# Molecular Dynamics and Related Computational Methods with Applications to Drug Discovery


**Jordane Preto, Francesco Gentile, Philip Winter, Cassandra Churchill, Sara Ibrahim Omar and Jack A. Tuszynski**

**Abstract** The main objective of this review chapter is to give the reader a practical toolbox for applications in quantitative biology and computational drug discovery. The computational technique of molecular dynamics is discussed, with special attention to force fields for protein simulations and methods for the calculation of solvation free energies. Additionally, computational methods aimed at characterizing and identifying ligand binding pockets on protein surfaces are discussed. Practical information about available databases and software of use in drug design and discovery is provided.

**Keywords** Molecular dynamics · Molecular docking · Structure-based drug design · Virtual screening · Scoring functions

J. Preto · P. Winter · S. I. Omar · J. A. Tuszynski (✉)
Department of Oncology, University of Alberta, Edmonton, AB T6G 1Z2, Canada
e-mail: jackt@ualberta.ca

J. Preto
e-mail: preto@ualberta.ca

P. Winter
e-mail: pwinter@ualberta.ca

S. I. Omar
e-mail: siomar@ualberta.ca

F. Gentile · C. Churchill · J. A. Tuszynski
Department of Physics, University of Alberta, Edmonton, AB T6G 2E1, Canada
e-mail: fgentile@ualberta.ca

C. Churchill
e-mail: churchil@ualberta.ca



L. L. Bonilla et al. (eds.), *Coupled Mathematical Models for Physical and Biological Nanoscale Systems and Their Applications*, Springer Proceedings in Mathematics & Statistics 232, https://doi.org/10.1007/978-3-319-76599-0_14

# 1 Introduction

Computational drug discovery is a conceptual approach to finding drug-like molecules by rational design, based on the information regarding their intended biomolecular target. A drug target is an important molecule, usually a protein, involved in a particular metabolic or signaling pathway that is specific to a disease condition. Most approaches attempt to inhibit the functioning of an aberrant or over-expressed pathway in the diseased state by interfering with the normal activity of the target. Medicinal compounds as candidate drugs can have their structures rationally designed at a molecular level in such a way as to optimize their binding to the active region of their target biomolecule in order to inhibit its activity and to simultaneously minimize their effects on other important biomolecules that may cause undesired side effects. Since many challenges are posed by the large chemical and biological spaces involved in designing drugs with high specificity and selectivity, serendipity has traditionally played an additional important role in finding potential new drugs. Conversely, structure-based drug design requires knowledge of the structure of the biomolecular target, and it utilizes 3D information about biomolecules obtained from techniques such as x-ray crystallography and NMR spectroscopy.

The first step in the rational drug design process is usually the identification and characterization of the biomolecular target, such as a protein or a DNA sequence. From here, computational techniques can be used to model a drug within the binding site of the biomolecular target, and this information can be used to design novel drug panels with enhanced activity. Of the computational techniques available, molecular dynamics (MD) is particularly important in the investigation of target characterization and drug-target interactions. In Sect. 2, an overview of the main aspects of MD simulations—including force field descriptions—and related methods intended to characterize drug-target binding is provided. In Sect. 3, other computational drug-discovery strategies, such as binding pocket prediction and molecular docking, are described. Virtual screening (VS) techniques are also discussed.

# 2 Molecular Dynamics

## *2.1 General*

Like most experimentally-measured properties of molecular systems, the binding affinity of a drug to its target is a thermodynamic quantity, i.e., an ensemble average over a representative statistical ensemble of a system. As a result, the knowledge of a single 3D structure of a given protein complex—obtained, e.g., from x-ray crystallography or cryo-electron microscopy—even if associated with a global energy minimum, is not enough to theoretically predict such macroscopic properties. Instead, it is necessary to generate a representative ensemble of conformations of the same system at a given (typically physiological) temperature. Two popular computational

methods may be applied to this end: molecular Monte Carlo simulations (MC) [1] and MD [2]. For the study of dynamic or non-equilibrium properties (e.g., the transport of molecules across biomembranes, chemical reactions, etc.), only the second method may be utilized. Although MC simulations are simpler than MD ones, they usually do not lead to any better statistics in a given amount of time [3]. That is why MD is generally preferred over MC. Popular MD engines include Amber [4], GROMACS [5], LAMMPS [6], NAMD [7].

MD simulations usually involve the numerical integration of Newton's equations of motion for a system of N interacting atoms representing the system of interest, possibly including the molecules of the surrounding solvent:

$$m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i, \quad i = 1 \ldots N, \tag{1}$$

where $\boldsymbol{r}_i$ is the position of atom $i$, $m_i$ its mass and $\boldsymbol{F}_i$ is the force acting on it, equal to the negative derivative of the molecular potential $U$, i.e., $\boldsymbol{F}_i = -\partial U / \partial \boldsymbol{r}_i$.

Using Newton's equations of motion automatically implies the use of classical physics, classical MD having the advantage of being far less computationally demanding than real quantum-dynamical simulations, which require solving the time-dependent Schrödinger equation for the system of interacting particles forming the molecule. However, because of classical approximations, standard MD simulations suffer from several limitations that the reader should be aware of. First, electronic motions are not considered per se. Instead, it is supposed that electrons are always in their ground state adjusting their dynamics instantly when atoms are moved (Born-Oppenheimer approximation). Secondly, most potential energy functions $U$ used to model atomic interactions, commonly referred to as *force fields* in chemistry and biology, are empirical thus approximate. They usually consist of a summation of bonded forces and non-bonded pair-additive forces. Such analytical potentials include free parameters (e.g., coupling constants, equilibrium bond lengths, van der Waals radii, etc.), which are estimated by fitting against detailed electronic calculations or experimental properties (e.g., spectroscopy measurements, elastic constants) in order to reproduce observed experimental equilibrium behaviors [8–10]. Typical classical MD force fields adopt the following functional form:

$$\begin{aligned} U = & \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{torsions}} K_\phi [\cos(n\phi + \varphi) + 1] + \sum_{i<j}^{N} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{i<j}^{N} \frac{q_i q_j}{4\pi \varepsilon_0 r_{ij}}. \end{aligned} \tag{2}$$

