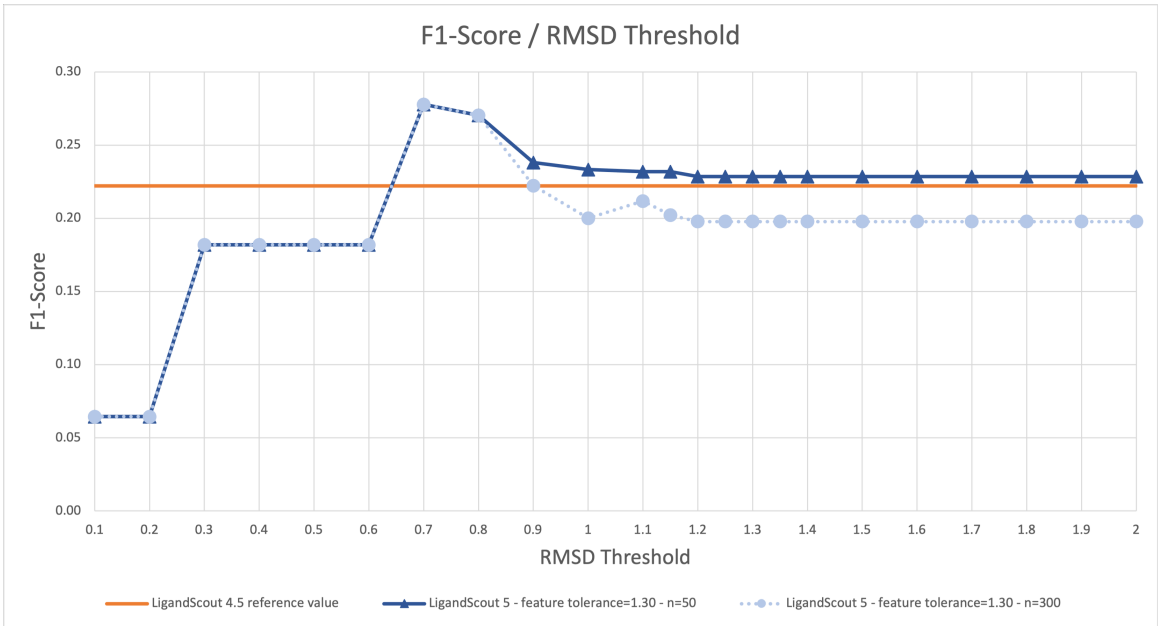


Figure 4.15: NDR-UV-PNS-M9-LB virtual screening results illustrating the F1-Scores with the best N settings (50, 300), feature tolerances at 1.30 and varying RMSD thresholds (0.10-2.00 in different increments).

In Figure 4.15, it is apparent that, with the established settings an improvement of the F1-Scores (compared to the LigandScout 4.5 reference, which lacks the function for RMSD thresholds) can be seen for N=50 throughout RMSD thresholds ranging from 0.70 to 2.00. The accurate setting N=300 reaches a maximum at RMSD value 0.70 and 0.90 before dropping just below the LigandScout 4.5 reference value for the remaining settings.

4.3.3 Conclusion of Optimized Parameters

The subsequent configurations and outcomes have been identified as the most promising after screening the NDR-UV-PNS-M9-LB model against the PNS-30-Neurotoxic-compds and PNS-Decoys-PCL databases.

LS Version	N	F	R	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	x	24	6	18	0.222
LS 5	50	1.30	1.10	39	8	31	0.232
LS 5	300	1.30	1.10	55	9	46	0.212
LS 5	50	1.30	-	40	8	32	0.229
LS 5	300	1.30	-	61	9	52	0.198

Table 4.9: NDR-UV-PNS-M9-LB virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; -=default setting; x=setting not available.

The experiments have demonstrated that by generally reducing the feature tolerances from 1.50 to 1.30 within LigandScout 5, it was possible to identify a greater quantity of active hits compared to LigandScout 4.5, indicating the improved accuracy of the new virtual screening

method. Especially, with feature tolerances of 1.30 and N=50, 8 active hits were obtained, 32 decoys, and a F1-Score of 0.229, outperforming LigandScout 4.5, which discovered 6 active hits, 18 decoys, resulting in a F1-Score of 0.222. The accurate setting N=300, further improved the results, retrieving 9 active hits and 52 decoys, despite slightly reducing the active-decoy ratio, leading to a F1-Score of 0.198. The findings in Table 4.9 also indicate that optimizing RMSD thresholds can enhance outcomes by maintaining the same number of active hits while reducing the number of decoys. For instance, with N=300, feature tolerances at 1.30, and an RMSD threshold of 1.10, a improved F1-Score of 0.212 was gained, alongside 9 active hits and only 46 decoys, in contrast to not applying RMSD thresholds, which resulted in a F1-Score of 0.198.

Nonetheless, throughout this experiments, it was found that the direct reduction of feature tolerances had a more substantial impact on the outcomes, than adjustments to RMSD thresholds. However, the potential for RMSD thresholds to acquire enhanced results in roughly filtering out decoys or inactives is not excluded, especially for researchers who prefer not to delve deeper into the modification of feature tolerances.

Based on the extensive tests and the resulting outcomes, following parameter sets are proposed for fast and and accurate settings for further investigations within LigandScout 5 in Table 4.10.

Setting	Number of alignments	Feature tolerance
fast	50	1.30
accurate	300	1.30

Table 4.10: Proposed parameter sets for further virtual screening investigations in LigandScout 5.

