The impact of in silico screening in the discovery of novel and safer drug candidates

Didier Rognan

The impact of in silico screening in the discovery of novel and safer drug candidates.

Didier Rognan

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, France

Corresponding author: Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 74 route du Rhin, 67400, Illkirch, France. Tel.: +33 3 68 85 42 35; fax: +33 3 68 85 43 10, email: rognan@unistra.fr

Abstract

Drug discovery is a multidisciplinary and multivariate optimization endeavor. As such, in silico screening tools have gained considerable importance to archive, analyze and exploit the vast and ever-increasing amount of experimental data generated throughout the process. The current review will focus on the computer-aided prediction of the numerous properties that need to be controlled during the discovery of a preliminary hit and its promotion to a viable clinical candidate. It does not pretend to the almost impossible task of an exhaustive report but will highlight a few key points that need to be collectively addressed both by chemists and biologists to fuel the drug discovery pipeline with innovative and safe drug candidates.

Keywords: screening, hit, target, profile, safety, pharmacokinetics

Table of Contents

1. Introduction

Drug discovery, as any other discipline, is accumulating experimental data at an exponential pace. Due to the sequential and multidisciplinary nature of drug discovery pipelines, archiving and efficient mining of key compound and target properties (e.g. structural, physicochemical, biochemical, pharmacological, toxicological) is crucial for a better understanding and prediction of the developability of a given compound. These good practices are supposed to reduce the overall attrition rates (Hay, Thomas, Craighead, Economides, & Rosenthal, 2014) and therefore lead to a significant decrease of the drug development costs (DiMasi, Grabowski, & Hansen, 2016).

Is is therefore not surprising that in silico methods have gained so much importance in drug discovery. This trend can be simply illustrated by the herein reported survey of several key descriptors for chemical/biological space and computing power (**Fig.1**). On the one hand, there are currently over 110 million chemicals registered by the Chemical Abstracts Service, out of which only 1.5% exhibit known biological activity (Gaulton, et al., 2012). On the other hand, about 11,000 pharmacological targets are known up to date (Gaulton, et al., 2012), giving rise to 125,000 different three-dimensional structures (Berman, et al., 2000). Both compound and target counts experience an exponential growth that mirrors the growth in computing power. This, expressed by the number of transistors in microprocessors, follows the well-known Moore's law stating that the count of the integrated circuits doubles approximately every two years. It is therefore not surprising that the application of in silico technologies in drug discovery literature also experiences an exponential growth with 4-5 PubMed citations every day (**Fig. 1**).

In silico technologies may be applied at any of the numerous possible stages of drug discovery and the review their overall applicability falls outside of the scope of the present article. Here, we will focus on any computerized method to assist chemists and biologists in the preclinical development of drug candidates, ranging from target validation, compound library design, hit identification, hit-to-lead optimization and preclinical candidate identification. To illustrate the integration of computational design in pharmaceutical companies, it is worth mentioning a recent report from Bayer HealthCare

(Hillisch, Heinrich, & Wild, 2015) stating that half of the 20 new chemical entities (NCEs) currently being tested in phase I clinical trials really benefited from computer-aided design methods.

2. Target validation

Target-related safety issues have recently been shown to be the major cause of attrition in clinical trials at a big pharmaceutical company (Cook, et al., 2014). It is therefore of utmost importance to carefully select the right target before entering costly compound screening processes. Considering validated targets as those to which FDA-approved drugs physically bind, we have progressively learned that: (i) targeting certain protein families (e.g. G protein-coupled receptors, protein kinases) reduces the probability of early closures (Hopkins & Groom, 2002; Rask-Andersen, Masuram, & Schioth, 2014), (ii) specific pockets to which launched drugs associate exhibit a well-defined range of physicochemical properties (e.g. hydrophobicity, accessibility, curvature) that are distinct from that of less druggable targets like protein-protein interfaces (Kuenemann, Bourbon, Labbe, Villoutreix, & Sperandio, 2014). However, there is still an urgent need for computational methods that would robustly reduce risks associated with a particular target selection. Of course, "druggability" is by far more complex than the simple propensity of a particular protein cavity to accommodate high-affinity bioavailable drug-like compounds. Other terms like "ligandability" (Edfeldt, Folmer, & Breeze, 2011) or "bindability" (Sheridan, Maiorov, Holloway, Cornell, & Gao, 2010) have recently been proposed since they better capture target property ranges (cavity volume, polarity and buriedness) known to be important for druggable targets (A. C. Cheng, et al., 2007). The most conservative way to define druggable target space is to identify those targets that do physically associate with approved small molecular-weight drugs. One of the most recent surveys (Rask-Andersen, Almen, & Schioth, 2011) identified 989 small molecular-weight drugs acting on 435 therapeutic effect-mediated human targets. In addition, drug-target interactions (Rask-Andersen, et al., 2014) suggest 475 potentially novel drug targets in addition those previously identified.

Altogether, three kinds of methods for predicting target druggability can be distinguished: methods based on the target's sequence, its three-dimensional structure, or its integration in more complex systems biology networks. Whatever the method, the first step is to define the instances (targets, drugs, networks) to which usually machine learning algorithms (Jordan & Mitchell, 2015) are applied in order to establish non-linear relationships between descriptors and the property to predict (**Fig.2**). Many specialized databases storing this information are freely accessible (**Table I**).

The most straightforward method to estimate target druggability relies on different amino acid sequence descriptors (e.g. amino acid composition, physicochemical properties) of known drug targets and putative non-drug targets (or targets still awaiting approved drugs). Such models usually report accuracies of 85-95% (Bakheet & Doig, 2009; Q. Li & Lai, 2007), but are optimistic because of an oversimplified definition of the large non-druggable target space (any target not explicitly defined as a drug target). As a consequence, sequence-based classification tends to reward entire protein subfamilies as potentially druggable (Q. Li & Lai, 2007) although experimental screening data usually indicates the opposite. Moreover, sequence-based models are hard to interpret and are not linked with any particular domain or pocket on which to focus hit identification efforts.

Structure-based methods are therefore much more popular to predict target druggability. They rely on 3D structural descriptors (polarity, hydrophobicity, buriedness, volume, curvature) of ligand-bound cavities in both druggable and undruggable targets to learn rules able to optimally distinguish both categories in a binary manner. Current state-of-the-art tools (Borrel, Regad, Xhaard, Petitjean, & Camproux, 2015; Desaphy, Azdimousa, Kellenberger, & Rognan, 2012; Krasowski, Muthas, Sarkar, Schmitt, & Brenk, 2011; Schmidtke & Barril, 2010; Volkamer, Kuhn, Grombacher, Rippmann, & Rarey, 2012) exhibit an accuracy of approximately 85% for conventional targets (GPCRs, kinases). The main advantage of such methods is their high interpretability in terms of pocket properties. For example, all methods agree on the most important properties for druggable cavities: a medium size (ca. 500 $Å^3$), highly buried (>75%), mostly apolar cavity with a few polar hotspots. Another advantage is that once

a potentially druggable pocket has been identified, it can be screened by various in silico tools (see section 4.1) to propose potential ligands for experimental validation and further optimization. A clear drawback is the limited applicability domain of this method to conventional targets (close to those on which mathematical models have been trained on) of known experimentally-determined structure. Moreover, only a static picture (X-ray structure) of the target is usually taken into account although consideration of structural flexibility to describe flexible and/or cryptic pockets has just been reported (Loving, Lin, & Cheng, 2014).

The overwhelming emergence of –omics data recently pushed bionformaticians to focus not only on drugs and their targets, but on more subtle systems biology approaches in which drug-target and/or target-target networks as well as gene and protein expression levels are explicitly considered (Kandoi, Acencio, & Lemke, 2015). Although druggable targets and their encoding genes have been shown to occupy well-defined (highly connected and central) regions of drug-target (Yildirim, Goh, Cusick, Barabasi, & Vidal, 2007), target-target and gene-gene networks (Yao & Rzhetsky, 2008), adding gene expression data to pure network measures (connectivity and centrality indices) significantly enhances the accuracy of network-based druggability predictions. Due to the absence of any gold standard protein or gene network, obtained accuracies are quite variable (between 60 and 90%), suggesting that this emerging field is still in evolution.

3. High-throughput screening

Since drug discovery initiatives often begin with an experimental medium to high-throughput (biochemical, biophysical, phenotypic or virtual) screening of a compound library, computational science has a major impact at both ends of the screening funnel: library selection and raw data analysis.

3.1. Compound library set-up

Chemical space described by all potential drug-like compounds is so huge that estimations to quantify it vary from 10^{23} to 10^{60} (Polishchuk, Madzhidov, & Varnek, 2013). Whatever the measure, this space is far greater than the currently known chemical space described by 115 million unique organic and inorganic chemical substance registered in the Chemical Abstract Service (**Fig. 3**). Out of this accessible chemical space, about 35 million are theoretically purchasable from a wide array of commercial and academic suppliers (**Table II**). These compounds are available as powders in variable amounts (usually from 1 to 20 mg) within 3-4 weeks. Originating from academic compound repositories, these libraries were primarily synthesized by combinatorial chemistry in order to guarantee high numbers at the cost of a low chemical diversity (Krier, Bret, & Rognan, 2006). In order to satisfy their customers, commercial libraries have now evolved towards a higher quality in terms of diversity, novelty, purity and analytical characterization. However, they are still significantly redundant. Only 1% of the currently available compounds (1.5 million) exhibit a biological activity, described mainly by in vitro binding assays. Archiving these data (**Fig. 3**), which have traditionally been the exclusive property of pharmaceutical companies, in publicly available databases like ChEMBL (Gaulton, et al., 2012) or Pubmed (PubMed, 2016) has a major impact on academic research and enables to better distinguish molecular properties of chemicals, drug-like, lead-like and approved drugs. Many molecular descriptors have been proposed to capture the main characteristics of drug-like compounds (Hann & Keseru, 2012) which can be used as filters to select the most promising compounds to screen. Interestingly, noticeable progress is made in defining a simple and interpretable metric to quantify drug-likeness (Bickerton, Paolini, Besnard, Muresan, & Hopkins, 2012). Assuming that we have a good idea of which molecules have to be kept, we also know which ones shall be removed. Public web resources (Villoutreix, Lagorce, Labbe, Sperandio, & Miteva, 2013) are available to filter out undesired molecules likely to interfere with bioassays, for instance aggregators (Irwin, et al., 2015) or promiscuous binders (Baell & Walters, 2014). Such filters present the advantage of being able to flag the affected compounds for future verifications in case of a positive screening result. However, they should not be used to strictly (e.g. completely excluding these compounds from a screening deck),

notably in screening previously unexplored target space. More important is to keep in mind that those compounds do exist but must be carefully validated in independent secondary assays.

If such filters may be easily applied to analyze existing compound collections, their usage in guiding the design of new compounds or libraries is less straightforward. Theoretically possible organic compounds can be generated from molecular graphs (Reymond, 2015) or a list of pre-defined organic chemistry reactions (Hartenfeller, et al., 2011) and therefore extend the current chemical space to novel structures. Since similar molecules are believed to share similar properties, any attempt to create novel chemistry ends with a quantitative measure of chemical (dis)similarity to existing compounds. (Dis)similarity estimation is probably the computational chemistry field with the most tremendous impact on drug discovery. Over 3000 molecular descriptors (Todeschini & Consonni, 2000) and dozens of similarity coefficients (Todeschini, et al., 2012) are available. The simplest descriptors (**Fig. 4**) encode molecular properties (e.g. atom and bond counts, molecular weight) but are usually not easily interpretable in terms of structure and medicinal chemistry. For example, quite different chemical structures may fall into the same group although they do not share common scaffolds and synthetic routes. Most descriptors rely on the two-dimensional (2D) molecular graphs (substructure, fingerprint) therefore enabling the fast comparison of millions of molecules. More comprehensive but computer-demanding properties may be addressed at the 3D level (fields, shapes, and pharmacophores) and requires the calculation of all low energy conformers for a particular molecule.

Having descriptors and a similarity metric in hand, many library design strategies are possible: (i) design general purpose compound libraries fulfilling drug-likeness filters that are chemically different from existing drug-like compounds (Horvath, et al., 2014), (ii) design scaffold (Rabal, Amr, & Oyarzabal, 2015) or target-focused (Naderi, Alvin, Ding, Mukhopadhyay, & Brylinski, 2016) libraries chemically similar to existing bioactive compounds , (iii) design innovative libraries of compounds irrespective of conventional drug-likeness considerations (Kirkpatrick, 2012). For example, small peptides or peptidomimetics (Verdine & Hilinski, 2012), macrocycles (Hoveyda, et al., 2011), or natural products

(Over, et al., 2012) may combine oral bioavailability and exquisite target selectivity; they are however still largely under-represented in current screening decks.

3.2. Screening data analysis

Many important paradigms in medicinal chemistry arise from a cheminformatics-based analysis of high-throughput screening data. Instead of focusing of individual molecules, the analysis is generalized to chemotypes (substructures, scaffolds, fragments) and derives more general rules about (in)activity. One of these very first concepts has been the notion of frequent hitters (Roche, et al., 2002), in other words compounds that are systematically selected independently of the assay and the target. Such compounds are either real promiscuous binders exhibiting target privileged substructures or fragments (Schnur, Hermsmeier, & Tebben, 2006), or interfere with bioassays because of peculiar physicochemical properties (aggregation, fluorescent emission). They can easily be detected upon analysis of several HTS data and converted into a set of molecular rules (Irwin, et al., 2015). In the same spirit, rational guides to identify latent hits (Mestres & Veeneman, 2003; Varin, Didiot, Parker, & Schuffenhauer, 2012) have also been proposed. Latent hits are compounds that would not have been selected based on a hard cut-off based analysis but that share chemical scaffolds having a significantly higher proportion of actives (scaffold recovery rate) than randomly-chosen scaffolds. Rescuing poorly active compounds by scaffold analysis enables the definition of otherwise masked structure-activity relationships and further hit to lead development.

Another very important concept that emerged from the interplay between computational chemistry and screening applied to low molecular-weight fragments is the notion of ligand efficiency (Hopkins, Groom, & Alex, 2004). Mathematically, ligand efficiency (LE) is the ratio of the Gibbs free energy (ΔG) to the number of heavy atoms of the ligand (HAC; **Eqn.1**)

$$LE = \frac{\Delta G}{HAC} \qquad \textbf{(1)}$$

Since it is a normalized metric, it permits to prioritize not necessarily the hits with the highest affinity in the primary screening assay but those with the higher developability. Lead efficiency is now a widely used concept in medicinal chemistry, notably in case a fragment hit has to be grown to generate a potent lead-like compound (Bembenek, Tounge, & Reynolds, 2009). Good starting fragments usually exhibit ligand efficiencies above 0.3 $kcal.mol^{-1}$ per heavy atom. Upon fragment growing and increase of the molecular weight, ligand efficiency should decrease as little as possible, keeping in mind that a 10 nM compound with a molecular weight of 500 Da presents a lead efficiency of 0.29 $kcal.mol^{-1}$ per heavy atom. Lead efficiency can therefore be monitored throughout the optimization process to ensure the smallest possible decrease as the series progresses and guaranty the best possible pharmacokinetic properties of final compounds.

Upon historical accumulation of HTS data on millions of compounds over hundreds of targets, Novartis scientists introduced the concept of HTS fingerprints (Petrone, et al., 2012) to provide an alternative measure of compound similarity. Each compound is characterized by a vector of reals, each describing a normalized percentage of inhibition towards a well-ordered series of targets. A survey of HTS fingerprints indicated that ca. 46% of 1.8 million compounds have never been active in any of the 230 screening assays (Petrone, et al., 2013). This so-called "dark chemical matter" was recently shown to be of great interest for identifying preliminary hits with very specific activity and selectivity profiles (Wassermann, et al., 2015).

Instead of focusing on chemical diversity, HTS fingerprints offer the possibility to bias library design towards maximum biological diversity (target coverage). HTS fingerprints obtained from a very large HTS data have been shown to outperform standard ligand-centric fingerprints in similarity-based virtual screening for maximum hit rates, scaffold rates and biodiverse plate selection (Petrone, et al., 2012).

4. Virtual screening for pharmacodynamics properties (Hit finding)

The computational equivalent of high-throughput screening to identify novel bioactive compounds with user-desired properties is by far the widest application of in silico technologies in preclinical drug discovery. A PubMed search with keywords directed to identify virtual screening reports (**Fig.5**) suggests that this precise computer usage represents 50% of the overall literature utilizing in silico technologies in drug discovery (compare **Fig.1** and **Fig. 5**). We will here describe three possible scenarios which are widely encountered in modern drug discovery: (i) hit identification (easiest and mostly occurring case), (ii) hit to lead optimization (most difficult case), and early safety profile (prediction of most probable targets).

Whatever the scenario, two categories of virtual screening methods exist (Heikamp & Bajorath, 2013; Sliwoski, Kothiwale, Meiler, & Lowe, 2014), based on either the structure of known ligands (ligand-based virtual screening, LBVS) or the structure of the target macromolecule (structure-based virtual screening, SBVS). The choice of which category to follow (**Table III**) depends on the context of the project and the preexisting knowledge. As a rule of thumb, LBVS methods are used in case many chemically diverse ligands of the desired properties (e.g. inhibitor of a particular enzyme) have already been described. On the other hand, SBVS algorithms are mainly utilized when the target 3D structure is available but with few known ligands. It is important to recall that virtual screening remains a multi-step endeavor that requires a lot of expertise and pragmatism and which is prone to many possible errors that the user must bear in mind (Scior, et al., 2012).