The first term in Eq. (2) represents the potential between two chemically-bound atoms, modeled as a simple harmonic potential, $b$ being the distance between the two atoms and $b_0$ the equilibrium bond length. The proximity between three atoms, which are connected via chemical bonds can be described with an angle. The second term in Eq. (2) stands for this angle-dependence involving three atoms and is also

modeled by a harmonic potential, θ being the angle between the three atoms in the structure and $\theta_0$ its equilibrium value. The third term represents the dihedral angle (torsion) potential and depends on four atom coordinates. Such a potential is periodic and is represented by a cosine function with $n$, the number of maxima and $\varphi$, the angular offset. The variable $\phi$ is obtained from dihedral angles in the structure. Noticeably, an additional term may be included in Eq. (2) to model out-of-plane bending motions, i.e., improper dihedral angles. This is usually done through a cosine or a harmonic function. The last two terms in Eq. (2) account for non-bonded interactions and are calculated pairwise between atoms $i$ and $j$. The fourth term is the van der Waals potential, which is typically represented by a Lennard-Jones 6-12 potential. The $1/r^6$ term is the attractive component while the $1/r^{12}$ term approximates Pauli repulsion. Parameters $A_{ij}$ and $B_{ij}$ are atom specific while $r_{ij}$ stands for the distance between atoms $i$ and $j$. The final term corresponds to the electrostatic potential between atoms, and is modeled as a Coulomb potential. Parameters $q_i$ and $q_j$ represent (fixed) charges on atoms $i$ and $j$, while the constant $\varepsilon_0$ is the vacuum permittivity. Electrostatic interactions dominate over van der Waals forces for long-range intermolecular interactions and they play a significant role in non-chemical binding.

MD simulations result in trajectories, which contain information about the changes of atomic positions over time, which can be analyzed in great detail to extract pertinent information regarding the dynamics of the system. This includes the root-mean-square deviation (RMSD) of ligand and protein atoms, supramolecular (non-covalent) interactions, changes in the potential energy of the system, short-lived reaction intermediates [11], conformational changes, flexibility, and optimum binding modes [12] among many various properties of the biomolecule and its environment. In a computer-aided drug design process, the mobility of crystal water molecules near proteins observed in MD simulations can help identify the amino acid residues that play an important role in ligand binding. MD simulations can also be used for studying ionic conductivity [13, 14], where the simulations provide atomic level insights into ionic mobility. In terms of particular applications, MD has been successfully used to study clinically important proteins such as HIV-1 gp120 [15], binding sites [16], drug resistance mechanisms [17], and protein folding [18, 19] to name but a few.

## *2.2 Polarizable Force Fields and Quantum Dynamics*

Most standard force fields do not incorporate atomic polarizability effects other than adjusting atomic partial charges obtained from quantum chemical computations. However, polarizable force fields include extra degrees of freedom in order to model electronic charges, usually attached to the nucleus by a spring as in the case of the shell model [20]. This allows for a dynamic redistribution of atomic dipoles, which responds to the local chemical environment. More realistic MD simulations, called ab initio molecular dynamics (AIMD), can be applied in order to reproduce electronic

dynamics more accurately [21, 22]. Instead of using a prescribed potential for $U$, AIMD implies solving the time-dependent Schrödinger equation for the many-body wave function of the electrons assuming the atom nuclei fixed (*Born-Oppenheimer* approximation). The Schrödinger equation is generally solved at each MD step using density functional theory (DFT) in order to get the potential energy as a function of the nuclear coordinates. The potential energy is then used to integrate the classical Newton's equations given by Eq. (1). Due to the cost of treating electronic degrees of freedom, the computational cost is far higher than classical MD, implying that AIMD is only applicable to small molecular systems and short time scales (picosecond). A good trade-off between accuracy and speed is achieved by hybrid quantum mechanics/molecular mechanics (QM/MM) force fields [23]. In such simulations, the region of the system in which quantum effects (e.g., bond breaking, quantum resonance …) take place is treated at an appropriate level of quantum chemistry theory, while the rest is described by a classical molecular mechanics force field [24]. Recently, machine-learning-based algorithms have been suggested as a way to accelerate ab initio methods [25]. A significant advantage of AIMD and QM/MM simulations is the ability to study reactions that involve breakage or formation of covalent bonds, which correspond to multiple electronic states. AIMD has also proved useful for reproducing typical dynamics and spectral features of liquids such as water [26]. Despite the accuracy provided by quantum methods, classical MD remains a reliable method to study biomolecular processes in large systems over long simulation times (up to a few milliseconds) including folding dynamics, conformational molecular changes as well as non-covalent bindings of drugs to their biological target.

## *2.3 Molecular Dynamics and Drug Discovery*

Molecular dynamics can be used together with other methods to solve a host of problems in biomolecular modeling [27, 28]. In the case of VS methods that involve large libraries of chemical compounds in order to identify a high-affinity small molecule that is expected to act as an enzyme inhibitor, or a protein-protein interaction blocker, the calculation of the binding energy of potential hits may help prioritize compounds for experimental testing.

While docking and scoring remain the most widely used computational techniques to predict the binding mode and affinity of a drug to its target due their low computational cost, these methods are not particularly accurate. More precise approaches utilize appropriate sampling of the molecular system generated beforehand with MD simulations as is required when estimating ensemble-averaged quantities like binding free energies. *End-point methods* such as linear interaction energy (LIE) and the molecular mechanics Poisson-Boltzmann Surface Area (MM/PBSA) technique, which rely only on appropriate samplings of the end states, i.e., the complex and possibly the free receptor and ligand, have intermediate efficiencies. The LIE method, originally introduced by Aqvist et al. [29], assumes that the binding free energy can be written as a linear combination of average interaction energies between the ligand

and the rest of the system (protein, water and ions). More explicitly, the binding free energy of the ligand is expressed as [30]:

$$\Delta G_{bind} = \alpha\langle \Delta E_{vdW}^{L-S} \rangle + \beta\langle \Delta E_{el}^{L-S} \rangle, \tag{3}$$

where $\langle \Delta E_{vdW}^{L-S} \rangle = \langle E_{vdW}^{L-S} \rangle_{bound} - \langle E_{vdW}^{L-S} \rangle_{unbound}$ refers to the change in van der Waals interactions between the bound and unbound states of the ligand. The averages stand for ensemble averages obtained from MD simulations whereas the L-S label indicates that the interaction energies are computed only between the ligand and the surroundings. Similarly, $\langle \Delta E_{el}^{L-S} \rangle = \langle E_{el}^{L-S} \rangle_{bound} - \langle E_{el}^{L-S} \rangle_{unbound}$ corresponds to the change in intermolecular electrostatic interactions between the bound and unbound states. Parameters α and β are generally obtained empirically using an appropriate fitting procedure.

