Recent Trends and Applications in 3D Virtual Screening

Leo Ghemtio\textsuperscript{1,2}, Violeta I. Pérez-Nunez\textsuperscript{1,#}, Vincent Leroux\textsuperscript{3}, Yasmine Asses\textsuperscript{1,#}, Michel Souchet\textsuperscript{4}, Lazaros Mavridis\textsuperscript{1,#}, Bernard Maigret\textsuperscript{1,#} and David W. Ritchie\textsuperscript{*1,#}

\textsuperscript{1}Université de Lorraine and \textsuperscript{*}INRIA Nancy Grand Est, LORIA–UMR 6503, Équipe-projet Orpailleur, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

\textsuperscript{2}Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, Finland

\textsuperscript{3}University of Bordeaux, IECB–UMR 5248, 2 rue Robert Escarpit, 33607 Pessac, France

\textsuperscript{4}Harmonic Pharma, Espace Transfert, 615 rue du Jardin botanique, 54600 Villers-lès-Nancy, France

Abstract: Virtual screening (VS) is becoming an increasingly important approach for identifying and selecting biologically active molecules against specific pharmaceutically relevant targets. Compared to conventional high throughput screening techniques, in silico screening is fast and inexpensive, and is increasing in popularity in early-stage drug discovery endeavours. This paper reviews and discusses recent trends and developments in three-dimensional (3D) receptor-based and ligand-based VS methodologies. First, we describe the concept of accessible chemical space and its exploration. We then describe 3D structural ligand-based VS techniques, hybrid approaches, and new approaches to exploit additional knowledge that can now be found in large chemogenomic databases. We also briefly discuss some potential issues relating to pharmacokinetics, toxicity profiling, target identification and validation, inverse docking, scaffold-hopping and drug re-purposing. We propose that the best way to advance the state of the art in 3D VS is to integrate complementary strategies in a single drug discovery pipeline, rather than to focus only on theoretical or computational improvements of individual techniques. Two recent 3D VS case studies concerning the LXR-β receptor and the CCR5/CXCR4 HIV co-receptors are presented as examples which implement some of the complementary methods and strategies that are reviewed here.

Keywords: 3D-QSAR, 3D shape matching, chemical libraries, chemical spaces, computer-aided drug design, conformational flexibility, docking, knowledge-based drug design, ligand-based drug design, molecular dynamics, pharmacophores, structure-based drug design, virtual screening.

INTRODUCTION

Virtual screening (VS) has become a useful technique for identifying and selecting chemical starting points from large chemical libraries in the endeavor to develop new drug candidates. Several VS "success stories" are now available in the literature [1, 2]. However, it is often difficult to assess the extent to which such leads become marketable drugs due to the long time periods involved and the proprietary nature of the molecules themselves. Nonetheless, examples such as Aggrastat, PRX-0023, PRX-03140, PRX-08066, SC12267, DMP450, and Cevoglitazar highlight the contribution VS has made to the discovery of compounds currently in clinical testing or which have reached the market [3]. The use of three-dimensional (3D) information in VS has now become well established [4-6]. Compared to conventional high throughput screening (HTS) techniques, in silico 3D VS is fast and inexpensive. It can tackle potentially very large numbers (i.e. several millions) of compounds, which would be prohibitively expensive in conventional experimental screening [5-8].

3D VS can be subdivided into two distinct methodological classes, namely structure-based VS (SBVS) and ligand-based VS (LBVS). In SBVS, the atomic coordinates of the protein target structure are either available from experimental measurements or are accessible through molecular modeling. In LBVS, prior knowledge of a reference set of compounds with high affinities for the target is used to derive relationships between the compounds’ properties and their affinities. This allows further similar molecules to be found which might be expected to bind to the target in a similar manner [9-11]. Like SBVS, LBVS depends on the way the molecular properties are modeled, and its usefulness is critically linked to the quality of this knowledge. LBVS queries may be formulated in terms of one-dimensional (1D), two-dimensional (2D), and 3D molecular descriptors. In the present work, we focus only on LBVS approaches that explicitly exploit 3D data (3D-LBVS), and we use the term 3DVS to refer to any 3D SBVS or LBVS method.

Several previous review articles discuss the various aspects of 3D-LBVS [11-16], SBVS [11, 17-23], and the combined use of these two classes of methods [10, 16, 24-27]. Of particular interest is an exhaustive review by Villoiturei et al. of available free resources that can support 3DVS experiments [28]. Several books dedicated to 3DVS are also available [29-31]. More recently, Polgár and Keserü
gave an overview of the major methodologies currently integrated in lead discovery protocols [32], and Sukumar et al. have described the main concepts behind most current SBVS and 3D-LBVS strategies [16]. These publications show that VS now plays an important role in modern drug discovery strategies.

In the present work, we focus on recent trends, developments, and challenges in 3DVS. We first describe the concept of "chemical space" which is important for understanding the scope of the different 3DVS techniques. We then describe the main state-of-the-art techniques in SBVS and 3D-LBVS, and we assess the strengths and limitations of each of them. We also describe hybrid 3DVS methods that bring new approaches and models from computational chemistry, and we discuss new emerging approaches that might be termed "knowledge-based drug discovery" (KBDD). Two recent 3D VS case studies concerning the LXR-β receptor and the CCR5/CXCR4 HIV co-receptors are presented as examples, which implement some of the complementary methods and strategies that are reviewed here.

THE CHEMICAL UNIVERSE

It has been estimated that in principle there could exist some $10^{60}$ distinct and chemically feasible small molecules (mass < 500 Daltons) [33]. This huge number far exceeds the size of current compound databases. For example, the CAS Registry, which is one of the largest databases, contains less than $10^8$ commercially available inorganic and small organic molecules [34]. Furthermore, since it may be estimated that the sun has some $10^{57}$ hydrogen atoms, the size of this theoretical "chemical universe" greatly exceeds the capability of even the most ambitious combinatorial chemist [35]. It is therefore safe to say that the ensemble of all currently known compounds probably do not represent a fair sampling of the chemical universe due to synthetic and biosynthetic tendencies to re-use existing scaffolds (see Fig. 1). Indeed, analyses of some existing compound collections have highlighted artificially frequent scaffolds or "chemical clichés" [36]. Furthermore, the existence of distinct signatures in corporate pharmaceutical catalogues has been demonstrated, suggesting that the "organizational influence" in current industrial drug discovery projects may bias our exploration of chemical space [37]. In any case, it is widely accepted that only a small fraction of the available chemical universe has been explored in drug discovery.

