Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

# Improving small molecule virtual screening strategies for the next generation of therapeutics

Bentley M Wingert and Carlos J Camacho

The new generation of post-genomic targets, such as protein–protein interactions (PPIs), often require new chemotypes not well represented in current compound libraries. This is one reason for why traditional high throughput screening (HTS) approaches are not more successful in delivering medicinal chemistry starting points for PPIs. *In silico* screening methods of an expanded chemical space are then potential alternatives for developing novel chemical probes to modulate PPIs. In this review, we report on the state-of-the-art pipelines for virtual screening, emphasizing prospectively validated methods capable of addressing the challenge of drugging difficult targets in the human interactome. Collectively, we show that optimal strategies for structure based virtual screening vary depending on receptor structure and degree of flexibility.

**Address**
Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA

Corresponding author: Camacho, Carlos J (ccamacho@pitt.edu)

**Current Opinion in Chemical Biology** 2018, **44**:87–92

This review comes from a themed issue on **Next generation therapeutics**

Edited by **Adrian Whitty** and **Peter J Tonge**

https://doi.org/10.1016/j.cbpa.2018.06.006

1367-5931/Published by Elsevier Ltd.

## Introduction

Small molecules remain an available and increasingly diverse source for new and repurposed drug compounds. As computational resources and algorithm quality have increased, Computer-Aided Drug Design (CADD) has become an integral part of the drug discovery process. With massive compound libraries available [1–3] and the ever increasing quantity and quality of receptor–ligand structures [4] and other biological data, more efficient algorithms and novel techniques will become increasingly necessary to take advantage of new data. In this review, we will discuss advances in computational drug discovery, including increased chemical diversity and virtual screening technologies.

Current libraries of compounds used for screening are mostly derived from historical medicinal-chemistry efforts by pharmaceutical companies. Thus, chemical phenotypes, or 'chemotypes', are dominated by past drug-discovery research into kinases, G-protein-coupled receptors, enzymes and other targets traditionally considered druggable [2,5]. New targets, such as protein–protein interactions, often require new chemotypes that are poorly sampled in chemical libraries [6]. Thus, expanding the diversity of compound libraries is essential in order to identify new chemical probes that could address the chemotypes required for new targets [7•].

Virtual small-molecule libraries provide access to an arbitrarily large and potentially more diverse chemical space. However, in order to be useful, these libraries must not only be available or readily synthesizable but also searchable for compounds likely to bind to the target. Many valuable technologies both commercial and open access exist to perform structure-based virtual screening of commercially available compounds [3,7•,8]. Of note, the Dömling and Camacho labs have recently developed breakthrough technologies that allow for drug discovery collaboration efforts to be performed in real time by screening millions of compound in seconds [7•]. These open access tools are not only capable of performing pharmacophore-based virtual screening of commercially available compounds [3], but can also screen chemical libraries specially designed to disrupt protein–protein interactions (PPIs) [7•]. The latter are a target class that has proven to be especially difficult to drug using traditional libraries. These anchor-biased libraries consist of multicomponent reactions (MCR)-derived compounds. MCR chemistry ('one step, one-pot') [9] is much faster than traditional multistep sequential synthesis, allowing for the timely experimental verification or falsification of virtual compounds [7•].

Critical in virtual screening is the prediction of accurate poses and the enrichment of active compounds. When evaluating ranking performance of new virtual screening methods, high correlation values between the predicted ranking of compounds by affinity and the actual rankings are commonly seen when evaluating on known targets [10]. However, these results don't stack up when methods are tested on prospective data sets, even when ample structural information is available [11•,12••]. In this review we discuss recent advances in both the software and strategies used for CADD. Much of these improvements has more to do with tuning the screening strategy

to the type of receptor structure, flexibility, and cofactors than the specific software platform or scoring function.

## Recent advances in virtual screening strategies

*Pose prediction*. Poses are usually predicted based on a two-step approach: ligand conformer generation followed by docking and scoring to the target. There are several efficient software tools used for conformer generation that can be described as deterministic or stochastic [13]. Although generally accurate, sampling of ring structures is still challenging and can sometimes impact the outcome. Docking programs combine conformer generation with pose scoring [14]. There are many docking programs both commercially and freely available, such as AutoDock Vina [15], Smina [16], Glide [17], and Gold [18]. Smina, for example, is a fork of AutoDock, which is not only faster but also facilitates the development of new scoring functions [16].