Alternatively, MM/PBSA [31], which is arguably the most popular end-point method, turned out to be successful in a number of drug-design case studies [32–34]. The method basically provides an estimate of the binding free energy as:

$$\Delta G_{bind} = \langle \Delta E_{MM} \rangle - T\Delta S + \Delta G_{solv}, \tag{4}$$

where $\langle \Delta E_{MM} \rangle - T\Delta S$ can be regarded as the change in the free energy of the system in vacuum (gas phase); it includes the change in the molecular mechanics energy due to the binding $\langle \Delta E_{MM} \rangle = \langle E_{MM} \rangle_{bound} - \langle E_{MM} \rangle_{unbound}$ and the change in the conformational entropy $\Delta S$, usually estimated from normal mode analysis (NMA) performed on the complex structure and on the free ligand and protein structures. As in the LIE method, every average quantity corresponds to an ensemble average obtained from output MD trajectories. The entropy contribution, which is relatively time-consuming and inaccurate to compute using NMA, can be neglected if a comparison of states of similar entropy is desired such as in the case of comparing two or more ligands binding to the same protein binding site. Finally, $\Delta G_{solv}$ stands for the difference of solvation free energies due to the binding, it is given as $\Delta G_{solv} = \Delta G_{solv}^{complex} - \Delta G_{solv}^{lig} - \Delta G_{solv}^{prot}$ where every term on the right-hand side is given as the sum of polar and nonpolar contributions. The polar parts are obtained by solving the Poisson-Boltzmann (PB) equation or by using the Generalized-Born (GB) model (as in the MM/GBSA method) whereas the nonpolar terms are estimated from a linear relation to the solvent accessible surface area (SASA). Despite the fact that MM/PBSA and MM/GBSA are computationally-inexpensive methods, they contain several crude and questionable approximations, e.g., due to the use of implicit solvent models to compute the solvation energies [35]. The capability of the MM/PBSA method to predict the correct binding free energy turns out to be more sensitive to the investigated system compared to the MM/GBSA method, the latter being more useful in multi-target comparisons [36]. Noticeably, the MM/PBSA and GBSA techniques can be used to perform per-residue-free-energy decompositions. The benefit of such decompositions is twofold: providing important information about residues which significantly contribute to the binding energy (hot spots) and

giving insights into the changes in binding free energies due to mutations, especially single point mutation.

## 2.4 Alchemical Free Energy Calculations

Another important category of MD-based methods used to estimate the binding free energy of ligand-protein complexes is called alchemical free energy methods, which include, for example, free energy perturbation (FEP) and thermodynamic integration (TI). Alchemical techniques make use of a parameter-dependent Hamiltonian to smoothly switch between the dynamics of systems with different chemical compositions. Applied to drug design, one may think about connecting the bound and unbound states of a given protein-ligand complex or transforming one ligand to another one within the same binding pocket. In the simplest case, such connections can be achieved by a linear combination of the corresponding Hamiltonians:

$$H(\boldsymbol{x}, \boldsymbol{p}; \lambda) = (1 - \lambda) H_a(\boldsymbol{x}, \boldsymbol{p}) + \lambda H_b(\boldsymbol{x}, \boldsymbol{p}), \tag{5}$$

where $\lambda$ is a parameter which varies from 0 to 1, and, $H_a$ and $H_b$ are the physical Hamiltonians associated with the two states $a$ and $b$, e.g., the bound and unbound systems. The FEP and TI approaches differ in the way of estimating the free energy difference $\Delta G_{a \to b}$ between states $a$ and $b$. Within the FEP framework, $\Delta G_{a \to b}$ is given by the following identity:

$$\Delta G_{a \to b} = -k_B T \ln \left\langle \exp\left( -\frac{(H_b(\boldsymbol{x}, \boldsymbol{p}) - H_a(\boldsymbol{x}, \boldsymbol{p}))}{k_B T} \right) \right\rangle_a, \tag{6}$$

where $\langle \ldots \rangle_a$ stands for an ensemble average over configurations representative of the initial state $a$. In contrast, the free energy difference in the TI approach is computed from:

$$\Delta G_{a \to b} = \int_0^1 \left\langle \frac{\partial H(\boldsymbol{x}, \boldsymbol{p}; \lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda. \tag{7}$$

Since both FEP and TI identities are based on Boltzmann equilibrium averages, which require extensive sampling of the complex and free ligand in solution, such techniques are generally computationally extensive. Other types of methods involving averages over non-equilibrium trajectories can, however, be used. This is the case of steered molecular dynamics (SMD) [37]. Applied to drug binding studies, SMD introduces a non-conservative force used to pull out the ligand from its binding site at constant speed [38]. The free energy difference during this non-equilibrium process can be obtained using the Jarzynski equality [39]:

$$\Delta G_{a\to b} = -k_B T \ \ln\left\langle \exp\left(-\frac{W_{a\to b}}{k_B T}\right)\right\rangle. \tag{8}$$

$W_{a\to b}$ represents the external work performed on the system during one steered simulation whereas $\langle \ldots \rangle$ corresponds to a non-equilibrium average performed over all the steered trajectories.

## 2.5 Enhanced Sampling Methods

The above-discussed alchemical methods including SMD can be integrated into a more general class of method called enhanced sampling methods [40]. As mentioned above, MD is a very useful and inexpensive tool to study the behavior of macromolecular systems in silico. However, the use of MD is limited by the computational time required to carry out a reasonable length of simulation. This, in turn, is dependent on the availability of computational hardware and time allocation for high-performance computing. Depending on the time required for a biomolecular system to reach equilibrium, MD can be run long enough to represent the evolution of a system from a few nanoseconds to a few microseconds. Long time-scale dynamical processes, such as slow conformational changes, protein assembly or ligand-protein binding processes, are notoriously difficult to model by MD.

In the past few decades, many computational methods have been developed based on MD simulations to explore relevant transitions between stables states of a system. In some of these techniques, a biased potential is added to the dynamics along one or a few collective variables. Such methods require to carefully choose the collective variables in such a way to facilitate the diffusion towards critical unexplored regions of the configuration space. The resulting trajectories are eventually unbiased to obtain the equilibrium distribution as a function of the collective variables used. Popular techniques performing MD with a biased potential along specific collective variables include metadynamics and umbrella sampling, both of them being related to a number of successful applications to biological systems.

In metadynamics, a positive Gaussian potential is added at regular time intervals to the free energy landscape of a system [41]. In this way, the system is discouraged to come back to the previous point, thus favoring the exploration of yet unvisited values of the collective variable. As more and more Gaussians are added to the true potential, the system is able to diffuse more freely along this variable. Papers on metadynamics research have reported a lot of successful applications to biological systems [42, 43]. As an important application, metadynamics—combined with nudged elastic band—was used to investigate the unbinding process of a ligand away from its specific protein target and the estimate of the associated binding energy [44].

