Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

Review

Molecular dynamics and related computational methods with applications to drug discovery

Jack A. Tuszynski[1,2*], T. Luchko[1], Philip Winter[2], Cassandra Churchill[3], Kamlesh Sahu[1], Francesco Gentile[1], Sara Ibrahim Omar[2], N. Nayebi[1], G. Hu, K. Wang[4] and J. Ruan[4]
[1]Department of Physics and [2]Department of Oncology, University of Alberta, Edmonton, Canada
[3]Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada
[4]College of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, P.R. China

[*]Correspondence Email: jackt@ualberta.ca

Received 15 January 2015; Revised 19 July 2017; Accepted 2 December 2017; Published 30 July 2018
Editor: Mohammad Ashrafuzzaman

Abstract
The computational technique of molecular dynamics is discussed, with special attention to force fields for protein simulations and to 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 databases and publicly-available software of use in drug design and discovery is provided. The main objective of this paper is to give the reader a practical toolbox for applications in quantitative biology and computational drug discovery.

Key words. Solvation free energy, ligand binding pocket, computation, molecular dynamics, database, software, drug design.

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

1. Introduction

Drug design is a conceptual and computational approach to find 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 signalling pathway that is specific to a disease condition. Most approaches attempt to inhibit the functioning of an aberrant pathway in the diseased state by interfering with the normal functioning of the target. Drugs may be designed at a molecular level that bind to the active region and inhibit this target molecule. However, these drugs should be designed in such a way as not to affect any other important biomolecules that may cause undesired side effects. Due to the complexity of the drug design process, serendipity has traditionally played a key role in finding potential new drugs since many challenges are posed by the large chemical and biological spaces involved. Conversely, rational drug design requires knowledge of the bio-molecular target. For example, structure-based drug design utilizes three-dimensional information about biomolecules obtained from techniques such as x-ray crystallography and NMR spectroscopy. Using such information, the effect of a drug bound to its binding site on the biomolecule can be studied. However, one must also consider whether a drug candidate molecule is “drug-like”, which is related to the necessary physical properties for effective absorption and biological action.

The first step in the rational drug design process is usually the identification and characterization of the bio-molecular 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 bio-molecular target, and this information can be used to design novel drug panels with enhanced activity. Of the computational techniques available, molecular dynamics is particularly important in the investigation of target characterization and drug-target interactions. Below, we provide an overview of protein structure characterization and molecular dynamics methods.

Protein morphology characterization is the first step in analyzing the properties of the drug target. For each protein of interest, we may obtain information about protein sequences, 3D structural data and binding pockets using these sources:

- Swiss-Prot (http://www.expasy.ch/sprot/sprot-top.html) and SWISS-MODEL Repository (http://swissmodel.expasy.org/repository/)
- X-ray crystallography, cryo-electron microscopy or nuclear magnetic resonance (NMR) spectroscopy (deposited in the Protein Data Bank (http://www.rcsb.org/pdb/))
- Sequence homology algorithms
- CASTp, (http://cast.engr.uic.edu/cast/)

3D structural data may be reported in the following formats: PDB, Cambridge, AMPAC/MOPAC, MDL Molfile e.g. SD files, Sybyl, MOL2, and xpdb-mini, JCAMP/CS, CIF, MDL RXN, RDF, XYZ, SMD4, SMD5, CTX, CACTVS, SMILES (including SMARTS subset), Compass, 441, Gaussian/Input, Gaussian/Archive, SCF, SHEL-X, XTEL, Cerius (Ascii exchange format of CeriusII Toolkit), Sharc (SHift ARChives format), Alchemy, Hyperchem, Molconn-Z, Sybyl, Sybyl2, SLN.