For the benefit of interested readers, a additional runtime comparison between LigandScout 4.5 and LigandScout 5, with a range of configurations for the number of alignment parameter, is offered.

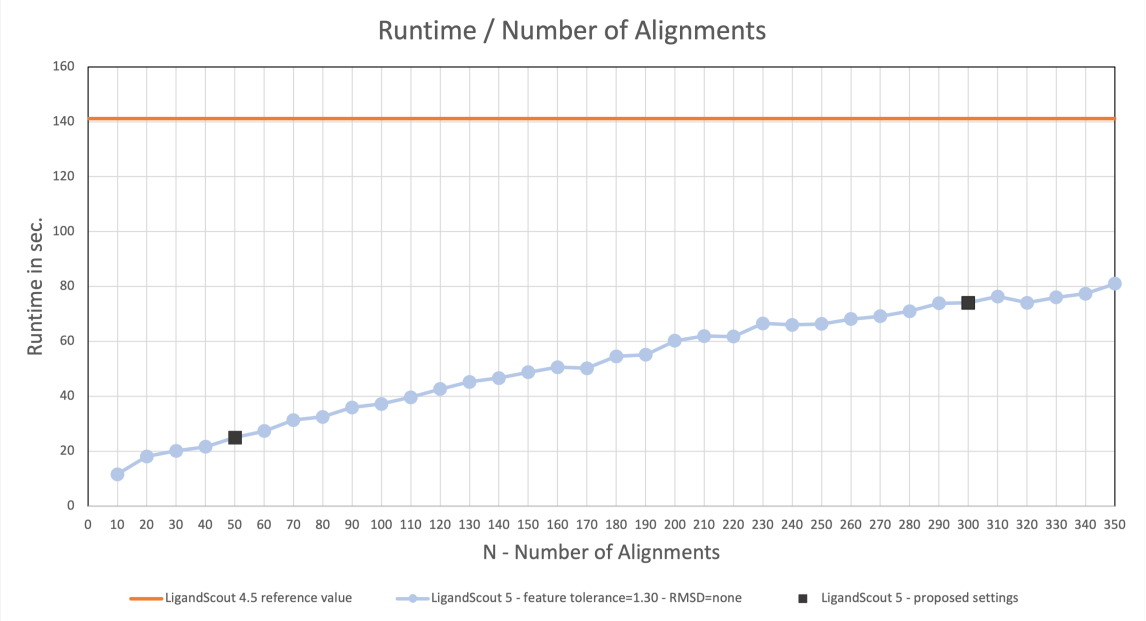


Figure 4.16: NDR-UV-PNS-M9-LB virtual screening results illustrating the runtimes with feature tolerance at 1.30 and varying N settings (10-350 in increments of 10).

To ensure dependable data, a total of six measurements were conducted for each number of alignments value. The respective initial run was designated as a warm-up run. The average value derived from the subsequent five measurements was employed for analysis. As can be seen in Figure 4.16, it is evident that as the value of parameter N increases, the runtime also rises, yet consistently outperforming LigandScout 4.5.

4.4 Proposed Settings on Different Datasets

Given that preliminary experiments for determining the optimal parameter set for virtual screening in LigandScout 5 exclusively employed a single dataset alongside a specific pharmacophore model, the ongoing objective is to assess the applicability of these settings to other models and datasets. The aim is to determine whether the findings derived from the NDR-UV-PNS-M9-LB model hold true in a broader condition.

To address this matter, a comparative analysis involving the proposed parameter sets, presented in Table 4.10, applied to various pharmacophore models and datasets was performed.

This included virtual screening with LigandScout 4.5, as well as direct assessments using the default settings for feature tolerances for the respective pharmacophores.

4.4.1 NDR-UV-PNS Models

Initially, screenings were conducted using two additional models, in addition to the NDR-UV-PNS-M9-LB model, on the same dataset to enable a direct comparison.

Screens were performed with the following dataset:

Databases:

- PNS-30-Neurotoxic-compds.ldb (30 actives)
- PNS-Decoys-PCL.ldb (1251 decoys)

Pharmacophore models:

- NDR-UV-PNS-M9-LB.pml
- NDR-UV-PNS-M18-LB.pml
- NDR-UV-PNS-snibs-LB.pml

NDR-UV-PNS-M9-LB

LS Version	N	F	T	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	141.20	24	6	18	0.222
LS 5	50	-	28.72	46	7	39	0.184
LS 5	300	-	59.44	70	8	62	0.160
LS 5	50	1.30	24.45	40	8	32	0.229
LS 5	300	1.30	65.77	61	9	52	0.198

Table 4.11: NDR-UV-PNS-M9-LB virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-PNS-M18-LB

LS Version	N	F	T	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	59.95	27	4	23	0.140
LS 5	50	-	16.99	36	5	31	0.152
LS 5	300	-	40.61	39	5	34	0.145
LS 5	50	1.30	15.67	35	5	30	0.154
LS 5	300	1.30	34.91	36	5	31	0.152

Table 4.12: NDR-UV-PNS-M9-LB virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-PNS-snibs-LB

LS Version	N	F	T	Hits	Actives	Decoys	F1-Score
LS 4.5	x	-	30.96	13	3	10	0.140
LS 5	50	-	8.01	12	3	9	0.143
LS 5	300	-	16.97	14	5	9	0.227
LS 5	50	1.30	9.07	14	3	11	0.136
LS 5	300	1.30	14.92	14	3	11	0.136

Table 4.13: NDR-UV-PNS-snibs-LB virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

4.4.2 NDR-UV-WP1 Models

In order to find out how the settings affect virtual screening within a different dataset, tests with additional databases and pharmacophore models were carried out:

