J Comput Aided Mol Des (2008) 22:213–228
DOI 10.1007/s10822-007-9163-6

Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes?

Johannes Kirchmair · Patrick Markt · Simona Distinto · Gerhard Wolber · Thierry Langer

Received: 9 October 2007 / Accepted: 17 December 2007 / Published online: 15 January 2008

Abstract Within the last few years a considerable amount of evaluative studies has been published that investigate the performance of 3D virtual screening approaches. Thereby, in particular assessments of protein–ligand docking are facing remarkable interest in the scientific community. However, comparing virtual screening approaches is a non-trivial task. Several publications, especially in the field of molecular docking, suffer from shortcomings that are likely to affect the significance of the results considerably. These quality issues often arise from poor study design, biasing, by using improper or inexpressive enrichment descriptors, and from errors in interpretation of the data output. In this review we analyze recent literature evaluating 3D virtual screening methods, with focus on molecular docking. We highlight problematic issues and provide guidelines on how to improve the quality of computational studies. Since 3D virtual screening protocols are in general assessed by their ability to discriminate between active and inactive compounds, we summarize the impact of the composition and preparation of test sets on the outcome of evaluations. Moreover, we investigate the significance of both classic enrichment parameters and advanced descriptors for the performance of 3D virtual screening methods. Furthermore, we review the significance and suitability of RMSD as a measure for the accuracy of protein–ligand docking algorithms and of conformational space sub sampling algorithms.

Keywords Virtual screening · Evaluation of computational methods · Pharmacophore modeling · Protein–ligand docking · Enrichment descriptors · Decoy selection · Virtual library design · Tautomerism · RMSD

J. Kirchmair · G. Wolber · T. Langer (✉)
Inte:Ligand Software-Entwicklungs- und Consulting GmbH, Clemens Maria Hofbauer-Gasse 6, 2344 Maria Enzersdorf, Austria
e-mail: Thierry.Langer@uibk.ac.at

P. Markt · T. Langer
Department of Pharmaceutical Chemistry, Institute of Pharmacy and Center for Molecular Biosciences (CMBI), University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria

S. Distinto
Dipartimento Farmaco Chimico Tecnologico, Universita' degli Studi di Cagliari, 09124 Cagliari, Italy

Introduction

Virtual screening (VS) techniques are well established tools in the modern drug discovery process and an almost unmanageable number of different 3D VS techniques are available today [1, 2]. Along with that, a plentitude of comparative performance assessments has been published in recent years in order to support computational chemists in finding the best software, particularly in the field of protein–ligand docking. However, the individual needs and aims of VS campaigns differ considerably [3], and so do the fields of application of different algorithmic approaches. Therefore, comparing different VS algorithms is challenging, particularly in case of protein–ligand docking [4].

A large part of the comparative studies suffers from shortcomings that are likely to decrease the significance of the results. Chen and co-workers state that it appears to be a general observation that '*only very rarely are independent workers able to match the docking success rates achieved and published by the vendors of docking programs.*' [3]. Similar issues are also reported by Kontoyianni et al. [5]. Cole et al. [4] have provided an overview of some of the

issues responsible for these difficulties, focusing on comparative studies and data analysis of docking campaigns. In this review, we want to highlight and discuss issues that emerge during the evaluation of 3D VS tools. We provide guidelines on how to evaluate 3D VS techniques—from study design to data analysis.

The review is organized according to the general workflow of an assessment of VS methods: We start with the build up of a representative test set and discuss the preparation of compounds for VS. Subsequently, we summarize issues that should be considered during protein target selection, and preparation. The review continues with the correct setup and setup definition of VS runs and general considerations. Next, we focus on data elaboration, assessment methods, and different benchmarks for the performance of VS techniques. Finally we provide conclusions and recommendations for high quality evaluations.

Library design for the assessment of virtual screening approaches

The performance of VS approaches is mostly measured in terms of their ability to discriminate between active and inactive compounds. Actives and inactives are thereby injected to the VS workflow either as one mixed library or as two dedicated collections. In the latter case, the resulting hit lists are combined before the actual analysis of the results. The collection and preparation of these compounds at the very beginning of an investigation represents the Achilles' heel of any evaluation of VS protocols. Shortages within this early stage of the VS workflow are considerably severe, but fortunately these issues can be excluded rather easily by considering some caveats discussed below.

Collection of active compounds

Profound conclusions on the performance of a certain VS protocol request a comprehensive set of active compounds. Of course the exhaustive collection of all known active compounds will meet this prerequisite in the best way. However, as described below, actives and inactives have to be in a reasonable quantitative relationship and therefore it may be better to consider only a certain amount of characteristic representatives for each relevant chemotype. Test sets should aim at diversity, which might be a challenge if there is only little knowledge on a target available. Therefore, authors of evaluative studies tend to use well-established targets where a fair amount of known actives and knowledge is available.

Another aspect crucial during the selection of actives for certain VS approaches is a common mechanism of binding of the ligand to the protein. Since, e.g., pharmacophore models in general are able to represent only one specific binding mode, in this case one has to make sure that either all compounds of the active set are sharing this binding mode or that all different binding modes are considered by dedicated pharmacophores in a parallel screening [6–8] approach. Aiming at maximum diversity may be in conflict with the consideration of compounds that show a common binding mode. Sometimes the binding mode is unknown or not confirmed by experimental structural data. In this case it is recommended to switch to another, more established target (if possible, of the same protein family).

Evaluations considering quantitative data (e.g., performance assessments of scoring functions and their ability to predict the protein–ligand affinity) require reliable activity data. All information should be obtained from the same bioassay and preferably also from the same laboratory. 3D approaches require that data gained in the laboratory be based on the pure compound, with defined stereochemistry. Assertions based on mixtures are in general ineligible, even if there is a vast surplus of a certain stereochemical configuration. In this case it cannot be excluded that the activity measured experimentally is caused by a highly active enantiomer that is present at a low concentration level. This problem is particularly evident in the case of CoMFA/CoMSIA studies [9]. Furthermore, quantitative evaluations require actives that cover a broad range of activity. For example, the quantitative pharmacophore model generation algorithm HypoGen [10], as implied to the software package Catalyst [11], requires an activity range of at least four orders of magnitude. Predictions based on training sets consisting of less than 16 actives cannot be considered reliable [10].

Collection of inactive compounds

Retrospective studies on the ability of VS approaches to separate active compounds from inactive ones are of course most reliable if based on a test set comprising known active and verified inactive compounds. In this case, a large pool of experimentally confirmed inactive compounds is needed, besides the collection of known actives. In particular academic research suffers from lacking data since inactive molecules are in general not available in the public domain, at least not in the required amount, and even in industrial research the generation of these data may be challenging and expensive. Therefore, it is common practice [12] to collect so-called decoys (i.e. molecules presumably inactive against the examined target) from structural pools or to create them with a virtual combinatorial library [13] generator like, e.g., ilib:diverse [14] and SmiLib [15].

It has been confirmed by several studies that the characteristics of the known inactives or decoys chosen for VS

assessments have significant impact on the enrichment of VS approaches [12, 16, 17]. Both known inactives and decoys are required to meet some essential prerequisites in order to achieve meaningful results. The most important need for decoys is the comparability of their physicochemical properties to the actives set. Probably one of the best examples for the hidden impact of decoy characteristics on the enrichment of VS techniques is the direct dependence of docking scores on the molecular weight of the ligands [3]. Verdonk et al. [12] investigate this issue by structure-based VS against neuraminidase, PTP1B, CDK2 and the estrogen receptor (ER) using GOLD [18, 19]. They demonstrate that docking campaigns conducted against smaller decoys than actives achieve significantly higher enrichment compared to larger decoys. Decoys of high molecular weight achieve higher docking scores on average and are therefore likely to obtain higher ranks than their smaller active counterparts. In fact, these docking results merely reflect the difference in 1D molecule properties and not the performance of the 3D VS approach [12]. This bias in lower dimensions is of course not only due to different molecular weight. If a scoring function for example considers hydroxy groups and their hydrogen bonding ability as a positive contribution to protein–ligand binding compounds containing more of these features will be promoted and considered in the hit list on a higher rank. In turn, inactives containing several hydroxy groups would be preferred over active compounds that contain less of these moieties. They provide evidence that it is not sufficient to just use a *random library* (e.g., subsets of public databases) for performance assessments, but it is essential to build up a so-called *focused library* that reflects the physicochemical properties of the actives set. Verdonk and co-workers therefore propose a simple and very efficient method for the selection of meaningful decoys based on only three basic 1D properties that assess the distance (D) between two molecules (i, j): (i) the number of hydrogen-bond donors (N_D); (ii) the number of hydrogen-bond acceptors (N_A); (iii) the number of nonpolar atoms (N_{NP}).