Properties of proteins: BioGeometry (http://biogeometry.cs.duke.edu/index.html) presents computational techniques for representing, storing, searching, simulating, analyzing, and visualizing biological structures. It contains a list of software to calculate the shape of biological structures. The following is a list of subroutines available and their functions:

- Almost-Delauanay Tetrahedra (AlmDel): is used for analyzing protein structure
- Alpha Shapes: is general purpose software for analyzing protein structures
- Ciel: can be used for generating interface surfaces from protein structural data
- Coreset: is used for computing shape descriptors of low-dimensional data sets
- Monte Carlo Simulation: For performing efficient Monte Carlo simulations of proteins, which is especially valuable in the case of intrinsically disorderd proteins where huge numbers of quasi-stable conformations exist
- ProShape: can be used for computing protein measures and their derivatives
- Skin Software: is found useful in triangulating the surface of a protein
- Stochastic Roadmap Simulation (SRS): is a good tool for computing ensemble properties of molecular motions (folding, binding)
- Writhe: is of use in computing the number of protein backbones
- CGAL: (CGAL.http://www.cgal.org/) is an Open Source Project to provide easy access to efficient and reliable geometric algorithms in the form of a C++ library.

2. Molecular dynamics

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

Molecular Dynamics (MD) simulations involve the numerical integration of Newton's equations of motion (EOM), calculating the force on each atom from a potential to evolve these atoms through time and space including forces acting on each atom and the effects of thermal noise. Generally, classical potentials are used, which provide accurate approximations of many properties, such as the inter-nuclear distance between bonded atoms and electrostatic interactions. However, not all properties of the system can be captured this way. It is standard practice to use one of several software packages that have been developed over the past 40 years. Some of the most popular packages are: Amber, CHARMM, NAMD, GROMACS, GROMOS, Martini etc.

Molecular dynamics 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. This includes the root-mean-square deviation (RMSD) of ligand and protein atoms, supramolecular (non-covalent) interactions, binding free energies, [7] changes in the potential energy of the system, short-lived reaction intermediates, [8] conformational changes, flexibility, and optimum binding modes [9] 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 by MD simulations can help identify the amino acid residues that play an important role in ligand binding (hot spots) [16]. MD simulations can also be used for studying ionic conductivity [10], 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 [11], binding sites [12], drug resistance mechanisms [13], and protein folding [14], [15] to name but a few.

In order to run an MD simulation on a crystal structure of a protein, several steps must be taken to ensure that the system is as physiologically-consistent as possible. The following general protocol is a balance between physiological accuracy and computational efficiency.

1. Remove solvent molecules, crystallization salts and other extraneous atoms.
2. Determine the protonation state of amino acids.
3. Convert the file of the system to be simulated to the appropriate format for MD software.
4. Add appropriate counter-ions to the system to produce a zero net charge in random positions that do no overlap with the protein.
5. Use Langevin Dynamics (LD), to simulate an aqueous environment, with a long electrostatic cutoff radius to approximate a Debye-Hückel distribution of the ions. The protein should be fixed in place to decrease computation time.
6. Solvate the system in a sphere of water, rectangular prism, truncated octahedron, etc. using a water sample equilibrated at standard pressure and temperature.
7. Minimize the solvated system using the Steepest Decent algorithm to remove energetically unfavorable conformations.

In x-ray crystallography, crystallization is induced through the use of salts and other molecules which may not represent physiological conditions. Tightly bound hydration water is also often found the in crystal structures. Thus, in Step 1 we need to remove these atoms as they interfere with physiological protein dynamics. Since proteins are composed of amino acids, they typically have some net charge. The net charge depends on the consistency of the solute, the pH and the local charge of the surrounding system. Several methods exist to determine the net charge of a protein, including the online program WHAT IF [3] where only histidine protonation is considered since it has a pK_a near 7.0. Since histidine has three protonation states (HSD - proton on ND1, HSE - proton on NE2, and HSP - proton on both ND1 and NE2) a hydrogen bond network analysis can determine what protonation state was used in the crystallographic assignment. However, this is typically set to HSP for all histidine residues, the least likely form. A second method uses Engh and Huber [4] geometries for histidine and used a statistical analysis on small molecule entries in the Cambridge Structural Database [4]. In Step 2 one should arrive at a topology that is consistent with both methods. Step 3 involves a "simple" file format conversion. Most protein structures are currently available in the Protein Data Bank [5] www.rcsb.org in the PDB format. In Step 4, one must ensure the net charge of the system is neutral. Neutralizing the system is accomplished by adding counter ions to the system, where the appropriate ions for the particular system must be selected. Typical concentrations of major ions in the mammalian cell are as follows [6]:

Ion	K^+	Na^+	Cl^-	Mg^{2+}	Ca^{2+}
Concentration	140 mM	10 mM	10 mM	0.5 mM	0.1 μM

In Step 5 we use a very large cut off radius to ensure that the ions are not merely diffusing. After neutralizing the system, the next most important step is hydration in Step 6. The addition of water to the system is completed using models pre-equilibrated to room temperature (300K) and atmospheric pressure. In Step 7, the structure is relaxed to eliminate any energetically-unfavourable positions, such as steric overlap, or overly-stretched bonds and angles. These instances would experience large forces when calculated as part of the simulation. As a result, unnatural vibrational modes may be found or extremely high atom velocities. By minimizing the system, these problems are addressed.