Umbrella sampling is an enhanced sampling technique where a series of biased MD simulations are conducted independently. This works by splitting the reaction coordinate into a series of windows and applying a harmonic potential which acts to force the reaction coordinate to remain close to the center of each window. Assuming

the biased trajectories overlap in the space defined by the collective variable, the resulting biased free energy landscapes can be used to compute the true free energy profile. This last step is performed by means of the weighted histogram analysis method (WHAM), which combines the information of the biased simulations so as to minimize the statistical error made on the resulting probability distribution [45]. In the context of drug-binding studies, umbrella sampling was used to estimate the potential of mean force for ion permeation and ligand binding to ion channels [46] and to predict protein–ligand binding structures in kinase systems [47].

## 3 Other Computational Drug-Discovery Methods

This section is intended to provide an account of other popular useful techniques for drug design and VS studies that can be used instead or in conjunction with MD-based methods discussed in the previous section.

### *3.1 Binding Pocket Prediction*

Identifying and characterizing a suitable binding pocket in a 3D protein structure is a central aspect of any drug discovery study. This step is also relevant to shed light on biomolecular functions as many proteins are biologically functional only after interacting with cofactors or other biological molecules.

A common way to define a binding pocket, if a ligand is already bound to it, is to introduce a distance cut-off. Binding pocket atoms are typically defined whenever their distance to the ligand is below 4–8 Å. Following is a list of physicochemical key properties of binding pockets:

- The solvent-accessible surface area (SASA) which is usually computed as the atoms of the pocket reachable by a solvent probe sphere rolling over the protein surface.
- The volume of the pocket and its depth which corresponds to the average distance of the pocket atoms to their nearest water molecules from bulk solvent [48].
- The pocket hydrophobicity which depends on polar and non-polar residues involved in the binding site.
- The number of hydrogen bond donors and acceptors on the pocket surface.
- The conservation of residues over similar binding pockets of other proteins, which is particularly relevant for functional sites.

In addition to experimental binding site detection techniques such as NMR-based methods [49], a number of computational methods can be found which may be helpful in the process of binding pocket identification of a molecular target. Pocket finder algorithms are usually tested and validated on protein and ligand datasets. Such tests are intended to estimate the reliability of identifying the correct binding pocket

within the first one to three hits provided by an algorithm. Two popular publicly-available databases are the Protein Data Bank (PDB) [50], which provides 3D protein structures for input into the pocket finding algorithms, and the PDBbind database [51], which contains bound protein structures filtered from the PDB database. The current version of PDBbind has around 3100 protein-ligand complexes, 1300 of these having been manually selected to form the refined set with the focus on the quality of structures and binding data. Due to its large size and manual curation, the refined set of the PDBbind database provides a suitable benchmark for most case studies. Further reduced from this, is the core set of 210 complexes. Optimal databases for pocket prediction testing should include high-resolution, diverse and non-redundant protein-ligand complexes. Pocket finder algorithms are generally split into two classes, namely, geometric-based and energetic-based approaches.

*Geometry-based algorithms* have the advantage of a low computational cost. The underlying assumption behind such methods is that the ligand binding pocket corresponds to the larger cleft within the protein structure [52, 53]. Therefore, geometrical criteria may be sufficient to identify the correct binding location on a protein. One such example, SURFNET [54] is an early-developed program which fits spheres between pairs of atoms so that they do not contain more than one atom. The binding pocket is defined as the volume containing the largest number of adjacent spheres. An improvement of the program, called SURFNET-ConSurf [55], refines the binding pocket prediction also considering the residue conservation within the binding site. The SURFNET-ConSurf algorithm was tested on a set of 244 non-redundant, diverse and representative ligand-protein complexes, obtained by a filtered version of the PDB database. A 75% rate of successfully recognized native ligand pockets is reported in the original paper about this method [55].

Another algorithm called VisGrid [56] is based on geometrical hashing and identifies cavities by considering the visibility of each point in a 3D grid, that is, the fraction of directions that are not blocked by protein atoms. In this way, a cluster of closely-located grid points with limited visibility indicates a pocket. VisGrid was compared with other pocket prediction methods, including SURFNET and LIGSITE, and the observed success rates on a set of bound and unbound structures were comparable with existing methods.

LIGSITE [57] uses a grid-based method in which points are either assigned to the solvent or protein category, and cavities are defined as groups of points in which solvent points are surrounded by protein points. Although the LIGSITE original validation identified the correct binding pocket in all the testing cases, a big limitation of this study was the reduced size of the dataset, with only ten ligand-receptor complexes. Its extension, LIGSITE$^{csc}$ [58], improves the original algorithm by calculating more accurately the contact between protein surface and solvent using the Connolly surface, and by re-ranking the identified pockets by their degree of residue conservation in homolog proteins. The LIGSITE$^{csc}$ testing process is more significant than the LIGSITE one, and a comparison with other geometry-based methods is also provided. The success rates calculated on a set of 210 non-redundant bound structures were 75% for LIGSITE$^{csc}$, 65% for LIGSITE and 42% for SURFNET.

The algorithm also showed good performances in recognizing the correct binding pocket in unbound structures.

Another class of methods dedicated at identifying binding pockets are *energy-based methods* which rely on the energetic properties of a binding site. A common approach of these methods is to use molecular probes to search for favorable interaction sites on a protein, and cluster them together to identify putative pockets. An early effort resulted in the GRID Fortran code [59]. The probes employed by this algorithm include water, methyl group, the hydroxyl, amine nitrogen and carboxy oxygen. Energetic contours are calculated with a function considering a 12-6 Lennard-Jones term, an electrostatic term and a hydrogen bond term, and negative energy levels indicate promising interaction sites for each probe.

Laurie and Jackson's Q-SiteFinder method [60] calculates the interaction energy of a methyl probe and the grid points generated on the protein structure. A clustering analysis step links favorable interaction sites to rank putative binding pockets based on their total interaction energy. Q-SiteFinder was tested on a diverse set of bound and unbound protein conformations, resulting in success rates of 74% and 71%, respectively.

EasyMIFs and SITEHOUND [61] are two complementary energy-based tools developed at the Sanchez lab. The first algorithm calculates the interaction energy between grid points and molecular probes using the GROMOS force field, while SITEHOUND recognizes putative binding sites by filtering and clustering the spatial variation of the interaction energy fields calculated by EasyMIFs or any other grid-based program. Multiple probes are used, as well as different site clustering algorithms. SITEHOUND's success rate was evaluated on a set of 77 complexes and it was reported as 95% (bound structures) and 79% (unbound structures) considering the binding pocket identified when present in the top three ranked sites.

Another similar energy-based algorithm, AutoLigand [62], was created by the developers of the popular molecular docking software Autodock. AutoLigand was reported to have a success rate of 73% when tested on a set of 187 bound structures and 80% when tested on 96 unbound structures.