Scoring functions often fall into one of three categories: force-field-based, knowledge-based, or empirical [14]. Force-field-based scoring functions use actual representations of forces between the receptor and ligand molecules. These are often based on existing molecular dynamics force field parameters such as the AMBER force field [19,20]. Knowledge-based scoring functions use simplified representations of atomic interactions in order to attempt to reproduce experimental structural data. Empirical scoring functions are generated by fitting parameters to experimental structural and affinity data. There have been continued improvements in scoring functions for docking applications, notably the development of the OPLS3 force field [21•]. This force field fit new parameters based on a data set consisting of small molecule and protein–ligand pairs which leads to better parameterization for analysis of protein–ligand interactions. Another recent development has been the use of convolutional neural nets (CNNs) [22•,23] which can be used for scoring. CNNs are a type of neural net architecture where connections between layers are spatially restricted, allowing each neuron to learn about nearby features. While neural nets have been used for receptor–ligand scoring previously [24], their use is pushing the boundaries of deep learning techniques by increasing the ability to learn from spatial interactions from known 3D co-crystal structures [22•,25,26].

*Receptor Flexibility*. Another important characteristic of docking programs is how they treat receptor flexibility. While it is not computationally feasible to simulate full protein flexibility when screening large numbers of ligands, various strategies have been developed to approximate receptor flexibility. For example, a common strategy is the application of ensemble docking [27–29], where docking is performed against multiple available receptor structures. Additionally, partial receptor flexibility has been modeled in a variety of ways, such as rotamer libraries [30], side chain flexibility [31], and full backbone flexibility near the binding site [32]. Because of these advances it is becoming increasingly feasible to account for protein flexibility in virtual screening. Recently the use of metadynamics [33] has been applied to protein–ligand binding [34]. Metadynamics is a method of enhanced sampling which introduces an extra variable into the system which is used to steer the simulation away from areas which have been previously sampled [33]. This method has allowed researchers to combine ideas from induced fit in docking.

## Lessons from prospective virtual screening predictions

Because the aforementioned developments are generally trained and tested *retrospectively*, it is difficult to fairly compare different methods. To that end, analysis of prospective community-wide experiments provides a unique opportunity to evaluate methods and identify problems with different approaches. The Drug Design Data Resource (D3R) project was started as a joint project between the NIH and UCSD with the goal of providing blinded datasets for *prospective* evaluation of drug discovery pipelines [11•,12••].

*Pose prediction*. Given compounds as SMILES strings [35], predictions for targets for which there are one or more publicly available co-crystal structures (Protein Data Bank (PDB) [4]), are generally performed using three major approaches: alignment-based [36••,37–40], standard docking as discussed above [36••,37–39,41–43], or simulation-based [37,41,44]. Alignment- and docking-based methods have been more consistent in prospective tests [11•,12••]. In the former, conformers of each compound are generated [45] and aligned to the ligand of an available co-crystal structure. Alignment metrics can involve chemical similarity measured by Tanimoto similarity [36••,37,38], 3D shape similarity [40], and hybrid 3D shape/pharmacophore feature similarity method [39]. Poses are then minimized and ranked. As expected, higher quality poses were generally correlated with

**Table 1**

**Best prospective pose prediction median RMSD from D3R Grand Challenges**

| Receptor | # Test compounds | # PDB structures | Best median RMSD [Å] | Prospective Best method |
|---|---|---|---|---|
| HSP90 | 5 | >200 | 0.3[a] | Align close |
| FXR | 35 | 27 | 1.17[b] | Dock close |
| Cathepsin S | 24 | 25 | 1.3[c] | Align close |
| MAP4K4 | 30 | 8 | 1.6[a] | Align close |

[a] [38].
[b] [11•].
[c] https://drugdesigndata.org/about/grand-challenge-3/cathepsin_s.