Performing an MD simulation requires an atomic-resolution model of the system being simulated. This model for biological macromolecules may be obtained from nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallographic data and electron microscopy. NMR or crystallographic structures of bio-macromolecules can be downloaded from the Protein Data Bank (http://www.pdb.org). In the absence of 3D experimental data of a desired target, the model can be obtained by homology modeling, which uses a known 3D structure of a homologous protein as a template, along with the amino acid sequence of the desired protein, to get a structure. However, the quality of results are dependent on the sequence identity of the target and template, which should be 40% or higher.

During an MD simulation the forces acting on each atom of the system are calculated and atoms are moved according to those forces. This type of simulation is advanced by a small time step (commonly on the order of a femtosecond) to obtain a trajectory. The net forces are a sum of contributions due to covalently-bonded interactions and non-bonded interactions. Parameter fitting is needed to calculate these interactions [17]. Non-bonded interactions include van der Waals' and electrostatic interactions modeled respectively by the Lennard-Jones (LJ) potential and Coulomb's law. Parameter fitting is done to reproduce the actual behavior of real molecules. This includes determining the van der Waals' radii, partial charges on atoms, bond lengths and bond angles. These parameters collectively define a force field. To simulate effects of solvent on biomolecules, one can use either explicit or implicit solvent models. While explicit solvent models attempt to provide a realistic model by including the solvent molecules in the system, implicit models use a mean field approach [18], [19]. A potential of mean force (PMF) is applied to approximate the behavior of many solvent molecules. Explicit solvent simulations are computationally expensive because enormous numbers of solvent molecules, such as TIP3P water molecules, are added to the system for realistic simulations reflecting the molecular complexity of the biomolecule and its environment. Implicit solvent models increase the speed of simulation because solvent effects, due to the presence of a huge number of explicit solvent molecules, have been represented by various empirical functions and no Newtonian equations of solvent molecules need to be solved.

Brownian dynamics utilizes the Langevin EOMs to simulate particles immersed in a solvent or in contact with a heat bath. The intended use of Langevin dynamics within CHARMM is as a heat bath when using a Stochastic Boundary Potential (SBP) or a Spherical Solvent Boundary Potential (SSBP). In both these formulations, the structure of interest is explicitly solvated in a sphere of water with radius r. Outside of this distance r, Langevin dynamics is used to implicitly simulate the effects of the water. Other than SBP and SSBP, Langevin dynamics is also used in fully-implicit solvent models. Such models use a dielectric constant (commonly $\varepsilon = 80$, representing water), although distance-dependent dielectrics may also be used. Since this method eliminates the need for explicit water molecules, it is significantly faster than utilizing an explicit fully-hydrated system. It may be particularly usefully for ion equilibration when preparing a system for simulation.

Regarding simulation efficiency and accuracy, several techniques described below are commonly used. First, one can restrain the fastest vibrations (involving hydrogen atoms). The SHAKE algorithm is one example. This allows a larger time step to be used during simulations [20]. Second, long-range Coulomb interactions are taken into account by approaches such as a cut-off based method [21] or the particle mesh Ewald (PME) approach [22]. If the cut-off range is shortened, computational time is reduced accordingly but this comes at the expense of computational accuracy. On the other hand, if the cut-off range is increased, accuracy improves at the cost of computational time. The researcher needs to select an optimal value in order to keep a balance between two.