The apparent lack of productivity of the pharmaceutical industry is fueling a heated debate [38, 39], and without doubt the current model of pharmaceutical research is facing a crisis. Fewer new drugs are being brought to the market due to high attrition rates in clinical trials [40], leading to a sharp rise of drug development costs [38]. As the so-called "patent cliff" (the pharmaceutical industry’s equivalent of the "peak oil" apocalypse) approaches [41], the whole industry is being reshaped [42, 43]. Among many plausible explanations for this situation, the very limited size of the space accessible for drug discovery investigations is often cited. In this context, devising new ways to explore chemical space more intelligently may lead to better and more rationally designed compound libraries. While this would not cure all of the industry’s ills, it could certainly lead to significant enhancements of the drug discovery process [44].

However, the in silico exploration of uncharted regions of chemical space requires new and improved ways to exploit recent advances in computing performance, and should be guided rationally by keeping in mind any pending constraints in the discovery pipeline such as pharmacokinetic profiles [45-47]. Reymond et al. reviewed the different aspects of exploring chemical space computationally and the

Fig. (1). Chemical space and high throughput VS.
resulting potential to improve the drug discovery process [48]. However, as in the bioinformatics arena, the increasing exploration of chemical space will also bring problems of "data deluge" [49]. For example, it has been shown that more than 26 million molecules may be constructed from just 11 C, N, O and F atoms [50]. As a typical drug molecule is twice as big [51], the size of the "combinatorial drug space" is in the region of some $10^{100}$ molecules [52, 53]. Nonetheless, several strategies are currently being explored to cover a more adequate and optimal sub-space, including better exploitation of natural products, combinatorial chemistry and compound collections, and the re-use of existing active molecules.

Natural Products

Throughout history, nature has been the source of new medicines. It is only recently that this source has been complemented by synthetic chemistry. Half of all anti-cancer molecules marketed from the 1940s to 2006 were natural products or derivatives [54]. Chemical diversity in nature is based on both biological and geographical diversity, but more importantly it is also the result of billions of years of evolution. On the other hand, genetic engineering and biosynthetic techniques are increasingly being used to generate natural product-like scaffolds. For example, improved synthetic methods can now build the complex ring structures of some natural products [55, 56]. Furthermore, recent diversity-oriented approaches can generate large collections of diverse and complex molecules from simple starting blocks [57]. A structural classification of natural products (SCONPs) has been defined on the basis of the underlying scaffolds present in natural products [58]. This new resource could do for natural product space what the Protein Data Bank [59] has done for the proteome.

Combinatorial Chemistry and Compound Collections

It is nowadays common practice in the pharmaceutical and agrochemical industry to design and synthesize arrays of structurally related molecules combinatorially [52, 60]. In this regard, combinatorial chemistry approaches have been described as "engines of diversity" [61, 62]. The use of combinatorial chemistry has led to constantly growing compound databases from both industrial and academic sources [63]. Among such resources (see Table 1), the CAS and ZINC databases [64] are the most well-known. The freely available ZINC database gathers over 13 million commercially available compounds from heterogeneous catalogs in a single convenient format suitable for use directly by most current 3DVS software [65]. In the academic world, efforts are also being made to gather and update digital molecular libraries in an approach comparable to corporate practices [66]. Additional valuable information on the chemical, pharmaceutical, and pharmacological status of molecules at various stages of compound development can also be obtained in several cases [67]. Koutsoukas et al. have reviewed the databases currently available and in development, and give an overview of their potential to predict protein targets for small molecule ligands [68]. The increasing availability of such large data sets now offers an opportunity to speed up the drug development process, and to learn from such data and thus improve our knowledge. However, treating very large data volumes is challenging, not least because they should be filtered by ADME-Tox (i.e. absorption, distribution, metabolism, excretion, and toxicity) properties before applying VS [45, 69].

While the quantity of data from compound databases continues to rise, this does not imply that data quality is also improving. Furthermore, the actual physical availability of chemical compounds is still often variable and unpredictable. One should always distinguish between compounds that have been synthesized and are available for experimental testing, and those compounds that should be easy to obtain or make. Even if the latter "virtual compounds" really are available, they may not have sufficient purity for experimental investigations [70]. For VS experiments, compounds that are readily at hand for biological testing or virtual combinatorial databases guided by the knowledge of reliable organic chemists should therefore be preferred.

Exploiting Existing Drugs

According to the 1994 Nobel Laureate Sir James Black, "the most fruitful basis for the discovery of a new drug is to start with an old drug" [71]. Several marketed drugs embody this principle, such as the β-blockers, the tricyclic antidepressants, the statins, and the 1,4-dihydropyridine calcium channel blockers. Such derivative "me too" drugs are produced in response to economic and competitive demands. No doubt, some are developed simply because it is far cheaper to do this than to develop a novel therapeutice compound from scratch. On the other hand, the need for drugs similar to existing ones but with a better pharmacokinetic and pharmacodynamic profile is a significant motivation [72]. Sir Black’s maxim also embodies the notion of drug re-purposing. For example, miltefosine was originally developed for breast cancer but it is now also used to treat visceral leishmaniasis [53, 73]. It should be noted that repositioning old drugs, either from rational observations or serendipity, further characterizes target knowledge and could thus facilitate the discovery of further novel active compounds.

There are various approaches for profiling existing drugs with known primary and secondary effects, such as "Selective Optimization of Side Activities" (SOSA) [74]. Thus, established strategies such as structural optimization by molecular modeling may eventually transform a "side effect" into a new primary activity. When such protocols are used in an industrial context, the notion of "patent space" also arises. Exploring the repurposing of a given drug while avoiding the patents of its manufacturer as well as those of the competitors may be particularly difficult, as the patent space around some molecules can be crowded. In any case, a more rational way of exploring chemical space than "blind" combinatorial search would be desirable.

Fragment Assembly

Amongst the rationally guided combinatorial chemistry strategies, fragment-based assembly [75-77] is undoubtedly the most popular approach, especially within an existing drug discovery pipeline. Fragment-based assembly assumes that molecular complementarily may be achieved using
small molecular fragments of up to 12 heavy atoms, with combinations of such fragments giving a wide coverage of chemical space [76]. Therefore, fragment-based drug design assumes that it is possible to map interesting drug-like properties from chemical space onto a small set of molecular building blocks. In order to maximize this possibility, a "rule of three" for "fragment drug-likeliness" was devised (i.e. \( M_r < 300 \text{ Da} \), \( c\text{LogP} < 3 \), H-bond donors < 3, acceptors < 3, rotatable bonds < 3, polar surface area < 60 Å) in analogy to the well-known Lipinski and Oprea rules for drug-like/lead-like molecules [47]. A further such "rule", might also be a requirement that the solubility of fragments should be sufficient (1 mM) to allow for high-concentration screening in aqueous buffer [77].