4.1. Choice of the compound library

Identifying ligands able to bind a particular target and/or elicit a particular functional effect (activation, inhibition) is the widest application of virtual screening. Starting from an in-house or a commercially available compound library, a virtual screening software will look for compounds fulfilling user-defined properties that will be further selected and experimentally tested for confirmation of the in silico hypothesis. Many compound libraries are commercially available and usually constitute the starting point of the screening process in academic environments. Initially arising from academic groups, such libraries have grown rapidly thanks to methodological developments in combinatorial organic synthesis. In most of the cases, there is no particular reason, beside practical considerations like purchase time, purity or price, to privilege one library over the others. The safest approach is to combine all of them, at least those arising from trustable suppliers (see **Table II**), or to start from a precompiled list of various sources like ZINC (Sterling & Irwin, 2015). In most instances, just a part of the full screening deck is retained for further evaluation to accelerate the screening. This filtering step is usually essential in limiting the risks of false positives due to the selection of compounds with problematic properties reviewed in section 3.1. The size of the screened library is therefore not an argument for its selection, as unfortunately still seen in some reports (Mirza, Salmas, Fatmi, & Durdagi, 2016). A recent survey of virtual screening papers in current literature (Zhu, et al., 2013) highlights the very high number (up to 60%) of initial virtual hits that can be flagged by various simple filters (chemical reactivity, toxicity, promiscuity) and for which medicinal chemistry optimization would be a waste of time. An adequate filtering of the full screening deck, or at least of the final hit list, is therefore key to avoid later problems with low-quality hits. Please note that libraries of approved drugs (e.g. the Prestwick chemical library) may be also used in case of drug repurposing attempts (Zeniou, et al., 2015). Although a large majority of virtual screening campaigns are realized with electronic libraries of physically existing compounds, libraries of virtual compounds still awaiting their synthesis may be used. For example, Reymond et al. reported a virtual library of 166 billion theoretically synthesizable compounds by assembling drug-like compounds made of less than 17 heavy atoms (Ruddigkeit, van Deursen, Blum, & Reymond, 2012). The major advantage of such virtual libraries is the unprecedented

large chemical space under investigation. However, many potential hits might be difficult (or impossible) to synthesize therefore essentially increasing the time between virtual screening and experimental testing.

4.2. Two-dimensional (2D) similarity search

2D similarity search is the method of choice in virtual screening since it combines both speed and accuracy. Many 2D molecular descriptors are available, out of which the most used are structural fingerprints (Willett, 2011) describing the relative orientation of atoms in a molecular graph in either a linear or circular manner (**Fig. 4**). Similarly to barcodes, linear fingerprints encode the presence (or the number) of key structural fragments (e.g. phenyl ring, alcohol, ketone, etc.) at precise positions of a vector. Alternatively, the respective location of these structural elements might also be encoded in circular fingerprints in which each atom is described as a function of its neighboring atoms in several iterative concentric shells (Rogers & Hahn, 2010). Whatever the descriptor, a metric is then needed to estimate the pairwise similarity between two molecules (**Fig. 6**). Very often, the Tanimoto coefficient is utilized to rank compounds by decreasing similarity to any known template. Defining a hit list by 2D similarity search is therefore as simple as gathering any library compound with a Tanimoto coefficient above a certain threshold (usually a value around 0.70). Many successful 2D similarity-based virtual screens have already been described in the literature (Ripphausen, Stumpfe, & Bajorath, 2012; Sliwoski, et al., 2014) and will not be described here in detail. We will just take one in-house example (Kellenberger, et al., 2007) to illustrate the strength and limitations of the method (**Fig. 7**). Starting from a previously identified chemokine CCR5 receptor agonist, a database of 60,000 commercially available compounds was screened for 2D fingerprint similarity to the template. Among the 100 selected virtual hits, seven molecules were confirmed experimentally in an in vitro competition assay out of which two compounds were more potent than the starting reference (**Fig. 7).** Despite their

modest in vitro potency, this ligand-based 2D screen is very representative of what might be achieved using this computational technique, irrespective of the chosen methods and molecular descriptors:

- Analysis of the chemical structures of the validated hits (**Fig. 7**) shows both structural analogies and differences to the template. Many key structural elements (5 or 6-membered nitrogen-containing unsaturated heterocyle, red; two aromatic rings, green; tertiary amine, blue) are shared between the reference and the hits but with different relative spatial locations.
- The hits exhibit clearly different chemotypes.
- The hits exhibit an in vitro potency relatively similar to that of the reference.
- Very little information (a single reference structure) is required to initiate a virtual screen.

However, one should be aware of the limitation of 2D similarity searches as well:

- The ligand stereochemistry is not considered here although ligand recognition is usually stereospecific.
- By definition, hits share key structural fragments with the reference, which may prohibit their patenting (to be examined on a case by case basis).
- Hits are dependent on the choice of the molecular descriptors, and of the reference if several chemically different actives are available. In the latter case, it is advised to repeat the virtual screen and to fuse obtained results (Hert, et al., 2006).
- Biological similarity may be target-dependent and does not always mirror chemical similarity. A retrospective analysis of 155 HTS data shows that the similarity principle holds true in only 30% of the cases (Martin, Kofron, & Traphagen, 2002). In fact, many chemical series exhibit activity cliffs (pairs or groups of structurally similar compounds with significant differences in potency) that impair simple 2D similarity searches (Maggiora, 2006).

2D similarity-based virtual screening is very often utilized when information on known actives is quite rich. Due to the large number of potential references, parallel virtual screening, although feasible, is very cumbersome. In this cases, machine learning algorithms (Jordan & Mitchell, 2015) are very powerful to discriminate actives from inactives. As defined in 1959 by Arthur Samuel (Samuel, 1959), machine learning is defined as "*a field of study that gives computers the ability to learn without being explicitly programmed*". Starting from a set of instances (actives and inactive molecules) and molecular descriptors, machine learning (ML) algorithms (e.g. Bayesian inference, support vector machines, random forest, Gaussian process) find the optimal separation between the two set of instances in the multidimensional descriptor hyperplane. When applied to hit identification, ML may be used in several modes: binary classification (actives vs. inactives), regression (prediction of binding affinities) and clustering (Meslamani, Bhajun, Martz, & Rognan, 2013). Of course, the accuracy of the prediction depends on the quality of the input data; among which the number of true actives, their chemical diversity, and a broad affinity range (Rognan, 2013b).

4.3. Three-dimensional (3D) similarity searches: Pharmacophore, shape and electrostatics

In contrast to 2D similarity search, the 3D conformation of compounds under investigation is explicitly taken into account in 3D similarity searches. Of course, this additional precision brings a higher level of complexity in the virtual screening since many conformations (up to a few hundreds) have to be either stored in advance or computed on the fly. For example, a molecule as simple as a pentapeptide may adopt thousands of possible conformations in solution, only one of which being selected by the receptor. In a virtual screen, all these possible conformations need to be inspected, sometimes with respect to a single template molecule although the number of references is usually higher (up to 15-20). 3D similarity search is therefore computationally much more demanding than 2D searches, and limited to smaller databases (usually less than 1 million). Among possible descriptors to depict 3D properties of a molecule are: the 3D atomic coordinates, the shape, steric and electrostatic fields, and

lastly the pharmacophore (**Fig. 4**) The pharmacophore concept was introduced in the late 60s by Kier (Kier, 1967) and much later officially defined by the IUPAC in 1998 as follows:

"A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" (Wermuth, Ganellin, Lindberg, & Mitscher, 1998).

Interestingly, the concept was formalized by medicinal chemists who immediately saw the true intuitive nature of the pharmacophore and its easy application to drug design. At this point, it should be recalled again that a pharmacophore is not a molecule but just a formal abstraction of key chemical features that are necessary for a compound to bind a particular target. In many representations, the pharmacophore is illustrated by a series of colored spheres and vectors (**Fig. 8**). The sphere center describes the optimal position of a ligand atom, the diameter relates to the tolerance about that position, while the color illustrates the property (hydrogen bond donor, hydrogen bond acceptor, hydrophobe, aromatic, negative charge, positive charge) the respective atom should bear. Since some interactions are directional (e.g. hydrogen bonds, aromatic stacking), vectors can be used to describe the direction from the ligand interaction atom (vector tail) towards the target protein (vector head). Sometimes, exclusion volume features are added to prevent the ligand occupying user-defined forbidden locations. A molecule fulfils a pharmacophore if at least one of its preferred conformations fits the pharmacophore; in other words, some ligand atoms are matching all pharmacophore spheres in terms of location, color and direction (**Fig. 8**). The quality of the match is usually inferred from a fitness value that can be used to rank database compounds in a pharmacophore-based virtual screening experiment. This strategy requires a priori knowledge of the bioactive 3D conformation of the reference molecule(s). If this information is missing, all possible pharmacophore combinations (e.g. 4 features pharmacophores)(Mason, et al., 1999) of a known bioactive compound can be stored as a 3D pharmacophore fingerprint (**Fig. 4**) in which each position registers the occurrence (or absence) of any specific combination. Database compounds are later compared to the reference with respect to

their 3D fingerprint, and rank-ordered from the most to the least similar. Please note that the complexity of these fingerprints rises exponentially with the number of selected features, their diversity and distance bins used to store all possible combinations. For example, four-point pharmacophores made of 7 possible feature types and 15 distance ranges can generate as many as 350 million potential 4-point 3D pharmacophores per molecule. Screening such fingerprints is therefore very computer demanding with respect to standard pharmacophore searches that only require a few seconds per molecule. Over the last 40 years, many successful usages of pharmacophore searches, starting from a single or many known actives, have been described in the literature (Leach, Gillet, Lewis, & Taylor, 2010; Sliwoski, et al., 2014). We will here illustrate the power of this virtual screening methodology by an old but very elegant study realized in 2002 at Novartis (Flohr, et al., 2002). The goal of the study was to identify non-peptide antagonists of the urotensin II receptor (**Fig. 9**). Urotensin-II (U-II) is a cyclopeptide with very potent vasoconstrictive properties. Alanine scanning and structure-activity relationships of U-II peptidic analogs rapidly identified 3 amino acids (Trp7, Lys8, and Tyr9) as the key residues responsible for the U-II receptor recognition. Determining the structure of the peptide in solution by NMR provided a starting pharmacophore definition that was further used to screen the Aventis compound collection. 500 potential hits fitting this pharmacophore were retained for experimental validation. Six different scaffold classes could be identified, antagonizing the biological activity of U-II in vitro, out of which one inhibitor (S6717) exhibits a nanomolar potency in a functional fluorometric imaging plate reader (FLIPR) assay (Flohr, et al., 2002). This example nicely illustrates the benefits of pharmacophore searches:

- The pharmacophore concept is both simple and very intuitive to interpret.
- Its definition is fuzzy enough to cope with uncertainties about the 3D conformational space accessible to a low molecular weight compound.
- Pharmacophore-based virtual screens are fast enough (a few seconds/molecule) to be applicable at a large scale (e.g. up to a few million compounds).
- The pharmacophore is stereospecific.

Of course, some drawbacks may also be highlighted among which:

- Limiting the 3D conformational space of flexible molecules (> 10 rotatable bonds) to a reasonable number of conformers (<1000) may lead to omission of the real bioactive conformation.
- The definition of a common features pharmacophore first requires a proper alignment of reference molecules. Alignment errors will then lead to incorrect pharmacophore queries.
- Pharmacophores dominated by hydrophobic features (e.g. PPI inhibitors) are generally non-specific and are likely to lead to many false positives.
- Pharmacophores are a property of the training set of compounds rather than reflecting a universal and absolute truth about a binding mode. Care should be taken when using them as filters.

Interestingly, the concept of pharmacophores may be applied to protein-ligand (Meslamani, et al., 2012) and protein-protein X-ray structures (Koes & Camacho, 2012) in order to assign pharmacophoric features to truly interacting atoms. Many other variations of the pharmacophore concept have emerged in the last 15 years, the most important being colored shapes and electrostatic fields (**Fig. 10**). Notably, shape matching methods are now increasingly popular (Diller, Connell, & Welsh, 2015; Hawkins, Skillman, & Nicholls, 2007; Kalaszi, Szisz, Imre, & Polgar, 2014; H. Li, Leung, Wong, & Ballester, 2016; Muegge & Zhang, 2015) for their pace and repeated success in identifying high quality and chemically diverse hits (Hevener, et al., 2012; Johnson & Karanicolas, 2016; Kilchmann, et al., 2016; Roy & Skolnick, 2015). The basic idea under this methodology is that molecular shape is the most conserved property among molecules sharing similar biological properties (Nicholls, et al., 2010). A key advantage of shape-based methods lies in their speed, since trillions of compounds can be screened using such methods with either massively parallel (Muegge & Zhang, 2015) or graphic card processing unit (GPU) architectures (Johnson & Karanicolas, 2016).

4.4. Protein-ligand docking

Molecular docking is a priori the most straightforward method to identify ligands for a target of known experimental structure (X-ray, NMR). When applied in the context of virtual screening, the method implies solving quickly (< 15 s /molecule) three enigma:

- What is the protein-bound conformation of the ligand?
- What is the relative orientation of the ligand with respect to the target protein?
- What is the absolute binding free energy (affinity) of the ligand?

Docking is generally limited to a user-defined pocket (catalytic site, known ligand-binding site) in order to avoid scanning of the entire protein surface, a procedure known as "blind docking" (Grosdidier, Zoete, & Michielin, 2009). The bioactive conformation of the ligand to dock may be deduced by several methods (Moitessier, Englebienne, Lee, Lawandi, & Corbeil, 2008): (i) storing a conformational ensemble for every compound and docking it in a rigid manner to the target protein, (ii) computing the protein-restrained conformational space accessible to the ligand thanks to stochastic methods (e.g. simulated annealing, Monte Carlo simulation, genetic algorithm, molecular dynamics simulation), (iii) using an incremental construction method building the ligand piece by piece.

Independent of the conformational sampling method, the location of the ligand with respect to the protein binding site is estimated assuming a steric and electrostatic complementarity principle governed by simple topological rules (Bohm, 1992). For example, a ligand hydrogen-bond donating atom will be placed in front of a protein hydrogen-bonding acceptor in a way that their respective locations are optimal for establishing a hydrogen bond. Lastly, every pose has to be scored by a fast

scoring function (Y. Li, Han, Liu, & Wang, 2014) that permits screening up to a few million compounds within a reasonable amount of time. The last step consists in sorting all database ligands by decreasing docking score, and picking of the top-ranked ones for experimental validation.

Many successful applications of docking for high-throughput virtual screening of compound libraries have led to thousands of hits over the last decade (Sliwoski, et al., 2014; Spyrakis & Cavasotto, 2015). Let us take the example of the beta2 adrenergic receptor to illustrate this screening method (**Fig. 11**). Docking of about 1 million commercially available compounds into the recently solved X-ray structure of the beta2 adrenergic receptor led to the selection of 25 potential hits showing a high docking score (Kolb, et al., 2009). Out of the 25 compounds, 6 could be confirmed experimentally via in vitro binding and functional assays. The obtained hit rate was excellent (24%) and provided novel chemotypes, out of which one potent hit exhibited an excellent potency (K_i = 9 nM).

The main advantage of docking over other screening methods lies in its very intuitive concept since every hit is provided with a putative binding mode and expected affinity to the target protein. The hit to lead optimization is therefore simplified by either adding chemical groups to unoccupied regions of the pocket or deleting substituents that may clash with the target. As docking does not require preexisting knowledge on known ligands, molecular docking can be applied to orphan targets and often yields novel and patentable chemical matter. Although starting from experimentally-determined protein structures is advised, there are now numerous examples of successful docking into homology models without any substantial decrease in hit rates (Spyrakis & Cavasotto, 2015).

Despite its very attractive foundation, molecular docking suffers from many drawbacks due to its complex parametrization level. Many sources of potential errors will dramatically affect the outcome of the screen (**Table IV**). The easiest to correct is the usage of an inappropriate set of protein atomic coordinates. For example, it is strongly advised to use whenever possible ligand-bound and not ligand-free (apo) protein structures since ligand-binding sites frequently adapt their shape to their bound

ligands. For ligands, many cheminformatics methods are available to standardize chemical structures (e.g. aromaticity, protonation and tautomeric states) in a uniform and coherent manner (Fourches, Muratov, & Tropsha, 2010). Please remember that very flexible ligands (e.g. small peptides) are much more difficult to dock than rigid heterocyclic ligands and it is generally advised to omit docking ligands with too many rotatable bonds. It should also be considered that unexpected binding modes not obeying standard molecular interaction rules will be probably missed by all docking algorithms. In suspicion of such a case, constrained docking forcing the ligand to match a predefined template location may be considered.

Among the most difficult problem to solve is the prediction of binding affinities. Scoring functions utilized by docking algorithms (Y. Li, et al., 2014) need to be fast enough to ensure the docking of hundred thousands of compounds within a couple of days. Their accuracy is therefore limited to ca. 1.5 log unit (7-10 kJ/mol) which is sufficient to discriminate namomolar from micromolar and from inactive compounds, but not enough to precisely rank order database compounds by decreasing affinity.