## *3.2 Ligand-Receptor Docking*

Molecular docking methods have been developed to predict how a given compound naturally binds to its biomolecular target, i.e., its binding mode, and to provide an estimate of its binding affinity. Docking software usually rely on optimization algorithms which include both a search algorithm and a scoring function. Such methods require at least one ligand structure and one target structure as inputs. The location of the targeted site should be provided although blind docking approaches [63, 64] can help deal with unknown binding locations in addition to the pocket prediction methods discussed in the previous section.

The search algorithms are dedicated to exhaustively explore the conformational space of the ligand within the targeted pocket. Three groups of such functions have

been described, namely matching, systematic and stochastic methods. *Matching algorithms* are based on shape complementarity between the ligand and the receptor site, and, possibly, chemical complementarity. *Systematic search algorithms* explore the degrees of freedom of the ligand in a progressive way. This class of methods can be divided in three subgroups:

- Exhaustive algorithms systematically rotate all ligand dihedral angles until an optimal solution is reached.
- Fragment-based methods break down the ligand into different fragments which are separately placed within the binding site and re-connected in the last step of the process.
- Ensemble-based methods pre-generate a large number of ligand conformations, which are then rigidly placed within the binding site.

The last class of search algorithms includes *stochastic methods* such as MC and evolutionary algorithms, which introduce random changes of the degrees of freedom of the ligands to rapidly reach an optimal solution.

In molecular docking, the binding free energy is calculated using a position-dependent scoring function. This is required not only to identify the correct binding pose corresponding to the lowest binding energy, but also to rank a set of tested compounds according to their affinity to a target. *Force-field-based scoring functions* used for docking are similar to MD force fields discussed in Sect. 2.1. In *empirical scoring functions*, the different contributions in the binding energies are weighted with coefficients, set beforehand to reproduce experimental dissociation constants of known ligand-receptor complexes. The Autodock4 scoring function [65] is an example of an empirical scoring function where a non-bonded interaction potential is calculated as

$$V = W_{vdW} \sum_{ij} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hb} \sum_{ij} E(\theta) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{ij} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{\frac{-r_{ij}^2}{2\sigma^2}}. \quad (9)$$

The first term represents van der Waals contributions, the second is a hydrogen bond term, the third is a Coulombic term for electrostatic interactions, and the last is a desolvation potential. $W$ represents the empirical coefficients, obtained from the training over 188 bound complexes from a PDB calibration subset. The coefficients $A$ and $B$ derived from the AMBER force field, while $C$ and $D$ are Autodock internal parameters. $E(\theta)$ depends on the angle of deviation $\theta$ from the ideal hydrogen bond geometry. $S$ and $V$ are a solvation parameter and the volume of the atoms surrounding one atom, respectively. $\sigma$ is distance-weighting factor of Autodock. The total Autodock score of a binding pose is calculated by summing the difference of intra-molecular energies between the bound and unbound forms of the ligand and the protein, then subtracting the difference of inter-molecular energies. A simple

entropic term is also included in the final score to model the variation of the system entropy upon binding. *Knowledge-based scoring functions* use the potential of mean force (PMF), derived from protein-ligand structures and calculated for each $ij$ ligand-receptor atom pair type as:

$$w_{ij}(r) = -k_B T ln(\rho(r_{ij})/\rho^*(r_{ij})), \tag{10}$$

where $k_B$ is the Boltzmann constant, $T$ is the temperature in Kelvin degrees, $\rho(r_{ij})$ is the number density of the $ij$ atom pair derived from the structural training set and $\rho^*(r_{ij})$ is the number density in a reference state. Although knowledge-based scoring functions do not provide a precise interaction energy potential due to difficulties arising from the reference state calculation, they directly connect the atomic interactions between ligand and protein to structural data instead of to kinetics as is the case for empirical methods. Knowledge-based scoring functions turn out also to be more computationally efficient than force-field-based methods [66, 67]. Finally, *consensus scoring methodologies* combine different scoring function outputs of the same ligand to obtain a single, consensual score. Different combination strategies can be employed, such as weighting and summing up the ranks or performing a regression analysis [68].

Accounting for the flexibility of the binding site is an important issue in molecular docking. As a result, different approaches have been proposed to address this problem. Soft docking algorithms use modified short-range repulsion parameters for the binding site atoms, which allow the ligand to slightly penetrate through the surface of the pocket to mimic the induced fit of the binding. Many algorithms also include the possibility to treat the side chains of pocket residues as flexible, although such methods still ignore backbone dynamics, while increasing noticeably the computational cost. Using multiple receptor conformations when performing molecular docking is a popular way to take into account the backbone flexibility [69]. This approach, called Relaxed Complex Scheme (RCS), relies on NMR or MD-derived conformational ensembles, which are used as molecular docking targets.

Testing of docking software is usually performed by evaluating the percentage of docking poses with small enough RMSD (typically 2 Å) compared to the co-crystallized poses extracted from a high-resolution structural database.

The original implementation of DOCK [70] is an example of geometry-matching algorithm, where the binding site and the ligand atoms are represented as spheres that are systematically matched using a shape-based routine. DOCK 6 [71], the latest version of the program, applies a fragment-based algorithm and a set of different force-field-based scoring functions which can be selected, such as PBSA, GBSA and Amber scoring methods. In addition, a minimization step is performed for the ligand in order to remove minor protein-ligand clashes and relax its internal geometry. The success rate of the latest DOCK release was estimated around 73% in reproducing crystallographic poses. The authors tested the algorithm on 1043 structures obtained from a ligand-receptor database designed as a benchmark for assessing docking software performances [72].

Autodock [73], probably the most popular docking software, uses a Lamarckian genetic algorithm to independently generate a large number of binding poses, scored with the empirical scoring function described in Eq. (9). A clustering algorithm can be optionally used to identify the most populated portion of the conformational space of the ligand, from which the lowest energy pose should match the native one. The success rate of the latest version, Autodock4 [74], was around 53% when the software was tested on the calibration structural set.

Autodock Vina [75] utilizes an iterated local search global optimizer searching method. The Vina scoring function combines aspects from knowledge-based and empirical potentials. Tested on the same set used for Autodock4, Vina was able to identify the correct binding pose in 78% of the cases. Noteworthy, Vina scoring function was trained with the PDBbind refined set, much bigger than the training set used for the Autodock scoring function.