Another related approach consists of deriving fragments from known drugs [78], although this might exacerbate the above problems of "chemical clichés" and "organizational influence" [36, 37]. In settings involving close collaborations between molecular modelers and organic chemists, a more focused fragment-based approach consists of starting from well-validated scaffolds with no more than 4 groups that may be substituted according to the chemist's knowledge. Typically several hundreds or thousands of possibilities are obtained, most of which are plausible from a chemical synthesis point of view. Provided the starting scaffolds are related to known active compounds, the chances of finding a hit using VS are good. Eventually, collaborations between the experimentalists and theoreticians may be pushed further as the chemists propose new compounds from on-going VS results, leading to further refinements until the predictions converge [79].

### 3D-LBVS

3D ligand-based (3D-LBVS) approaches assume that structurally-related molecules will have similar biological activities [80]. While this assumption is often reasonable, sometimes small structural or conformational changes can lead to large differences in activity. Furthermore, structurally similar active molecules may exhibit different binding modes on a given target, and this can limit the usefulness of 3D-LBVS techniques [81, 82]. Hence, it is important to consider carefully the number and relevance of any computationally generated molecular conformation [83]. The following sections discuss some concepts, issues, and advances in 3DVS techniques that are currently of interest. Table 2 lists a selection of related approaches.

### Similarity-Based Approaches

Similarity searches use descriptors to match members of a compound virtual library with the corresponding descriptors of a query molecule [84-86]. The growing
interest in molecular similarity techniques mirrors the growth of available structural databases. While simple metrics such as the Tanimoto coefficient are often considered to be sufficient for comparing descriptors, there is no universally accepted way to define the descriptors themselves, especially for 3D information. Hence many different ways to compare molecular shapes and their properties have been proposed [87-89]. Unfortunately, different molecular descriptors capture different aspects of molecular structures, and therefore can rank a given set of molecules differently in VS. This suggests using principal component analysis (PCA) or other data fusion techniques to combine multiple descriptors in prospective VS [90].

Some examples of such approaches are described further in the following sections.

### 3D Fingerprints

As with human fingerprints, the fingerprint calculated from a molecule’s geometry may be sufficiently detailed that comparing two molecular fingerprints may be almost equivalent to comparing the molecules themselves. Several popular molecular fingerprint approaches such as Daylight, Unity, and MACCS use long bit strings to encode 1D or 2D projections of molecular topology along with macroscopic properties such as pKa and logP [91]. This information can be complemented with additional structural descriptors. For example, a hybrid descriptor composed of the MACCS key coupled with Ballester and Richards' shape recognition data has demonstrated its ability to identify active molecules [92-94].

### 3D Pharmacophore Models

A pharmacophore may be defined using a set of distance constraints to relate the relative positions of common chemical features (e.g. H-bond donors/acceptors, aromatic rings, partial charges, hydrophobic regions) in a series of active molecules with respect to one or more targets. Designing a pharmacophore model involves identifying a spatial consensus of key interaction sites on the assumption that it will match the conformation of a bound ligand within the target receptor binding site [95, 96]. One advantage of this strategy is that it can be used without requiring knowledge of a target structure, or even its identity. Superposing conformational ensembles of known active compounds may be sufficient to suggest common features that explain their activity [97]. The process of building pharmacophore models has been automated in several computer programs (e.g. Catalyst [98], PHASE [99], and Sybyl [100]), although user intervention is often still necessary to pick and validate good models.

<table>
<thead>
<tr>
<th>Class</th>
<th>Name</th>
<th>Methodology/Model</th>
<th>Spatial Exploration Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-LBVS</td>
<td>3D fingerprints-based screening</td>
<td>Alignment/comparison of molecular fingerprints encompassing structural information in addition to molecular properties</td>
<td>Diverse</td>
</tr>
<tr>
<td></td>
<td>Molecular fields comparison</td>
<td>Superimposition/comparison of molecular fields modeled from the projection on an encompassing 3D grid of the properties of molecule conformations(s)</td>
<td>Focused</td>
</tr>
<tr>
<td></td>
<td>Ligand shape matching</td>
<td>Superimposition/comparison of molecular surfaces (various kinds available, e.g. grid-, Gaussian-, spherical harmonics-based)</td>
<td>Diverse</td>
</tr>
<tr>
<td></td>
<td>3D pharmacophore matching</td>
<td>Screening for a ligand conformation matching a pharmacophore built from the conformational space of reference compounds</td>
<td>Focused</td>
</tr>
<tr>
<td></td>
<td>3D-QSAR</td>
<td>Implementation of information derived from the ligands conformational space into QSAR models</td>
<td>Focused</td>
</tr>
<tr>
<td>SBVS</td>
<td>Protein-ligand docking</td>
<td>Exploration of the ligand conformational space in order to find a relevant bound conformation on more or less flexible (depending on model) target(s). Computation of the binding affinity is approximate.</td>
<td>Diverse</td>
</tr>
<tr>
<td></td>
<td>Binding site similarity</td>
<td>Comparison of simplified representations of macromolecular binding sites, followed by focus on active molecules for similar receptors, assuming they may also be actives on the query.</td>
<td>Focused</td>
</tr>
<tr>
<td></td>
<td>Molecular dynamics</td>
<td>Whole target structure and ligand atomic-level properties represented using pre-defined molecular mechanics forcefield parameters, then complex system dynamics computed iteratively.</td>
<td>Focused</td>
</tr>
<tr>
<td>Other 3DVS Approaches</td>
<td>Fragment-Based VS</td>
<td>Identification of small chemical fragments, which may bind only weakly to the biological target, and then growing them or combining them to produce a lead with a higher affinity.</td>
<td>Diverse/Focused</td>
</tr>
<tr>
<td></td>
<td>Chemogenomic</td>
<td>Identification of novel drugs and drugs target, embracing multiple early phase drug discovery technologies ranging from target identification and validation, through compound design and chemical synthesis, to biological testing and ADME profiling.</td>
<td>Diverse/Focused</td>
</tr>
<tr>
<td></td>
<td>Knowledge-Based Drug Discovery</td>
<td>Data modeling and knowledge extraction from protein and/or ligand information’s store in the database or available in literature to guided new drug discovery.</td>
<td>Diverse/Focused</td>
</tr>
</tbody>
</table>
Comparing Molecular Fields

The first 3D molecular similarity measures to be developed typically matched groups of atoms using a variety of geometric criteria. However, such approaches may miss topologically different molecules that exhibit similar biological effects. For example, the opioids have a rigid polycyclic structure but they bind to the same receptors as the natural endorphin ligands which are oligopeptides [101]. Thus, two molecules can possess similar biological properties despite having fundamentally different chemical formulae and covalent structures. Therefore, there has been considerable interest in developing techniques to compare 3D molecular field properties such as the molecular electrostatic potential (MEP) as a way to avoid the limitations of atomistic models [102-104]. Most current field-based approaches use 3D grids to sample and compare MEPs, but this can be computationally expensive compared to other 3D-LBVS techniques. Furthermore, the global superpositions obtained can sometimes be misleading due to the pitfalls of using atomistic root-mean-square deviations (RMSDs) to define the initial structural superposition [105], and because a global superposition may not correspond to the best overlap in a specific region of interest [103].