Three main approaches have been followed to rescue the inability of fast scoring functions to prioritize the best docking poses: (i) develop more sophisticated first-principle scoring functions, (ii) use supervised machine learning (ML) algorithms to predict the likelihood of docking poses, (iii) apply knowledge-based (chemical and topological) rules to filter out unreliable solutions. The first approach uses CPU-intensive energy calculations (e.g. MM-PBSA, MM-GBSA) to refine early docking results. Unfortunately, the benefit of this extra computational cost is controversial as it appears to be target-dependent and hardly predictable (Hou, Wang, Li, & Wang, 2011; B. Kuhn, Gerber, Schulz-Gasch, & Stahl, 2005; Virtanen, Niinivehmas, & Pentikainen, 2015). The second approach consists in training machine learning algorithms (e.g. support vector machines (L. Li, Wang, & Meroueh, 2011), random forests (Ballester, Schreyer, & Blundell, 2014; Zilian & Sotriffer, 2013)) with 3D protein-ligand structural descriptors in order to discriminate good from bad poses. If remarkable results in predicting binding

affinities from protein-ligand X-ray structures have been recently published (Ballester, et al., 2014), such scoring functions have rarely been applied to prospective virtual screening campaigns and their true utility in virtual screening remains unknown. In any case, docking/ML combinations (Khamis, Gomaa, & Ahmed, 2015) must be regarded with great care due to the tendency of machine learning methods to be overtrained (Gabel, Desaphy, & Rognan, 2014). The third strategy, which is currently experiencing a revival, utilizes various knowledge-based approaches to rescore docking poses. The main idea is to use non-energetical topological criteria to address the quality of docking poses, notably by comparing docking solutions with protein-ligand complexes of known X-ray structures. Among the knowledge-based approaches, we can clearly distinguish those methods aimed at constraining the docking algorithms towards expected poses (pharmacophore-constrained docking (Hindle, Rarey, Buning, & Lengauer, 2002), shape-guided docking (Kelley, Brown, Warren, & Muchmore, 2015; Kumar & Zhang, 2015), template matching (C. Gao, Thorsteinson, Watson, Wang, & Vieth, 2015)) from computational protocols that just restrain the analysis of docking poses to reward user-defined features. Both methods have proven useful in many examples for enhancing the quality of top-ranked poses as well as enriching virtual hit lists in true actives. Constrained docking may however be dangerous in forcing known inactive compounds to properly dock in a binding site. It is therefore common practice to conduct a completely free docking calculation and further apply simple cheminformatics descriptors (1D fingerprints (Deng, Chuaqui, & Singh, 2004), 3D similarity (Anighoro & Bajorath, 2016)) to enable the selection of docking solutions that look the most similar to experimentally-determined poses of known ligands. For example, several years ago, we (Marcou & Rognan, 2007) and others (Deng, et al., 2004; Kelly & Mancera, 2004; Mpamhanga, Chen, McLay, & Willett, 2006) proposed the concept of molecular interaction fingerprints (IFPs) (Marcou & Rognan, 2007) to post-process docking data and pick poses producing IFPs similar to that of known actives. Computing IFPs from docking poses is a robust and very efficient manner to predict ligand binding modes (Chalopin, et al., 2010), propose reliable scaffold hops (Venhorst, Nunez, Terpstra, & Kruse, 2008), and enrich virtual hits in true actives upon docking a compound library (de Graaf, Kooistra, et

al., 2011; de Graaf, Rein, Piwnica, Giordanetto, & Rognan, 2011). The success of this post-processing approach is based on the fact that true ligands of a same target often share key interactions with key anchoring residues and thereby produce relatively similar IFPs.

The next problem to solve is the potential role of bound water molecules to mediate ligand binding. Despite the existence of many algorithms to predict bound waters in a protein cavity (Spyrakis & Cavasotto, 2015), there are no general rules to consider whether a water molecule has to be kept or removed upon ligand binding. Even though the problem could eventually be addressed for a few compounds and has led to remarkable successes (Chen, Xu, Wawrzak, Basarab, & Jordan, 1998; Liu, et al., 2005) , it is hardly applicable to the docking of a screening deck. Although water molecules may be switched on or off in a ligand-dependent manner, the benefit of considering bound waters is often target-dependent and hardly predictable. Only a deep knowledge of the system itself can guide the user with the best possible choice.

In most cases, the protein is considered as a rigid bogy during docking, although molecular recognition frequently implies modest to large conformational changes of the target. Whereas moderate flexibility (side chain) is relatively easy to handle, larger modifications are much more difficult to predict. Two solutions to this problem exist: (i) start from multiple protein conformations (experimental or simulated) and repeat the docking as many times as there are input structures (ensemble docking), (ii) use the protein conformation as a docking variable (4D docking). As previously highlighted for the role of bound water molecules, the benefit of explicitly considering protein flexibility is often target-dependent (Moitessier, et al., 2008) and should be considered in the light of known experimental data.

Last, it should be recalled that docking a ligand to a protein requires the presence of a structurally druggable cavity at its surface. In case of flat surfaces (e.g. protein-protein interfaces), docking is unlikely to yield a small molecular weight protein-protein interface modulator.

As a summary, protein-ligand docking is a very powerful computational technique, very often not used under optimal conditions (Rognan, 2013a). By contrast to the above described and much simpler 2D

ligand-based similarity methods, docking is often considered to yield inferior results. In most of the cases, discrepancies are observed because the user has not been able to solve the many problems associated with docking. Only experienced users well aware of the limitations of the methods will repeatedly provide biologists with reliable hit lists. Notably, docking hits, whatever the methodology followed for selection, must be carefully visualized within the target's binding site, therefore giving the chance to rescue compounds located well down the scoring list.

4. 5. De novo ligand design

De novo design methods aim at constructing novel chemical matter by assembling structural pieces (atoms, fragments) until the desired properties (synthetic accessibility, potency towards the main target, avoidance of off-targets, good pharmacokinetic properties) are achieved (Segall, 2014). Initiated in the late 80s during the hype of structure-based design (Hol, 1988), the first generation of de novo design methods were almost structure-based and supposed to deliver ideal molecules with a perfect complementarity to the protein binding site to be occupied. After a few years of practice, it turned out that the designed molecules were usually chemically complex, difficult to synthesize, often requiring human intervention to simplify their structure, and with micromolar potencies far beyond the initial expectations (Babine, et al., 1995). A second generation of computational tools has emerged that learned from the early failures. These novel methods are almost exclusively ligand-centric and reaction-driven, and generate drug-like compounds from a set of building blocks and popular organic reactions (Besnard, et al., 2012; Hartenfeller, et al., 2012; Vinkers, et al., 2003). Using 1,000 building blocks and 50 bi-molecular organic reactions (e.g. reductive amination) in a 5-step synthesis scheme provides an unprecedented vast chemical space of 3.10^{26} compounds. By opposition to the early methods, theoretically possible compounds are now filtered to adopt desired properties thanks to a

series of ligand-based machine learning models for predicting desired and off-targets as well as several pharmacokinetics properties.

Many impressive reports of de novo designed bioactive compounds have been published over the last decade with user-controlled properties aimed at optimizing potency, selectivity or multi-target profiles (Schneider & Schneider, 2016). Contrarily to early hopes, these methods express their full potential when sufficient ligand knowledge is available but cannot be applied to identify the very first ligands of still orphan targets.

4.6. Target fishing: Early safety profile and polypharmacology

Whereas the above-described virtual screening methods (similarity search, pharmacophore mapping, protein–ligand docking) have proven useful to predict novel ligands for a single target, profiling a single ligand against a set of heterogeneous targets has long been neglected. Scientific and economic pressure to design drugs with controlled selectivity profiles (Hopkins, Mason, & Overington, 2006; Morphy, 2010) as well as the boost of drug repurposing (Ekins, Williams, Krasowski, & Freundlich, 2011), led to the development of in silico ligand-profiling methods (Rognan, 2010; Westermaier, Barril, & Scapozza, 2015) aimed at (i) predicting potential targets (and thus a mechanism of action) from phenotypic screening hits, (ii) identifying off-targets potentially responsible for side effects and adverse reactions, and (iii) proposing novel targets for existing drugs. Several of these methods are freely accessible as webservers (**Table V**) where the user draws first the ligand structure, runs the virtual screening engine and finally saves a list of putative targets.

From a conceptual point of view, there are three possible approaches to predict novel targets for a known ligand (**Fig. 12**). At the simplest level of theory is the concept that similar ligands bind to similar targets. Estimating the similarity between a ligand of interest and target-annotated compounds is thus

an easy way to predict novel target–ligand associations (Keiser, et al., 2009; Reker, et al., 2014) and even to some extent binding affinities (Vidal, Garcia-Serna, & Mestres, 2011). Ligand-centric profiling methods are, however, restricted to targets for which sufficient ligand information is available. For example, the similarity ensemble approach (SEA) (Keiser, et al., 2009) only applies to 246 targets annotated by more than 100 ligands. Likewise, we designed a hybrid profiling method (Profiler) relying on public ChEMBL binding affinity data and predicting either binding constants for 141 human targets or only binding (yes/no answer) for 661 additional targets (Meslamani, et al., 2013). Target predictions by ligand-based 2D similarity methods are successful in approximatively 50% of the cases and have led to the prediction of (i) toxic liabilities and side effects for known drugs (Lounkine, et al., 2012), (ii) main targets of phenotypic screening hits (Laggner, et al., 2012), novel drug usages (Keiser, et al., 2009), detailed polypharmacological profiles (Besnard, et al., 2012) and targets for complex natural products (Reker, et al., 2014). Interestingly, 2D similarity methods have recently been shown to be effective for identifying main targets, whereas three-dimensional 3D similarity methods were better suited for proposing off-targets (Yera, Cleves, & Jain, 2011).

A second group of methods relies on the concept that similar ligands bind to similar binding sites. Binding site similarity either at the sequence (Surgand, Rodrigo, Kellenberger, & Rognan, 2006) or at the structure level (Xie & Bourne, 2008) can thus be used as a means to pair an existing ligand (with a known binding site) to a novel target sharing a similar binding pocket (Ehrt, Brinkjost, & Koch, 2016). Binding site-based comparisons show a great potential in computer-assisted target identification because of their large applicability domain. Starting from sequence-based approaches, the method could in theory be applied to any of the 71 million amino acid sequences registered in the UniProt database. On a structure-based scale, the applicability domain is of course smaller but still covers 125,000 3D structures of macromolecular targets stored in the Protein Data Bank. There are, however, many possible druggable cavities on the surface of each of these macromolecule (Kufareva, Ilatovskiy, & Abagyan, 2012), and the combinatorics are even higher if protein–protein interfaces are considered (M. Gao & Skolnick, 2012). The number of druggable pockets is therefore much larger than the number

of unique protein structures in the PDB. Up to now, binding site comparisons have been used to discover off-targets for known bioactive compounds (Ehrt, et al., 2016; Rognan, 2010).

At the highest level of theory are methods focusing on protein–ligand complexes that can be described either as either simple one-dimensional (1D) fingerprints (van Westen, Wegner, Ijzerman, van Vlijmen, & Bender, 2011), protein–ligand-derived pharmacophores (Meslamani, et al., 2012) or protein–ligand-docking poses (Y. Y. Li, An, & Jones, 2011). Chemogenomic (or proteochemometric) approaches (Cortes-Ciriano, et al., 2015) take into account both ligand and target 1D descriptors to derive machine learning models that can be further used to predict any new binary association (F. Wang, et al., 2011). Such methods have a wide applicability domain due to the large body of data already available in bioactivity databases (**Table 1**), but cannot predict binding constants. Ligand-based (Rollinger, et al., 2009) and protein-ligand pharmacophores (Lei, Liu, Peng, & Xiao, 2015) have also been used to predict the main targets of natural compounds. These methods first require the set-up of a collection of pharmacophores (Meslamani, et al., 2012) and then the fitting of the compound under scrutiny to any of these pharmacophores to select the best matches. Last, the identification of novel targets accounting for main or secondary effects has been reported in numerous reverse docking studies (Durrant, et al., 2010; Muller, et al., 2006; Yang, Chen, & He, 2009; Yang, et al., 2011) despite notorious deficiencies of empirical scoring functions to rank-order target–ligand complexes by increasing binding free energies. Of course, structure-based methods are restricted in their applicability domain to targets of known X-ray structures. Although hybrid profiling methods combining ligand-based and structure-based methods have been described (Meslamani, et al., 2013), ligand-based methods usually outperform structure-based approaches in target fishing experiments (Meslamani, et al., 2013) for the simple reason that there are much more target-annotated ligands (ca. 2 millions) than unique target X-ray structures (ca. 40,000). A list of representative ligands (known drugs, preclinical candidates, and screening hits) for which successful computational target assignments have been confirmed experimentally is given **Fig.13.**

We should recall that affinity is only one component to a compound's pharmacology and the PK/PD relationship may have a much greater role in determining the observed effect in vivo. This is doubly true for off-target effects, so experimental validation of any target prediction is paramount.

4.7. Which method to use?

Among the myriad of software and different methods amenable to virtual screening, it remains still difficult to prioritize one particular solution with respect to others. A frequently observed error is to believe that the quality of the results will be dependent on the sophistication level of the method. Virtual screening is a pragmatic exercise aimed at integrating all experimental knowledge to guide the choice of the best method. For example, identifying inhibitors for an orphan target at the computational level requires a 3D structure of the target, meaning there is no need to use ligand-based methods. Conversely, fine-tuning the polypharmacological profile of a biogenic amine GPCR ligand does not require protein structures, but chemical structures and binding constants of known ligands. Machine learning algorithms will be excellent in proposing ligands with user-controlled profiles.

Several years ago, it was quite frequent to see workflows combining all these methods in a serial funnel, starting with the simplest ones (e.g. 2D similarity search) and ending with the most complex (e.g. protein-ligand docking). A decade of prospective applications suggests doing the opposite. LBVS and SBVS methods tend to yield to overlapping sets of hits (Kruger & Evers, 2010). It is therefore now good practice to combine hit lists from different methods in order to optimize both hit rates and potencies (Ripphausen, et al., 2012). A systematic survey of successful virtual screening reports in the literature (Ripphausen, et al., 2012; Zhu, et al., 2013) indicates that the mean hit rate obtained by prospective VS is really excellent (ca. 13%) and much higher than that expected by HTS (0.01-0.1%). This analysis also pinpoints that knowledge, intuition and experience still plays a decisive role in selecting the right hits

5. Virtual screening for ADMET properties (Hit to lead optimization)

Although hits are usually easy to find irrespective of the method (in silico or experimental screening), advancing a hit to a viable lead is a much more difficult enterprise since many parameters (potency, selectivity, pharmacokinetics, toxicity) have to be optimized simultaneously. All previously-described methods can be in principle applied to the hit-to-lead optimization, though with much more difficulties since predicting potency (in other words binding constants), for example, is still an unsolved problem (Y. Li, et al., 2014). Virtual screening by quantitative structure-property relationships (QSPR) or machine learning models trained on compound properties (measured or predicted) can nevertheless be applied to many steps of the hit optimization phase (**Table V**) provided that a single-source set of homogenous data is available.

5.1. Physicochemical properties

Since physicochemical properties of drug candidates (e.g. pKa, aqueous solubility, octanol/water partition coefficient or logP, melting point) strongly influence their pharmaceutical developability, many QSAR /machine learning models have been proposed to predict these key properties (Y. Wang, et al., 2015). As a rule of thumb, logP and pKa values are the easiest properties to predict, the boiling point the most difficult, whereas aqueous solubility predictions represent a moderately difficult problem (Hughes, Palmer, Nigsch, & Mitchell, 2008). Interestingly, a recent report from a pharmaceutical company (Hillisch, et al., 2015) shows that the overall accuracy of these models is constantly increasing thanks to more homogeneous and numerous high-quality experimental data (Fraczkiewicz, et al., 2015). Some pharmaceutical companies even agree to share raw data to improve the quality of the resulting models (C.S. Fishburn, 2013). However, poor performance can still be

observed in daily applications for two major reasons: (i) compounds are located outside the applicability domain of the prediction model; (ii) the property to predict is depending on a too complex mechanistic behavior.

5.2. ADMET properties

Predicting the absorption of a drug by the human intestine or the simplified Caco-2 cell model has led to many QSAR models with mitigated results. Instead of predicting absolute values, it is advised to predict a value normalized with respect to a reference standard (Larregieu & Benet, 2013). Since Caco-2 cells markedly differ from human intestinal cells with respect to the expression of several transporters, the prediction of highly permeable hydrophilic compounds is classically underrated by such models. Some gastrointestinal drug absorption models are commercially available (Sjogren, Thorn, & Tannergren, 2016) and are useful to guide drug development, but still fail in predicting the intestinal absorption of incompletely absorbed molecules (Sjogren, et al., 2016). For compounds aimed at targeting the central nervous system (CNS), it is important to predict the brain-plasma partitioning in order to allow the compound to reach its cellular target but also to avoid peripheral side effects. QSPR classification models relying on compound properties reach accuracies close to 80% and still need to be improved, notably by considering the rate and extent of brain penetration as well as the plasma and brain tissue binding strengths (Lanevskij, Japertas, & Didziapetris, 2013).

Once the compound has been absorbed by the intestine, it is important to know the fraction that will be bound to plasmatic proteins (e.g. serum albumin) since this parameter will drastically influence the pharmacokinetic properties such as the volume of distribution, clearance and elimination, as well as the pharmacological effect of the drug. Plasmatic protein binding is known to heavily depend on the hydrophobicity (the higher the better) and can be modeled with a reasonable accuracy (average error of ca. 15%) using decision trees and random forest models (Ghafourian & Amin, 2013). More generally, machine learning models have been set up to predict oral bioavailability (Kim, Sedykh, Chakravarti,

Saiakhov, & Zhu, 2014). While predicting absolute values remains difficult, notably due to the binding of some compounds to human intestinal transporters, binary classification models (F% > 50% or F% < 50%) reach accuracies close to 75-80% (Kim, et al., 2014).

Metabolic stability is another very important criterion to guide drug development. Due to the numerous X-ray structures of cytochrome P450 enzymes (CYPs) that are currently available (notably 1A2, 2C9, 2C19, 2D6 and 3A4), predicting binding to CYPs as well as the site of metabolism is one of the very rare structure-based applications to predict ADMET properties (Sliwoski, et al., 2014). Contrarily to hit finding approaches in which the highest possible affinity to the target is desired, docking substrates to CYPs is conceptually different as loose binding is requested here. Using a combination of docking and machine learning, experimentally observed sites of metabolisms could be confirmed within the two top-ranked predicted positions for 86% and 83 % of the 261 unique 1A2 and 100 different 2A6 substrates, respectively (Huang, Zaretzki, Bergeron, Bennett, & Breneman, 2013). In addition to docking based methods, simpler reactivity rules have been embedded in freely available web servers to predict the likelihood of every ligand atom to be recognized by major CYPs (Rydberg, Gloriam, & Olsen, 2010).