Databases:

- NDR-WP1-actives-272005.ldb (63 actives)
- NDR-WP1-inactives-272005.ldb (29 inactives)

Pharmacophore models:

- NDR-UV-WP1-M44-172009.pml
- NDR-UV-WP1-M50-172009.pml
- NDR-UV-WP1-M53-172009.pml
- NDR-UV-WP1-M54-172009.pml
- NDR-UV-WP1-M55-172009.pml
- NDR-UV-WP1-M58-172009.pml

NDR-UV-WP1-M44-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	18.63	9	9	0	0.250
LS 5	50	-	3.78	9	8	1	0.222
LS 5	300	-	5.03	12	11	1	0.293
LS 5	50	1.30	5.09	9	8	1	0.222
LS 5	300	1.30	6.49	12	11	1	0.293

Table 4.14: NDR-UV-WP1-M44-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-WP1-M50-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	27.17	11	10	1	0.270
LS 5	50	-	5.32	11	10	1	0.270
LS 5	300	-	7.52	14	13	1	0.338
LS 5	50	1.30	8.12	10	10	0	0.274
LS 5	300	1.30	10.48	12	11	1	0.293

Table 4.15: NDR-UV-WP1-M50-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-WP1-M53-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	57.85	8	8	0	0.225
LS 5	50	-	8.07	11	11	0	0.297
LS 5	300	-	18.82	12	11	1	0.293
LS 5	50	1.30	14.96	10	10	0	0.274
LS 5	300	1.30	20.68	10	10	0	0.274

Table 4.16: NDR-UV-WP1-M53-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-WP1-M54-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	32.92	5	5	0	0.147
LS 5	50	-	10.27	8	6	2	0.169
LS 5	300	-	14.36	8	6	2	0.169
LS 5	50	1.30	11.39	10	6	4	0.164
LS 5	300	1.30	18.58	12	7	5	0.187

Table 4.17: NDR-UV-WP1-M54-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-WP1-M55-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	39.73	6	5	1	0.145
LS 5	50	-	5.96	8	6	2	0.169
LS 5	300	-	8.61	8	6	2	0.169
LS 5	50	1.30	8.34	9	8	1	0.222
LS 5	300	1.30	10.84	13	10	3	0.263

Table 4.18: NDR-UV-WP1-M55-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

NDR-UV-WP1-M58-172009

LS Version	N	F	T	Hits	Actives	Inactives	F1-Score
LS 4.5	x	-	137.78	4	4	0	0.119
LS 5	50	-	8.06	3	3	0	0.091
LS 5	300	-	13.61	4	4	0	0.119
LS 5	50	1.30	10.04	4	3	1	0.222
LS 5	300	1.30	16.74	5	4	1	0.118

Table 4.19: NDR-UV-WP1-M58-172009 virtual screening results for best parameter sets. N=number of alignments; F=feature tolerance; R=RMSD threshold; T=runtime in seconds; -=default setting; x=setting not available.

4.4.3 Conclusion of Tests with Different Datasets

This section, evaluated the best parameter sets over various pharmacophore models. The analyses encompassed different datasets, revealing that the novel virtual screening method generally demonstrated superior accuracy in identifying suitable hits, particularly for active hits, for each model.

Regarding the number of alignments, it was observed that N=50 and N=300 proved to be reliable default settings, with the exception of models NDR-UV-WP1-M44-172009 and NDR-UV-WP1-M58-172009, which exhibited a reduction to 8 identified actives at N=50 compared to LigandScout 4.5 (9 actives).

Within the first dataset, in Section 4.4.1, the findings highlighted that it was possible to enhance discovered active hits, active-decoy ratios and consequently the F1-Scores for all models. Further improving results for NDR-UV-PNS-M9-LB and NDR-UV-PNS-M18-LB by reducing all feature tolerances to 1.30. Conversely, in the case of the NDR-UV-PNS-snibs-LB model, improved results were found only for the default tolerance settings. Specifically, when employing the fast N=50 configuration, it achieved an equivalent number of active hits as LigandScout 4.5 alongside a better active-decoy ratio and F1-Score. The accurate N=300 setting, successfully identified an even greater number of active compounds further optimizing the outcomes. The second dataset, in section 4.4.2 revealed, that modifying feature tolerances to a value of 1.30 was only beneficial for NDR-UV-WP1-M54-172009 and NDR-UV-WP1-M55-172009. All other models performed more effectively when maintaining their default tolerance settings. NDR-UV-WP1-M44-172009 showed a minor reduction in active

hits (8) at N=50 but demonstrated a significant increase (11 actives) at N=300, surpassing the virtual screening results of LigandScout 4.5 (9 actives). In contrast to LigandScout 4.5, NDR-UV-WP1-M50-172009 identified the same quantity of active hits (10) with N=50 but increased the number (13 actives) again with N=300.

A noteworthy exception appeared in the case of NDR-UV-WP1-M58-172009. The Model encountered a marginal reduction in active hits (3) with default feature tolerances at N=50 but identified the same hits as LigandScout 4.5 (4 actives) with N=300. This discovery suggests that, in specific occasions, a thorough reevaluation of the pharmacophore model and their distinct pharmacophoric features may be necessary and potential refinements should be considered to maximise the potential of the new method.