Molecular Shape Matching

Molecular shape plays a significant role in protein–ligand interactions, and many researchers agree that shape-matching algorithms show promising potential among state-of-the-art 3DVS techniques [9, 106]. Most drugs bind to their target through non-covalent van der Waals interactions, and hence desolvation effects dominate the binding free energy. Unlike hydrogen bonds and electrostatic interactions, the strength of these hydrophobic interactions is mainly governed by the proximity of the respective electron densities. Therefore, it is often assumed that molecules that share similar 3D surface shapes are likely to share similar drug–like properties. Thus, 3D LBVS shape comparison techniques could offer a “scaffold-hopping” way to find novel or unexpected drug leads [107-109]. Nonetheless, there still remains the problem of how to take molecular flexibility into account. For example, using just one single calculated conformation per candidate [110] could neglect important conformational changes that might be induced by binding. On the other hand, re-using a crystallographically determined target-bound conformation can lead to similar problems because a given molecule may bind to different targets in different ways. Nonetheless, if crystal structures are available, they can provide a useful way to validate computationally generated ensembles because at least one member of the ensemble should resemble the crystallographic conformation. Therefore, it is normally desirable to generate a diverse ensemble of low-energy conformers despite the consequent computational cost of comparing multiple conformations [111].

Gaussian Shape-Density Representations

Grant et al. developed an elegant and efficient approach to represent and compare molecules using Gaussian density functions [112]. As implemented in the ROCS program, each molecule is first aligned with the Cartesian coordinate axes by diagonalizing a matrix of second order steric multipoles calculated from atom-centred Gaussian functions. By then minimizing the sum of atomistic Gaussian overlaps, ROCS can superpose pairs of molecules at a rate of about $10^3$ per second per central processor unit (CPU) [113].

Spherical Harmonic Molecular Surfaces

The use of spherical harmonics (SH) functions for VS [114-116] is a more recent approach [117, 118]. The ParaSurf and ParaFiT programs [114] were developed for 3D-LBVS, while the MSSH and SHEF programs were developed for fast shape-based protein-ligand docking [119]. These approaches use SH polynomial expansions to represent the surface shape and other molecular surface properties, and they exploit the special rotational properties of SH functions to perform fast Fourier-based correlations [23]. Consequently, the speed of ParaFiT is similar to ROCS [117], and the high efficiency of SHEF in high-throughput receptor-based VS has been demonstrated [120].

3D-QSAR Approaches

3D-QSAR methods aim to correlate the chemical structures and biological activities of a series of related compounds [121]. They use a number of location-dependent measures that describe molecular properties without any explicit calculation of how they might interact with a target. 3D-QSAR models may be derived from different combinations of structural and spatial descriptors using linear and non-linear optimization techniques such as genetic algorithms, neuronal networks, and multivariate analysis. This often involves calculating the autocorrelation of ligand surface MEPs and receptor-ligand MIPs, for example [122]. One of the oldest but still widely used 3D-QSAR approaches is comparative molecular field analysis (CoMFA). Using the same principal as MEP comparisons, CoMFA uses a variety of atomic probes to compute and map molecular properties on a 3D grid [123]. CoMFA mainly compares the enthalpic and structural contributions to protein-ligand binding on the assumption that the entropic contributions to the binding free energy will be similar for ligands having similar probe potentials [124].

However, using sharp Lennard-Jones potentials directly in CoMFA can introduce errors in scaling, alignment, and interpretation [125]. In order to overcome such problems, the comparative molecular similarity indices approach (CoMSIA) was developed. Current CoMSIA models are less sensitive to alignment variations and use potentials which model better the entropic aspects of binding [124, 126]. The use of 3D graphics then allows the most interesting features of chemical fragments to be identified, and helps the relationships between CoMFA/CoMSIA and pharmacophore models to be visualised [127] and refined [25, 99, 128, 129]. However, the initial alignment step in CoMFA/CoMSIA is often time-consuming and can introduce bias in the subsequent QSAR modeling step. Hence, further approaches such as ALMOND have been developed which use grid independent descriptors calculated by the GRIND program [130].

STRUCTURE-BASED VS

Structure-based VS (SBVS) approaches use the 3D structure of the target to guide the selection of active
compounds from a molecular database. If it is not already known, the location of the active site can be inferred from mutagenesis data, by homology to known structures, or by using a pocket-finding algorithm such as PocketFinder [17, 131, 132]. SBVS techniques mainly involve protein-ligand docking studies, but molecular dynamics (MD) simulations may also be used as a more precise investigation tool.

**Binding Site Similarity**

If it is assumed that similar ligands might bind in similar ways to a given receptor. It might equally be supposed that similar binding sites might bind similar ligands. Therefore, several binding site comparison algorithms have been developed. Some examples include SiteBase [133] (atomic coordinates comparison), PIPSA [134] (molecular interaction fingerprints), GRID/PCA [135] (principal component analysis on grid representations of receptors), and FLAP [136] (grids + molecular interaction fingerprints). More generally, the assumption that similar binding sites provide similar selectivity profiles to an ensemble of small molecules is equivalent to the assumption that similar ligands bind to their targets in similar ways. Thus, the pharmacophore concept can be applied in a structure-based fashion, starting from receptor conformations rather than from ligand conformations. Receptor conformational space is much larger than small molecules space, requiring significant computational power in the VS context. Consequently, binding site similarity-based SBVS is a relatively new approach [17].

**Protein-Ligand Docking**

Protein-ligand docking techniques evaluate the potential activity of a library of compounds positioned in the binding site of the receptor. Docking allows the orientation and conformation that a ligand adopts in the binding site to be identified. It can also be used to identify binding sites (blind docking) [137], but this is less common. Docking algorithms are often characterized by the combination of the method used to sample the ligand-receptor conformational space and the scoring function used to evaluate the strength of the interaction [20]. It is acknowledged that most docking scoring functions cannot accurately predict free energies even if applied to correct binding poses [18, 138-140]. A recent community benchmark demonstrated that none of the current empirical, knowledge-based, or force field scoring functions could accurately calculate the binding energies of all members of a diverse set of crystallographic protein-ligand complexes [141]. On the other hand, it has been shown that developing target-specific scoring functions can still give some significant improvements [142].