For each active compound the distance to the nearest other active compound is assessed using Eq. 1. Subsequently, the minimum distances are averaged over all active compounds (D_{min}). Now, for each active compound of the test set a certain number of compounds is randomly chosen from a pool of inactive compounds. These compounds must not exceed the maximum average distance D_{min}. This procedure allows efficient collecting of high quality decoy sets for the evaluation of VS methods.

The largest public available database of decoys that considers comparable 1D properties is the *Directory of Useful Decoys* (DUD) [16], available from http://dud.docking.org/. The DUD is a collection of 36 decoys for each of the 2,950 collected actives of 40 different targets (95,316 in total, after duplicate removal). The compounds represent a subset of the ZINC database with physical properties (e.g., molecular weight, calculated LogP) [20].

Decoys should be topologically dissimilar with respect to the active compounds. Else it is likely that a number of decoys might be identified as actives, which they actually could be. This would lead to a significant amount of false positive hits, an unwanted bias. The DUD also meets this requirement.

In concordance with the actives set also the inactive compounds should be based on diverse scaffolds in order to allow conclusions on the overall reliability of a VS method. Last but not least, an adequate ratio of actives and decoys is necessary for both statistical reasons and in order to satisfy comparability to real-life application scenarios [21]. Only evaluations using a 'high-quality haystack' allow profound conclusions on the performance of VS methods.

Besides the considerations mentioned above, there are some further needs to be met for the evaluation of VS programs. Current studies show a clear tendency toward using high affinity actives for performance assessment. This does not reflect the scenario of a realistic VS campaign, where we are looking for new hits that can be transformed later on into *leads* [3]. These structures are in general considerably smaller than high affinity drug molecules and may therefore achieve different (i.e. most of the time inferior) enrichment rates during VS. In the last few years the trend toward screening for smaller lead structures has been even intensified.

Preparing compounds for virtual screening

Once a test set suitable for VS evaluations is available, the molecular library has to be conditioned to meet the individual needs of the investigated VS approach. There are several pitfalls to consider during compound preparation in order to avoid artificial enrichment and bias.

File formats, import/export issues, and hybridization states

Everyday experience with computational tools, regardless if published recently or if being well established or

$$D(i,j) = \sqrt{(N_D(i) - N_D(j))^2 + (N_A(i) - N_A(j))^2 + (N_{NP}(i) - N_{NP}(j))^2} \quad (1)$$

commercial, teaches computational chemists to be aware of file formats. Every read in or read out of structural data, every file conversion is a possible risk for information loss or data misinterpretation. Special care should be taken on less conspicuous issues like, e.g., stereochemistry. File formats may have different flavors that are not correctly interpreted by all software tools. Depending on the software settings, the stereochemical information may be derived from the 3D input geometry and the chirality flags might be ignored or vice versa. Also the representation and handling of aromatic moieties differs and some programs (e.g., feature-based pharmacophore modeling programs) may have difficulties if the input data does not include appropriate aromatic annotations.

Furthermore, great care should be taken especially on ligand structures gained from the PDB [22]. The PDB file format does not contain hybridization states of the ligands and bond types. Therefore, the correct structure is derived from the atom coordinates during data import. However, inaccuracies in experimental data complicate the interpretation of the molecular structure. Advanced algorithms for the automated identification are available, yet these tools need manual validation of the ligand structures. LigandScout [23] is a pharmacophore model generator based on a sophisticated and customizable ligand-macromolecule complex interpretation algorithm. The software extracts and interprets ligands from PDB structural data and automatically generates structure-based pharmacophore models for VS in several different screening platforms. Besides correct connectivity, the correct placement of hydrogen atoms may be required in order to define the hybridization state (e.g., for protein–ligand docking with GOLD). Hydrogen atoms added in 2D space may thereby affect the 3D conformation [24].

Ionization states

The accurate representation of the ionization states of ligands and the target are of extraordinary importance for VS, especially for docking applications, since the vast majority of docking programs shows considerably high impact of ionization states on the results. Thereby, the correct ionization state may be fairly difficult to assign, since it heavily depends on the microenvironment within the binding site. Thus, it may be more efficient to systematically enumerate all relevant ionization states [25]. Pharmacophore modeling programs require correct assignment of ionization states only in part: MOE [26] pharmacophore screening is depending on charges; Catalyst does not rely on charges and allows to treat relevant scaffolds as being *ionizable*.

Tautomerism

Tautomerism is most of the time a still underestimated issue that may heavily affect VS campaigns [27, 28]. It influences molecular descriptors (e.g., CLOGP; especially fragment-based methods, which depend on the way fragments are produced, their number, size, and the training sets) and similarity searches [29, 30] (e.g., divergent Tanimoto similarity indices of different tautomeric forms). Moreover, tautomerism is likely to affect substructure search and molecular alignment. While most issues listed above do have impact primarily on data examination in the case of VS evaluations (e.g., analyses on the diversity of hit lists) there are even more important issues to consider during the actual VS screening process.

Protein–ligand binding interactions heavily depend on the tautomeric form of both the small organic molecule and the protein target. Thereby, the tautomeric form of the ligand bound to the protein may differ from the favorable tautomer observed in the aqueous phase. In particular hydrogen bonding between protein and ligand may be possible only for certain tautomers. Prominent examples are available, e.g., for barbiturate-based matrix metalloproteinase (MMP-8) inhibitors [31] and for pterin-based inhibitors of the Ricin Toxin A-chain (RTA) [32]. Tautomerism is a serious problem, e.g., for histamines; on the protein side, the correct interpretation of the histidine side chains is particularly important [33].

Both molecular docking and pharmacophore modeling [28] heavily rely on the correct tautomerization of the ligand and the protein. However, the calculation of the accurate tautomeric form is very time consuming. In general it is therefore better efficient to sample all relevant tautomeric forms and to consider all as individual molecules during VS. Taking into account all relevant tautomers during VS enlarges the number of degrees of freedom as well as the chemical space covered by molecular collections. Thereby, enriching databases with tautomers raises the chance for detecting a hit [27]. Popular software tools for tautomer enumeration are, e.g., TAUTOMER [34] and AGENT [35].

Seed structures and conformational space sub-sampling

3D VS approaches heavily rely on the accurate representation of the bioactive conformation. Most well-known 3D pharmacophore modeling tools use pre-calculated databases for screening that include 3D conformational models of all ligands. Today there are highly efficient conformational model generators available (e.g., CAESAR [36], Catalyst FAST/BEST [37, 38], and Omega [38]) that are able to represent the bioactive conformation in a quality that is suitable for VS most of the time. Other screening

tools perform on-the-fly conformation calculation. However, pre-calculating conformational databases is likely to be the most efficient way for handling molecular flexibility since the CPU power demanding calculation step is only executed once.

Most protein–ligand docking programs start from a single 3D seed structure. During the docking process only the torsion angles of this structure are alternated; the bond angles and bond lengths are in general kept rigid. Therefore it is crucial to provide an energetically favorable seed structure to the docking protocol.

Regardless of the dimensions of data input, the ligand seed structure presented to conformational space sub-sampling and docking algorithms [39] significantly affects the outcome of VS campaigns. Furthermore, the number of conformers and docking poses calculated, as well as using different system platforms and different random seeds that are fed to the programs affect screening results considerably.

A frequently used workflow to overcome this seed structure bias is to derive the canonical SMILES code of a compound (i.e. a unique identifier based on a simplistic connection table w/o any 3D structural information), generate a primary 3D structure using a 3D structure generator like CORINA [40], and minimize this structure before injection to the VS process [3]. Alternatively, ligands represented in SMILES code may be directly injected to the docking process. However, it is important to use a well-defined (reproducible) canonical SMILES flavor, as Knox et al. [25] and Carta et al. [41] demonstrated that SMILES code permutations are likely to significantly bias the VS process. This bias can even be used in order to explore conformational space more efficiently and more exhaustively.

Protein target selection and preparation

It is a common observation that the performance of docking algorithms is highly depending on the target [42–46]. This experience goes along with the fact that docking algorithms are usually calibrated and validated using small protein–ligand data sets [47]. For meaningful comparative assessments it is therefore crucial to investigate a large number of target structures of high diversity. Only the investigation of a representative sample of targets from different protein families assures valuable, global conclusions on the performance of a specific docking algorithm [3, 5]. Warren et al. [46] have published the most comprehensive investigation of docking techniques based on a variegated target set and provide so far the most profound insight to the performance of state-of-the-art approaches.