Glide [76] is a docking software included within the Schrödinger molecular modeling package. It is based on an exhaustive systematic search algorithm used to sample the ligand conformational space, followed by a minimization step. An optimal choice for the scoring method [76] is given as a combination of a force-field-based function, an empirical function (GlideScore) and the strain energy of the ligand conformation. The pose success rate was reported around 66% when tested on 282 ligand-receptor complexes selected from the PDB database.

GOLD [77] is another popular docking program. The software maps together complementary chemical features of the ligand and the receptor within the binding site. A genetic algorithm is then used to explore different binding modes. Three main scoring functions are available, namely, Goldscore (force-field-based), Chemscore (empirical) and Astex Statistical Potential [78] (knowledge-based). Testing of GOLD on the CCDC/Astex database [79], a PDB subset designed to test docking software, resulted in success rates up to 87% depending on the scoring function used. The correlation coefficients ($R^2$) between experimentally-measured and GOLD binding affinities were reported between 0.51 and 0.55.

### *3.3 Virtual Screening*

The discovery of a new drug is an expensive and long process. It is estimated that up to two and half billion dollars and twenty years are required to bring a new product from the bench to the clinic [80, 81]. Consequently, efforts are made to shorten the process and reduce the cost. Some of the time and funding savings are expected to result from a wider use of computational techniques applied to drug discovery. One of such techniques is called virtual screening (VS). VS refers to an in silico active compound search against biomolecular targets [82]. It has the advantage of being fast and inexpensive compared with traditional high throughput screening. Nowadays, libraries including billions of compounds can be virtually screened depending on the available computational resources.

A typical VS workflow consists of sequential series of filtering and scoring steps aimed at providing a set of promising compounds for experimental validation. VS methods can be divided in ligand-based VS (LBVS) and structure-based VS (SBVS). LBVS approaches are computationally faster but they do not provide any estimate of the binding energy of the ligand. On the other hand, SBVS methods are more computationally expensive but they enable to rank potential hits based on their predicted binding affinity. Regardless of the chosen approach, a compound database is always required as starting point for VS. Examples of extensive small molecule repositories are PubChem [83, 84], ZINC [85] and the National Cancer Institute databases [86]. Such collections usually include millions of compounds that can be downloaded for screening purposes [87].

LBVS techniques rely solely on the 2D or 3D structure of ligands, ignoring the biological target. The main assumption behind these methods is that structurally related compounds share similar activities [88]. Therefore, the structure of at least one known active compound should be available as template for the computational search, and a measure of the distance between structures needs to be computed. Simple ways to represent the chemical structure of a compound in a computer-readable format are chemical fingerprints or pharmacophore representations. Chemical fingerprints [89] are binary strings in which each bit codes for the presence or the absence of particular chemical groups. A widely-used way to compare two fingerprints is to use the Tanimoto index, given by

$$T_{A,B} = \frac{c}{a + b - c}, \tag{11}$$

where A and B are the two fingerprints, c is the number of bits set to 1 at the same position in both the fingerprints, and, a and b are the total number of bits set to 1 in A and B, respectively. A publicly-available package which can be used for fingerprint-based VS is chemfp [90].

Another way to perform LBVS is to use a pharmacophore model of the active compounds [91], which provides a representation of the ligand from its spatially-distributed chemical features (hydrogen bond acceptor, hydrogen bond donor, hydrophobic moiety, ring structure, polar or charged residue) including the distances between centers forming a chemical structure. In pharmacophore-based VS, the distance between two ligand structures is usually calculated as the RMSD between the superposed pharmacophore points. The main benefit of this approach is to identify molecules with different chemical groups but similar generic features, providing novel scaffolds to medicinal chemistry. Contrary to most chemical fingerprints, pharmacophore models also include 3D properties of the ligand.

Data mining and machine learning methods including support vector machines, neural networks, Bayesian networks and decision trees, are also utilized for LBVS [92]. LBVS methods are useful in case a 3D structure of the target is not available, but they can also be used to clean up large databases in order to generate focused libraries [93]. Indeed, these structurally-related subsets are designed to interact with a specific target and they are built by screening larger and diverse databases. Focused

libraries have a limited size compared with the parent databases. Therefore, they can be rapidly and efficiently screened with SBVS or experimental techniques. LBVS methods have led to the discovery of novel and promising compounds with low-range potency [94]. Examples of such successes are the discovery of anti-cancer tubulin dimerization inhibitors [95], inhibitors of the 17β-HSD2 enzyme for osteoporosis treatment [96] and novel scaffolds for the inhibition of the HIV-1 integrase [97].

SBVS methods provides a ranking of the screened compounds based on their computed binding affinities. Therefore, one or multiple structures of the target are required. SBVS always relies on molecular docking methods, which are used to place the compounds within the targeted pocket, and to estimate the binding affinity from the resulting binding poses. We have already discussed all the docking-related aspects in Sect. 3.2. The docking scoring functions are designed to quickly estimate the binding energy from a ligand pose, therefore they often do not lead to very accurate results. Several strategies have been developed to deal with this, including the already mentioned consensus scoring, MD simulations of the complex structures and/or more accurate scoring functions (e.g., MM/PBSA or GBSA) [98, 99]. Recently, SBVS techniques were successfully applied to the discovery of DNA repair inhibitors [98, 100–103], anti-malarian compounds [104], kinase inhibitors [105] and HIV-1 inhibitors [106, 107].

## 4 Conclusions

This review chapter provides introductory information regarding the computational tools currently used in the drug design and discovery process. We have given an overview of molecular dynamics methods that are very useful in biomolecular target characterization for drug action. We have also given practical information regarding the identification of binding pockets for putative inhibitors of proteins, as well as an overview of molecular docking techniques that are based on the protein-ligand interactions. These interactions and their ranking involving the binding free energy of the ligand-target pair are used in massive searches for specific and selective inhibitors of particular protein, a methodology referred to as virtual screening. The latter methodology relies on large and diverse databases of pharmacologically-acceptable compounds. Lists of databases and software packages used in all stages of computational drug design have been presented in this chapter to assist in practical aspects of research in this area.

**Acknowledgements** The authors are grateful to the Natural Sciences and Engineering Research Council of Canada, the Li Ka Shing Institute of Applied Virology and the Alberta Cancer Foundation for funding support.