Although it clearly seems desirable to develop better docking potentials, this alone will not necessarily ensure better VS performance because there is always a risk that docking searches might become trapped in local minima. The close interdependence between the search algorithm and the scoring function [143] complicates the development and evaluation of docking algorithms [144, 145]. It might even be supposed that using a generic scoring function for compound selection is so case-specific that it could be no more rational than random picking [146]. On the other hand, perhaps a more sensible way to use protein-ligand docking programs is to expect less from them. More specifically, if it is assumed that a docking program might only suggest several plausible binding modes, and then it might be better to use docking only to eliminate infeasible candidates. The surviving candidates may then be refined using increasingly precise scoring functions in a hierarchical manner [147] using force field-based molecular mechanics calculations [148-150], projections into interaction fingerprints [105], clustering techniques [151], compact descriptors for representing favorable protein interaction spots within the protein binding [152], interactive analysis using molecular modeling software [153], and consensus scoring techniques [154, 155]. For example, it has been shown that taking the intersection of top-scoring ligands from two or more scoring functions can often be advantageous [156, 157].

**Dealing with Receptor Flexibility**

Current flexible protein-ligand docking programs include GOLD [158, 159], GLIDE [160, 161], FlexX [162, 163] AutoDock [164, 165] and Vina [166, 167]. Although these approaches can explore ligand flexibility extensively, they generally assume that the protein receptor is rigid. Several comparative studies of these approaches have been carried out [140, 168-170]. In principle, exploring the torsional degrees of freedom in both binding partners should be able to predict receptor induced fit effects [171-174], but this is easier said than done. Firstly, introducing flexibility in the receptor dramatically increases the size of conformational space that must be searched. Furthermore, small structural changes in the binding site can sometimes have a large effect on the docking result [175, 176]. Nonetheless, some recent algorithms are beginning to address such challenges [177].

However, if designing relevant scoring functions in the semi-flexible context is already an intrinsic limitation of docking approaches, the situation gets even worse if the receptor flexibility is allowed. Recently, new docking protocols aiming to tackle the “dynamic personality” of proteins [178] have been developed which use ensembles of receptor conformations [179] that may be considered equivalent in terms of internal energy and can therefore be evaluated concurrently [180]. The advantages and drawbacks of ensemble-based methodologies have been discussed by Amaro et al., especially for receptors with high degree of flexibility [181]. Initial investigations of such models suggested that docking performance may be significantly improved, but that care is required during the analysis of results as more false-positives seem to be generated [182].

**Entropic Contributions to Protein-Ligand Binding**

The rigorous evaluation of entropic fluctuations during macromolecular associations is a long-standing challenge in molecular modeling [183]. If receptor flexibility is a major concern for docking search algorithms, evaluating the entropic contribution to the binding free energy is equally important at the scoring stage. However, performing rigorous free energy simulations is non-trivial [184] because it requires both an accurate molecular mechanics force field and extensive sampling using specialized procedures [185, 186], both of which require a good degree of expertise and
computational power. The sampling requirement is crucial because entropy is physically rooted in statistical thermodynamics. In this sense, protein flexibility and the entropic effects are intimately related, especially if ligand-binding changes the number of available torsional degrees of freedom. Hence free energy calculations may be useful for giving appropriate weight to individual receptor conformations, as well as to providing better free energy estimates than docking based scoring functions. Ongoing methodological developments [187, 188] and systematic validation against calorimetric experimental data [189, 190] may eventually allow for a more routine use of free energy calculations.

**Molecular Dynamics**

MD simulations [191, 192] are mainly used to study the conformational space of biological macromolecules and small compounds. Current MD programs includes AMBER [193], CHARMM [194], GROMACS/GROMOS [195, 196] and NAMD [197, 198]. Such programs may also be used to take into account explicitly the effect of solvent molecules on protein-ligand complexes [199] and receptor flexibility, which are both difficult in conventional docking programs [18, 138, 139]. MD may also be used during the refinement of experimentally derived structures [200, 201].

MD is nowadays a well established molecular modeling technique. Interestingly, MD and the associated molecular mechanics theoretical framework constituted a vivid landscape 20 years ago [202], much as can be observed regarding current 3DVS approaches [203]. Although the core concepts in MD are now mature, there continue to be improvements in force field models [204] and other computing related innovations [205, 206]. With recent advances in high performance computing and graphic processor units, the computational demands of MD are no longer a limiting factor. However, it still requires highly skilled staff to set up a MD simulation and to analyse the results. Indeed, the man-power effort required to simulate a single protein-ligand system using MD roughly corresponds to that of screening thousands of compounds against a given receptor using docking techniques. In this sense, MD may be considered as the last and heaviest weapon in the computer-aided drug discovery arsenal. However, a possible limitation of MD is the shortage of parameters for drug-like molecules. The CHARMM General force field (CGenFF) is in active development, and currently includes a good number of drug-like fragments [207, 208]. The general AMBER force field (GAFF) provides a useful alternative, as its less specialised parameters can be used on a wider range of molecules [209, 210]. Additionally, several public web servers such as Automated Topology Builder [211] (for GROMACS, available at http://compbio.biosci.uq.edu.au/atk/), ParamChem/CGenFF [204] (for CHARMM, http://www.paramchem.org) and SwissParam [212] (for CHARMM/GROMACS, http://www.swissparam.ch) can provide force field parameters for new molecules.

**Water and Solvation Effects in 3DVS**

Water is one of the most important yet least well known biological components [213]. Water molecules and solvation effects play a key role in influencing the structure and dynamics of a binding site and in mediating protein–ligand recognition [11, 214]. However, despite its chemical simplicity, accurately simulating interactions involving water is difficult even when using quantum mechanical levels of theory [215], and the hydrophobic effect in bio-molecular processes is still not well understood [216]. The importance of explicitly considering the involvement of water molecules for successful SBDD has been summarized well by Ladbury [217] and a recent review on this topic is also available [218]. In the context of protein-ligand docking, it is generally assumed that a potent binder should displace all water molecules initially present at the receptor surface, restoring them back to the bulk solvent. But a water molecule making multiple hydrogen bonds may provide an enthalpic contribution that out-weighs the entropic cost of its immobilization. Hence such permanently bound waters should be identified before using a docking program. Several programs are dedicated to this task [219]. Among those, DOWSER [220] "docks" waters into cavities, which is particularly useful as a first step for water box solvation prior to explicit-water MD equilibration [201]. AcquaAlta [221] is based on knowledge of an extensive mining of X-ray data from the Cambridge Structural Database [222] along with ab initio calculations. JAWS (for "Just Add Water molecules") [223] uses Monte-Carlo equilibration to find optimal water positions on a grid. On the other hand, favorably bound receptor waters should not be considered as unmovable for three reasons. Firstly, a ligand may still displace them if it provides sufficient energetic compensation upon binding. Secondly, bridging waters identified from X-ray structural data of protein-ligand complexes may be a by-product of crystallization conditions. Thirdly, a water molecule may be trapped in an energetically unstable position (e.g. near hydrophobic residues and not making many hydrogen bonds) if it locks a favorable protein conformation or simply fills a void [224]. Such waters define highly druggable hot spots that should be prioritized for desolvation when designing potent binders [225]. Some docking programs like GOLD can incorporate multiple possible water positions and turn them on or off as the calculation evolves [226]. This approach may lead to predictive improvements, but it requires human expertise and intervention for best results [227, 228]. Other current computational approaches can assess the impact of water molecules on ligand binding using only empirical or implicit methods. Some examples include SuperStar [229], knowledge-based approach [230], the HINT and Rank algorithm [231], Multivariate statistical analysis [232], and CS-Map [233]). Such approaches allow the effect of solvation on ligand binding to be estimated without the human and computational expense of MD.