In summary, it is evident that the new virtual screening approach employed in LigandScout 5 demonstrated its superior performance in comparison to the current method. Altering feature tolerances in a general sense did not consistently exceed outcomes across various datasets, likely due to fact that most of the models examined had already undergone substantial refinement with specific features having more precise tolerance setting. Nevertheless, it was illustrated that, for certain models, improvements of the screening results were possible, by downsizing feature tolerances. Furthermore, the potential for even greater enhancements by selectively reducing or modifying feature tolerance spheres for particular models is acknowledged, as the new screening algorithm offers the opportunity to identify suitable hits with even greater precision.

4.5 Discussion

In the context of this thesis, detailed in Chapter 4, the aim was to investigate potential improvements in virtual screening within LigandScout by the implementation of G3PS, a novel alignment algorithm.

The extensive tests, featured in Sections 4.2 and 4.3, revealed that, for the NDR-UV-PNS-M9-LB pharmacophore model, under thorough examination, a general reduction of its respective feature tolerance spheres from 1.50 to 1.30 yielded favorable outcomes. This adjustment led to a significant increase in the discovery of active compounds and active-decoy ratios when compared to the prior LigandScout 4.5 version and also exhibited enhanced performance to the model's default state in LigandScout 5. Investigations also highlighted the potential benefit of using values of 50 for a fast, and 300 for an accurate setting, regarding the number of alignments parameter.

Additionally, the practicability of applying RMSD thresholds was explored in Section 4.3.3 and demonstrated, that they have the potential to lead to superior results in certain cases, as seen in Table 4.9. However, this findings suggested that a direct modification of the pharmacophores feature tolerances provided a more pronounced and favorable impact on the outcomes.

To validate this hypotheses, virtual screening on diverse datasets and pharmacophore models was conducted, as delineated in Section 4.4. After employing fast (N=50) and accurate (N=300) settings in LigandScout 5, the respective results denoted the superiority over the previous LigandScout 4.5 method. Nevertheless, it was not possible to manifest a broad reduction of feature tolerances leading to universal improvements, as it applied only to particular instances.

However, with minor exceptions, improvements over LigandScout 4.5 were discovered in active hits, active-decoy or active-inactive ratios, F1-Scores, and computational runtimes in

the investigated pharmacophore models within default feature tolerance settings in LigandScout 5.

In conclusion, the analysis underlines the distinct enhancement of virtual screening in LigandScout 5, as it consistently identified more hits, especially active hits, aligned with the associated pharmacophore models in their default feature tolerance settings. The predefined number of alignment settings, N=50 and N=300, emerged as respectable defaults for a fast and accurate configuration. While the application of RMSD thresholds as coarse filters may be useful, it is evident that feature tolerance modifications constitute a more effective approach and general reductions of feature tolerance spheres can lead to improved screening outcomes in specific cases.

Nonetheless, the results confirm the advanced precision and accuracy of the new alignment algorithm and suggest the potential for even greater benefits through the targeted modification of individual feature tolerances, capitalizing on the method's opportunity to define stricter models that can identify highly precise hits suitable for corresponding pharmacophore models.

Chapter 5

Conclusion

Pharmacophore-based virtual screening has proven as an indispensable tool in modern drug development. Pharmacophores, serving as abstract models, utilize pharmacophoric features to delineate electrostatic and steric relationships between bioactive molecules and their respective target structures. These models allow the screening of extensive molecular databases to identify potential medicinal candidates [6, 9].

This thesis has elucidated how the integration of a novel alignment algorithm Greedy 3-Point Search (G3PS) can enhance the virtual screening method within an updated version of the software LigandScout. To investigate this, virtual screening experiments were conducted using a preview version of LigandScout 5, spanning a variety of datasets and pharmacophore models. The resultant findings were then compared with the performance of the current LigandScout 4.5 version.

The outcomes of the investigation, as presented in Chapter 4, underscore that the new virtual screening method exhibits superior performance across multiple datasets and pharmacophore models, identifying a greater quantity of accurate hits while also enhancing computational efficiency.

Furthermore, suitable default values for the number of alignment parameter were found, for both a fast and an accurate setting. In the context of RMSD thresholds, the findings indicate their practical utility for a coarse filtering approach. However, the enhanced precision offered by the new screening algorithm offers the potential for even more significant benefits through the deliberate modification of feature tolerance spheres. This allows for the creation of stricter pharmacophore models, enabling the identification of suitable hits with enhanced accuracy.

The evaluations conducted using the preview version of LigandScout 5 provided an initial insight into enhancements for virtual screening using the novel method. Additional assessments, alongside the software's final development phase, will yield a more definitive perspective on optimal parameter configurations and the newfound capabilities.

Appendices

Appendix A

Abstract

A.1 English abstract

Pharmacophore-based virtual screening has established to take an indispensable role in modern drug development. Pharmacophores represent abstract constructs that capture the electrostatic and steric interactions between biologically active molecules and their respective targets, defined by specific pharmacophoric features. With the utility of pharmacophores an exploration of vast molecular databases is made possible, enabling the selective filtration of searched compounds. This strategic filtration, performed in the early stages of drug discovery, can lead to substantially reducing time and expense resources [2, 5, 6, 9, 11].

Conventionally, virtual screening alignment algorithms employed in diverse software packages have primarily concentrated on the minimization of root mean square deviation (RMSD) or the maximization of volumetric overlap through the utilization of Gaussian spheres. However, such alignment strategies do not necessarily align with the fundamental principles of pharmacophores in finding the optimal alignment. Addressing this issue, the introduction of the Greedy 3-Point Search (G3PS) algorithm specifically aims to maximize the count of matching feature pairs, aligning more closely with pharmacophore-based objectives [13].

By incorporating the G3PS algorithm into the LigandScout software, dedicated to pharmacophore modeling and virtual screening, this research endeavors to ascertain whether this novel alignment approach has the potential to enhance the existing virtual screening method.