There are several issues to be considered for the structural data input. Firstly, structural data itself contains experimental errors and uncertainties. There is a trend toward increased docking accuracy with higher resolution of the protein complexes, which suggests that to a certain extend inferior docking results may be caused by issues of the structural data input [48]. This indicates that discarding low quality protein–ligand complexes might be advantageous for the evaluation of docking algorithms [49]. However, this may not reflect the application scenario of a VS screening campaign [3, 5]. Especially for novel targets there may be only low quality structures available. Yet in this early period of drug discovery molecular docking is of particular interest in order to discover new lead structure candidates. Thus, analyzing the performance of docking algorithms with structural data from low-resolution complexes is also of certain interest. There is still only little awareness about conformational changes of the binding site caused by crystal packing [48]. This quite common issue is usually caused by ligands and protein chains that are found in the vicinity of the binding site [4, 48]. Some docking approaches are highly sensitive on fine details of the X-ray structure; marginal changes of the target structure may change the outcome considerably. Special account should be taken on correct protonation and tautomeric states of the amino acids. Refinements of the protein may be needed in order to balance structural shortcomings and to meet the individual requirements of a docking algorithm.

Evaluations of structure-based VS approaches should take into account different active site topologies and cover large, voluminous binding sites and tight pockets, as well as surface exposed sites [3]. Authors would do the scientific community a great favor if they use crystal structure data that are available in the public domain for evaluative studies in order to make results traceable [3]. This is of course also true for the protein and the ligand test set structures. The re-use of public available test sets [16, 19, 48] allows to compare the outcome of different studies.

Metals show very specific binding characteristics and are usually handled with dedicated features by VS tools. The accurate prediction of metal binding interactions is of high importance during VS and should be considered for comprehensive assessments [5].

Virtual screening setup

Reproducibility should be the highest maxim of any scientific publication and particularly of evaluative studies. Only a precisely defined software setup allows following up investigations. However, publications frequently suffer from insufficiently described parameters and scoring functions. Hitherto, there are only very few publications available aiming to overcome these issues. Kirstam et al. [50] provide

an extensive guide for the description of the Catalyst software setup; Cole et al. [4] provide an overview on the difficulties of protein–ligand docking. The problem of how to define the steps of a VS study concerns all parts of an investigation. A lot of VS publications relate the results obtained from screening to the flexibility of the investigated ligands. The number of rotatable bonds reflects flexibility, however, there are different definitions of rotatable bonds and therefore different software tools may obtain different results. In cases where the accurate definitions are not available at least the software used for calculations should be precisely defined. Results obtained from adapted scoring functions are not traceable without the source code [4].

The correct handling of water molecules is crucial for the performance of docking programs. Usually, water molecules are removed from the active site, except for such molecules that are known to bind very tightly to the protein or are known to be essential for drug action [3]. Re-docking of ligands to the binding site with depleted waters is likely to obtain inferior accuracy. The ligand may thereby be placed at positions that are occupied by water molecules. This is also true in the case of co-factors. If a co-factor is present in the vicinity of the protein–ligand site it will significantly affect the docking results and it may be crucial for obtaining an accurate ligand pose.

Most established docking programs provide individual tools to define the parts of a protein relevant for docking [3]. For the sake of direct comparability of different docking algorithms, the definition of the protein–ligand binding site of all investigated protocols should be identical. It is insufficient to describe the active site as the entirety of all amino acids (waters and cofactors) in a certain distance (usually 5–12 Å [3, 5, 39, 47, 51]) around the ligand: Some tools cut precisely at the borderline; some tools include moieties that have at least one atom within this area. Again other tools define the binding site with a certain radius around the centroid of the ligand and others by a certain distance from every ligand atom. If one considers e.g., small, circular flavonoids versus long fatty acids one would obtain absolutely different binding site definitions.

The definition of the hardware setup used for screening should at least consider the processor architecture and memory resources, the definition of the software setup the operating system version and the software versions used for the assessment. Especially for time measurements a more precise definition of the computing environment is usually required.

But even if all these data are provided to the community, results may still be not entirely comprehensible. Several VS algorithms use a random number during screening and therefore the results may not be accurately reproducible. Onodera and co-workers have recently published an investigation on this issue [39].

Comparing different approaches—general considerations

In contrast to rapid VS methods like pharmacophore modeling, the performance of docking methods is always a trade-off between computational demands and accuracy. This explains that there is a plentitude of docking approaches available that aim at different fields of application: Incremental construction approaches (e.g., FlexX [52]), shape-based algorithms (e.g., DOCK [53, 54]), genetic algorithms (e.g., GOLD [19]), systematic search (e.g., Glide [55, 56]), Monte Carlo simulations (e.g. LigandFit [57]), and surface-based molecular similarity methods (e.g., Surflex [58]). Most exhaustive algorithms focus on the accurate prediction of a binding pose, more efficient algorithms on the docking of small ligand databases within reasonable time, and rapid algorithms on the virtual high-throughput screening of millions of compounds. Today, however, CPU power is usually no longer a limiting factor for protein–ligand docking due to recent technologic advances and dropping hardware costs.

There is a strong affinity to using default settings for the evaluation of VS screening programs [39]. Nevertheless, different requirements of these docking techniques make direct comparisons difficult [47], since it is obviously problematic to directly relate the accuracy of GOLD using program defaults (i.e. the most exhaustive settings) with e.g., FRED. Besides the other field of applications these methods are based on completely different algorithmic approaches. While GOLD represents a genetic algorithm, FRED [59] is based on rapid shape complementarily and pharmacophoric feature mapping. The other way round, it is questionable whether more CPU time demanding approaches are necessary to obtain a certain degree of accuracy or if equivalent results are also achieved by a more efficient software setup. In fact, we have recently shown that in the exact opposite is the case for pharmacophore modeling, where fastest settings for both conformational model generation as well as screening—overall—achieve highest enrichment rates [60]. We think that—where possible—it is worth to investigate setups that require similar CPU power for direct performance comparison, or—at least—to consider and clearly point out the possible impact of different settings on VS speed and data accuracy.

The direct comparison of protein–ligand docking considering target flexibility [61] (e.g., FlexE [62]) to docking approaches that keep the target structure rigidly (e.g., FlexX [52]) is problematic. By consideration of protein flexibility the number of degrees of freedom increases.

Our personal experience is that this may lead to inferior enrichment during VS in certain cases. It seems that—statistically speaking—active compounds fit into the protein binding site even if the receptor is kept rigid, while probability that there is a receptor conformation present that is able to house inactive compounds increases with target flexibility. Moreover, a time problem arises and docking algorithms are forced to make further approximations about the docking process, which is also likely to lead to inferior docking accuracy.

As already mentioned above, the disclosure of the data material used for publication is of great value for the community. Thereby it may not be sufficient to provide IDs of public available compounds and protein structures; it is absolutely favorable to provide the definitive structures used for the VS assessment. The preparation of docking studies offers virtually indefinite options that cannot be knocked down into a simple process description. For example, the often-read statement, that the ligand structures used for VS have been charged at pH 7, leaves a lot of questions open. If the data cannot be made available for public an offer to work with the data within the author's research facilities may be an alternative to overcome this bottleneck. The use of propriety data for VS evaluations should require direct reasons.

There is consensus in the scientific community that VS methods need experts on both the biological system and the program in order to achieve maximum performance, especially protein–ligand docking [46]. For obvious reasons it is quite improbable that even mid sized research groups are in the lucky situation to have experts for ten different docking programs on twenty protein targets.

Hitherto, the statistical relevance of comparative docking studies has been considered in only a few cases. The problem is that VS assessments are highly expensive in terms of manpower and therefore most of the time ends up in a trade-off between significance and budget situation. Moreover, under normal conditions it is tedious to obtain software licenses for a lot of very promising VS tools. This implies that studies published are biased toward availability and focused on new methods.

Assessment modes and methods

Protein–ligand docking mainly consists of two separate parts: Pose prediction and the estimation of affinity (scoring). While it seems that today's docking algorithms are promising tools for the prediction of the correct ligand binding pose, the usefulness of state-of-the-art scoring functions is controversial [46]. There are four ways to analyze the outcome of a docking campaign: (i) The accuracy of the binding pose prediction, (ii) the accuracy of the affinity prediction, (iii) the enrichment rates obtained by virtual screening, (iv) the diversity of the hit list. A comprehensive reference collection on this topic is provided by Kellenberger et al. [47]. Here we summarize major modes of performance assessment; the merits of VS descriptors are discussed in the consecutive section.

Assessing the success of pose prediction (protein–ligand docking)

The accuracy of pose prediction is usually determined by re-docking of a ligand into the binding site. This evaluation method itself suffers from a significant bias since it neglects the changes in protein conformations during ligand binding [4]. However, it is currently the most applicable approach available. The bias has been proven and examined by, e.g., Murray and co-workers, who achieved inferior—in part considerably lower—enrichment rates using a cross-docking approach [63].