**HYBRID 3DVS APPROACHES**

In addition to SBVS approaches, several hybrid approaches have been proposed to improve the performance of 3D-LBVS when structural knowledge of the receptor is also available. As binding modes are often conserved amongst different ligands, it is natural to use information extracted from complexes co-crystallized with other ligands. This approach is known as direct guided docking [234]. Another approach is the construction of 3D ligand pharmacophore models augmented by structural information...
from the active site of the receptor in other co-crystalized ligand-protein complexes. For example, LigandScout [235] builds 3D pharmacophores from crystallographic structures using seven types of chemical features and volume constraints [236]. An excluded volume sphere can be used to mimic the shape of the binding site. Receptor pharmacophores produced using software such as LigandScout [235], Catalyst [98], Pocket v.2 [237], and GBPM [238] are useful for establishing 3D patterns for 3DVS databases of ligand conformers. Furthermore, the pharmacophore type constraints in the receptor may be included as a constraint during docking by FlexX-Pharm, for example [239]. 3D-QSAR models can also be derived this way, but achieving good structural alignments of the ligands can be challenging when dealing with structurally diverse or highly flexible compounds.

The use of alignment techniques based on binding site geometry and minimization within binding sites can be an effective alternative to conformational searches or superposition of a 3D map of molecular properties. Pérez-Nueno et al. recently introduced the notion of SH-based “consensus shapes” to help deal with these problems [118]. Several recent studies have shown that some protein targets bind different ligands in different ways [240, 241]. Hence it is worth using VS approaches, which can help to associate specific sub-sets of ligands with their corresponding receptor binding sites. In addition, the consensus shape approach allows one or more “pseudo-molecules” to be created and used as VS query structures. In a previous study on 15 diverse families of CCR5 inhibitors, which could not all be superposed together, Perez-Nueno et al. found that the ligands may be clustered into four main super-consensus families and might bind within three sub-sites in the CCR5 extra-cellular pocket. These results were consistent with experimental site-directed mutagenesis information and other computational studies [117]. Thus, consensus clustering may offer a straightforward way to understand how multiple ligands might distribute themselves within a given binding site, using computational algorithms much faster than alternate protein-ligand docking.

Another new and fast hybrid approach is to calculate protein-ligand interaction fingerprints. This method encodes a 1D representation of protein-ligand interactions or predicted docking poses. Such fingerprints can therefore be used as templates for designing or filtering compound collections and for rescoring docked ligand poses. For example, the structural interaction fingerprint (SIFt) approach is one of the first methods for representing and analyzing 3D protein-ligand binding interactions [242, 243]. More recently, similar approaches have been combined with machine learning techniques [244, 245]. However, these methods often require human expertise to select the ligand and protein complexes and the target space to be considered [19].

**OTHER 3DVS APPROACHES**

**Fragment-Based VS**

Fragment-based VS (FBVS) is based on the identification of small molecular fragments from libraries. To date, this approach is well established in both academic and industrial research, with many success stories demonstrating it as a solid complement to HTS. However, as discussed above, the quality of the screening libraries is important because selected hits can sometimes turn out to be unavailable. FBVS approaches are able to overcome this problem by their ability to adapt a weaker ligand into an optimized hit. Despite a relatively low throughput, this is a valuable strategy with a particular ability to quickly explore chemical space. The approach needs a library of interaction fragments and the 3D structure of the target with a clear and well-defined binding surface. Fragment-based assembly techniques can be extended when a biological macromolecule is used as a template on which molecular fragments assemble and form covalent bonds to make the active molecule [246]. This is sometimes called “click chemistry”. One example of this approach was the synthesis of a femto-molar AChE inhibitor from the smaller tacrine and phenanthridinium motifs [247]. The recent review of Congreve et al. discusses several cases in which FBVS was able to propose more compounds at the hit prioritization stage, and therefore increase the chances of success in the following stages [75]. This highlights the potential of FBVS to complement the use of HTS in drug discovery.

**Chemogenomic Approaches**

Chemogenomic approaches aim to profile a small molecule against a large collection of macromolecular structures to identify proteins which may bind to it (target fishing), or determine its global pharmacological profile (ligand profiling) as complementary means to increase drug discovery productivity [69]. The chemogenomic application of several 3DVS approaches has been reviewed by Rognan [19], Ekins et al. [248], and Bajorath et al. [249] Most approaches aim to identify new ligands for a given biological target. On the other hand, in inverse VS, a single ligand is compared against a collection of targets, and much effort has been made to make target fishing and ligand profiling fruitful for drug discovery [22, 138, 244]. Developing chemogenomic approaches will certainly help to identify new molecular targets for existing drugs or for compounds in clinical trials. This is often called “drug repositioning” or “drug repurposing”. For example, in a recent large-scale study of some 3655 approved drugs, Keiser et al. found 23 confirmed new drug-target interactions of which 5 had high affinities (< 100 nM) [250]. This study is clearly a milestone in drug repositioning, and it further highlights the corresponding need to develop new approaches which can aggregate large volumes of polypharmacological data. Hence, building a framework to integrate knowledge from diverse sources would be highly desirable. Aiming to answer this need, semantic web technology has recently been used to develop a drug-disease ontology in order to link to different diseases [251]. In addition, one of the most common reasons that drug candidates are abandoned is their poor pharmacokinetic and ADME-Tox properties [252, 253]. Hence, chemogenomic methods are now being used at an early stage for modeling and predicting pharmacokinetic properties of compounds [19, 248, 249]. These properties can be profiled to focus screening and testing on only the most promising compound to reduce the risk of late-stage attrition and thus to reduce overall costs [254]. The increasing availability of experimental data on small...
Compounds will be used with \textit{in silico} 3D based methods to provide models that allow the prediction of toxicity and several crucial ADME-Tox properties such as clearance, volume of distribution, various aspects of metabolism, transporters, and plasma protein binding, for example [255, 256].