In section 4.6, we showed that many computational methods can be used to predict the potential main and off-targets of drug candidates. Provided that direct activation or inhibition of some targets is inherently linked to well-defined side effects (e.g. dry mouth for muscarinic M3 receptor blockade) (Bowes, et al., 2012), chemical-based similarity approaches have been widely used to infer side-effects from compound structures (Lounkine, et al., 2012) or predicted compound-target interactomes (Simon, et al., 2012). Of course, in vitro binding to some off-targets may be harmless if the compound cannot reach the target in vivo. The pharmaceutical industry has agreed on a set of targets which should be avoided whenever possible (Bowes, et al., 2012) and that can be systematically screened for determining early in vitro safety profiles. Among these targets, the hERG channel is certainly the most investigated macromolecule since its inhibition causes a drug-induced QT syndrome and a severe

cardiac adverse reaction (torsades de pointes) potentially leading to sudden death. A myriad of ligand-based similarity and pharmacophore models as well as numerous docking attempts have been undertaken to predict both hERG blockade and rationally designed structural modifications to avoid it (Sliwoski, et al., 2014).

On the biological side, pathway-based approaches have been developed using molecular networks (drug-target, drug-drug, target-target and drug-side effects connections) to infer potential side effects for new drugs (Campillos, Kuhn, Gavin, Jensen, & Bork, 2008; F. Cheng, et al., 2013; Shaked, Oberhardt, Atias, Sharan, & Ruppin, 2016). Classifiers usually report area under the ROC curve (AUC) values in the 0.7-0.9 range for most prominent side effects (Perez-Nueno, Souchet, Karaboga, & Ritchie, 2015). These methods are strongly dependent on the availability of high quality databases with uniform ontologies to describe diseases (Altman, 2007), side effects (M. Kuhn, Letunic, Jensen, & Bork, 2016), biological processes (Kanehisa, et al., 2014) and drugs (Law, et al., 2014).

In addition, many toxic liabilities (e.g. carcinogenicity, mutagenicity, genotoxicity, skin sensitization, teratogenicity) may be inferred independently on any target/gene-drug association notably by using structural alerts (toxicophores), a set of rules that link chemical fragments/substructures to well-defined toxicity events with a probability (Raies & Bajic, 2016). In such predictions, high false positive rates can be accepted to be sure to remove any problematic compound. These models are usually inspected with respect to the Cohen Kappa coefficient, a statistical measure of inter-rater agreement for categorical items, e.g. mutagen/non-mutagen (Modi et al., 2012).

6. Impact of rational drug design on the discovery of marketed drugs

Has computer-aided design any significant impact on the productivity of the pharmaceutical industry? In terms of marketed drugs, there are several reports of significant contribution of computer-aided drug design in the discovery of launched drugs (Alex & Millan, 2012; Hillisch & Hilgenfeld, 2003;

Kubinyi, 2006). Some of these success stories are summarized in **Table VI** and the exact role of computer simulations in each specific context is highlighted. In most if not all of these cases, computer-aided design was applied in an iterative cycle with protein and protein-ligand X-ray structure elucidation, but was restricted to target families (e.g. proteases, kinases) amenable to structural determinations. Key advances have often been registered when several ligands could be co-crystallized with the target of interest in order to exploit atomic details (additional subpockets, specific ligand structural features) enabling to fine-tune ligand recognition (e.g. neuraminidase inhibitors, carbonic anhydrase inhibitors, protein kinase inhibitors). Remarkable technological advances in recent times enable us to apply the same structure-based strategies to targets previously considered as difficult, like G-protein coupled receptors (Ghosh, Kumari, Jaiman, & Shukla, 2015), ion channels (Kuang, Purhonen, & Hebert, 2015) or large macromolecular assemblies (Garreau de Loubresse, et al., 2014). Computer-aided drug design is nowadays used, as many other technologies (e.g. mass spectrometry) in most drug discovery projects. Quantifying its impact in the discovery of new drugs remains hard but has been attempted in a recent report from a major pharmaceutical company (Loughney, Claus, & Johnson, 2011). By classifying the influence of computer-aided design on projects over a period of four years (2006-2009) as 'none', 'data provided', 'significant' or 'project enabling', a clear trend could be observed with a continuously increasing percentage of significant contributions (from 25 to 40%) whereas the proportion of projects with no contribution at all constantly decreased (from 25 to 5%). Likewise, another company acknowledges in the annual report of its CADD group a direct contribution of computational methods to the discovery of two molecules under clinical evaluation, eight clinical candidates, 37 hit/lead series, 18 lead optimization programs and over 70 examples of screening data analysis (Green, Leach, & Head, 2012).

Computer-aided design has already experienced four out of the five famous steps of the hype cycle: (i) the technology trigger in the 80s (Langridge, Ferrin, Kuntz, & Connolly, 1981), (ii) the peak of inflated expectations in the 90s (Hol, 1988), (iii) the trough of disillusionment at the beginning of the century, and (iv) the slope of enlightenment (Seddon, et al., 2012). We are coming closer and closer to the last

step of the cycle, the plateau of productivity (Green, et al., 2012). In silico predictions currently stand at almost all steps of the drug discovery pipeline, albeit with different accuracy levels. Hit identification is by far the easiest task, thanks to the enormous amount of in vitro data that have been gathered both on bioactive ligands and their targets, at a very precise molecular level. Despite the fact that identifying viable hits by in silico screening is really doable for most targets, optimizing them into efficient leads still remains cumbersome. First, we are still lacking a computational method to accurately predict binding constants for a large set of chemically diverse compounds. Then, many ADMET properties that need to be optimized during the hit-to-lead transition step cannot be simplified to a single molecular event, and are therefore more difficult to predict. Nevertheless, the simplicity, high-throughput and low cost of most in silico prediction methods has definitely placed them in the arsenal of all academic and private drug discovery institutions to routinely guide drug discovery. These methods are not going to solve the attrition rate problem in drug discovery. However, the interplay of computational and experimental data is key to many issues including (i) a better representation of chemical and target space, (ii) a simplified analysis of very complex high-throughput data, (iii) pinpointing areas where experimental data are still insufficient in number, diversity and quality, (iv) saving time and money to design novel experiments.

At the era of big data, systems biology and translational research (C. S. Fishburn, 2013), computational methods are more than ever required to guide experimentalists towards the best possible tracks. Although we still have not reached the paradise anticipated 20 years ago (Hol, 1988; van de Waterbeemd & Gifford, 2003), in silico methods have already modified the mindset of the next generation of scientists and will deeply influence drug discovery in the next decade. We notably hope to better understand complex mechanisms on a systems level in order to deliver better predictions with enhanced applicability domains and refined confidence levels.

Conflict of interest

"The author declares that there are not conflicts of interest"

Acknowledgements

D.R. thanks the Agence Nationale de la Recherche under 'Programme d'investisssement d'avenir' for funding the LABEX ANR-10-LABX-0034 Medalis, and Dr. M Drwal for critical comments on the manuscript.

Table I. Drugs, targets and target-drugs databases

Name	Instance	Content	URL
DrugBank	Drugs, targets	Drug entries: 8602 Targets: 4,333	http://www.drugbank.ca/
ChEMBL	Ligands, drugs, targets	Targets: 11,019 Compounds: 1,592,191 Bioactivities: 13,967,816 Publications: 62,500	https://www.ebi.ac.uk/chembl/
PubChem	Ligands, drugs, targets	BioAssays: 1,154,431 Tested compounds: 2,145,625 Protein Targets: 9,961 Gene Targets: 19,517	https://pubchem.ncbi.nlm.nih.gov/
PDB	Targets	Three-dimensional structures: 110,000	http://www.rcsb.org
TTD	Targets, drugs	therapeutic protein and nucleic acid targets: 2,025 targeted diseases pathway information corresponding drugs:17,816	http://bidd.nus.edu.sg/group/cjttd/
DGIdb	Drugs, Genes, targets	Genes: 6,761 Drugs: 6,307 Gene-drug interactions: 14,144	http://dgidb.genome.wustl.edu/
STITCH	Drugs, targets, drug-target networks	Small molecules: 300,000 Proteins: 9.6 million Interactions: 128 million	http://stitch.embl.de/
STRING	Target networks	Proteins: 9.6 million Interactions: 184 million	http://string-db.org/

Table II. Main commercial suppliers of screening compound libraries

Libraries	Compounds	website
Individual suppliers		
Abamachem	1,500,000	http://www.abamachem.net
Alinda	186,000	http://www.alinda.ru
AnalytiCon Discovery	32, 0000	http://www.ac-discovery.com/
Asinex	600,000	http://www.asinex.com/
BCH-research	1,500,000	http://bchresearch.com
Bionet	79,200	http://www.keyorganics.net
Chembridge	1,120,000	http://www.chembridge.com
ChemDiv	1,500,000	http://www.chemdiv.com
Enamine	2,000,000	http://www.enamine.net
InterBioScreen	525,000	http://www.ibscreen.com/
LifeChemicals	412,000	http://www.lifechemicals.com/
Maybridge	53,000	http://www.maybridge.com/
Otava	290,000	http://www.otavachemicals.com/
Pharmeks	340,000	http://www.pharmeks.com/
Princeton Biomolecular Research	1,300,000	http://www.princetonbio.com/
Specs	460,000	http://www.specs.net
TimTec	891,350	http://www.timtec.net/
Vitas-M	1,500,000	http://www.vitasmlab.com/
Uorsy	600,000	http://www.ukrorgsynth.com/
Web portals (many suppliers)		
ChemNavigator	60,000,000	http://www.chemnavigator.com
ChemSpider	34,000,000	http://chemspider.com
e-molecules	7,000,000	http://www.emolecules.com
ZINC	19,000,000	http://zinc15.docking.org

Table III. Virtual screening strategies and software[a]

Library	Method	Descriptor	Metric	Strengths	Weaknesses
up to 10^9	2D similarity search	2D fingerprint	Similarity score or QSAR equation	Speed, applicable to any target class, accuracy increases with number of known actives	Structural novelty, patentability, choice of the reference(s)
up to 10^7	3D similarity search	Pharmacophore 3D Fingerprint Shape & electrostatics	Fitness Similarity score Overlap score	id. 2D methods, easy interpretation by medicinal chemists	Calculation of most plausible conformers, not efficient for very flexible compounds (e.g. peptides)
up to 10^6	Docking De novo design	Protein-ligand coordinates	Binding energy	Binding mode proposition, Chemical novelty, patentability	Scoring and ranking, applicability to few targets (known 3D structure)

[a] for an exhaustive list of available software and references, see click2drug and VLS3D (Villoutreix, et al., 2013) on-line resources.

Table IV. Sources of errors of increasing complexity in protein-ligand docking

Error	What to do
Wrong set of coordinates (protein, ligands)	Change protein coordinates, correct ligand structures
Ligand flexibility	Remove highly flexible ligands (> 10 rotatable bonds)
Unexpected binding mode	Template matching
Scoring function	Rescore by more accurate methods (e.g. MM-PBSA)
Bound waters	Use switchable waters
Protein flexibility	Ensemble docking, 4D docking
Flat pocket	Use another method

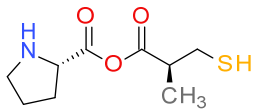
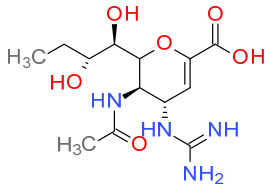
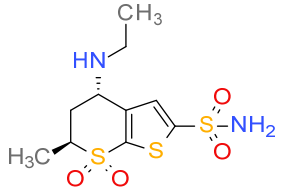
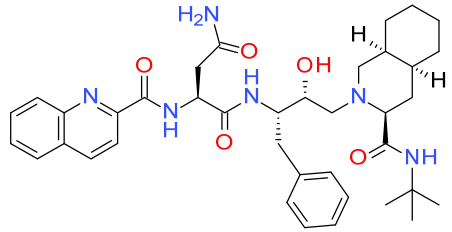
Table V. Freely available web servers as potential target fishing tools

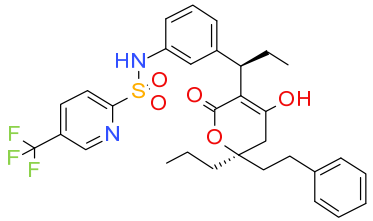
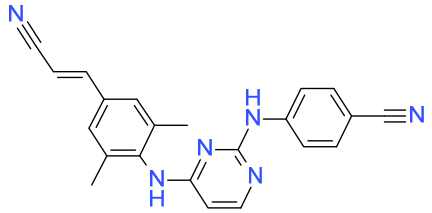
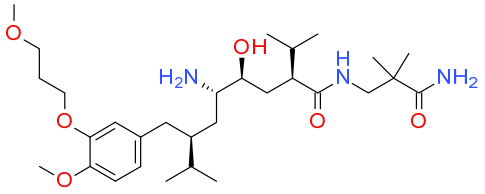
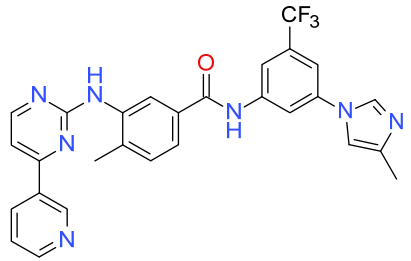
Name	Principle	website
Ligand-based		
BindingDB	2D similarity	http://www.bindingdb.org/bind/chemsearch/marvin/FMCT.jsp
HitPick	2D similarity	http://mips.helmholtz-muenchen.de/hitpick/
PASS online	2D similarity	http://www.pharmaexpert.ru/passonline/
SEA	2D similarity	http:// sea.bkslab.org/
SPiDER	Topological pharmacophores & physicochemical properties	http://modlab-cadd.ethz.ch/software/spider/
SuperPred	2D and 3D similarity	http://prediction.charite.de
SwissTargetPrediction	2D & 3D similarity	http://www.swisstargetprediction.ch/
Protein-Ligand based		
DRAR-CPI	Protein-ligand docking	https://cpi.bio-x.cn/drar/
IdTarget	Protein-ligand docking	http://idtarget.rcas.sinica.edu.tw/
PharmMapper	3D Pharmacophores	http://59.78.96.61/pharmmapper/
TargetHunter	1D proteochemometrics	http://www.cbligand.org/TargetHunter/
TarFisDock	Protein-ligand docking	http://www.dddc.ac.cn/tarfisdock/

Table VI. *In silico* models for predicting ADMET properties

	Predicted property	Prediction	Reference
Absorption	Caco-2 influx and efflux	Relative permeability	(Larregieu & Benet, 2013)
	Gastrointestinal absorption	Numerical	(Sjogren, et al., 2016)
	Blood brain barrier permeation	Classifier	(Lanevskij, et al., 2013)
Distribution	Plasmatic protein binding	numerical	(Ghafourian & Amin, 2013)
	Oral bioavailability	Classifier	(Kim, et al., 2014)
Metabolism	Cytochrome metabolism	Docking model	(Huang, et al., 2013)
		Reactivity score	(Rydberg, et al., 2010)
	Microsomal stability	Probabilities	(Aliagas, et al., 2015)
Side effects	Off-targets	Numerical & classifier	(Lounkine, et al., 2012)
	Networks (drug-target, drug-drug, drug-side effects, metabolic)		(Perez-Nueno, et al., 2015)
Toxicity	Carcinogenicity, mutagenicity, genotoxicity, skin sensitization, teratogenicity	Classifier	(Raies & Bajic, 2016)

Table VII. List of marketed drugs for which computer-aided design played a decisive role

Compound	Structure	Target	Indication	Role of CADD	References
Captopril		Angiotensin-converting enzyme	hypertension	Active site modelling by homology to carboxypeptidase A	(Cushman, Cheung, Sabo, & Ondetti, 1977)
Zanamivir		Influenza Neuraminidase	Flu	De novo design suggests replacing the hydroxyl group of a transition state analog by a basic guanidine, which dramatically enhances potency and selectivity	(von Itzstein, et al., 1993)
Dorzolamide		Carbonic anhydrase	Glaucoma	Energy calculations suggest adding a methyl group on the 6-membered ring to optimize the complementarity to the X-ray structure of the enzyme active site	(Greer, Erickson, Baldwin, & Varney, 1994)
Saquinavir		HIV-1 protease	AIDS	Structure-based optimization of lipophilic substituents filling the 4 lipophilic subpockets (S2', S1', S1, S2) of the substrate binding site	(Craig, et al., 1991)

Tiprinavir		HIV-1 protease	AIDS	Replacement of a coumarine ring (HTS hit) by a 4-hydroxypyrone and further incorporation of a sulfonamide in the meta position to gain additional interactions	(Turner, et al., 1998)
Rilpivirine		HIV reverse transcriptase	AIDS	Adaptation to the allosteric non-nucleoside binding site and targeting of the conserved Trp299	(de Bethune, 2010)
Aliskiren		Renin	Hypertension	Structure-based optimization of the size of the hydrophobic group filling the large S1-S3 cavity. The methoxyalkoxy sidechain of the inhibitor is essential for strong binding	(Wood, et al., 2003)
Nilotinib		Abl-Kinase	Chronic myeloid leukemia	Replacement of imatinib N-methyl piperazine by a methylindole that optimally fits a subpocket of the kinase inactive structure.	(Weisberg, et al., 2005)