Usually, the accuracy of docking poses is quantified by calculation of the RMSD between the experimentally determined ligand structure (as it is bound to the protein) and the pose calculated by the docking algorithm. Whatever measure is selected, there are at least three approaches to consider docking poses for the assessment: (i) Considering the number one ranked docking pose, (ii) considering each docking pose, (iii) considering all plausible docking poses. Approach (iii) is to a certain extend subjective and therefore problematic. More detail on benchmarks is provided in the next section.

Assessing the success of affinity prediction (protein–ligand docking)

The accuracy of affinity prediction is in general measured by analyzing the correlation of the experimentally determined affinity and the docking score. It presumes that experimental affinity data are available, measured for the binding as observed by X-ray crystallography or NMR spectroscopy. The correlation can be examined for a single scoring function or several combined scoring functions (consensus scoring) [12, 43, 44, 64, 65]. The discussion about the usefulness of state-of-the-art scoring functions currently divides the scientific community and there are data available that demonstrate no strong correlations of a multitude of scoring functions on several protein targets [46]. The consensus scoring approach is thought as a tool to reduce the number of false positives and to decrease the errors in scores. However, this issue is again discussed controversially and examples show that in some cases the performance of a single scoring function is superior [42, 66]. Results based on consensus scoring are usually quantified in *rank-by-number* (average score after normalizing), *rank-by-rank* (hit list based on the

average score), or *rank-by-vote* (compounds are considered in the final hit list if within top x% of the ranked lists of each of the individual scoring functions) mode. For more information the reader is referred to [12, 64].

Rescoring docking results with alternative scoring functions may be a suitable approach for improving the outcomes, however, it implies further pose preparation for the individual needs of the scoring function. Some authors state that this represents an additional complicating step and a further challenge for the VS assessment and therefore avoid this technique [3].

Assessing the enrichment of VS runs

Enrichment is usually considered the key benchmark for the success of VS. It quantifies the number of active compounds found in the hit list, with respect to the fraction of inactives. The success of VS is correlated with its ability to rank active compounds at high positions of the hit list, since only the first fraction of a hit list will be screened experimentally. We will discuss this problem in detail below, together with the plethora of descriptors available to characterize the success of a VS approach.

Assessing the diversity of hits lists obtained from VS experiments

In our opinion the diversity of hit lists obtained by VS is—so far—underestimated. VS methods should not be characterized exclusively by their ability to rank as many active compounds as possible at high positions; diversity of the obtained hit list is of similar importance for lead identification. It is favorable to find a few valuable representatives per scaffold instead of a large amount of actives based on only a few compounds. Methods that allow identifying diverse scaffolds are much more useful than those which achieve high enrichment for the sake of plurality. The diversity of hits can be quantified, e.g., by calculation of the Tanimoto index and clustering by scaffold similarity.

Descriptors for the performance of virtual screening protocols

Descriptors for the accuracy of binding mode prediction (protein–ligand docking)

RMSD as a benchmark for the accuracy of docking poses

Currently, the RMSD between a generated docking pose and the experimental ligand conformation represents the most established benchmark for the ability of docking algorithms to predict the protein-bound ligand conformation. In the vast majority of publications only heavy atoms are considered during RMSD assessment. Symmetry detection is needed, e.g., for carboxylates and phosphates in order to allow adequate consideration of symmetric moieties [47, 67]. There are two modes of application available for RMSD: The absolute RMSD and the relative RMSD. The absolute RMSD measures the distance between corresponding atom pairs of two conformers without coordinate translation or rotation. This measure is primarily used for docking evaluations. The relative RMSD implies an additional alignment step of the molecules before the actual RMSD calculation. This mode of RMSD assessment is especially useful for the investigation of the accuracy of conformational model generators [37, 38]. The characteristics of RMSD that make it the current standard approach for the definition of the docking accuracy is objectivity, high responsiveness, and its easy automated calculation [68]. Nevertheless, RMSD suffers from serious problems. First of all, it implies no information on the quality of representation of the complex interactions of a ligand with the protein. Furthermore, differences in the force fields applied during docking variations in the predicted pose may result in relatively large RMSD values without changing the principal interactions of the ligand with the protein. Molecules may comprise flexible side chains that are not important for binding at all. Even if the core structures of such ligands were placed accurately, high RMSD levels would suggest an inaccurate pose prediction. Moreover, large, almost symmetric molecules may be swapped in the binding site during docking. In this case the binding mode may be predicted correctly, if the essential interaction features are detected, however, RMSD would be at a very high level and would suggest wrong placement [68]. RMSD is depending on the molecular weight of compounds. Small compounds can easily achieve low RMSDs even when placed randomly. Cole and co-workers found that in the case of arabinose random rotation about the center of gravity obtains RMSD < 2 Å in 10–15% of all placements [4]. Even more, some very high RMSD levels may dominate the average accuracy; a total failure with RMSD, e.g., around 8 Å may push conclusions into a wrong direction. It is highly questionable whether a docking algorithm achieving RMSD 5 Å is twice as accurate as a competitor achieving a RMSD value of 10 Å; a statement, that both algorithms are wrong implies more information on their accuracy [4]. A rather unknown benchmark related to the RMSD is the RDE (Relative Displacement Error) [69]. RDE helps to soften the impact of large discrepancies on the average benchmark value but it still suffers from the other insufficiencies mentioned

above. Therefore, Kroemer et al. [68] proposed a new benchmark for the quality of docking poses based on visual inspection, as discussed below.

Classification based on visual inspection of the ligand poses

Assigning docking poses to bins according to their accuracy in terms of representing the protein-bound ligand conformation is not a new approach. Eleven years ago it was used by Jones and co-workers for the validation of their genetic docking algorithm [19]. Visual inspection for the classification of docking poses was also, e.g., used by Kontoyianni et al. [5]. Kroemer et al., however, were the first that systematically investigated the usefulness and merits of visual inspection and introduced their Interactions-Based Accuracy Classification (IBAC) [68]. IBAC is not a standardized protocol for the accuracy assessment; the criteria for this benchmark are depending on the protein–ligand binding mode to be inspected. In order to define IBAC criteria for a certain binding mode, Kroemer and co-workers analyze the experimentally determined protein–ligand complex. Thereby, the ligand is divided into two major areas: A core comprising essential features for binding to the protein, and the peripheral area of lower priority for binding. Correctly docked ligands show both correct core and peripheral interactions with the protein. If the core interactions are represented accurately but the peripheral interactions differ moderately, the docking pose is considered 'nearly correct'. This is the overall IBAC idea, however, as already mentioned above, Kroemer and co-workers analyze different classification protocols with different criteria. During data analysis they found a fairly good correlation between RMSD and IBAC, however, in a significant number of cases both benchmarks obtained differing results. They were able to identify poses of low RMSD that do not represent the key interactions with the protein. The study demonstrates that RMSD is obviously not a universally applicable benchmark for the assessment of docking poses and may lead to wrong conclusions on the accuracy of docking algorithms. So far there is no automation for IBAC available, which is a severe disadvantage compared to RMSD. In order to overcome this problem, an automated protocol for the detection of core interactions would be necessary.

Considering the severe bottlenecks of RMSD, there is strong need for the development of a more reliable benchmark for the accuracy of docking poses. It is highly recommended to support the data assessment based on RMSD by information gained from IBAC-like inspections.

Descriptors for enrichment

The general aim of VS methods is to retrieve a significant larger fraction of true positives from a molecular database than a random compound selection. If a VS method selects n molecules from a database with N entries, the selected hit list comprises active compounds (true positive compounds, TP) and decoys (false positive compounds, FP). Active molecules that are not retrieved by the VS method are defined false negatives (FN), whereas the unselected database decoys represent the true negatives (TN) (Fig. 1) [70].