## References


1. A. Vitalis, R.V. Pappu, Annu. Rep. Comput. Chem. **5**, 49 (2009)
2. M. Karplus, Acc. Chem. Res. **35**, 321 (2002)
3. W.L. Jorgensen, J. Tirado-Rives, J. Phys. Chem. **100**, 14508 (1996)
4. D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, P. Kollman, Comput. Phys. Commun. **91**, 1 (1995)
5. S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, Bioinformatics **29**, 845 (2013)
6. S. Plimpton, J. Comput. Phys. **117**, 1 (1995)
7. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, J. Comput. Chem. **26**, 1781 (2005)
8. W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995)
9. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, Proteins **65**, 712 (2006)
10. A.D. MacKerell, D. Bashford, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, J. Phys. Chem. B **102**, 3586 (1998)
11. M. Fujihashi, T. Ishida, S. Kuroda, L.P. Kotra, E.F. Pai, K. Miki, J. Am. Chem. Soc. **135**, 17432 (2013)
12. G. Tiwari, D. Mohanty, PLoS One **8** (2013)
13. J.A. Tuszynski, C. Wenger, D.E. Friesen, J. Preto, Int. J. Environ. Res. Public Health **13** (2016)
14. D. Zahn, J. Mol. Model. **17**, 1531 (2011)
15. N.T. Wood, E. Fadda, R. Davis, O.C. Grant, J.C. Martin, R.J. Woods, S.A. Travers, PLoS One **8** (2013)
16. A. Chaudhuri, I. Sarkar, S. Chakraborty, J. Biomol. Struct. Dyn. **32**, 1969 (2014)
17. G. Leonis, T. Steinbrecher, M.G. Papadopoulos, J. Chem. Inf. Model. **53**, 2141 (2013)
18. T. Yoda, Y. Sugita, Y. Okamoto, Biophys. J. **99**, 1637 (2010)
19. T. Yoda, Y. Sugita, Y. Okamoto, Proteins-Struct. Funct. Bioinform. **82**, 933 (2014)
20. B.G. Dick, A.W. Overhauser, Curr. Contents/Phys. Chem. Earth Sci. **24** (1985)
21. B. Kirchner, P.J. di Dio, J. Hutter, Multiscale Mol. Methods Appl. Chem. **307**, 109 (2012)
22. D. Marx, J. Hutter, Mod. Methods Algorithms Quantum Chem. **1**, 141 (2000)
23. A. Warshel, M. Levitt, J. Mol. Biol. **103**, 227 (1976)
24. E. Brunk, U. Rothlisberger, Chem. Rev. **115**, 6217 (2015)
25. V. Botu, R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015)
26. A.A. Hassanali, J. Cuny, V. Verdolino, M. Parrinello, Philos. Trans. A Math. Phys. Eng. Sci. **372**, 20120482 (2014)
27. S.Y. Hu, H.J. Yu, Y.J. Liu, T. Xue, H.B. Zhang, J. Mol. Model. **19**, 3087 (2013)
28. W. Hu, S.W. Deng, J.Y. Huang, Y.M. Lu, X.Y. Le, W.X. Zheng, J. Inorg. Biochem. **127**, 90 (2013)
29. J. Aqvist, C. Medina, J.E. Samuelsson, Protein Eng. **7**, 385 (1994)
30. H. Guitierrez-de-Teran, J. Aqvist, Comput. Drug Discov. Des. **819**, 305 (2012)
31. P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, Acc. Chem. Res. **33**, 889 (2000)
32. F. Chen, H. Liu, H.Y. Sun, P.C. Pan, Y.Y. Li, D. Li, T.J. Hou, Phys. Chem. Chem. Phys. **18**, 22129 (2016)
33. J.M. Sanders, M.E. Wampole, M.L. Thakur, E. Wickstrom, PLoS One **8** (2013)
34. T. Zhu, H. Lee, H. Lei, C. Jones, K. Patel, M.E. Johnson, K.E. Hevener, J. Chem. Inf. Model. **53**, 560 (2013)

35. S. Genheden, U. Ryde, Expert Opin. Drug Discov. **10**, 449 (2015)
36. H.Y. Sun, Y.Y. Li, S. Tian, L. Xu, T.J. Hou, Phys. Chem. Chem. Phys. **16**, 16719 (2014)
37. B. Isralewitz, M. Gao, K. Schulten, Curr. Opin. Struct. Biol. **11**, 224 (2001)
38. J.S. Patel, A. Berteotti, S. Ronsisvalle, W. Rocchia, A. Cavalli, J. Chem. Inf. Model. **54**, 470 (2014)
39. S. Park, F. Khalili-Araghi, E. Tajkhorshid, K. Schulten, J. Chem. Phys. **119**, 3559 (2003)
40. R.C. Bernardi, M.C.R. Melo, K. Schulten, Biochim. Biophys. Acta Gen. Subj. **1850**, 872 (2015)
41. A. Laio, F.L. Gervasio, Reports. Prog. Phys. **71**, 126601 (2008)
42. A. Barducci, R. Chelli, P. Procacci, V. Schettino, F.L. Gervasio, M. Parrinello, J. Am. Chem. Soc. **128**, 2705 (2006)
43. F.L. Gervasio, A. Laio, M. Parrinello, J. Am. Chem. Soc. **127**, 2600 (2005)
44. D. Branduardi, F.L. Gervasio, M. Parrinello, J. Chem. Phys. **126**, 54103 (2007)
45. M. Souaille, B. Roux, Comput. Phys. Commun. **135**, 40 (2001)
46. T. Başţuğ, P.-C. Chen, S.M. Patra, S. Kuyucak, J. Chem. Phys. **128**, 155104 (2008)
47. H. Kokubo, T. Tanaka, Y. Okamoto, J. Chem. Theory Comput. **9**, 4660 (2013)
48. S. Chakravarty, R. Varadarajan, Structure **7**, 723 (1999)
49. P.J. Hajduk, J.R. Huth, S.W. Fesik, J. Med. Chem. **48**, 2518 (2005)
50. H.M. Berman, Nucleic Acids Res. **28**, 235 (2000)
51. R. Wang, X. Fang, Y. Lu, C.Y. Yang, S. Wang, J. Med. Chem. **48**, 4111 (2005)
52. R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein Sci. **5**, 2438 (1996)
53. X. Zheng, L. Gan, E. Wang, J. Wang, AAPS J **15**, 228 (2013)
54. R.A. Laskowski, J. Mol. Graph. **13**, 323 (1995)
55. F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski, J.M. Thornton, Proteins Struct. Funct. Bioinform. **62**, 479 (2005)
56. B. Li, S. Turuvekere, M. Agrawal, D. La, K. Ramani, D. Kihara, Proteins Struct. Funct. Bioinform. **71**, 670 (2008)
57. M. Hendlich, F. Rippmann, G. Barnickel, J. Mol. Graph. Model. **15**, 359 (1997)
58. B. Huang, M. Schroeder, BMC Struct. Biol. **6**, 19 (2006)
59. P.J. Goodford, J. Med. Chem. **28**, 849 (1985)
60. A.T.R. Laurie, R.M. Jackson, Bioinformatics **21**, 1908 (2005)
61. D. Ghersi, R. Sanchez, Bioinformatics **25**, 3185 (2009)
62. R. Harris, A.J. Olson, D.S. Goodsell, Proteins Struct. Funct. Bioinform. **70**, 1506 (2007)
63. C. Hetényi, D. van der Spoel, Protein Sci. **11**, 1729 (2002)
64. C. Hetényi, D. van der Spoel, FEBS Lett. **580**, 1447 (2006)
65. R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, J. Comput. Chem. **28**, 1145 (2007)
66. P.D. Thomas, K.A. Dill, J. Mol. Biol. **257**, 457 (1996)
67. Z. Zheng, K.M. Merz Jr., J. Chem. Inf. Model. **53**, 1073 (2013b)
68. M. Feher, Drug Discov. Today **11**, 421 (2006)
69. R.E. Amaro, R. Baron, J.A. McCammon, J. Comput. Aided Mol. Des. **22**, 693 (2008)
70. I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, J. Mol. Biol. **161**, 269 (1982)
71. W.J. Allen, T.E. Balius, S. Mukherjee, S.R. Brozell, D.T. Moustakas, P.T. Lang, D.A. Case, I.D. Kuntz, R.C. Rizzo, J. Comput. Chem. **36**, 1132 (2015)
72. S. Mukherjee, T.E. Balius, R.C. Rizzo, J. Chem. Inf. Model. **50**, 1986 (2010)
73. G. Morris, D. Goodsell, J. Comput. Chem. 1639 (1998)
74. G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, J. Comput. Chem. **30**, 2785 (2009)
75. O. Trott, A.J. Olson, J. Comput. Chem. **31**, 455 (2010)
76. R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, J. Med. Chem. **47**, 1739 (2004)
77. M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Proteins **52**, 609 (2003)
78. W.T.M. Mooij, M.L. Verdonk, Proteins Struct. Funct. Bioinform. **61**, 272 (2005)