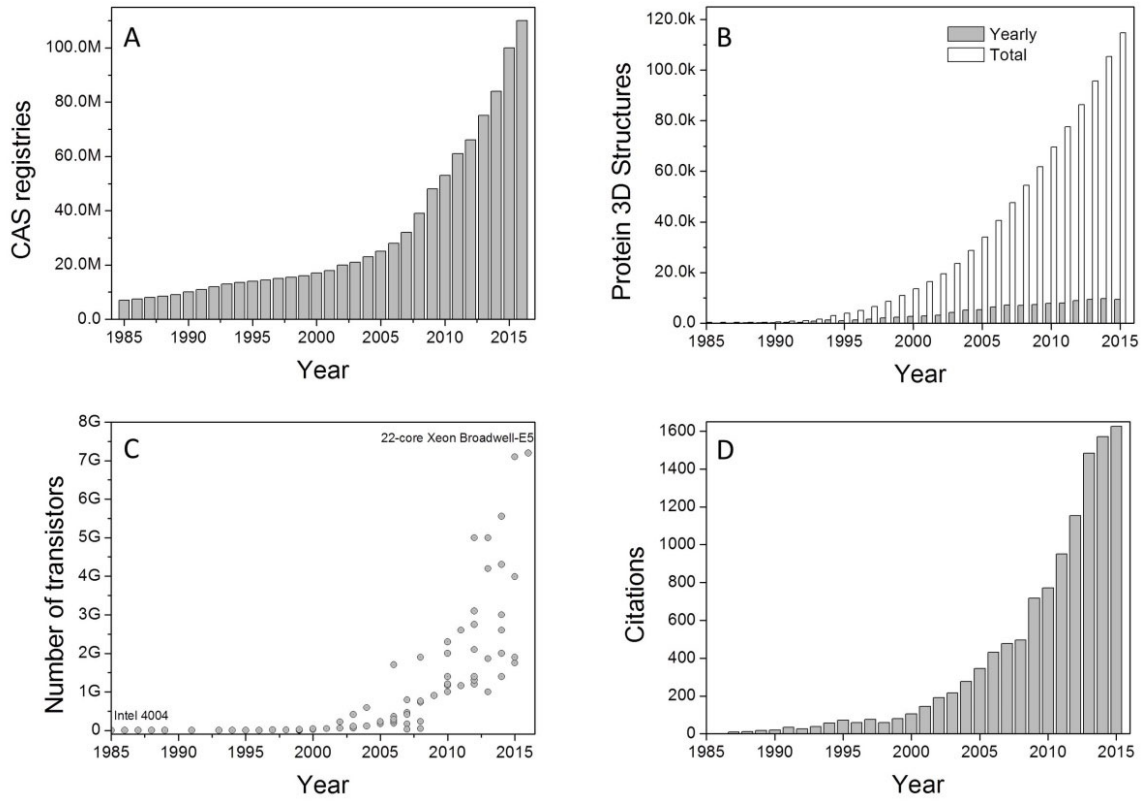


Fig. 1. Variation over time of a few key properties relevant for rational drug discovery. **A)** Registered substances in the Chemical Abstract Service (http://www.cas.org) **B)** Entries in the Protein Data Bank (Berman, et al., 2000) ; **C)** Transistor count in microprocessors (https://en.wikipedia.org/wiki/Transistor_count); **D)** Citations with the following combination of keywords "*in silico*" AND ("*drug discovery*" OR "*drug design*") in the PubMed resource (http://www.ncbi.nlm.nih.gov/pubmed).

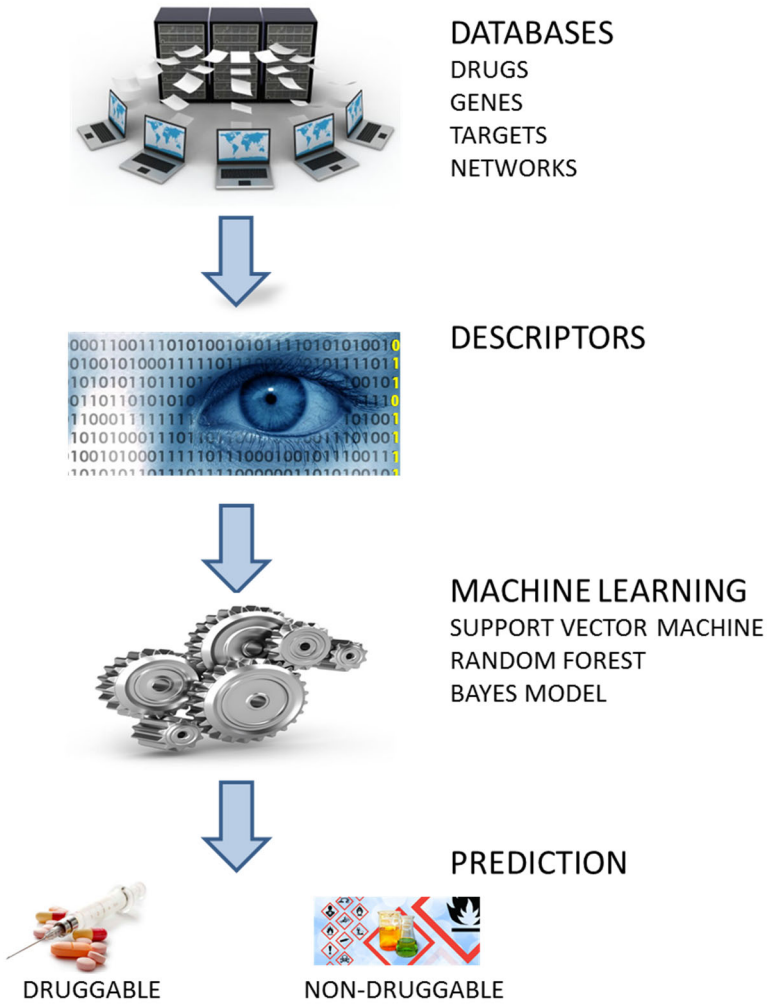


Fig.2 Target druggability prediction by machine learning algorithms. Databases of drugs, targets, genes or drug-target and target-target networks are mined to retrieve positive (druggable) and negative (non-druggable) instances which are represented by various descriptors (yellow digit describes the druggability status: undruggable,0; druggable, 1). Supervised machine learning algorithms are trained to discriminate druggable from non-druggable instances and further predict the status of novel targets.

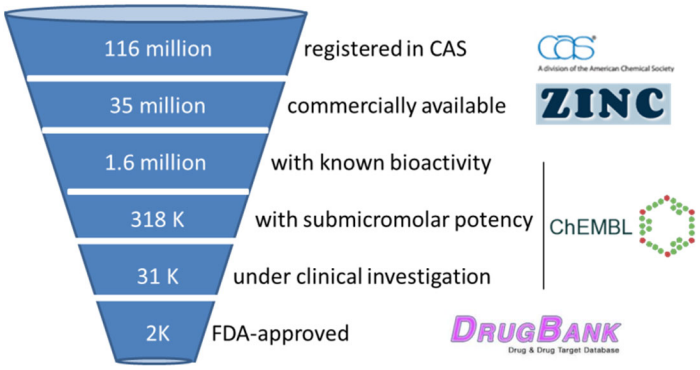


Fig.3. Number of chemical structures registered in chemistry (CAS), bioactivity (ChEMBL) and drug (DrugBank) databases.

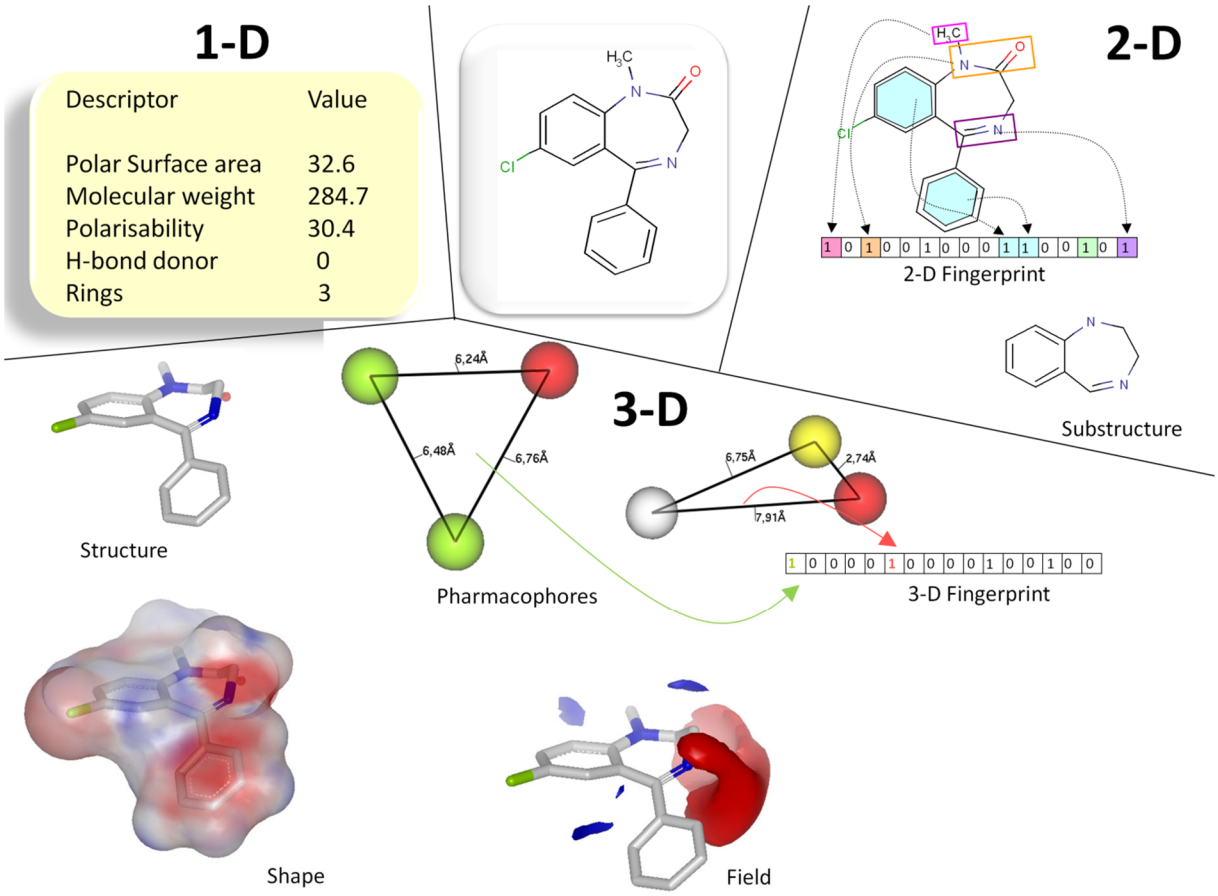


Fig. 4. Possible representations of a bioactive compound (diazepam). 1D descriptors encode simple property counts. 2D descriptors are based on the molecular graph and are represented by substructure or fingerprints accounting for the presence/absence of particular features across the graph. 3D fingerprints take into account the conformational freedom of the compound encoded as a pharmacophore, 3D fingerprint, shape or field.

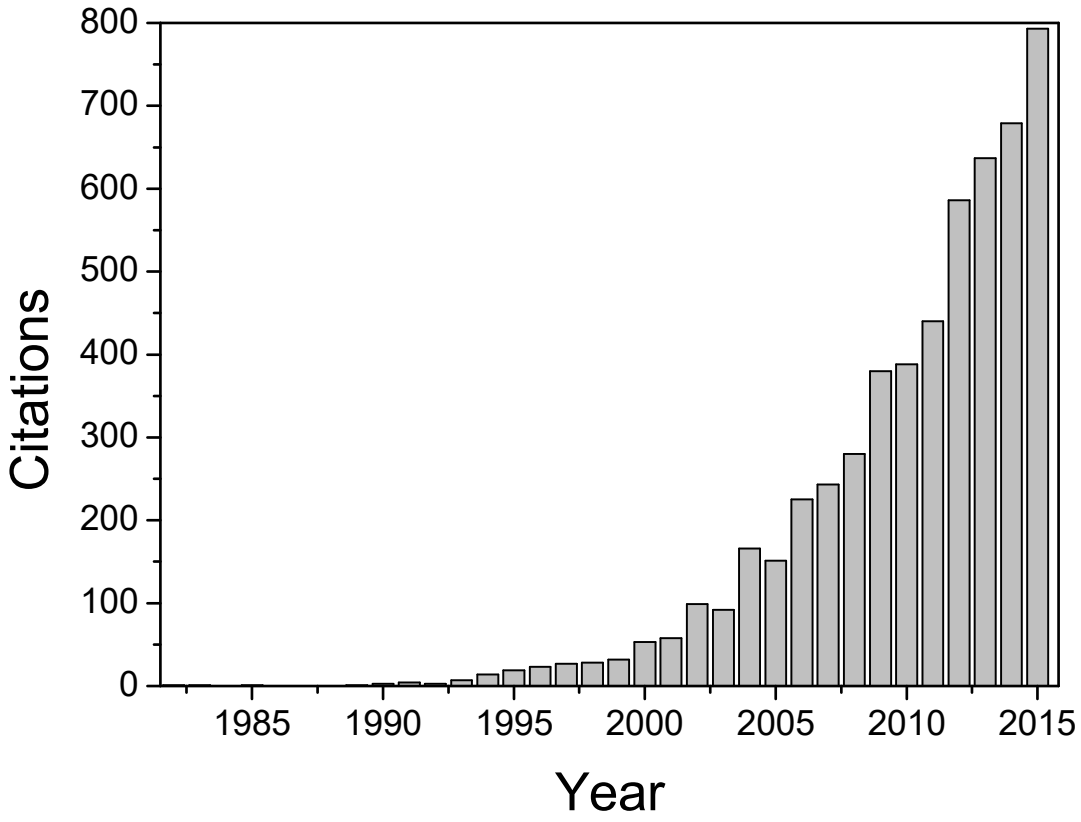


Fig.5. Number of citations in PubMed with the following keywords combination: ("in silico screening" OR "virtual screening" OR "ligand-based" or "structure-based" OR "receptor-based") AND ("agonist" OR "antagonist" OR "hit" OR "ligand") AND ("discover" OR "discovery" OR "identify" OR "identification" OR "confirm" OR "conformation" OR "validate" OR "validation" OR "experiment" OR "experimental")

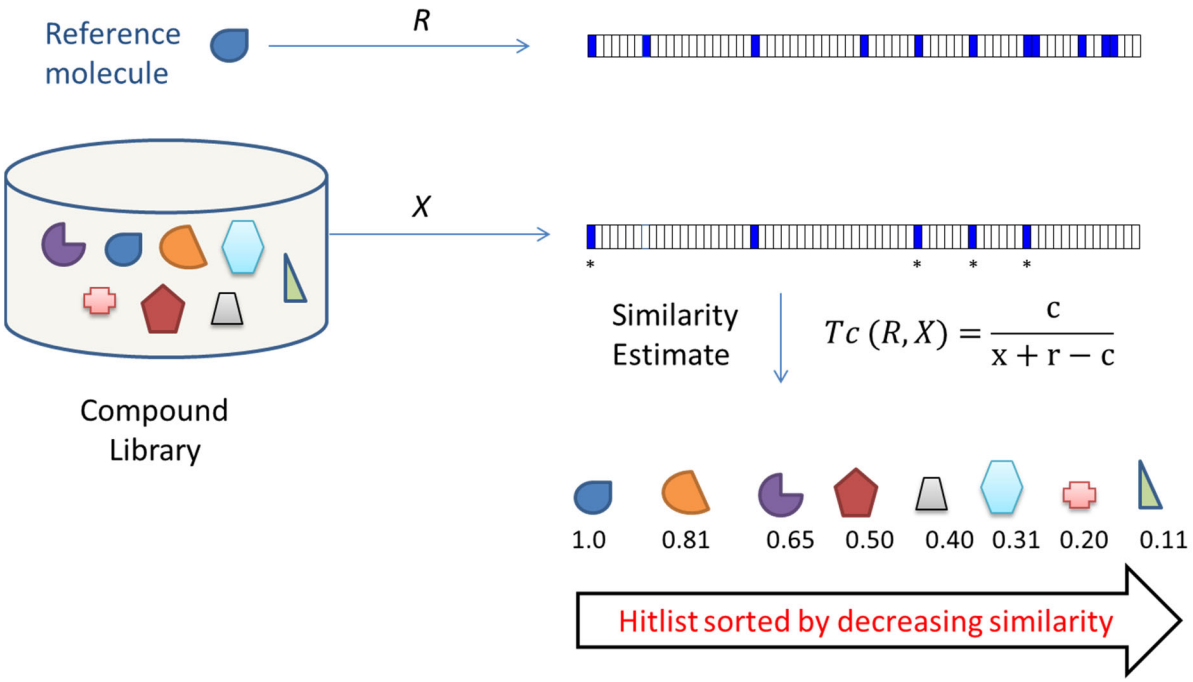


Fig.6. 2D similarity search principle. Molecule X from a compound library is compared to a known active (reference R) by computing their 2D structural fingerprints (bit strings) registering the presence or absence of key structural fragments. The similarity between X and R is estimated by the Tanimoto coefficient Tc which is a function of the number of common bits (c) and the number of bits unique to X (x) and R(r). Molecules are usually considered as chemically similar if Tc > 0.7.