Descriptors that assess the enrichment of active molecules from a database containing active molecules and decoys seem to be a rational approach to evaluate the VS performance. Most of the commonly used enrichment descriptors are based on two values. The first value is the sensitivity (Se, true positive rate, Eq. 2), which describes the ratio of the number of active molecules found by the VS method to the number of all active database compounds [70, 71].

$$Se = \frac{N \text{ selected actives}}{N \text{ total actives}} = \frac{TP}{TP + FN} \tag{2}$$

The second value is the specificity (Sp, false positive rate, Eq. 3), which represents the ratio of the number of inactive compounds that were not selected by the VS protocol to the number of all inactive molecules included in the database [71].

$$Sp = \frac{N \text{ discarded inactives}}{N \text{ total inactives}} = \frac{TN}{TN + FP} \tag{3}$$

Most enrichment descriptors do not include a weight for the rank that is assigned to the active molecule by the VS algorithm. Therefore, only the fraction of actives recouped from the database using the VS method is taken into account by these descriptors. The positions of the active molecules in a list ordered by the VS rank are disregarded. However, only about 0.1–10% of the molecules retrieved by a VS method are investigated, e.g., by biological testing

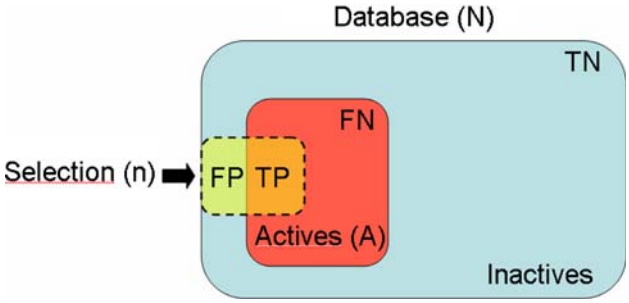

Fig. 1 Selection of n molecules from a database containing N entries by a VS protocol

[3]. Therefore, it is not only important to have a VS protocol that performs well in discriminating actives from decoys, but also to have a VS workflow that is able to rank the actives at the beginning of a rank-ordered list. A VS algorithm that is able to retrieve numerous active molecules, but that ranks actives at random positions of the ordered list, is useless. On that account, new enrichment descriptors were derived with respect to this so-called "early recognition problem" in VS practice [72]. In this review, we categorized enrichment descriptors into classic and advanced descriptors. Classic descriptors do not take into account the "early recognition problem", whereas advanced descriptors possess a weight that favors active molecules ranked in high positions over low-ranked actives.

Classic enrichment descriptors

In this section, some of the classic enrichment descriptors, including the well-known yield of actives and the enrichment factor are summarized.

Accuracy. The accuracy *Acc* describes the percentage of molecules which are correctly classified by the screening protocol (Eq. 4) [73–75].

$$Acc = \frac{TP + TN}{N} = \frac{A}{N} \cdot Se + \left(1 - \frac{A}{N}\right) \cdot Sp \tag{4}$$

Efficiency. The analysis of efficiency *AE* evaluates the screening performance if the database includes molecules with unknown activity (Eq. 5). U_s is the number of compounds with unknown activity selected by the screening protocol, whereas U_{total} represents the number of all database molecules for which no activity data exists [76].

$$AE = \frac{1}{2} \cdot (Se + Sp) \cdot \left(1 - \frac{U_s}{U_{total}}\right) \tag{5}$$

Balanced labeling performance. Equation 6 represents the balanced labeling performance l_{bal}. If all the active and inactive molecules are correctly identified by the screening method, this weighted accuracy descriptor has the value 1 [77, 78].

$$l_{bal} = \frac{1}{2} \cdot Se + \frac{1}{2} \cdot Sp \tag{6}$$

Ford's M. Ford's *M* is described by Eq. 7. It represents another descriptor, which is based on *Se* and *Sp*. The descriptor includes an adjustable weighting coefficient ω [79]. The accuracy, the analysis of efficiency, the balanced labeling performance, and the Ford's *M* are equal as long as $\omega = A/N = 1/2$ and $U_s = 0$ [70].

$$M = \omega \cdot Se + (1 - \omega) \cdot Sp \tag{7}$$

Discrimination ratio. The discrimination ratio *DR* represents another combination of *Se* and *Sp* that describes screening performance (Eq. 8) [80].

$$DR = \frac{TP/A}{TN/(N - A)} = \frac{Se}{Sp} \tag{8}$$

Information content. Equation 9 shows the information content *I*, another descriptor which has been used in the validation of pharmacophore-based screening methods [81].

$$I = TP \cdot \log\left(\frac{TP}{FP}\right) + FN \cdot \log\left(\frac{FN}{TN}\right) \tag{9}$$

"Matthews" correlation coefficient. The "Matthews" correlation coefficient *C* is described by Eq. 10. In the ideal case of a screening protocol which discriminates all actives from all inactive molecules, the "Matthews" correlation coefficient is 1 [82, 83].

$$C = \frac{TP \cdot TN - FN \cdot FP}{((TN + FN) \cdot (TN + FP) \cdot (TP + FN) \cdot (TP + FP))^{1/2}} \tag{10}$$

Goodness of hit list. The "Goodness of hit list" *GH* was designed by Güner and Henry for evaluation of the discriminatory power of pharmacophore models (Eq. 11). With respect to the presence of active molecules that bind to another site of the target which cannot be represented by the pharmacophore model, the descriptor favors the *Ya* over *Se* [10].

$$GH = \left(\frac{3}{4}Ya + \frac{1}{4}Se\right) \cdot Sp \tag{11}$$

Screening percentage. Chen et al. [3] introduced the screening percentage, which is defined as the fraction of a database that has to be screened in order to retrieve a certain percentage of molecules with known activity.

Yield of actives. One of the most popular descriptors for evaluating VS methods is the yield of actives *Ya* (Eq. 12). This descriptor quantifies the probability that one of *n* selected compounds is active. In other words, it represents the hit rate that would be achieved if all compounds selected by the VS protocol would be tested for activity [10, 70, 73]. However, it contains no information about the consistence of the database and the increase of the ratio of active molecules to decoys within a VS compound selection compared to a random compound selection. For instance, a value of 0.3 could be caused by a VS protocol that performs comparably to a random molecule selection from a database containing 30% active molecules. On the other hand, the database could include only 3% actives. Therefore, a value of 0.3 would describe a VS algorithm that performs ten times better than a random selection.

$$Ya = \frac{TP}{n} \tag{12}$$

Enrichment factor. Another frequently used evaluation descriptor is the enrichment factor EF (Eq. 13). This descriptor takes into account the improvement of the hit rate by a VS protocol compared to a random selection [70, 73, 84, 85].

$$EF = \frac{TP/n}{A/N} \tag{13}$$

One disadvantage of the EF is its high dependency on the ratio of active molecules of the screened database [71, 72]. This descriptor can be used to decide which VS method possesses the best performance if the same database of actives and decoys is utilized for evaluation. In contrast to that, comparisons of EFs derived from VS workflow evaluations using compound sets with different ratios of active molecules are less reliable [72]. Another disadvantage is that all actives contribute equally to the value. On that account, the EF does not distinguish high ranked active molecules from actives ranked at the end of a rank-ordered list. In other words, two VS methods that differ in the ability of ranking the highest scored active molecules at the beginning of such an ordered list, but show the same enrichment for active molecules, would be assessed to perform equal [72]. Thus, the EF belongs to the classic enrichment descriptors that do not consider this "early recognition problem".

Statistical significance of the enrichment. In relation to EF, the statistical significance of the enrichment given by Eq. 14 is used to assess VS performance (Eq. 14). It describes the probability that a random selection of molecules contains an equal or higher number of active compounds than a molecule selection derived by a VS protocol [86].

$$S = \sum_{k=TP}^{A} \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} \tag{14}$$

Receiver operating characteristic (ROC) curve analysis. The ROC curve method describes *Se* for any possible change of *n* as a function of (1−*Sp*) [71]. If all molecules scored by a VS protocol with sufficient discriminatory power are ranked according to their score, starting with the best-scored molecule and ending with the molecule that got the lowest score, most of the actives will have a higher score than the decoys. Since some of the actives will be scored lower than decoys, an overlap between the distribution of active molecules and decoys will occur, which will lead to the prediction of false positives and false negatives [87].

The selection of one score value as a threshold strongly influences the ratio of actives to decoys and therefore the validation of a VS method. The ROC curve method avoids the selection of a threshold by considering all *Se* and *Sp* pairs for each score threshold, which represents another advantage of this method [71].

A ROC curve is plotted by setting the score of the active molecule as the first threshold. Afterwards, the number of decoys within this cutoff is counted and the corresponding *Se* and *Sp* pair is calculated. This calculation is repeated for the active molecule with the second highest score and so forth, until the scores of all actives are considered as selection thresholds. Figure 2 shows a theoretical distribution for actives and decoys according to their scores. One selection threshold S represents one point on the ROC graph, which in turn stands for one *Se* and *Sp* pair [71].

The ROC curve representing ideal distributions, where no overlap between the scores of active molecules and decoys exists, proceeds from the origin to the upper-left corner until all the actives are retrieved and *Se* reaches the value of 1. Thereafter, only decoys can be found using the VS method. Thus, the ideal ROC curve continues as a horizontal straight line to the upper-right corner where all actives and all decoys are retrieved, which corresponds to $Se = 1$ and $Sp = 0$. In contrast to that, the ROC curve for a set of actives and decoys with randomly distributed scores tends towards the $Se = 1-Sp$ line asymptotically with increasing number of actives and decoys. Finally, ROC curves between the random graph and the ideal curve are plotted for VS workflows which are able to score more active molecules higher than decoys and cause overlapping distributions which represents the usual case in VS (Fig. 3). [87].