A.2 Deutsche Zusammenfassung

Pharmakophorbasiertes Virtual Screening hat in der modernen Entwicklung von neuen Wirkstoffen eine unverzichtbare Rolle eingenommen. Pharmakophore sind abstrakte Konstrukte, die elektrostatische und sterische Beziehungen zwischen bioaktiven Molekülen und deren Zielstrukturen beschreiben und durch chemische, pharmakophorische Charakteristika definieren.

Mithilfe von Pharmakophoren ist es möglich, große Moleküldatenbanken nach gewünschten Strukturen zu durchsuchen und zu durchfiltern, um damit die potentiellen Kanditaten für neue Medikamente in der Anfangsphase der Entwicklung enorm zu reduzieren und dabei vor allem Zeit- und Geld Ressourcen einzusparen [2, 5, 6, 9, 11].

Virtual Screening Algorithmen verschiedenster Software Programme waren bisher vorallem darauf fixiert, Alignments zu ermitteln, die entweder eine minimale Abweichung der *Root Mean Square Deviation* (RMSD) oder eine möglichst große volumetrische Überschneidung von Gauß´schen Sphären aufweisen. Diese Idee steht allerdings nicht unbedingt im Einklang mit dem eigentlichen Konzept eines Pharmakophors, das optimale Alignment zu finden. Mit dem Greedy 3-Point Search (G3PS) wurde ein neuer Alignment Algorithmus vorgestellt, der sich genau diesem Problem widmet und sich darauf fokussiert, die maximale Anzahl an tatsächlich passenden Feature Paaren zu finden [13].

Durch die Implementierung von G3PS in die Pharmakophor und Virtual Screening Software LigandScout, soll im Verlauf dieser Arbeit herausgefunden werden, ob die bestehende Virtual Screening Methode mit dem neuen Alignment Algorithmus verbessert werden kann.

Bibliography

[1] D. Giordano, C. Biancaniello, M. A. Argenio, and A. Facchiano. Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel)*, 15(5), 2022. Giordano, Deborah Biancaniello, Carmen Argenio, Maria Antonia Facchiano, Angelo eng Review Switzerland 2022/05/29 Pharmaceuticals (Basel). 2022 May 23;15(5):646. doi: 10.3390/ph15050646.

[2] X. Lin, X. Li, and X. Lin. A review on applications of computational methods in drug screening and design. *Molecules*, 25(6), 2020. 1420-3049 Lin, Xiaoqian Orcid: 0000-0002-4633-1281 Li, Xiu Lin, Xubo Orcid: 0000-0002-4417-3582 21903002/National Natural Science Foundation of China/ Journal Article Review Switzerland 2020/03/22 Molecules. 2020 Mar 18;25(6):1375. doi: 10.3390/molecules25061375.

[3] X. Lu, H. Yang, Y. Chen, Q. Li, S. Y. He, X. Jiang, F. Feng, W. Qu, and H. Sun. The development of pharmacophore modeling: Generation and recent applications in drug discovery. *Curr Pharm Des*, 24(29):3424–3439, 2018. 1873-4286 Lu, Xin Yang, Hongyu Chen, Yao Li, Qi He, Si-Yu Jiang, Xueyang Feng, Feng Qu, Wei Sun, Haopeng Journal Article Research Support, Non-U.S. Gov't Review United Arab Emirates 2018/08/14 Curr Pharm Des. 2018;24(29):3424-3439. doi: 10.2174/1381612824666180810162944.

[4] David Schaller, Dora Šribar, Theresa Noonan, Lihua Deng, Trung Ngoc Nguyen, Szymon Pach, David Machalz, Marcel Bermudez, and Gerhard Wolber. Next generation 3d pharmacophore modeling. *WIREs Computational Molecular Science*, 10(4):e1468, 2020.

[5] Thomas Seidel, Oliver Wieder, Arthur Garon, and Thierry Langer. Applications of the pharmacophore concept in natural product inspired drug design. *Molecular Informatics*, 39(11):2000059, 2020.

[6] Teresa Kaserer, Katharina R. Beck, Muhammad Naveed Akram, Alex Odermatt, and Daniela Schuster. Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20:22799 – 22832, 2015.

[7] T. Langer and G. Wolber. Pharmacophore definition and 3d searches. *Drug Discov Today Technol*, 1(3):203–7, 2004. Langer, T Wolber, G Journal Article England 2004/12/01 Drug Discov Today Technol. 2004 Dec;1(3):203-7. doi: 10.1016/j.ddtec.2004.11.015.

[8] Xiao-Yu Qing, Xiao Yin Lee, Joren De Raeymaecker, Jeremy Tame, Kam Zhang, Marc De Maeyer, and Arnout Voet. Pharmacophore modeling: Advances, limitations, and current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research*, 7:81–92, 11 2014.

[9] Thomas Seidel, Gökhan Ibis, Fabian Bendix, and Gerhard Wolber. Strategies for 3d pharmacophore-based virtual screening. *Drug Discovery Today: Technologies*, 7(4):e221–e228, 2010. 3D Pharmacophore Elucidation and Virtual Screening.

[10] A. B. Gurung, M. A. Ali, J. Lee, M. A. Farah, and K. M. Al-Anazi. An updated review of computer-aided drug design and its application to covid-19. *Biomed Res Int*, 2021:8853056, 2021. Gurung, Arun Bahadur Ali, Mohammad Ajmal Lee, Joongku Farah, Mohammad Abul Al-Anazi, Khalid Mashay eng Review 2021/07/15 Biomed Res Int. 2021 Jun 24;2021:8853056. doi: 10.1155/2021/8853056. eCollection 2021.