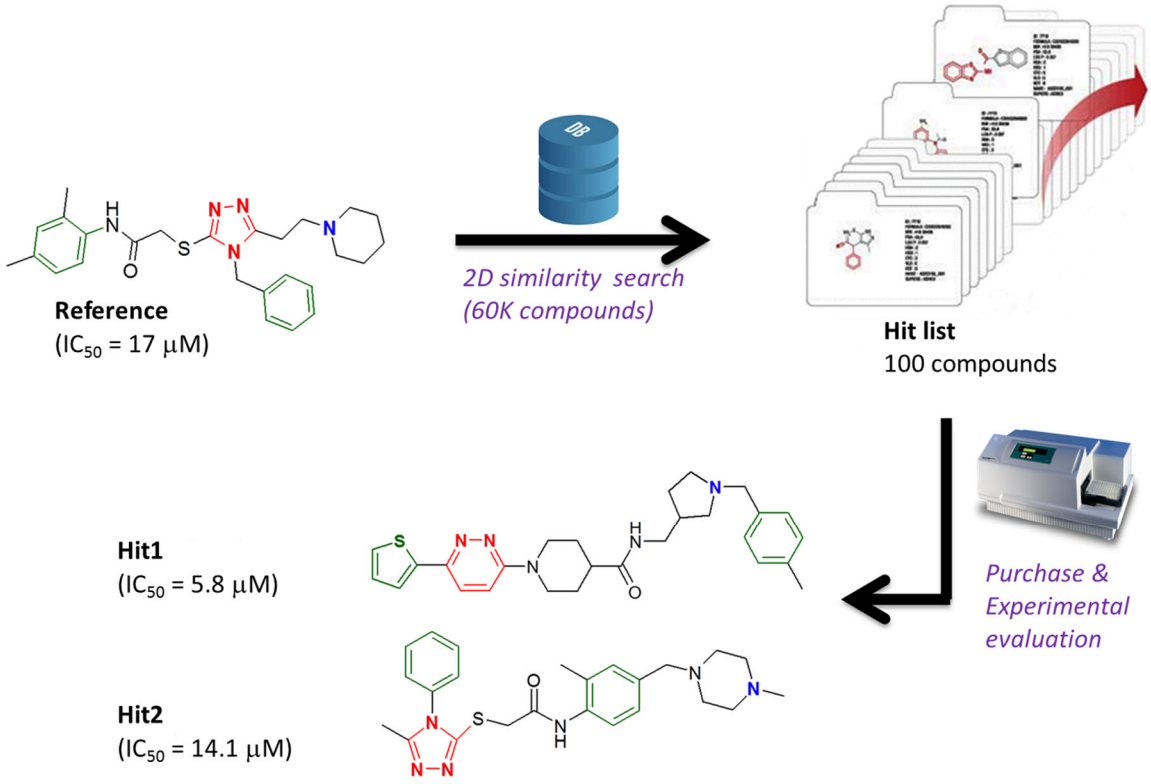


Fig.7. Identification of novel CCR5 receptor agonists by 2D similarity search (Kellenberger, et al., 2007)

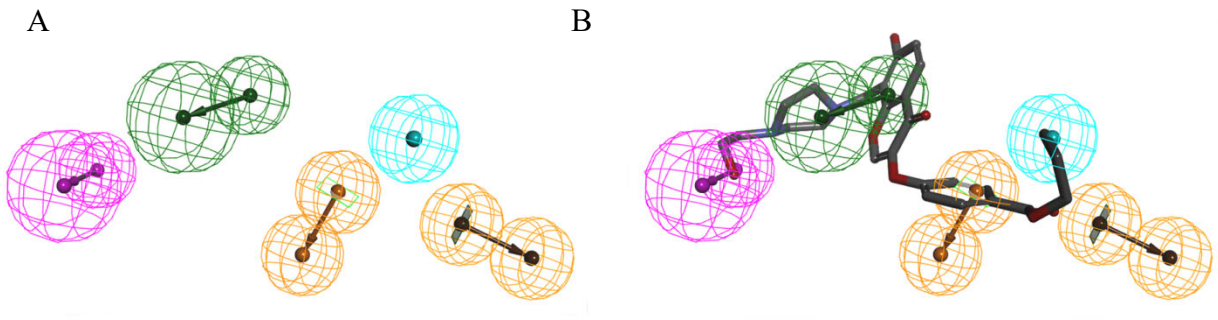


Fig.8. A) Representation of a pharmacophore as a collection of physicochemical properties (hydrophobic, cyan; aromatic, orange; hydrogen-bond donor, magenta; hydrogen bond acceptor, green) with a specific spatial orientation. The larger sphere (donor and acceptor features) indicates the position of the ligand atom whereas the smallest one describes the position of the protein atom to which it interacts. Vectors indicate the directionality of the interaction. **B)** Optimal fit of a molecule to the pharmacophore.

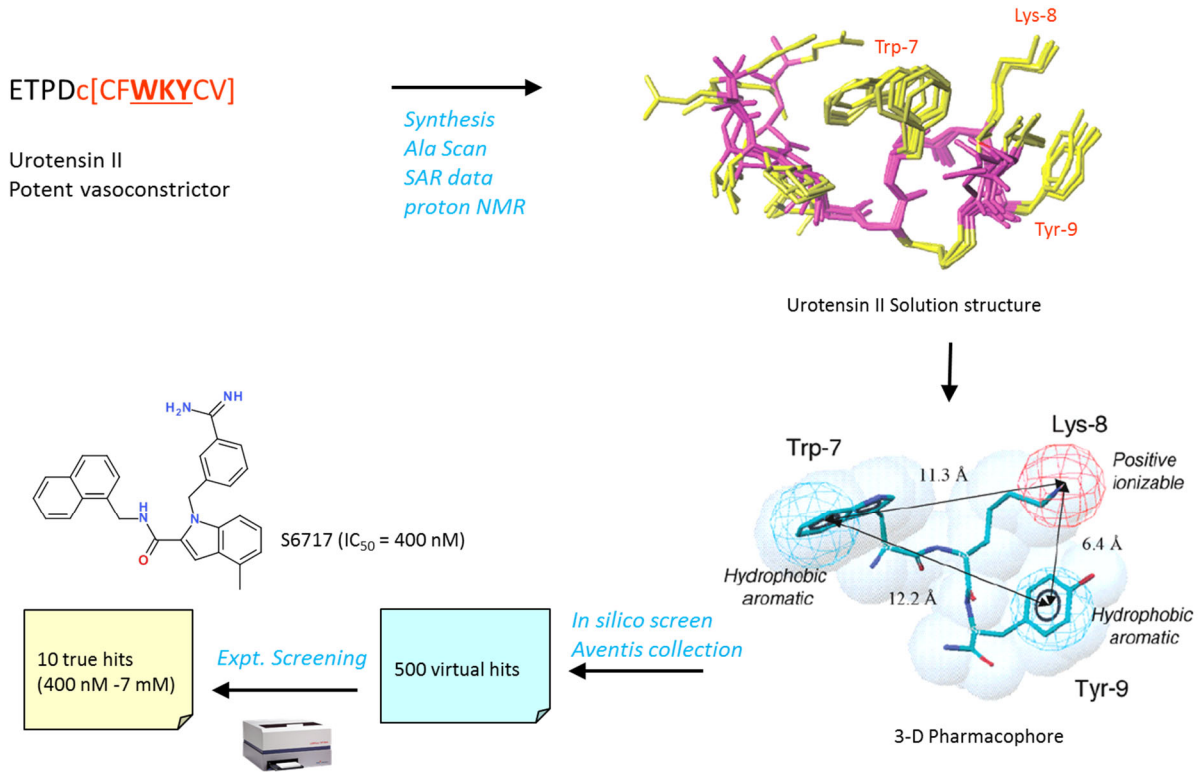


Fig.9. Example of a pharmacophore-based virtual screen.

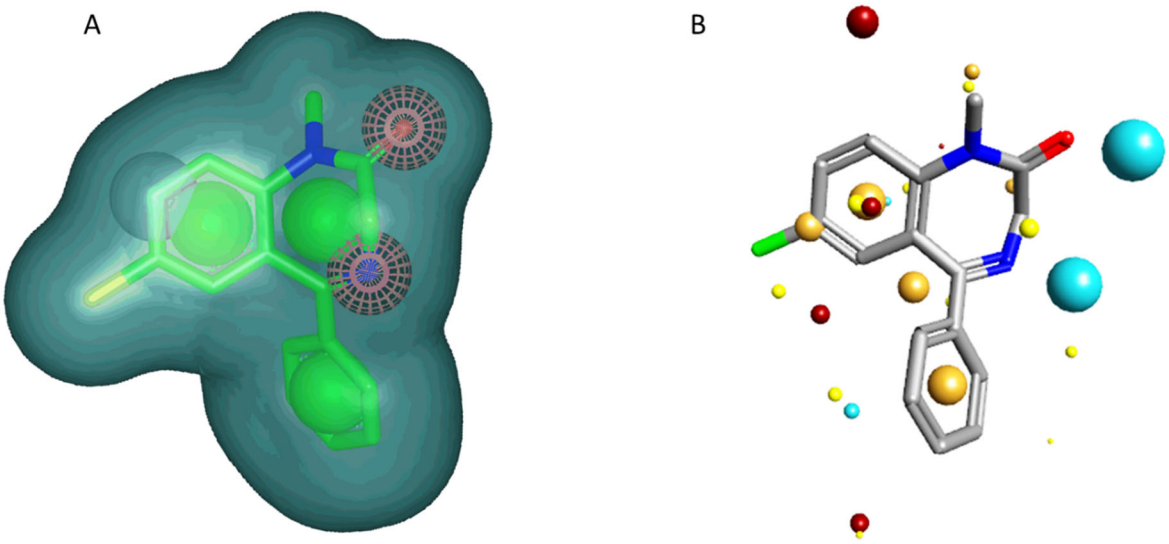


Fig.10. Alternative pharmacophore representations of the anxiolytic diazepam. **A)** Shape (green transparent surface) colored by pharmacophoric properties (aromatic, green; hydrogen-bond acceptor, red) from the ROCS software (OpenEye Scientific Software, Santa Fe, U.S.A). **B)** Steric and electrostatic fields, from TorchLite (Cresset Biomolecular Discovery, Ltd., Litlington, U.K.). Field minima corresponding to the tightest possible interactions are displayed as spheres (hydrogen bond acceptor, cyan; aromatic to H, red; hydrophobe, orange; van des Waals, yellow)

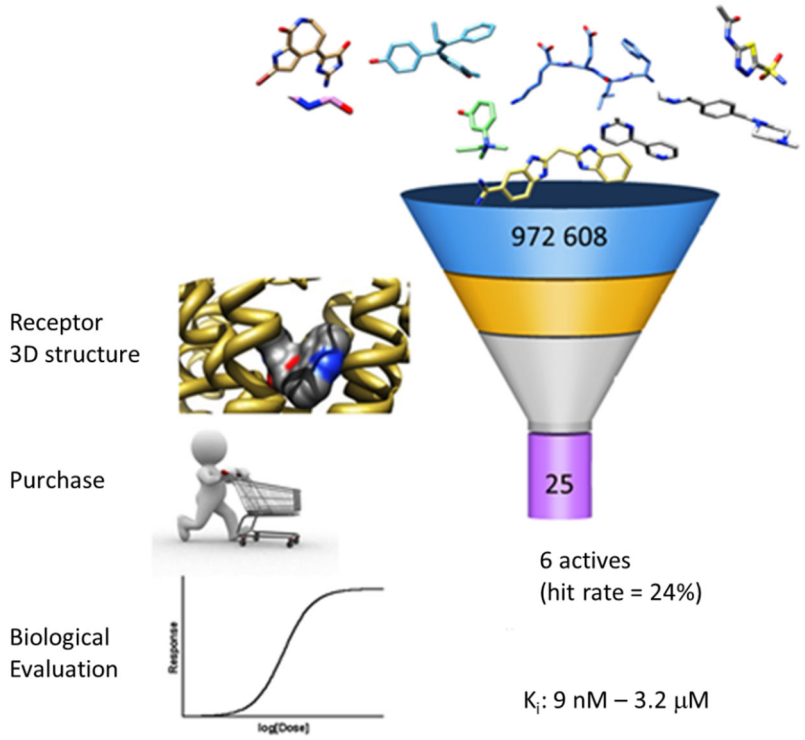


Fig. 11. Docking-based virtual screening for beta2 receptor antagonists.

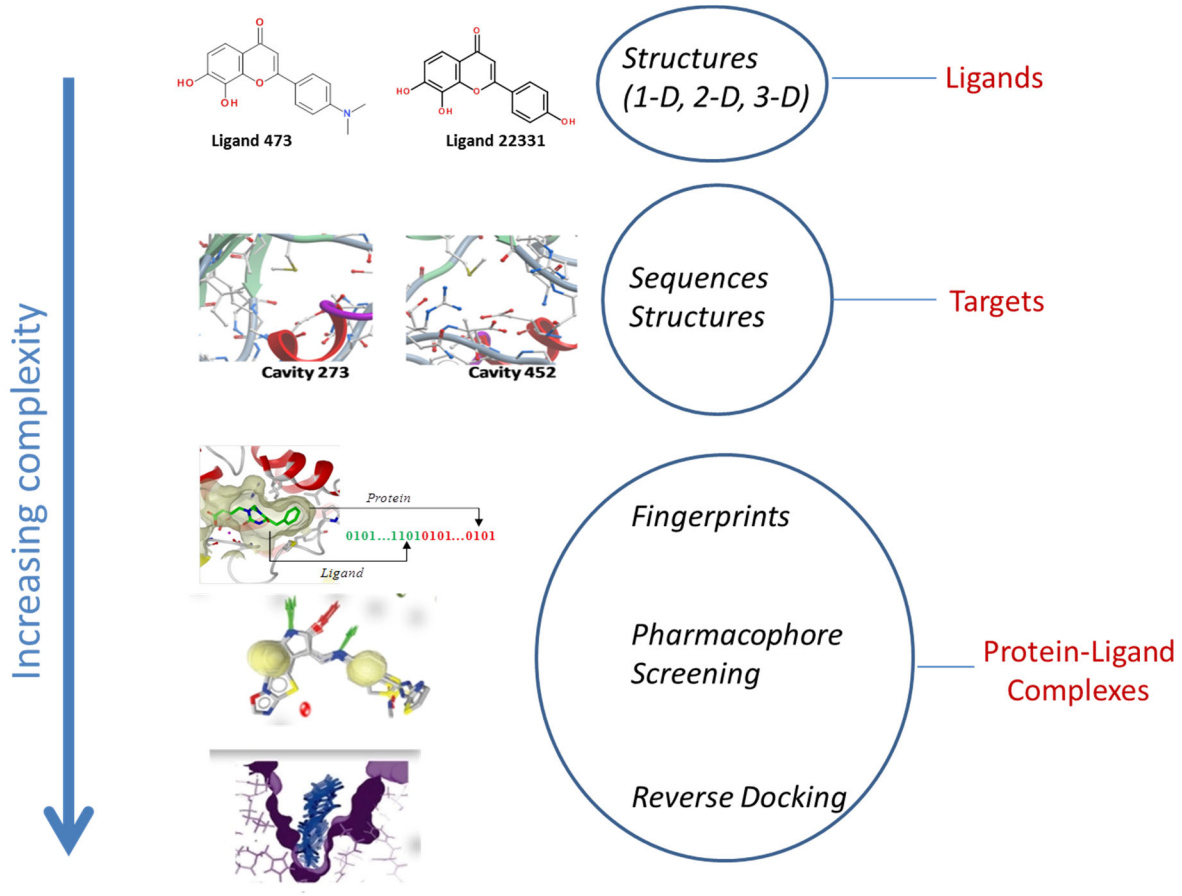


Fig.12. Computational approaches for target prediction.

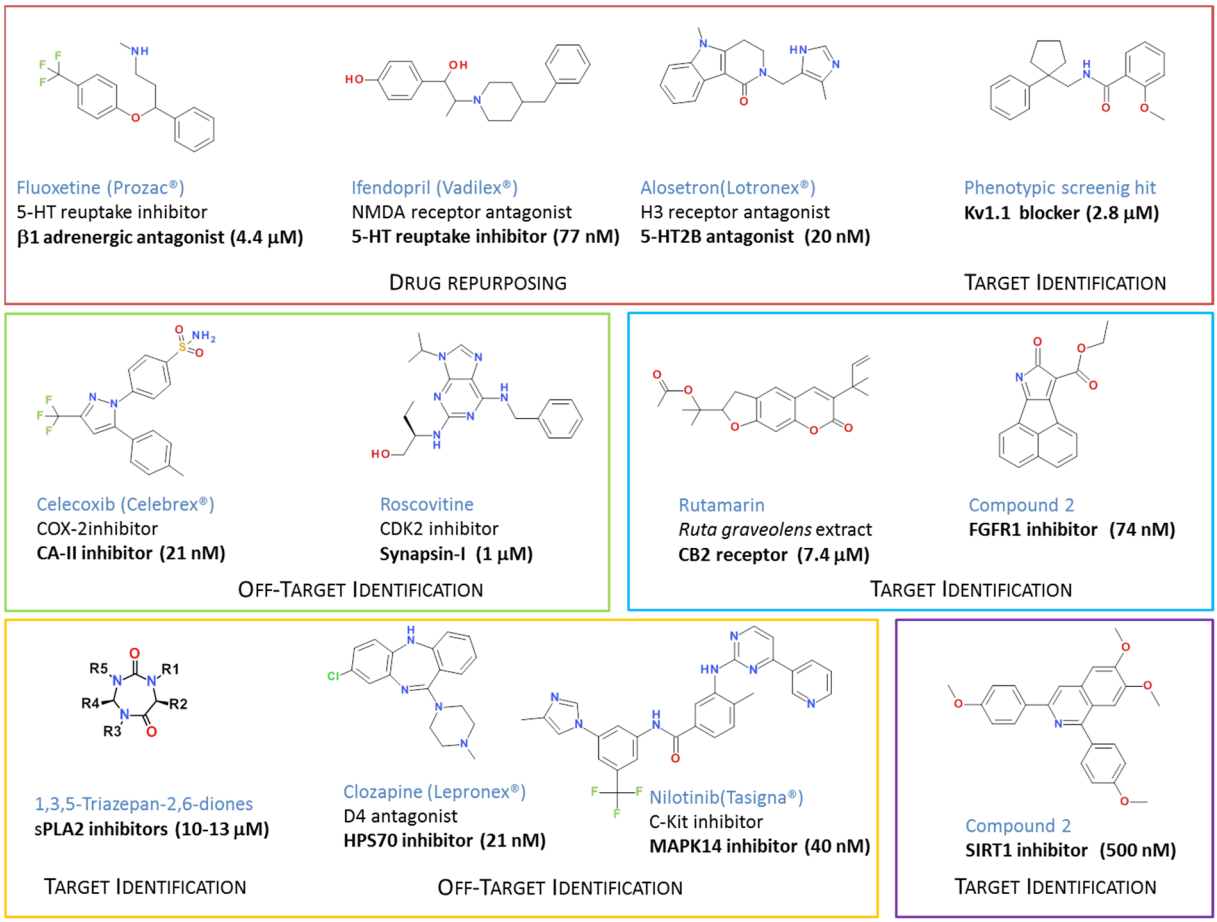


Fig.13. List of representative ligands for which main or secondary targets have been computationally predicted and experimentally confirmed. Predictions based on different methods are enclosed in colored boxes: Ligand-centric predictions (red), binding site-based predictions (green), pharmacophore-based predictions (cyan), docking-based predictions (yellow) and proteochemometrics-based predictions (violet). For each compound, the main target (when known) as well as the newly predicted target (in bold) with the corresponding binding constant are given.