79. J.W.M. Nissink, C. Murray, M. Hartshorn, M.L. Verdonk, J.C. Cole, R. Taylor, Proteins Struct. Funct. Bioinform. **49**, 457 (2002)
80. J. Avorn, N. Engl, J. Med. **372**, 1877 (2015)
81. M. Dickson, J.P. Gagnon, Discov. Med. **4**, 172 (2009)
82. A. Lavecchia, C. Di Giovanni, Curr. Med. Chem. **20**, 2839 (2013)
83. E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, Annu. Rep. Comput. Chem. **4**, 217 (2008)
84. X.-Q. Xie, Expert Opin. Drug Discov. **5**, 1205 (2010)
85. T. Sterling, J.J. Irwin, J. Chem. Inf. Model. **55**, 2324 (2015)
86. NCI Compound Sets (The National Cancer Institute (NCI), Bethesda, MD, United States of America, 2017), https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets. Accessed 20 July 2015
87. S. Cosconati, S. Forli, A.L. Perryman, R. Harris, D.S. Goodsell, A.J. Olson, Expert Opin. Drug Discov. **5**, 597 (2010)
88. M.A. Johnson, G.M. Maggiora, *Concepts and Applications of Molecular Similarity* (Wiley, 1990)
89. P. Willett, Drug Discov. Today **11**, 1046 (2006)
90. A. Dalke, J. Cheminform. **5**, 36 (2013)
91. D. Horvath, Methods Mol. Biol. **672**, 261 (2011)
92. J. Melville, E. Burke, J. Hirst, Comb. Chem. High Throughput Screen. **12**, 332 (2009)
93. C.J. Harris, R.D. Hill, D.W. Sheppard, M.J. Slater, P.F.W. Stouten, Comb. Chem. High Throughput Screen. **14**, 521 (2011)
94. P. Ripphausen, B. Nisius, J. Bajorath, Drug Discov. Today **16**, 372 (2011)
95. Y.-K. Chiang, C.-C. Kuo, Y.-S. Wu, C.-T. Chen, M.S. Coumar, J.-S. Wu, H.-P. Hsieh, C.-Y. Chang, H.-Y. Jseng, M.-H. Wu, J.-S. Leou, J.-S. Song, J.-Y. Chang, P.-C. Lyu, Y.-S. Chao, S.-Y. Wu, J. Med. Chem. **52**, 4221 (2009)
96. A. Vuorinen, R. Engeli, A. Meyer, F. Bachmann, U.J. Griesser, D. Schuster, A. Odermatt, J. Med. Chem. **57**, 5995 (2014)
97. A. Kurczyk, D. Warszycki, R. Musiol, R. Kafel, A.J. Bojarski, J. Polanski, J. Chem. Inf. Model. **55**, 2168 (2015)
98. K.H. Barakat, L.P. Jordheim, R. Perez-Pineiro, D. Wishart, C. Dumontet, J.A. Tuszynski, PLoS ONE **7**, e51329 (2012)
99. D.C. Thompson, C. Humblet, D. Joseph-McCarthy, J. Chem. Inf. Model. **48**, 1081 (2008)
100. F. Gentile, J.A. Tuszynski, K.H. Barakat, J. Mol. Graph. Model. **65**, 71 (2016)
101. F. Gentile, J.A. Tuszynski, K.H. Barakat, Curr. Pharm. Des. **22**, 3527 (2016)
102. L.P. Jordheim, K.H. Barakat, L. Heinrich-Balard, E.-L. Matera, E. Cros-Perrial, K. Bouledrak, R. El Sabeh, R. Perez-Pineiro, D.S. Wishart, R. Cohen, J. Tuszynski, C. Dumontet, Mol. Pharmacol. **84**, 12 (2013)
103. E.M. McNeil, K.R. Astell, A.-M. Ritchie, S. Shave, D.R. Houston, P. Bakrania, H.M. Jones, P. Khurana, C. Wallace, T. Chapman, M.A. Wear, M.D. Walkinshaw, B. Saxty, D.W. Melton, DNA Repair (Amst). **31**, 19 (2015)
104. R.R. Nunes, M.D.S. Costa, B.D.R. Santos, A.L. da Fonseca, L.S. Ferreira, R.C.R. Chagas, A.M. da Silva, F.P. de Varotti, A.G. Taranto, Mem. Inst. Oswaldo Cruz **111**, 721 (2016)
105. D. Bajusz, G. Ferenczy, G. Keserű, Curr. Top. Med. Chem. **17**, 2235 (2017)
106. W.-G. Gu, X. Zhang, J.-F. Yuan, AAPS J **16**, 674 (2014)
107. N. Li, R.I. Ainsworth, B. Ding, T. Hou, W. Wang, J. Chem. Inf. Model. **55**, 1400 (2015)