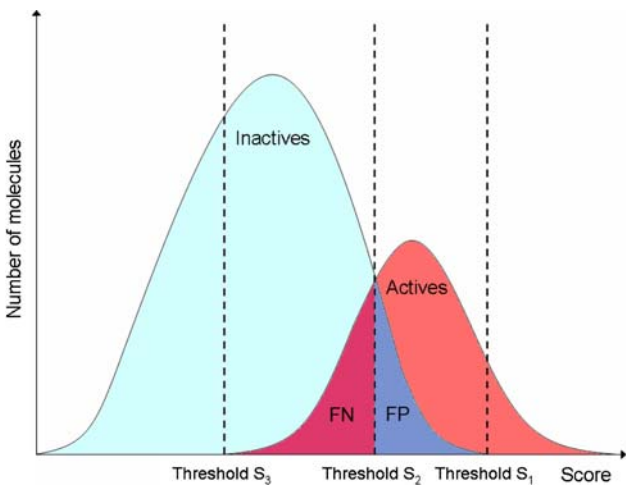

Fig. 2 Theoretical distributions for active molecules and decoys according to their score. Due to distributions overlap, different ratios of false positives (*FP*) and false negatives (*FN*) are retrieved, depending on the selection threshold *S*

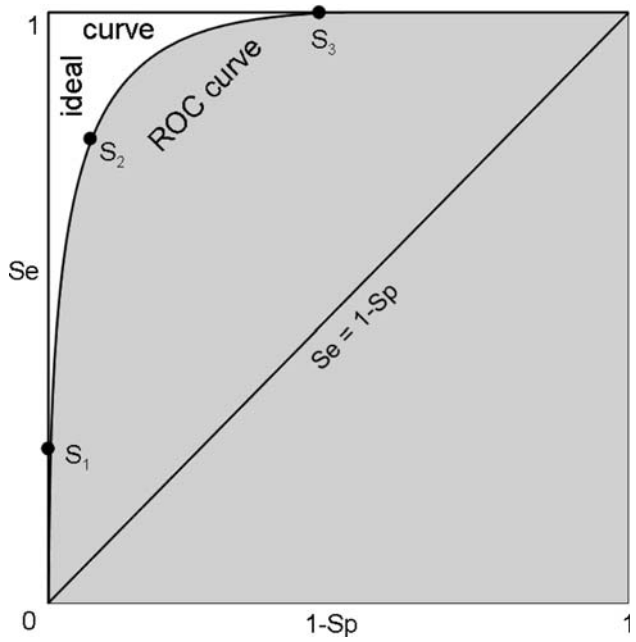

Fig. 3 The ROC curves for ideal and overlapping distributions of actives and decoys. The three ROC curve points S_1, S_2, S_3 are representing the corresponding thresholds displayed in Fig.2. A random distribution causes a ROC curve which tends towards the $Se = 1 - Sp$ line asymptotically with increasing number of actives and decoys

If the ROC curves do not cross each other, the curve that is located closer to the upper-left corner represents the VS workflow with the better performance in discriminating actives from decoys. On that account, ROC curves allow an intuitive visual comparison of the discriminatory power of different VS methods over the whole spectrum of *Se* and *Sp* pairs [72].

Another way of interpreting the results of ROC curves is the area under the ROC curve. The area under the curve (AUC) can be calculated as the sum of all rectangles formed by the *Se* and 1-*Sp* values for the different thresholds. Threshold S_i is the score of the *i*th active molecule (Eq. 15) [87].

$$\mathrm{AUC} = \sum_i [(Se_{i+1})(Sp_{i+1} - Sp_i)] \quad (15)$$

For ideal distributions of actives and decoys an AUC value of 1 is obtained; random distributions cause an AUC value of 0.5. VS workflows that perform better than a random discrimination of actives and decoys retrieve an AUC value between 0.5 and 1, whereas an AUC value lower than 0.5 represents the unfavorable case of a VS method that has a higher probability to assign the best scores to decoys than to actives. For instance, a randomly selected decoy is ranked higher by the VS workflow than a randomly selected active molecule 7 times out of 10 if the AUC value is 0.3 [87].

The ROC curve and the AUC value are useful and easily manageable evaluation techniques for determining the discriminatory power of VS methods, which—in contrast to the EF—do not depend on the ratio of actives to decoys in a database [87]. However, the AUC value itself suffers from the disadvantage that two VS methods cannot be discriminated according to their ability of recognizing actives at the beginning of an ordered list. For example, an identical AUC value for two different VS workflows does not mean that both workflows are equal in scoring the actives of a database. As displayed in Fig. 4a, VS method 1 retrieves more actives at the beginning of a list ordered by the score than VS method 2 [88].

Thus, VS method 1 addresses the "early recognition" problem better than VS method 2. Only their overall discriminatory performance and therefore their AUC values are identical [89]. Another example for VS workflows that achieve AUC values, which do not correlate with the "early recognition" performance of the VS workflows, is displayed in Fig. 4b.

In Fig. 4b workflow 1 retrieves more actives at the beginning of the ordered list, but has a significant lower AUC than VS workflow 2. However, early-recognized actives contribute more area to the AUC than actives at the end of an ordered list. This becomes even more obvious if horizontal (instead of vertical) rectangles—like in a Lebesgue integration scheme—are used to calculate the AUC. On that account, actives contribute according to their rank in an ordered list to the AUC. Therefore, if only the area under the beginning of the ROC curve, e.g. the highest ranked 10% of the screened database is considered, VS protocols could be evaluated with respect to the "early recognition" problem.

Truchon et al. showed that comparisons between AUC values derived from different evaluation studies should be performed with a reasonable size of the ratio of actives ($R_a \ll 1$) [72]. This is important, because like for other enrichment metrics the accuracy in measuring the AUC, but not the AUC itself depends on the ratio of actives.

To sum up, the AUC represents somehow a transition state between classical and advanced enrichment descriptors, as it weights actives according to their ranks in an ordered list such as an advanced enrichment descriptor, but the AUC value itself describes the overall performance of the VS method. If a certain threshold is set, the area under the beginning of the ROC curve allows addressing the early recognition problem. Therefore, the AUC is a useful and well-established enrichment metric for comparing the performance of VS workflows.

Advanced enrichment descriptors

As mentioned above, classic enrichment descriptors, such as the EF, cannot discriminate between a VS algorithm that

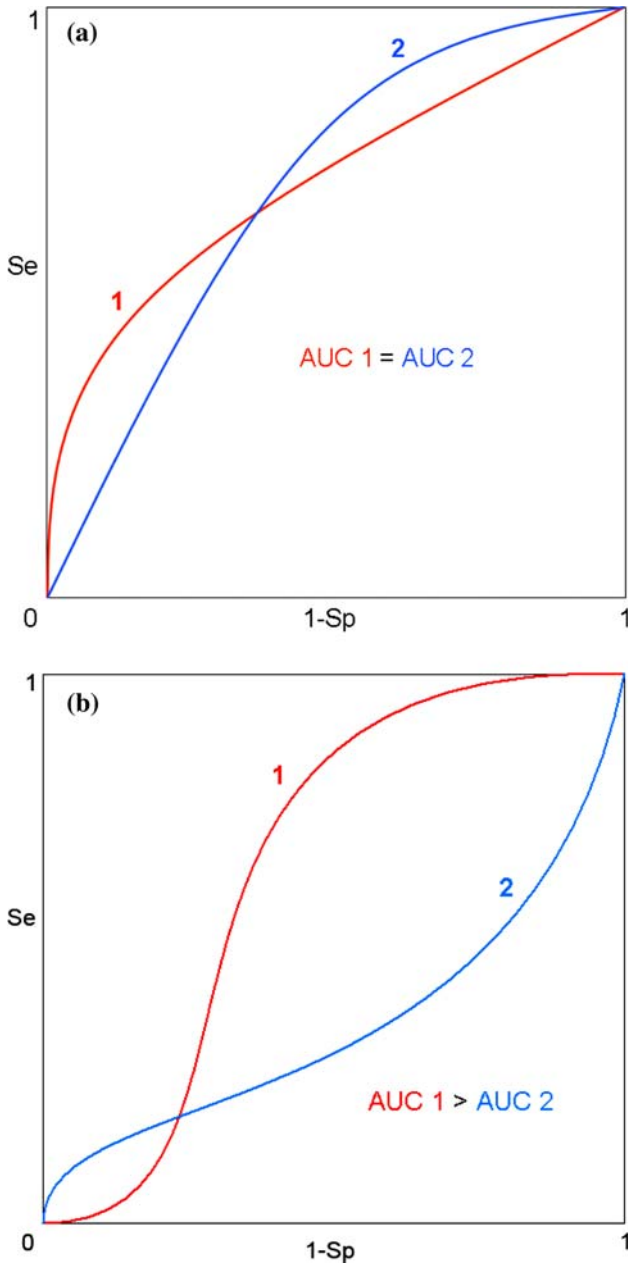

Fig. 4 (**a**) ROC curves for two different VS methods that possess an equal AUC. Notwithstanding both VS application examples are equal in the overall performance, VS method 1 (red line) performs better in discriminating actives from decoys in an early part of an ordered score list than VS method 2 (blue line). (**b**) Although VS workflow 2 (blue line) addresses the "early recognition" problem better than VS workflow 1 (red line), the AUC value for VS workflow 1 is significant larger than for VS workflow 2

ranks half of the actives at the beginning of the ordered list and the other half at the end and a VS protocol that ranks all actives at the beginning of the list. This "early recognition problem" of VS methods is addressed by only a very small amount of existing evaluation descriptors such as the robust initial enhancement (RIE), the Boltzmann-enhanced discrimination of ROC (BEDROC) descriptor and—as already explained above—by the AUC.