[11] T. Kainrad, S. Hunold, T. Seidel, and T. Langer. Ligandscout remote: A new user-friendly interface for hpc and cloud resources. *J Chem Inf Model*, 59(1):31–37, 2019. 1549-960x Kainrad, Thomas Orcid: 0000-0003-3251-4447 Hunold, Sascha Orcid: 0000-0002-5280-3855 Seidel, Thomas Orcid: 0000-0002-9815-6577 Langer, Thierry Orcid: 0000-0002-5242-1240 Journal Article United States 2018/12/13 J Chem Inf Model. 2019 Jan 28;59(1):31-37. doi: 10.1021/acs.jcim.8b00716. Epub 2018 Dec 27.

[12] Muhammed Tilahun Muhammed and Esin Aki. Pharmacophore modeling in drug discovery: Methodology and current status. *Journal of the Turkish Chemical Society Section A: Chemistry*, 8:759–772, 06 2021.

[13] C. Permann, T. Seidel, and T. Langer. Greedy 3-point search (g3ps)-a novel algorithm for pharmacophore alignment. *Molecules*, 26(23), 2021. 1420-3049 Permann, Christian Orcid: 0000-0002-3574-0899 Seidel, Thomas Orcid: 0000-0002-9815-6577 Langer, Thierry Orcid: 0000-0002-5242-1240 Journal Article Switzerland 2021/12/11 Molecules. 2021 Nov 27;26(23):7201. doi: 10.3390/molecules26237201.

[14] G. Wolber and T. Langer. Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model*, 45(1):160–9, 2005. Wolber, Gerhard Langer, Thierry eng 2005/01/26 J Chem Inf Model. 2005 Jan-Feb;45(1):160-9. doi: 10.1021/ci049885e.

[15] Thomas Seidel, Sharon D. Bryant, Gökhan Ibis, Giulio Poli, and Thierry Langer. *3D Pharmacophore Modeling Techniques in Computer-Aided Molecular Design Using LigandScout*, chapter 20, pages 279–309. John Wiley Sons, Ltd, 2017.

[16] O. F. Guner and J. P. Bowen. Setting the record straight: the origin of the pharmacophore concept. *J Chem Inf Model*, 54(5):1269–83, 2014. Guner, Osman F Bowen, J Phillip eng 2014/04/22 J Chem Inf Model. 2014 May 27;54(5):1269-83. doi: 10.1021/ci5000533. Epub 2014 Apr 18.

[17] John Van Drie. Monty kier and the origin of the pharmacophore concept. *Internet Electronic Journal of Molecular Design*, 6:271–279, 09 2007.

[18] C. G. Wermuth, C. R. Ganellin, P. Lindberg, and L. A. Mitscher. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). *Pure and Applied Chemistry*, 70(5):1129–1143, 1998.

[19] S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw, and R. A. Friesner. Phase: a new engine for pharmacophore perception, 3d qsar model development, and 3d database screening: 1. methodology and preliminary results. *J Comput Aided Mol Des*, 20(10-11):647–71, 2006. Dixon, Steven L Smondyrev, Alexander M

Knoll, Eric H Rao, Shashidhar N Shaw, David E Friesner, Richard A Journal Article Netherlands 2006/11/25 J Comput Aided Mol Des. 2006 Oct-Nov;20(10-11):647-71. doi: 10.1007/s10822-006-9087-6. Epub 2006 Nov 24.

[20] D. Barnum, J. Greene, A. Smellie, and P. Sprague. Identification of common functional configurations among molecules. *J Chem Inf Comput Sci*, 36(3):563–71, 1996. Barnum, D Greene, J Smellie, A Sprague, P Journal Article United States 1996/05/01 J Chem Inf Comput Sci. 1996 May-Jun;36(3):563-71. doi: 10.1021/ci950273r.

[21] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, and J. S. Mason. A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (flap): theory and application. *J Chem Inf Model*, 47(2):279–94, 2007. Baroni, Massimo Cruciani, Gabriele Sciabola, Simone Perruccio, Francesca Mason, Jonathan S Journal Article Research Support, Non-U.S. Gov't United States 2007/03/27 J Chem Inf Model. 2007 Mar-Apr;47(2):279-94. doi: 10.1021/ci600253e.

[22] T. J. Cheeseright, M. D. Mackey, and R. A. Scoffin. High content pharmacophores from molecular fields: a biologically relevant method for comparing and understanding ligands. *Curr Comput Aided Drug Des*, 7(3):190–205, 2011. 1875-6697 Cheeseright, Timothy J Mackey, Mark D Scoffin, Robert A Comparative Study Journal Article Review United Arab Emirates 2011/07/06 Curr Comput Aided Drug Des. 2011 Sep 1;7(3):190-205. doi: 10.2174/157340911796504314.

[23] M. N. Drwal and R. Griffith. Combination of ligand- and structure-based methods in virtual screening. *Drug Discov Today Technol*, 10(3):e395–401, 2013. 1740-6749 Drwal, Malgorzata N Griffith, Renate Journal Article Research Support, Non-U.S. Gov't Review England 2013/09/21 Drug Discov Today Technol. 2013 Sep;10(3):e395-401. doi: 10.1016/j.ddtec.2013.02.002.

[24] S. Y. Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today*, 15(11-12):444–50, 2010. Yang, Sheng-Yong eng Review England 2010/04/07 Drug Discov Today. 2010 Jun;15(11-12):444-50. doi: 10.1016/j.drudis.2010.03.013. Epub 2010 Apr 1.