References

Alex, A. A., & Millan, D. S. (2012). Contribution of structure-based drug design tothe discovery of marketed drugs. In D. J. Livingstone, Davis, A.M. (Ed.), *Drug design strategies: Quantitative approaches* (pp. 108-150): RSC publishing.

Aliagas, I., Gobbi, A., Heffron, T., Lee, M.-L., Ortwine, D. F., Zak, M., & Khojasteh, S. C. (2015). A probabilistic method to report predictions from a human liver microsomes stability QSAR model: a practical tool for drug discovery. *J. Comput-Aided Mol Des, 29*, 327-338.

Altman, R. B. (2007). PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet, 39*, 426-426.

Anighoro, A., & Bajorath, J. (2016). Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J Chem Inf Model*.

Babine, R. E., Bleckman, T. M., Kissinger, C. R., Showalter, R., Pelletier, L. A., Lewis, C., Tucker, K., Moomaw, E., Parge, H. E., & Villafranca, J. E. (1995). Design, synthesis and X-ray crystallographic studies of novel FKBP-12 ligands. *Bioorg Med Chem Lett, 5*, 1719-1724.

Baell, J., & Walters, M. A. (2014). Chemistry: Chemical con artists foil drug discovery. *Nature, 513*, 481-483.

Bakheet, T. M., & Doig, A. J. (2009). Properties and identification of human protein drug targets. *Bioinformatics, 25*, 451-457.

Ballester, P. J., Schreyer, A., & Blundell, T. L. (2014). Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model, 54*, 944-955.

Bembenek, S. D., Tounge, B. A., & Reynolds, C. H. (2009). Ligand efficiency and fragment-based drug discovery. *Drug Discov Today, 14*, 278-283.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*, 235-242.

Besnard, J., Ruda, G. F., Setola, V., Abecassis, K., Rodriguiz, R. M., Huang, X. P., Norval, S., Sassano, M. F., Shin, A. I., Webster, L. A., Simeons, F. R., Stojanovski, L., Prat, A., Seidah, N. G., Constam, D. B., Bickerton, G. R., Read, K. D., Wetsel, W. C., Gilbert, I. H., Roth, B. L., & Hopkins, A. L. (2012). Automated design of ligands to polypharmacological profiles. *Nature, 492*, 215-220.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nat Chem, 4*, 90-98.

Bohm, H. J. (1992). LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des, 6*, 593-606.

Borrel, A., Regad, L., Xhaard, H., Petitjean, M., & Camproux, A. C. (2015). PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J Chem Inf Model, 55*, 882-895.

Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov, 11*, 909-922.

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science, 321*, 263-266.

Chalopin, M., Tesse, A., Martinez, M. C., Rognan, D., Arnal, J. F., & Andriantsitohaina, R. (2010). Estrogen receptor alpha as a key target of red wine polyphenols action on the endothelium. *PLoS One, 5*, e8554.

Chen, J. M., Xu, S. L., Wawrzak, Z., Basarab, G. S., & Jordan, D. B. (1998). Structure-based design of potent inhibitors of scytalone dehydratase: displacement of a water molecule from the active site. *Biochemistry, 37*, 17735-17744.

Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C., & Huang, E. S. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol, 25*, 71-75.

Cheng, F., Li, W., Wang, X., Zhou, Y., Wu, Z., Shen, J., & Tang, Y. (2013). Adverse drug events: database construction and in silico prediction. *J Chem Inf Model, 53*, 744-752.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., & Pangalos, M. N. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat Rev Drug Discov, 13*, 419-431.

Cortes-Ciriano, I., Ain, Q., Subramanian, V., Lenselink, E. B., Mendez-Lucio, O., Ijzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T. E., van Westen, G. J. P., & Bender, A. (2015). Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm, 6*, 24-50.

Craig, J. C., Duncan, I. B., Hockley, D., Grief, C., Roberts, N. A., & Mills, J. S. (1991). Antiviral properties of Ro 31-8959, an inhibitor of human immunodeficiency virus (HIV) proteinase. *Antiviral Res, 16*, 295-305.

Cushman, D. W., Cheung, H. S., Sabo, E. F., & Ondetti, M. A. (1977). Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry, 16*, 5484-5491.

de Bethune, M. P. (2010). Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). *Antiviral Res, 85*, 75-90.

de Graaf, C., Kooistra, A. J., Vischer, H. F., Katritch, V., Kuijer, M., Shiroishi, M., Iwata, S., Shimamura, T., Stevens, R. C., de Esch, I. J., & Leurs, R. (2011). Crystal structure-based virtual screening for fragment-like ligands of the human histamine H(1) receptor. *J Med Chem, 54*, 8195-8206.

de Graaf, C., Rein, C., Piwnica, D., Giordanetto, F., & Rognan, D. (2011). Structure-based discovery of allosteric modulators of two related class B G-protein-coupled receptors. *ChemMedChem, 6*, 2159-2169.

Deng, Z., Chuaqui, C., & Singh, J. (2004). Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem, 47*, 337-344.

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

Diller, D. J., Connell, N. D., & Welsh, W. J. (2015). Avalanche for shape and feature-based virtual screening with 3D alignment. *J Comput Aided Mol Des, 29*, 1015-1024.

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

Durrant, J. D., Amaro, R. E., Xie, L., Urbaniak, M. D., Ferguson, M. A., Haapalainen, A., Chen, Z., Di Guilmi, A. M., Wunder, F., Bourne, P. E., & McCammon, J. A. (2010). A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput Biol, 6*, e1000648.

Edfeldt, F. N., Folmer, R. H., & Breeze, A. L. (2011). Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov Today, 16*, 284-287.

Ehrt, C., Brinkjost, T., & Koch, O. (2016). Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J Med Chem, 59*, 4121-4151.

Ekins, S., Williams, A. J., Krasowski, M. D., & Freundlich, J. S. (2011). In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov Today, 16*, 298-310.

Fishburn, C. S. (2013). Attenuating attrition. *SciBX, 6*.

Fishburn, C. S. (2013). Translational research: the changing landscape of drug discovery. *Drug Discov Today, 18*, 487-494.

Flohr, S., Kurz, M., Kostenis, E., Brkovich, A., Fournier, A., & Klabunde, T. (2002). Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II. *J Med Chem, 45*, 1799-1805.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model, 50*, 1189-1204.

Fraczkiewicz, R., Lobell, M., Goller, A. H., Krenz, U., Schoenneis, R., Clark, R. D., & Hillisch, A. (2015). Best of both worlds: combining pharma data and state of the art modeling technology to improve in Silico pKa prediction. *J Chem Inf Model, 55*, 389-397.

Gabel, J., Desaphy, J., & Rognan, D. (2014). Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J Chem Inf Model, 54*, 2807-2815.

Gao, C., Thorsteinson, N., Watson, I., Wang, J., & Vieth, M. (2015). Knowledge-Based Strategy to Improve Ligand Pose Prediction Accuracy for Lead Optimization. *J Chem Inf Model, 55*, 1460-1468.

Gao, M., & Skolnick, J. (2012). The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proc Natl Acad Sci U S A, 109*, 3784-3789.

Garreau de Loubresse, N., Prokhorova, I., Holtkamp, W., Rodnina, M. V., Yusupova, G., & Yusupov, M. (2014). Structural basis for the inhibition of the eukaryotic ribosome. *Nature, 513*, 517-522.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res, 40*, D1100-1107.

Ghafourian, T., & Amin, Z. (2013). QSAR models for the prediction of plasma protein binding. *Bioimpacts, 3*, 21-27.

Ghosh, E., Kumari, P., Jaiman, D., & Shukla, A. K. (2015). Methodological advances: the unsung heroes of the GPCR structural revolution. *Nat Rev Mol Cell Biol, 16*, 69-81.

Green, D. V., Leach, A. R., & Head, M. S. (2012). Computer-aided molecular design under the SWOTlight. *J Comput Aided Mol Des, 26*, 51-56.

Greer, J., Erickson, J. W., Baldwin, J. J., & Varney, M. D. (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J Med Chem, 37*, 1035-1054.

Grosdidier, A., Zoete, V., & Michielin, O. (2009). Blind docking of 260 protein-ligand complexes with EADock 2.0. *J Comput Chem, 30*, 2021-2030.

Hann, M. M., & Keseru, G. M. (2012). Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nat Rev Drug Discov, 11*, 355-365.

Hartenfeller, M., Eberle, M., Meier, P., Nieto-Oberhuber, C., Altmann, K. H., Schneider, G., Jacoby, E., & Renner, S. (2011). A collection of robust organic synthesis reactions for in silico molecule design. *J Chem Inf Model, 51*, 3093-3098.

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., & Schneider, G. (2012). DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol, 8*, e1002380.

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

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nat Biotechnol, 32*, 40-51.

Heikamp, K., & Bajorath, J. (2013). The future of virtual compound screening. *Chem Biol Drug Des, 81*, 33-40.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., & Schuffenhauer, A. (2006). New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model, 46*, 462-470.

Hevener, K. E., Mehboob, S., Su, P. C., Truong, K., Boci, T., Deng, J., Ghassemi, M., Cook, J. L., & Johnson, M. E. (2012). Discovery of a novel and potent class of F. tularensis enoyl-reductase (FabI) inhibitors by molecular shape and electrostatic matching. *J Med Chem, 55*, 268-279.

Hillisch, A., Heinrich, N., & Wild, H. (2015). Computational Chemistry in the Pharmaceutical Industry: From Childhood to Adolescence. *ChemMedChem, 10*, 1958-1962.

Hillisch, A., & Hilgenfeld, R. (2003). The role of protein 3D structures in the drug discovery process. In A. Hillisch & R. Hilgenfeld (Eds.), *Modern methods in drug discovery* (pp. 157-182): Birkhäuser Verlag.

Hindle, S. A., Rarey, M., Buning, C., & Lengauer, T. (2002). Flexible docking under pharmacophore type constraints. *J Comput Aided Mol Des, 16*, 129-149.

Hol, W. G. J. (1988). Protein cristallography and computer-graphics towards rational drug design. *Angew Chem Int Ed Engl, 25*, 767-778.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nat Rev Drug Discov, 1*, 727-730.

Hopkins, A. L., Groom, C. R., & Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discov Today, 9*, 430-431.

Hopkins, A. L., Mason, J. S., & Overington, J. P. (2006). Can we rationally design promiscuous drugs? *Curr Opin Struct Biol, 16*, 127-136.

Horvath, D., Lisurek, M., Rupp, B., Kuhne, R., Specker, E., von Kries, J., Rognan, D., Andersson, C. D., Almqvist, F., Elofsson, M., Enqvist, P. A., Gustavsson, A. L., Remez, N., Mestres, J., Marcou, G., Varnek, A., Hibert, M., Quintana, J., & Frank, R. (2014). Design of a general-purpose European compound screening library for EU-OPENSCREEN. *ChemMedChem, 9*, 2309-2326.

Hou, T., Wang, J., Li, Y., & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model, 51*, 69-82.

Hoveyda, H. R., Marsault, E., Gagnon, R., Mathieu, A. P., Vezina, M., Landry, A., Wang, Z., Benakli, K., Beaubien, S., Saint-Louis, C., Brassard, M., Pinault, J. F., Ouellet, L., Bhat, S., Ramaseshan, M., Peng, X., Foucher, L., Beauchemin, S., Bherer, P., Veber, D. F., Peterson, M. L., & Fraser, G. L. (2011). Optimization of the potency and pharmacokinetic properties of a macrocyclic ghrelin receptor agonist (Part I): Development of ulimorelin (TZP-101) from hit to clinic. *J Med Chem, 54*, 8305-8320.

Huang, T. W., Zaretzki, J., Bergeron, C., Bennett, K. P., & Breneman, C. M. (2013). DR-predictor: incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism. *J Chem Inf Model, 53*, 3352-3366.

Hughes, L. D., Palmer, D. S., Nigsch, F., & Mitchell, J. B. (2008). Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J Chem Inf Model, 48*, 220-232.

Irwin, J. J., Duan, D., Torosyan, H., Doak, A. K., Ziebart, K. T., Sterling, T., Tumanian, G., & Shoichet, B. K. (2015). An Aggregation Advisor for Ligand Discovery. *J Med Chem, 58*, 7076-7087.

Johnson, D. K., & Karanicolas, J. (2016). Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein-Protein Interactions. *J Chem Inf Model, 56*, 399-411.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255-260.

Kalaszi, A., Szisz, D., Imre, G., & Polgar, T. (2014). Screen3D: a novel fully flexible high-throughput shape-similarity search method. *J Chem Inf Model, 54*, 1036-1049.

Kandoi, G., Acencio, M. L., & Lemke, N. (2015). Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review. *Front Physiol, 6*, 366.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res, 42*, D199-205.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L., Edwards, D. D., Shoichet, B. K., & Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature, 462*, 175-181.

Kellenberger, E., Springael, J. Y., Parmentier, M., Hachet-Haas, M., Galzi, J. L., & Rognan, D. (2007). Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J Med Chem, 50*, 1294-1303.

Kelley, B. P., Brown, S. P., Warren, G. L., & Muchmore, S. W. (2015). POSIT: Flexible Shape-Guided Docking For Pose Prediction. *J Chem Inf Model, 55*, 1771-1780.

Kelly, M. D., & Mancera, R. L. (2004). Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J Chem Inf Comput Sci, 44*, 1942-1951.

Khamis, M. A., Gomaa, W., & Ahmed, W. F. (2015). Machine learning in computational docking. *Artif Intell Med, 63*, 135-152.

Kier, L. B. (1967). Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. *Mol Pharmacol, 3*, 487-494.

Kilchmann, F., Marcaida, M. J., Kotak, S., Schick, T., Boss, S. D., Awale, M., Gonczy, P., & Reymond, J. L. (2016). Discovery of a Selective Aurora A Kinase Inhibitor by Virtual Screening. *J Med Chem*.

Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., & Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm Res, 31*, 1002-1014.

Kirkpatrick, P. (2012). An audience with Chris Lipinski. *Nat Rev Drug Discov, 11*, 900-901.

Koes, D. R., & Camacho, C. J. (2012). PocketQuery: protein-protein interaction inhibitor starting points from protein-protein interaction structure. *Nucleic Acids Res, 40*, W387-392.

Kolb, P., Rosenbaum, D. M., Irwin, J. J., Fung, J. J., Kobilka, B. K., & Shoichet, B. K. (2009). Structure-based discovery of beta2-adrenergic receptor ligands. *Proc Natl Acad Sci U S A, 106*, 6843-6848.

Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., & Brenk, R. (2011). DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *J Chem Inf Model, 51*, 2829-2842.

Krier, M., Bret, G., & Rognan, D. (2006). Assessing the scaffold diversity of screening libraries. *J Chem Inf Model, 46*, 512-524.

Kruger, D. M., & Evers, A. (2010). Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem, 5*, 148-158.

Kuang, Q., Purhonen, P., & Hebert, H. (2015). Structure of potassium channels. *Cell Mol Life Sci, 72*, 3677-3693.

Kubinyi, H. (2006). Sucess stories of computer-aided design. In S. Ekins (Ed.), *Computer applications in pharmaceutical research and development* (pp. 377-424): John Wiley & SOns, Inc.

Kuenemann, M. A., Bourbon, L. M., Labbe, C. M., Villoutreix, B. O., & Sperandio, O. (2014). Which three-dimensional characteristics make efficient inhibitors of protein-protein interactions? *J Chem Inf Model, 54*, 3067-3079.

Kufareva, I., Ilatovskiy, A. V., & Abagyan, R. (2012). Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res, 40*, D535-540.

Kuhn, B., Gerber, P., Schulz-Gasch, T., & Stahl, M. (2005). Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem, 48*, 4040-4048.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res, 44*, D1075-1079.

Kumar, A., & Zhang, K. Y. (2015). Application of Shape Similarity in Pose Selection and Virtual Screening in CSARdock2014 Exercise. *J Chem Inf Model*.

Laggner, C., Kokel, D., Setola, V., Tolia, A., Lin, H., Irwin, J. J., Keiser, M. J., Cheung, C. Y., Minor, D. L., Jr., Roth, B. L., Peterson, R. T., & Shoichet, B. K. (2012). Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat Chem Biol, 8*, 144-146.

Lanevskij, K., Japertas, P., & Didziapetris, R. (2013). Improving the prediction of drug disposition in the brain. *Expert Opin Drug Metab Toxicol, 9*, 473-486.