Robust initial enhancement RIE. The RIE was developed by Sheridan et al. [88] to provide a descriptor that does not suffer from large value variations if only a small number of actives are investigated. For these purposes, they related the rank for the *i*th active molecule to the number of scored compounds investigated *a*. To get a weight of approximately 1 for the active molecule which is located at the beginning of the list and to retrieve decreased weights for increasing ranks of the actives, an exponential function described in Eq. 16 was utilized [88].

$$S = \sum_{i=1}^{\text{actives}} \exp(-\text{rank}(i)/a) \tag{16}$$

The sum of all weights for all active molecules S is then related to the mean sum $\langle S \rangle$, which is derived from calculations where the active molecules get randomly selected ranks. This leads to the final RIE descriptor (Eq. 17) [88].

$$RIE = \frac{S}{\langle S \rangle} \tag{17}$$

The RIE descriptor describes for how many times the distribution of the ranks for active molecules caused by a VS protocol is better than a random rank distribution. If the VS method is able to score more active molecules higher than a random distribution, the RIE value is greater than 1. In addition, a RIE value of 1 indicates a random rank distribution of the active molecules. On that account, a distribution of all active molecules at the beginning of a rank-ordered list will retrieve a higher RIE value than a distribution with half of the actives ranked at the beginning and the other half ranked at the end of the list. Therefore, the RIE descriptor takes into account the "early recognition problem". However, like the EF the RIE descriptor still suffers from high variability if the ratio of actives in the list changes. Thus, comparisons between RIE values are impaired by the different ratios of active molecules in the databases [72, 88]. Moreover, the RIE descriptor, as well as the EF and other descriptors consider only one selection threshold, whereas the AUC describes the VS workflow performance for all possible thresholds [72, 87].

Boltzmann-enhanced discrimination of ROC (BEDROC). In order to derive a new descriptor that addresses the early recognition problem like the RIE descriptor, but also possesses the advantages of AUC, such as values limited by 0 and 1 or a measurement of the performance above all thresholds, Truchon et al. [72] created the Boltzmann-enhanced discrimination of ROC (BEDROC) descriptor. The BEDROC descriptor is a generalized AUC descriptor that includes a decreasing exponential

weighting function that focuses on active molecules ranked at the beginning of the ordered list. Equation 18 displays the formula of the BEDROC descriptor [72].

$$BEDROC = RIE \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_a)} + \frac{1}{1 - e^{\alpha(1-R_a)}} \approx \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}}, \quad \text{if } \alpha R_a \ll 1 \text{ and } \alpha \neq 0 \tag{18}$$

In order to obtain comparable BEDROC values from VS workflows with different underlying distributions of actives and decoys, αR_a should be smaller than 1. If $\alpha R_a \ll 1$, the BEDROC descriptor is independent from the ratio of actives and can be described by Eq. 10. In this case, the BEDROC descriptor represents the probability that a randomly selected active molecule ranked by a VS workflow will be retrieved before a randomly selected compound from a hypothetical probability distribution function following an exponential of the early recognition parameter α. A high value for α corresponds to a high weighting of the early part of the ordered list. Truchon et al. [72] analytically derived a formal relationship between α and the percentage θ of the total score at z percent of the normalized rank (Eq. 19).

$$0 = \theta(1 - e^{-\alpha}) + 1 - e^{-\alpha z} - 1 \tag{19}$$

With respect to Eq. 19, Truchon et al. [72] recommend a value of 20 for α if the BEDROC descriptor is used for evaluation of VS methods. This means that the first 8% of the relative rank contribute to 80% of the BEDROC value ($z = 8\%$, $\theta = 80\%$) [72].

The accumulation curve produced by the exponential distribution of α serves as comparator for the accumulation curve caused by the VS protocol. Since the limiting values of the BEDROC descriptor are 0 and 1, a BEDROC value of 0.5 indicates that both accumulation curves are equal. In other words, the enrichment of the evaluated VS workflow is equal to the baseline enrichment from which a workflow is useful in VS practice. Taking this into consideration, BEDROC values greater than 0.5 are received for VS workflows that perform better than a workflow, which is useful in solving the respective screening problem of finding a certain amount of actives within a certain number of compounds that are tested for activity. Accumulation curves for distributions with BEDROC values of 0.5 (red line), 0.5–1 (yellow line), and 1 (blue line) are displayed in Fig. 5.

To conclude, the newly developed BEDROC descriptor is supposed to unify the advantages of RIE and AUC [72, 90]. It represents an advanced enrichment descriptor that determines the usefulness of the VS workflow for a specific screening problem. As long as the same early recognition parameter α is utilized and the condition $\alpha R_a \ll 1$ is met,

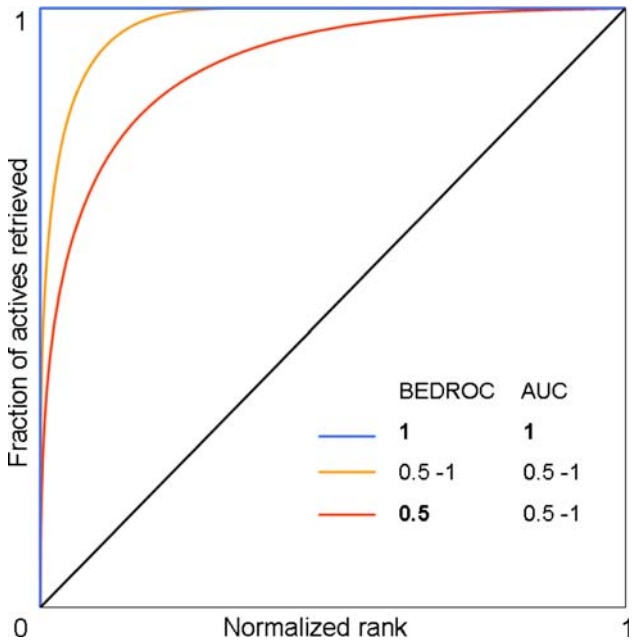

Fig. 5 BEDROC and AUC values for different accumulation curves. An AUC value of 0.5 is received for a random distribution of actives. The accumulation curve for this distribution will tend toward the black line asymptotically with increasing number of actives and decoys. The exponential distribution representing a defined baseline enrichment expected by a useful VS method corresponds to a BEDROC value of 0.5 (red line)

the BEDROC values for different VS workflows are comparable [72]. Therefore, the BEDROC descriptor seems to be a promising evaluation technique for comparing the performances of different VS methods in a screening problem. Recently, McGaughey et al. [90] used the EF, the AUC, the RIE and the BEDROC descriptor for evaluating different VS methods. However, the RIE and BEDROC descriptors did not lead to dramatically different conclusions about the performance of the VS methods [90]. Therefore, the advantage of the new BEDROC metric over the well-known AUC and other classic enrichment factors in screening practice needs to be proven in future studies.

Conclusions

Numerous as the publications on the performance assessment of 3D VS tools are the issues and caveats that we are currently facing during the evaluation of computational methods. However, there are a lot of efforts going on in order to overcome bottlenecks that limit the significance of evaluative performance assessments. We provide guidelines for good evaluation practices based on the investigation of recent literature. All steps of a VS assessment are considered, starting from the selection of a representative VS test set to data analysis of VS campaigns.

Our survey renders the test set collection and preparation for VS as being solved, considering several caveats. Lacking information on the exact software setup frequently prevents required reproducibility and also the multitude of possible manipulations on the data input calls for publications with extended supporting material. We are facing controversies in particular during the analysis of the VS results. Both enrichment parameters and benchmarks for the accuracy of the binding mode prediction are in large part not satisfactory for characterizing the performance of VS approaches and still offer a lot of space for further developments.

We have outlined a list of recommendations that we hope will help authors of future studies to obtain information-rich, representative results and we strongly encourage researchers to publish issues that they are facing during such studies in order to overcome these bottlenecks.