[25] C. Barillari, G. Marcou, and D. Rognan. Hot-spots-guided receptor-based pharmacophores (hs-pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J Chem Inf Model*, 48(7):1396–410, 2008. Barillari, Caterina Marcou, Gilles Rognan, Didier Journal Article Research Support, Non-U.S. Gov't Validation Study United States 2008/06/24 J Chem Inf Model. 2008 Jul;48(7):1396-410. doi: 10.1021/ci800064z. Epub 2008 Jun 21.

[26] T. Sato, T. Honma, and S. Yokoyama. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J Chem Inf Model*, 50(1):170–85, 2010. 1549-960x Sato, Tomohiro Honma, Teruki Yokoyama, Shigeyuki Journal Article United States 2009/12/30 J Chem Inf Model. 2010 Jan;50(1):170-85. doi: 10.1021/ci900382e.

[27] José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis. Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, 2018. PMID: 29309725.

[28] Miha Skalic, Alejandro Varela-Rial, José Jiménez, Gerard Martínez-Rosell, and Gianni De Fabritiis. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. *Bioinformatics*, 35(2):243–250, 07 2018.

[29] D. R. Koes and C. J. Camacho. Pharmer: efficient and exact pharmacophore search. *J Chem Inf Model*, 51(6):1307–14, 2011. 1549-960x Koes, David Ryan Camacho, Carlos J R21 GM087617/GM/NIGMS NIH HHS/United States R01 GM097082/GM/NIGMS NIH HHS/United States R01GM097082-01/GM/NIGMS NIH HHS/United States R01 GM097082-01/GM/NIGMS NIH HHS/United States 1R21GM087617-01A1/GM/NIGMS NIH HHS/United States R21 GM087617-01A1/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural United States 2011/05/25 J Chem Inf Model. 2011 Jun 27;51(6):1307-14. doi: 10.1021/ci200097m. Epub 2011 Jun 2.

[30] Y. Kurogi and O. F. Güner. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem*, 8(9):1035–55, 2001. Kurogi, Y Güner, O F Journal Article Review United Arab Emirates 2001/07/27 Curr Med Chem. 2001 Jul;8(9):1035-55. doi: 10.2174/0929867013372481.

[31] Chemical Computing Group, Molecular Operating Environment (MOE). `https://www.chemcomp.com/index.htm`. Accessed: 2023-10-20.

[32] J. Taminau, G. Thijs, and H. De Winter. Pharao: pharmacophore alignment and optimization. *J Mol Graph Model*, 27(2):161–9, 2008. 1873-4243 Taminau, Jonatan Thijs, Gert De Winter, Hans Journal Article United States 2008/05/20 J Mol Graph Model. 2008 Sep;27(2):161-9. doi: 10.1016/j.jmgm.2008.04.003. Epub 2008 Apr 11.

[33] D. Schneidman-Duhovny, O. Dror, Y. Inbar, R. Nussinov, and H. J. Wolfson. Pharmagist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Res*, 36(Web Server issue):W223–8, 2008. 1362-4962 Schneidman-Duhovny, Dina Dror, Oranit Inbar, Yuval Nussinov, Ruth Wolfson, Haim J 1UC1AI067231/AI/NIAID NIH HHS/United States ImNIH/Intramural NIH HHS/United States N01-CO-12400/CO/NCI NIH HHS/United States N01CO12400/CA/NCI NIH HHS/United States UC1 AI067231/AI/NIAID NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't England 2008/04/22 Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W223-8. doi: 10.1093/nar/gkn187. Epub 2008 Apr 19.

[34] J. Chen and L. Lai. Pocket v.2: further developments on receptor-based pharmacophore modeling. *J Chem Inf Model*, 46(6):2684–91, 2006. Chen, Jing Lai, Luhua Journal Article United States 2006/11/28 J Chem Inf Model. 2006 Nov-Dec;46(6):2684-91. doi: 10.1021/ci600246s.

[35] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. Feng, S. Ganesan, D. S. Goodsell, S. Ghosh, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, C. L. Lawson, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Persikova, C. Randle, A. Rose, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, Y. P. Tao, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, and M. Zhuravleva. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and

applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*, 49(D1):D437–d451, 2021. 1362-4962 Burley, Stephen K Bhikadiya, Charmi Bi, Chunxiao Bittrich, Sebastian Chen, Li Crichlow, Gregg V Christie, Cole H Dalenberg, Kenneth Di Costanzo, Luigi Duarte, Jose M Dutta, Shuchismita Feng, Zukang Ganesan, Sai Goodsell, David S Ghosh, Sutapa Green, Rachel Kramer Guranović, Vladimir Guzenko, Dmytro Hudson, Brian P Lawson, Catherine L Liang, Yuhe Lowe, Robert Namkoong, Harry Peisach, Ezra Persikova, Irina Randle, Chris Rose, Alexander Rose, Yana Sali, Andrej Segura, Joan Sekharan, Monica Shao, Chenghua Tao, Yi-Ping Voigt, Maria Westbrook, John D Young, Jasmine Y Zardecki, Christine Zhuravleva, Marina R01 GM133198/GM/NIGMS NIH HHS/United States DBI-1832184/National Science Foundation/International DE-SC0019749/US Department of Energy/International Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England 2020/11/20 Nucleic Acids Res. 2021 Jan 8;49(D1):D437-D451. doi: 10.1093/nar/gkaa1038.

[36] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012. PMID: 22716043.