\textbf{Knowledge-Based Drug Discovery}

The enormous quantities of data in current biological and chemical databases can only be managed and analysed using automated techniques. Knowledge-based drug discovery (KBDD) approaches aim to extract novel and useful "knowledge units" from large data sets [257]. Data mining is the core of the KBDD process, and involves applying algorithms to explore the data, and to discover significant patterns that can be turned into knowledge units [257-260]. This is illustrated in Fig. (2). Recent work by Ghose \textit{et al.} has demonstrated how a KBDD approach was able to help medicinal chemists select and optimize a better lead for a central nervous system (CNS) disease by using a classification tree to differentiate between CNS and non-CNS oral drugs [259]. Thus, by applying machine-learning techniques to 3D information about proteins, ligands, and their interactions, KBDD offers a way to enhance 3DVS [229, 261-265]. For example, Ghemtio \textit{et al.} applied association rules to a 3D database of the LXR-\(\beta\) protein and its ligands to define ligand-based VS queries [266]. In this study, the KBDD approach gave better results than conventional 3D pharmacophore and shape-based approaches. However, to achieve this, the KBDD process had to be trained on a carefully selected data set representing active and inactive compounds, as it is not clear how best to automate the training step. Several recent studies showed how KBDD could be used to guide new drug discovery [267, 268]. Furthermore, KBDD could be used with protein-ligand interaction to design target focused libraries [263], scoring functions to improve prediction power of SBVS approaches [265, 269, 270] or on databases of known active ligands to derive structural, physicochemical and ADME–Tox property profiles of successful ligands [271]. Finally, prediction of the 3D structures of proteins to aid the design of drugs may be guided with KBDD approach by identification of similarity or ligand interactions between a protein of interest and those of known 3D structures [262].

\textbf{COMBINED USE OF DIVERSE 3DVS STRATEGIES IN A GLOBAL DISCOVERY PIPELINE}

VS protocols are often adapted as they generate new knowledge. In prospective VS, a protocol is devised to select interesting candidates out of a larger data set [25, 113, 272]. Existing knowledge (either pre-existing data for validation, experimental results gathered in the meantime, or experimental testing of VS predictions) may also be used retrospectively to refine the VS parameters. Thus, VS can become a cyclic process, with each cycle leading to more accurate data [22, 117, 157, 272]. This strategy has proven to be successful, especially when it involves collaboration between organic chemists and molecular modelers [79]. VS strategies also often involve applying a hierarchy of filters of increasing selectivity in order to reduce rationally a large virtual library of known or potential molecular structures to a small set of from around 10 to 1000 candidate molecules for experimental evaluation. This may be considered as a kind of "filtering funnel" in which fast but crude filters are used first, and then more sensitive but more computationally expensive filters are applied to distinguish the remaining candidates [120, 273]. As illustrated in Fig. (3), a database may first be pre-screened to select molecules with good drug-like features and ADME-Tox properties in order to optimize both potency and pharmacokinetics. Then, 2D or 3D similarity filters are applied to select compounds similar to a known active query structure. The next step is usually to perform 3D pharmacophore modeling to propose compounds in a more focused way according to specific desired structural chemical features. If the protein target structure is available, docking approaches with MD refinement may also be

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Knowledge discovery from databases.}
\end{figure}
applied to the candidate ligands to predict their binding modes and to suggest ways to optimize their substituents. The resulting lead molecules may be clustered and analyzed to suggest candidates for synthesis and testing in the laboratory. Finally, the test results may be re-cycled to refine the 3D QSAR models and ADME-Tox predictions, and to guide further lead optimization.

**3DVS Workflow Examples**

One practical way to scan very large compound libraries is to use multiple levels of filters of varying sensitivities. In other words, fast filters (including ADME-Tox filters) are applied first, and followed by finer grained filters using more accurate but more expensive scoring functions [92, 273, 274]. Such hierarchical screening approaches usually combine fast ligand-based VS methods with more expensive receptor-ligand docking, for example. The VSM-G (virtual screening manager for computational grids) approach has been developed by us as a proof-of-concept demonstrator for this strategy [120]. The VSM-G platform consists of a series of LBVS and SBVS methods organized as a sequential filter. The available techniques range from fast SH and pharmacophore searches to more sophisticated docking and MD simulations (Fig. 4).

As well as VSM-G, several more advanced workflow management systems have been developed to make these tasks more accessible and to provide an intuitive and flexible framework that can be used by scientists themselves to build complex workflow and analysis pipelines [275]. These systems were designed to integrate and execute complex
multi-strategy approaches in order to handle the more specific needs in different areas of life science like drug discovery, genomics, proteomics and system biology [92, 273, 274]. Such integrated workflow systems can speed up the life sciences research [276]. Three examples of scientific workflow management systems are KNIME [277] (GPL licence), Pipeline Pilot (®Accelrys) [278], and InforSense (®ID Business Solutions) [279]. KNIME is an open-source platform that enables modular data exploration [277]. The user can visually create data flows or pipelines, execute selected analysis steps, and later investigate the results through interactive views on data and models. Pipeline Pilot provides a robust and scalable environment and integrates a set of applications to address the modeling and simulation needs of both research informatics and businesses intelligence [276, 280]. InforSense helps the user to analyse complex chemistry data and to apply predictive and statistical modeling techniques. The InforSense environment includes several specialized extensions such as BioSense (bioinformatics solutions) and ChemSense (chemoinformatics solutions) for specific analyses [275, 276] The capabilities of workflow systems are growing rapidly, and several now include integrated web services, grid processing, and domain-specific tools and ontologies [276].

CASE STUDIES

The remaining sections present and discuss two 3D VS case studies involving LXR-β and the CCR5/CXCR4 HIV entry-inhibitors. Conventional 3D virtual screening techniques are compared with the performance of our own SH-based shape matching approach. In both cases, receptor flexibility of the modeled targets is studied. Multiple conformers are generated and used as VS queries to compare their SH shapes, and a consensus shape clustering technique is applied to the diverse CCR5 ligands. This helps to deal with cases in which different ligands bind to different subsites of a protein pocket.

Case Study 1: LXR-β

LXR-β is of therapeutic interest in metabolic disorders. Several X-ray structures are available which provide insights about the flexibility of the protein binding-site [281]. We have carried out several VS studies on LXR-β. The first study used a VS funnel with two main filters: (i) the fast SHEF shape-based filter using SH representations of the target and the ligands, and (ii) the Glide or GOLD semi-flexible docking programs [120]. Our results confirmed the utility of this two-stage funnel for selecting promising ligand candidates. In the second study, we aimed to enrich the VSMG funnel with a knowledge-based filter. This filter was designed as a decision tree model using various ligand and ligand-target descriptors from SHEF. This allowed us to predict the "dockability" of a given molecule with respect to the target, and hence to rapidly exclude molecules which could not possibly fit in the target binding site [282]. Our results showed that these filters are complementary and that aggregating their outputs led to a satisfactory set of candidate ligands for LXR-β [266].