Langridge, R., Ferrin, T. E., Kuntz, I. D., & Connolly, M. L. (1981). Real-time color graphics in studies of molecular interactions. *Science, 211*, 661-666.

Larregieu, C. A., & Benet, L. Z. (2013). Drug discovery and regulatory considerations for improving in silico and in vitro predictions that use Caco-2 as a surrogate for human intestinal permeability measurements. *AAPS J, 15*, 483-497.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., & Wishart,

D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res, 42*, D1091-1097.

Leach, A. R., Gillet, V. J., Lewis, R. A., & Taylor, R. (2010). Three-dimensional pharmacophore methods in drug discovery. *J Med Chem, 53*, 539-558.

Lei, Q., Liu, H., Peng, Y., & Xiao, P. (2015). In silico target fishing and pharmacological profiling for the isoquinoline alkaloids of Macleayacordata (Bo Luo Hui). *Chin Med, 10*, 37.

Li, H., Leung, K. S., Wong, M. H., & Ballester, P. J. (2016). USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res, 44*, W436-441.

Li, L., Wang, B., & Meroueh, S. O. (2011). Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. *J Chem Inf Model, 51*, 2132-2138.

Li, Q., & Lai, L. (2007). Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics, 8*, 353.

Li, Y., Han, L., Liu, Z., & Wang, R. (2014). Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J Chem Inf Model, 54*, 1717-1736.

Li, Y. Y., An, J., & Jones, S. J. (2011). A computational approach to finding novel targets for existing drugs. *PLoS Comput Biol, 7*, e1002139.

Liu, C., Wrobleski, S. T., Lin, J., Ahmed, G., Metzger, A., Wityak, J., Gillooly, K. M., Shuster, D. J., McIntyre, K. W., Pitt, S., Shen, D. R., Zhang, R. F., Zhang, H., Doweyko, A. M., Diller, D., Henderson, I., Barrish, J. C., Dodd, J. H., Schieven, G. L., & Leftheris, K. (2005). 5-Cyanopyrimidine derivatives as a novel class of potent, selective, and orally active inhibitors of p38alpha MAP kinase. *J Med Chem, 48*, 6261-6270.

Loughney, D., Claus, B. L., & Johnson, S. R. (2011). To measure is to know: an approach to CADD performance metrics. *Drug Discov Today, 16*, 548-554.

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Cote, S., Shoichet, B. K., & Urban, L. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature, 486*, 361-367.

Loving, K. A., Lin, A., & Cheng, A. C. (2014). Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility. *PLoS Comput Biol, 10*, e1003741.

Maggiora, G. M. (2006). On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model, 46*, 1535.

Marcou, G., & Rognan, D. (2007). Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model, 47*, 195-207.

Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J Med Chem, 45*, 4350-4358.

Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., & Labaudiniere, R. F. (1999). New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J Med Chem, 42*, 3251-3264.

Meslamani, J., Bhajun, R., Martz, F., & Rognan, D. (2013). Computational profiling of bioactive compounds using a target-dependent composite workflow. *J Chem Inf Model, 53*, 2322-2333.

Meslamani, J., Li, J., Sutter, J., Stevens, A., Bertrand, H. O., & Rognan, D. (2012). Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J Chem Inf Model, 52*, 943-955.

Mestres, J., & Veeneman, G. H. (2003). Identification of "latent hits" in compound screening collections. *J Med Chem, 46*, 3441-3444.

Mirza, S. B., Salmas, R. E., Fatmi, M. Q., & Durdagi, S. (2016). Virtual screening of eighteen million compounds against dengue virus: Combined molecular docking and molecular dynamics simulations study. *J Mol Graph Model, 66*, 99-107.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., & Corbeil, C. R. (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol, 153 Suppl 1*, S7-26.

Morphy, R. (2010). Selectively nonselective kinase inhibition: striking the right balance. *J Med Chem, 53*, 1413-1437.

Mpamhanga, C. P., Chen, B., McLay, I. M., & Willett, P. (2006). Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J Chem Inf Model, 46*, 686-698.

Muegge, I., & Zhang, Q. (2015). 3D virtual screening of large combinatorial spaces. *Methods, 71*, 14-20.

Muller, P., Lena, G., Boilard, E., Bezzine, S., Lambeau, G., Guichard, G., & Rognan, D. (2006). In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J Med Chem, 49*, 6768-6778.

Naderi, M., Alvin, C., Ding, Y., Mukhopadhyay, S., & Brylinski, M. (2016). A graph-based approach to construct target-focused libraries for virtual screening. *J Cheminform, 8*, 14.

Nicholls, A., McGaughey, G. B., Sheridan, R. P., Good, A. C., Warren, G., Mathieu, M., Muchmore, S. W., Brown, S. P., Grant, J. A., Haigh, J. A., Nevins, N., Jain, A. N., & Kelley, B. (2010). Molecular shape and medicinal chemistry: a perspective. *J Med Chem, 53*, 3862-3886.

Over, B., Wetzel, S., Grutter, C., Nakai, Y., Renner, S., Rauh, D., & Waldmann, H. (2012). Natural-product-derived fragments for fragment-based ligand discovery. *Nat Chem, 5*, 21-28.

Perez-Nueno, V. I., Souchet, M., Karaboga, A. S., & Ritchie, D. W. (2015). GESSE: Predicting Drug Side Effects from Drug-Target Relationships. *J Chem Inf Model, 55*, 1804-1823.

Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., Deng, Z., Davies, J. W., Jenkins, J. L., & Glick, M. (2012). Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem Biol, 7*, 1399-1409.

Petrone, P. M., Wassermann, A. M., Lounkine, E., Kutchukian, P., Simms, B., Jenkins, J., Selzer, P., & Glick, M. (2013). Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discov Today*.

Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des, 27*, 675-679.

PubMed. (2016). US National Library of Medicine, National Institutes of Health, Bethesda MD, 20894 U.S.A. (http://www.ncbi.nlm.nih.gov/pubmed). In.

Rabal, O., Amr, F. I., & Oyarzabal, J. (2015). Novel Scaffold FingerPrint (SFP): applications in scaffold hopping and scaffold-based selection of diverse compounds. *J Chem Inf Model, 55*, 1-18.

Raies, A. B., & Bajic, V. B. (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci, 6*, 147-172.

Rask-Andersen, M., Almen, M. S., & Schioth, H. B. (2011). Trends in the exploitation of novel drug targets. *Nat Rev Drug Discov, 10*, 579-590.

Rask-Andersen, M., Masuram, S., & Schioth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu Rev Pharmacol Toxicol, 54*, 9-26.

Reker, D., Perna, A. M., Rodrigues, T., Schneider, P., Reutlinger, M., Monch, B., Koeberle, A., Lamers, C., Gabler, M., Steinmetz, H., Muller, R., Schubert-Zsilavecz, M., Werz, O., & Schneider, G. (2014). Revealing the macromolecular targets of complex natural products. *Nat Chem, 6*, 1072-1078.

Reymond, J. L. (2015). The chemical space project. *Acc Chem Res, 48*, 722-730.

Ripphausen, P., Stumpfe, D., & Bajorath, J. (2012). Analysis of structure-based virtual screening studies and characterization of identified active compounds. *Future Med Chem, 4*, 603-613.

Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E. M., Rogers-Evans, M., Neidhart, W., Stalder, H., Dillon, M., Sjogren, E., Fotouhi, N., Gillespie, P., Goodnow, R., Harris, W., Jones, P., Taniguchi, M., Tsujii, S., von der Saal, W., Zimmermann,

G., & Schneider, G. (2002). Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J Med Chem, 45*, 137-142.
Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *J Chem Inf Model, 50*, 742-754.
Rognan, D. (2010). Structure-based approaches to target fishing and ligand profile. *Mol Inform, 29*, 167-187.
Rognan, D. (2013a). Proteome-scale docking: myth and reality. *Drug Discov Today Technol, 10*, e403-409.
Rognan, D. (2013b). Towards the Next Generation of Computational Chemogenomics Tools. *Mol Inform, 32*, 1029-1034.
Rollinger, J. M., Schuster, D., Danzl, B., Schwaiger, S., Markt, P., Schmidtke, M., Gertsch, J., Raduner, S., Wolber, G., Langer, T., & Stuppner, H. (2009). In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens. *Planta Med, 75*, 195-204.
Roy, A., & Skolnick, J. (2015). LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics, 31*, 539-544.
Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J. L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J Chem Inf Model, 52*, 2864-2875.
Rydberg, P., Gloriam, D. E., & Olsen, L. (2010). The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics, 26*, 2988-2989.
Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev., 3*, 210-229.
Schmidtke, P., & Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem, 53*, 5858-5867.
Schneider, P., & Schneider, G. (2016). De Novo Design at the Edge of Chaos. *J Med Chem, 59*, 4077-4086.
Schnur, D. M., Hermsmeier, M. A., & Tebben, A. J. (2006). Are target-family-privileged substructures truly privileged? *J Med Chem, 49*, 2000-2009.
Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martinez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., & Agrafiotis, D. K. (2012). Recognizing pitfalls in virtual screening: a critical review. *J Chem Inf Model, 52*, 867-881.
Seddon, G., Lounnas, V., McGuire, R., van den Bergh, T., Bywater, R. P., Oliveira, L., & Vriend, G. (2012). Drug design for ever, from hype to hope. *J Comput Aided Mol Des, 26*, 137-150.
Segall, M. (2014). Advances in multiparameter optimization methods for de novo drug design. *Expert Opin Drug Discov, 9*, 803-817.
Shaked, I., Oberhardt, M. A., Atias, N., Sharan, R., & Ruppin, E. (2016). Metabolic Network Prediction of Drug Side Effects. *Cell Syst, 2*, 209-213.
Sheridan, R. P., Maiorov, V. N., Holloway, M. K., Cornell, W. D., & Gao, Y. D. (2010). Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J Chem Inf Model, 50*, 2029-2040.
Simon, Z., Peragovics, A., Vigh-Smeller, M., Csukly, G., Tombor, L., Yang, Z., Zahoranszky-Kohalmi, G., Vegner, L., Jelinek, B., Hari, P., Hetenyi, C., Bitter, I., Czobor, P., & Malnasi-Csizmadia, A. (2012). Drug effect prediction by polypharmacology-based interaction profiling. *J Chem Inf Model, 52*, 134-145.
Sjogren, E., Thorn, H., & Tannergren, C. (2016). In Silico Modeling of Gastrointestinal Drug Absorption: Predictive Performance of Three Physiologically Based Absorption Models. *Mol Pharm, 13*, 1763-1778.
Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W., Jr. (2014). Computational methods in drug discovery. *Pharmacol Rev, 66*, 334-395.
Spyrakis, F., & Cavasotto, C. N. (2015). Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch Biochem Biophys, 583*, 105-119.
Sterling, T., & Irwin, J. J. (2015). ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model, 55*, 2324-2337.

Surgand, J. S., Rodrigo, J., Kellenberger, E., & Rognan, D. (2006). A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins, 62*, 509-538.

Todeschini, R., & Consonni, V. (2000). *Handbook of Molecular Descriptors*: Wiley-VCH.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J Chem Inf Model, 52*, 2884-2901.

Turner, S. R., Strohbach, J. W., Tommasi, R. A., Aristoff, P. A., Johnson, P. D., Skulnick, H. I., Dolak, L. A., Seest, E. P., Tomich, P. K., Bohanon, M. J., Horng, M. M., Lynn, J. C., Chong, K. T., Hinshaw, R. R., Watenpaugh, K. D., Janakiraman, M. N., & Thaisrivongs, S. (1998). Tipranavir (PNU-140690): a potent, orally bioavailable nonpeptidic HIV protease inhibitor of the 5,6-dihydro-4-hydroxy-2-pyrone sulfonamide class. *J Med Chem, 41*, 3467-3476.

van de Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov, 2*, 192-204.

van Westen, G. J. P., Wegner, J. K., Ijzerman, A. P., van Vlijmen, H. W. T., & Bender, A. (2011). Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm, 2*, 16-30.

Varin, T., Didiot, M. C., Parker, C. N., & Schuffenhauer, A. (2012). Latent hit series hidden in high-throughput screening data. *J Med Chem, 55*, 1161-1170.

Venhorst, J., Nunez, S., Terpstra, J. W., & Kruse, C. G. (2008). Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J Med Chem, 51*, 3222-3229.

Verdine, G. L., & Hilinski, G. J. (2012). Stapled peptides for intracellular drug targets. *Methods Enzymol, 503*, 3-33.

Vidal, D., Garcia-Serna, R., & Mestres, J. (2011). Ligand-Based Approaches to In Silico Pharmacology. In J. Bajorath (Ed.), *Chemoinformatics and Computational Chemical Biology* (pp. 489-502). Totowa, NJ: Humana Press.

Villoutreix, B. O., Lagorce, D., Labbe, C. M., Sperandio, O., & Miteva, M. A. (2013). One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discov Today, 18*, 1081-1089.

Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F., Heeres, J., Koymans, L. M., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., & Janssen, P. A. (2003). SYNOPSIS: SYNthesize and OPtimize System in Silico. *J Med Chem, 46*, 2765-2773.

Virtanen, S. I., Niinivehmas, S. P., & Pentikainen, O. T. (2015). Case-specific performance of MM-PBSA, MM-GBSA, and SIE in virtual screening. *J Mol Graph Model, 62*, 303-318.

Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., & Rarey, M. (2012). Combining global and local measures for structure-based druggability predictions. *J Chem Inf Model, 52*, 360-372.

von Itzstein, M., Wu, W. Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., Van Phan, T., Smythe, M. L., White, H. F., Oliver, S. W., & et al. (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature, 363*, 418-423.

Wang, F., Liu, D., Wang, H., Luo, C., Zheng, M., Liu, H., Zhu, W., Luo, X., Zhang, J., & Jiang, H. (2011). Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J Chem Inf Model, 51*, 2821-2828.

Wang, Y., Xing, J., Xu, Y., Zhou, N., Peng, J., Xiong, Z., Liu, X., Luo, X., Luo, C., Chen, K., Zheng, M., & Jiang, H. (2015). In silico ADME/T modelling for rational drug design. *Q Rev Biophys, 48*, 488-515.

Wassermann, A. M., Lounkine, E., Hoepfner, D., Le Goff, G., King, F. J., Studer, C., Peltier, J. M., Grippo, M. L., Prindle, V., Tao, J., Schuffenhauer, A., Wallace, I. M., Chen, S., Krastel, P., Cobos-Correa, A., Parker, C. N., Davies, J. W., & Glick, M. (2015). Dark chemical matter as a promising starting point for drug lead discovery. *Nat Chem Biol, 11*, 958-966.

Weisberg, E., Manley, P. W., Breitenstein, W., Bruggen, J., Cowan-Jacob, S. W., Ray, A., Huntly, B., Fabbro, D., Fendrich, G., Hall-Meyers, E., Kung, A. L., Mestan, J., Daley, G. Q., Callahan, L., Catley, L., Cavazza, C., Azam, M., Neuberg, D., Wright, R. D., Gilliland, D. G., & Griffin, J. D.

(2005). Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell, 7*, 129-141.

Wermuth, C. G., Ganellin, C. R., Lindberg, P., & Mitscher, L. A. (1998). Glossary of terms used in medicinal chemistry (IUPAC recomendations 1998). *Pure Appl Chem, 70*, 1129-1143.

Westermaier, Y., Barril, X., & Scapozza, L. (2015). Virtual screening: an in silico tool for interlacing the chemical universe with the proteome. *Methods, 71*, 44-57.

Willett, P. (2011). Similarity searching using 2D structural fingerprints. *Methods Mol Biol, 672*, 133-158.

Wood, J. M., Maibaum, J., Rahuel, J., Grutter, M. G., Cohen, N. C., Rasetti, V., Ruger, H., Goschke, R., Stutz, S., Fuhrer, W., Schilling, W., Rigollier, P., Yamaguchi, Y., Cumin, F., Baum, H. P., Schnell, C. R., Herold, P., Mah, R., Jensen, C., O'Brien, E., Stanton, A., & Bedigian, M. P. (2003). Structure-based design of aliskiren, a novel orally effective renin inhibitor. *Biochem Biophys Res Commun, 308*, 698-705.

Xie, L., & Bourne, P. E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A, 105*, 5441-5446.

Yang, L., Chen, J., & He, L. (2009). Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Comput Biol, 5*, e1000441.

Yang, L., Wang, K., Chen, J., Jegga, A. G., Luo, H., Shi, L., Wan, C., Guo, X., Qin, S., He, G., Feng, G., & He, L. (2011). Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome--clozapine-induced agranulocytosis as a case study. *PLoS Comput Biol, 7*, e1002016.

Yao, L., & Rzhetsky, A. (2008). Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res, 18*, 206-213.

Yera, E. R., Cleves, A. E., & Jain, A. N. (2011). Chemical structural novelty: on-targets and off-targets. *J Med Chem, 54*, 6771-6785.

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., & Vidal, M. (2007). Drug-target network. *Nat Biotechnol, 25*, 1119-1126.

Zeniou, M., Feve, M., Mameri, S., Dong, J., Salome, C., Chen, W., El-Habr, E. A., Bousson, F., Sy, M., Obszynski, J., Boh, A., Villa, P., Assad Kahn, S., Didier, B., Bagnard, D., Junier, M. P., Chneiweiss, H., Haiech, J., Hibert, M., & Kilhoffer, M. C. (2015). Chemical Library Screening and Structure-Function Relationship Studies Identify Bisacodyl as a Potent and Selective Cytotoxic Agent Towards Quiescent Human Glioblastoma Tumor Stem-Like Cells. *PLoS One, 10*, e0134793.

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

Zilian, D., & Sotriffer, C. A. (2013). SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J Chem Inf Model, 53*, 1923-1933.