[37] N. Triballeau, F. Acher, I. Brabet, J. P. Pin, and H. O. Bertrand. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem*, 48(7):2534–47, 2005. Triballeau, Nicolas Acher, Francine Brabet, Isabelle Pin, Jean-Philippe Bertrand, Hugues-Olivier Journal Article Research Support, Non-U.S. Gov't United States 2005/04/02 J Med Chem. 2005 Apr 7;48(7):2534-47. doi: 10.1021/jm049092j.

[38] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020. 1471-2164 Chicco, Davide Orcid: 0000-0001-9655-7142 Jurman, Giuseppe Orcid: 0000-0002-2705-5728 Journal Article England 2020/01/04 BMC Genomics. 2020 Jan 2;21(1):6. doi: 10.1186/s12864-019-6413-7.

[39] Enamine Compound Libraries. `https://enamine.net`. Accessed: 2023-10-27.

[40] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res*, 47(D1):D1102–d1109, 2019. 1362-4962 Kim, Sunghwan Chen, Jie Cheng, Tiejun Gindulyte, Asta He, Jia He, Siqian Li, Qingliang Shoemaker, Benjamin A Thiessen, Paul A Yu, Bo Zaslavsky, Leonid Zhang, Jian Bolton, Evan E Journal Article Research Support, N.I.H., Intramural England 2018/10/30 Nucleic Acids Res. 2019 Jan 8;47(D1):D1102-D1109. doi: 10.1093/nar/gky1033.

[41] Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 11 2016.

[42] T. Sterling and J. J. Irwin. Zinc 15–ligand discovery for everyone. *J Chem Inf Model*, 55(11):2324–37, 2015. 1549-960x Sterling, Teague Irwin, John J R01 GM071896/GM/NIGMS NIH HHS/United States GM71896/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural United States 2015/10/20 J Chem Inf Model. 2015 Nov 23;55(11):2324-37. doi: 10.1021/acs.jcim.5b00559. Epub 2015 Nov 9.

[43] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):D901–6, 2008. 1362-4962 Wishart, David S Knox, Craig Guo, An Chi Cheng, Dean Shrivastava, Savita Tzur, Dan Gautam, Bijaya Hassanali, Murtaza Journal Article Research Support, Non-U.S. Gov't England 2007/12/01 Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6. doi: 10.1093/nar/gkm958. Epub 2007 Nov 29.

[44] Ligandscout Tutorial Cards. Inte:Ligand Software-Entwicklungs und Consulting GmbH. `http://www.inteligand.com/download/LigandScout-4.2-Tutorial-Cards.pdf`. Accessed: 2023-10-19.

[45] A. S. Karaboga, J. M. Planesas, F. Petronin, J. Teixidó, M. Souchet, and V. I. Pérez-Nueno. Highly specific and sensitive pharmacophore model for identifying cxcr4 antagonists. comparison with docking and shape-matching virtual screening performance. *J Chem Inf Model*, 53(5):1043–56, 2013. 1549-960x Karaboga, Arnaud S Planesas, Jesús M Petronin, Florent Teixidó, Jordi Souchet, Michel Pérez-Nueno, Violeta I Comparative Study Journal Article United States 2013/04/13 J Chem Inf Model. 2013 May 24;53(5):1043-56. doi: 10.1021/ci400037y. Epub 2013 Apr 25.

[46] Giulio Poli, Thomas Seidel, and Thierry Langer. Conformational sampling of small molecules with icon: Performance assessment in comparison with omega. *Frontiers in Chemistry*, 6, 2018.

[47] G. Wolber, A. A. Dornhofer, and T. Langer. Efficient overlay of small organic molecules using 3d pharmacophores. *J Comput Aided Mol Des*, 20(12):773–88, 2006. Wolber, Gerhard Dornhofer, Alois A Langer, Thierry Journal Article Netherlands 2006/10/20 J Comput Aided Mol Des. 2006 Dec;20(12):773-88. doi: 10.1007/s10822-006-9078-7. Epub 2006 Oct 19.

[48] H. W. Kuhn. Variants of the hungarian method for assignment problems. *Naval Research Logistics Quarterly*, 3(4):253–258, 1956.

[49] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[50] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978.

[51] E. Nazarshodeh and S. Gharaghani. Toward a hierarchical virtual screening and toxicity risk analysis for identifying novel ca xii inhibitors. *Biosystems*, 162:35–43, 2017. 1872-8324 Nazarshodeh, Elmira Gharaghani, Sajjad Journal Article Ireland 2017/09/14 Biosystems. 2017 Dec;162:35-43. doi: 10.1016/j.biosystems.2017.09.005. Epub 2017 Sep 9.

[52] N. Moussa, A. Hassan, and S. Gharaghani. Pharmacophore model, docking, qsar, and molecular dynamics simulation studies of substituted cyclic imides and herbal medicines as cox-2 inhibitors. *Heliyon*, 7(4):e06605, 2021. 2405-8440 Moussa, Nathalie Hassan, Ahmad Gharaghani, Sajjad Journal Article England 2021/04/24 Heliyon. 2021 Apr 1;7(4):e06605. doi: 10.1016/j.heliyon.2021.e06605. eCollection 2021 Apr.

[53] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer. 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(3-4):213–28, 2008. Kirchmair, Johannes Markt, Patrick Distinto, Simona Wolber, Gerhard Langer, Thierry Comparative Study Journal Article Review Netherlands 2008/01/16 J Comput Aided Mol Des. 2008 Mar-Apr;22(3-4):213-28. doi: 10.1007/s10822-007-9163-6. Epub 2008 Jan 15.