References

1. Klebe G (2006) Drug Discovery Today 11:580–594
2. Schneider G, Bohm H-J (2002) Drug Discovery Today 7:64–70
3. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) J Chem Inf Model 46:401–415
4. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Proteins: Struct Funct Bioinf 60:325–332
5. Kontoyianni M, McClellan LM, Sokol GS (2004) J Med Chem 47:558–565
6. Steindl TM, Schuster D, Wolber G, Laggner C, Langer T (2006) J Comput Aided Mol Des 20:703–715
7. Steindl TM, Schuster D, Laggner C, Chuang K, Hoffmann RD, Langer T (2007) J Chem Inf Model 47:563–571
8. Steindl TM, Schuster D, Laggner C, Langer T (2006) J Chem Inf Model 46:2146–2157
9. Sheng C, Zhang W, Ji H, Zhang M, Song Y, Xu H, Zhu J, Miao Z, Jiang Q, Yao J, Zhou Y, Zhu J, Lue J (2006) J Med Chem 49:2512–2525
10. Güner OF (2000) Pharmacophore perception, development and use in drug design. International University Line, La Jolla, CA
11. Catalyst; Accelrys: San Diego, CA
12. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P (2004) J Chem Inf Comput Sci 44:793–806
13. Weber L (2005) QSAR Comb Sci 24:809–823
14. Wolber G, Langer T (2000) 13th European symposium on quantitative structure-activity relationships, Duesseldorf, Germany, Aug 27 - Sept 1, 2000
15. Schueller A, Haehnke V, Schneider G (2007) QSAR Comb Sci 26:407–410
16. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49: 6789–6801
17. Pan Y, Huang N, Cho S, MacKerell AD Jr (2003) J Chem Inf Comput Sci 43:267–272
18. Jones G, Willett P, Glen RC (1995) J Mol Biol 245:43–53
19. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
20. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45:177–182
21. Edwards BS, Bologa C, Young SM, Balakin KV, Prossnitz ER, Savchuck NP, Sklar LA, Oprea TI (2005) Mol Pharmacol 68:1301–1310
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
23. Wolber G, Langer T (2005) J Chem Inf Model 45:160–169
24. Miller MA (2002) Nat Rev Drug Discovery 1:220–227
25. Knox AJS, Meegan MJ, Carta G, Lloyd DG (2005) J Chem Inf Model 45:1908–1919
26. Moe; Chemical Computing Group: Montreal, QC, Canada
27. Pospisil P, Ballmer P, Scapozza L, Folkers G (2003) J Recept Signal Transduction 23:361–371
28. Oellien F, Cramer J, Beyer C, Ihlenfeldt W-D, Selzer PM (2006) J Chem Inf Comput Sci 46:2342–2354
29. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Comput Sci 38:983–996
30. Flower DR (1998) J Chem Inf Comput Sci 38:379–386
31. Brandstetter H, Grams F, Glitz D, Lang A, Huber R, Bode W, Krell H-W, Engh RA (2001) J Biol Chem 276:17405–17412
32. Yan X, Hollis T, Svinth M, Day P, Monzingo AF, Milne GWA, Robertus JD (1997) J Mol Biol 266:1043–1049
33. Nederkoorn PHJ, Vernooijs P, Donne-Op den Kelder GM, Baerends EJ, Timmerman H (1994) J Mol Graph 12:242–256
34. Tautomer; Molecular Networks GmbH: Erlangen, Germany
35. Pospisil P, Ballmer P, Scapozza L, Folkers G (2004) 15th European symposium on Structure-Activity Relationships (QSAR) and molecular modelling, Istanbul, Turkey, Sept 5–10, 2004
36. Li J, Ehlers T, Sutter J, Varma-O'Brien S, Kirchmair J (2007) J Chem Inf Model 47:1923–1932
37. Kirchmair J, Laggner C, Wolber G, Langer T (2005) J Chem Inf Model 45:422–430
38. Kirchmair J, Wolber G, Laggner C, Langer T (2006) J Chem Inf Model 46:1848–1861
39. Onodera K, Satou K, Hirota H (2007) J Chem Inf Model 47:1609–1618
40. Corina; Molecular Networks GmbH: Erlangen, Germany
41. Carta G, Onnis V, Knox AJS, Fayne D, Lloyd DG (2006) J Comput Aided Mol Des 20:179–190
42. Stahl M, Rarey M (2001) J Med Chem 44:1035–1042
43. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB (2002) J Mol Graph Model 20:281–295
44. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) J Med Chem 42:5100–5109
45. Schulz-Gasch T, Stahl M (2003) J Mol Mod 9:47–57
46. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912–5931
47. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Proteins: Struct, Funct, Bioinf 57:225–242
48. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R (2002) Proteins: Struct, Funct, Genet 49:457–471
49. Goto J, Kataoka R, Hirayama N (2004) J Med Chem 47: 6804–6811
50. Kristam R, Gillet VJ, Lewis RA, Thorner D (2005) J Chem Inf Model 45:461–476
51. Kramer B, Rarey M, Lengauer T (1999) Proteins 37:228–241
52. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470–489
53. Ewing TJA, Makino S, Skillman AG, Kuntz ID (2001) J Comput Aided Mol Des 15:411–428
54. Ewing TJA, Kuntz ID (1997) J Comput Chem 18:1175–1189
55. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) J Med Chem 47:1739–1749
56. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) J Med Chem 47:1750–1759

57. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) J Mol Graph Model 21:289–307
58. Jain AN (2003) J Med Chem 46:499–511
59. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Biopolymers 68:76–90
60. Kirchmair J, Ristic S, Eder K, Markt P, Wolber G, Laggner C, Langer T (2007) J Chem Inf Model 47:2182–2196
61. Teodoro ML, Kavraki LE (2003) Curr Pharm Des 9:1635–1648
62. Claussen H, Buning C, Rarey M, Lengauer T (2001) J Mol Biol 308:377–395
63. Murray CW, Baxter CA, Frenkel AD (1999) J Comput Aided Mol Des 13:547–562
64. Wang R, Wang S (2001) J Chem Inf Comput Sci 41:1422–1426
65. Terp GE, Johansen BN, Christensen IT, Jorgensen FS (2001) J Med Chem 44:2333–2343
66. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11–26
67. Zavodszky MI, Sanschagrin PC, Kuhn LA, Korde RS, Kuhn LA (2003) J Comput Aided Mol Des 16:883–902
68. Kroemer RT, Vulpetti A, McDonald JJ, Rohrer DC, Trosset J-Y, Giordanetto F, Cotesta S, McMartin C, Kihlen M, Stouten PFW (2004) J Chem Inf Comput Sci 44:871–881
69. Abagyan RA, Totrov MM (1997) J Mol Biol 268:678–685
70. Langer T, Hoffmann RD (2006) Pharmacophores and pharmacophore searches. Wiley-VCH, Weinheim
71. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) J Med Chem 48:2534–2547
72. Truchon J-F, Bayly CI (2007) J Chem Inf Model 47:488–508
73. Jacobsson M, Liden P, Stjernschantz E, Bostroem H, Norinder U (2003) J Med Chem 46:5781–5789
74. Shepherd AJ, Gorse D, Thornton JM (1999) Protein Sci 8:1045–1055
75. Gao H, Williams C, Labute P, Bajorath J (1999) J Chem Inf Comput Sci 39:164–168
76. Martineau E, Aman AM, Kong X (2004) Accelrysworld. Accelrys, Inc., San Diego, CA
77. Guha R, Jurs Peter C (2005) J Chem Inf Model 45:65–73
78. Weston J, Perez-Cruz F, Bousquet O, Chapelle O, Elisseeff A, Schoelkopf B (2003) Bioinformatics 19:764–771
79. Ford MG (2003) 2nd international symposium on computational methods in toxicology and pharmacology integrating internet resources, Thessaloniki, Greece, 2003
80. Bradley EK, Miller JL, Saiah E, Grootenhuis PDJ (2003) J Med Chem 46:4360–4364
81. Bradley EK, Beroza P, Penzotti JE, Grootenhuis PDJ, Spellmeyer DC, Miller JL (2000) J Med Chem 43:2770–2774
82. Matthews BW (1975) Biochim Biophys Acta 405:442–451
83. Frimurer TM, Bywater R, Nrum L, Lauritsen LN, Brunak S (2000) J Chem Inf Comput Sci 40:1315–1324
84. Hecker EA, Duraiswami C, Andrea TA, Diller DJ (2002) J Chem Inf Comput Sci 42:1204–1211
85. Diller DJ, Li R (2003) J Med Chem 46:4638–4647
86. Diller DJ, Merz KM Jr (2001) Proteins: Struct, Funct, Genet 43:113–124
87. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) J Med Chem 48:2534–2547
88. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) J Chem Inf Comput Sci 41:1395–1406
89. Park SH, Goo JM, Jo CH (2004) Korean J Radiol 5:11–18
90. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) J Chem Inf Model 47:1504–1519