Our third study concerned an alternative VS workflow in which several filters are run in parallel on the same LXR-β compound library (Fig. 5). The results show that these filters are complementary and that aggregation of their outputs leads to a satisfactory set of candidates for the LXR-β (Fig. 6). Given that each filtering method has its own merits and drawbacks, the influence of protein flexibility during SBDD processes is highlighted. Additionally, there are also several practical factors that have to be considered when working with large data volumes, such as computing time, data management, and analysis. This use case shows how different approaches (SBVS and 3D-LBVS) may be combined hierarchically or in parallel to improve hit identification. In future work, we believe it would be useful to include better target sampling and MD simulations to mimic protein flexibility as additional filtering steps modules after shape matching and flexible docking. We also believe it will be useful to optimize the data management and analysis processes that constituted bottlenecks in the above studies.

Case Study 2: HIV Entry Blockers

HIV infection is initiated by fusion of the virus with the target cell through binding of the viral gp120 protein with the CD4 cell surface receptor protein and the CXCR4 or CCR5 co-receptors. There is currently considerable interest in developing novel ligands that can modulate the conformations of these co-receptors and hence ultimately block virus-cell fusion [118]. In this study a detailed comparison of receptor-based and ligand-based virtual screening approaches was performed to find CXCR4 and CCR5 antagonists that could potentially serve as HIV entry inhibitors. Because no crystal structures for these proteins were available at the time, homology models were built using bovine rhodopsin as the template [283]. For LBVS, several shape-based and property-based molecular comparison approaches were compared using known high affinity ligands as query molecules. These methods were compared by virtually screening a custom library consisting of 602 known CXCR4 and CCR5 inhibitors and ca. 4700 decoys (analogue, presumably inactive). For each receptor, the library was queried using known binders, and the enrichment factors and diversity of the resulting virtual hit lists were analyzed [118].

Overall, we found that ligand-based shape matching searches yielded higher enrichments than receptor-based docking, especially for CXCR4 [284]. The results obtained for CCR5 suggested the possibility that different active scaffolds might bind in different ways within the CCR5 pocket [117]. This hypothesis was quantitatively explored by constructing different SH consensus shape pseudo-molecules and by measuring their VS utility against our CCR5 inhibitor database. This study found four shape clusters whose members were predicted to bind to three different but somewhat overlapping sites within the CCR5 pocket. Pseudo-molecules corresponding to these shape clusters were docked into the CCR5 pocket, and the locations of these positions were related to the locations predicted by previous docking studies. Several allosteric inhibitors within each shape cluster had experimentally supported or computationally predicted binding modes, which were consistent with our docked poses. Therefore, our consensus shape clustering technique provides strong supporting evidence for the hypothesis that the CCR5 pocket has
multiple ligand-binding subsites, and it helps to give a better picture of how the allosteric HIV-1 entry inhibitor modulators of CCR5 are probably distributed in the receptor pocket.

For both receptor targets, receiver-operator-characteristic (ROC) curves were plotted to compare the different LBVS approaches (ROCS and Parafit shape matching and MOE pharmacophore modelling) and SBVS (Autodock, GOLD, Hex, and FRED docking) [117]. Retrospective validation was done using our database of CXCR4 and CCR5 antagonists. Then, prospective virtual screening was applied to a large virtual combinatorial library of candidate CXCR4 antagonists designed by us to identify new anti-HIV compounds [272, 285]. The above screening approaches were used to select five compounds for synthesis and testing, and which ultimately gave activities in the 20 to 0.008 μg/ml range. Experimental binding assays confirmed that the compounds’ mode of action was to block the CXCR4 receptor. This new knowledge was used for the development of ligand-based QSAR models in order to use them to predict activity of hitherto unsynthesized molecules. Prospective virtual screening (using the same protocol as in retrospective screening analysis) was then used to guide the selection of other molecules from the virtual combinatorial library. Molecules found at the first positions of the consensus ranked hit list showed activity values in the 4 to 0.022 μg/ml range [117]. Overall, this study demonstrates how a combination of 3D-LBVS and SBVS techniques may overcome the lack of an initial receptor structural model.

**CONCLUSIONS**

3DVS techniques are becoming increasingly important in drug discovery, as demonstrated by several success stories reported for a variety of targets and therapeutic areas. However, it is still important to understand the limitations of computational approaches in order to select the most appropriate protocol for a given target. This preparation phase should take into account available instruments, the economic context, and the extent of initial available knowledge. However, different methods will probably suggest different drug candidate subsets for a given target. No single modeling technique will be best in all cases. Some documented successes link very heterogeneous approaches to even more diverse targets, and a method that was successful for a one target is not guaranteed to work well for...
the next. Nonetheless, a common factor in the success stories covered here is that the initial knowledge was carefully evaluated to derive an optimal computational approach. Given the sheer size of the chemical universe, it would seem that brute-force approaches will almost inevitably fail. Hence, even if there is insufficient knowledge to devise a single optimal strategy, it can still be useful to employ multiple VS approaches and to combine their results using consensus scoring techniques [286-288].

There are many bottlenecks and areas of uncertainty in current 3DVS approaches, including the reliable scoring and ranking of test compounds, ligand flexibility, modeling protein flexibility and plasticity, assignment of protonation states, solvation effects, and knowledge management, to name but a few. Hence the 3DVS field could be said to be still in its infancy. On the other hand, the known limitations of current 3DVS approaches present some exciting research challenges [11, 18, 20, 106, 138, 234, 248, 256, 289]. While there is no recipe to guarantee success, it is now clear that drug discovery can be guided by the rational use of multiple parallel approaches. In the current post-genomic era, closer collaborations between theoreticians and experimentalists will be critical. In our opinion, drug discovery research should involve a more informed consideration of genomics, proteomics, and chemistry in order to identify the causes of pathologies and to implement focused treatments. We also believe that future developments in 3DVS will come not so much from individual algorithmic improvements, but more from how different techniques might be better combined to exploit the available knowledge. As many current 3DVS algorithms appear to be complementary, we therefore propose that one way to advance the state of the art is to develop better knowledge-based ways to integrate complementary drug discovery strategies.

ACKNOWLEDGEMENTS

Leo Ghemtio is grateful for support from the Drug Discovery and Chemical Biology Consortium of Finland, CDR of University of Helsinki, the French Institut National de Recherche en Informatique et en Automatique (INRIA) and Centre National de la Recherche Scientifique (CNRS). Violeta Pérez-Nueno is grateful for a European Marie Curie fellowship, grant reference DOVSA-254128.

REFERENCES


[98] Catalyst Molecular Simulations Inc, San Diego, CA, USA.


[100] SYBYLUNITY Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.


[246] Lewis, W. G.; Green, L. G.; Grynszpan, F.; Radič, Z.; Carlier, P. R.; Taylor, P.; Finn, M. G.; Sharpless, K. B. Click chemistry in situ: acetylcholinesterase as a reaction vessel for the selective assembly of a femtomolar inhibitor from an array of building blocks. Angewandte Chemie, 2002, 41, 1053-1057.


[260] Pipeline Pilot Accelrys Inc., San Diego, CA, USA.

[261] Inforsense ID Business Solutions Ltd., Guildford, Surrey, UK.