MD can provide valuable insights into how the conformation of a protein may change with time. The binding of ligands is also affected by protein flexibility. Receptor flexibility can be handled by a structure ensemble approach. When a small molecule approaches a receptor protein, the receptor is in continuous motion and both the

receptor and small molecule adjust their conformations to fit each other. Consequently, the corresponding binding energy can be calculated more accurately as the average binding energy of an ensemble of snapshots obtained from MD trajectories [23]–[25]. Using multiple trajectories obtained from same initial structures improves the accuracy of binding energy calculations [26]. The improvement of accuracy is due to denser sampling of the conformational space by utilizing multiple trajectories. Increasing the time duration of a simulation allows even more conformational sampling and may result in more accurate binding energy calculations.

Decomposition of binding free energies obtained using frames from MD trajectories is an important way to obtain information about the residues that significantly contribute to binding affinity. Residue-wise decomposition also gives insights into the changes in binding free energies due to mutations, especially single point mutations [27]. The molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) method of estimating the binding free energy has been widely used to calculate and decompose binding energies [28].

Improvements in computer hardware and molecular modeling techniques have resulted in great improvements in the ability to study 3D models of molecules. Molecular dynamics is a very useful and inexpensive tool to study behavior of molecules *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 bio-molecular system to reach equilibrium, MD can be run long enough to represent evolution of a system from a few nanoseconds to a few microseconds. Long time-scale dynamical processes, such as slow conformational changes or protein assembly processes, are notoriously difficult to model by MD. Enhanced sampling can be achieved by metadynamics [29] to explore the transition regions between stable states of a system. Replica-exchange molecular dynamics (REMD) also enhances sampling by overcoming barriers between stable states. This is accomplished by allowing systems of similar potential energies to sample conformations at different temperature. In addition to sampling, MD is limited by the accuracy of the force field selected for simulations. Further improvement of force fields to reproduce, *in silico*, the behavior of real molecules is still important, especially as improvements in computer hardware provide the ability to run longer simulations.

Molecular dynamics is used together with other methods to solve a host of problems in bio-molecular modeling [32]–[38]. Still, the accuracy of force fields and the treatment of solvent effects are two key areas where significant scope for improvement exists. In the case of virtual screening 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. This will first require an MD protocol to be validated. Validating a protocol can be done with the help of inhibitors of the same enzyme that have an experimentally-determined activity (i.e. positive controls). The better the correlation between the calculated binding free energy and the known activity, the higher the confidence in the predicted binding energies of potential hits. Longer simulations using multiple trajectories are computationally expensive but may aid in calculating more accurately the respective binding energies and may result in a better correlation with experimental data. Including the ions and co-factors present in the system for MD simulation with correct parameters is a major step towards improved accuracy, especially if the ions or cofactors are close to the binding site of a receptor. Good MD sampling positively influences the process of virtual screening. Accurate modeling of physiologically relevant conformations is essential to structure-based drug design. Several studies indicate that virtual screening can be improved by taking into account the conformational freedom of a protein [39].

As mentioned above, in a classical MD simulation, the forces on atoms are calculated from Newton's second law is $m\ddot{x} = -\nabla V$, where V is the potential energy function. The function that describes the potential energy is called a force field. Today, several force fields are available, and they are divided in three groups: (a) all-atoms force fields (parameters are considered for every atoms), (b) united atoms force fields (aliphatic hydrogens are represented implicitly) and (c) coarse-grained force fields (groups of atoms are treated as super atoms). Here, we focus on all-atom and united atom force fields.

In order to study large systems with MD, a potential must be both simple and physically accurate. Potentials derived from classical mechanics called forcefields are largely empirical,

i.e. they consist of functions depending on atom coordinates that are parameterized based on experimental and *ab initio* data to reproduce observed experimental equilibrium behaviours. Classical forcefields generally adopt the following form:

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

The first term represents the potential between two chemically-bound atoms, modeled as a simple harmonic potential. The parameters are K_b (related to force constant) and b_0 (related to equilibrium bond length), while the variable, b, is the distance between atoms.
The proximity between three atoms that are related via chemical bonds can be described with an angle. The second term in Equation (1) represents this angle-dependence involving three atoms and is also modeled by a harmonic potential, where θ_0 (equilibrium angle) and K_θ (force constant) are parameters and θ is the angle between the three atoms in the structure. The third term in Equation (1) represents dihedral angle (torsion) potential and depends on four atom coordinates. This potential is periodic and is usually represented by a cosine function and involves parameters K_ϕ, (barrier for rotation), n (number of maxima) and φ (angular offset). The variable, ϕ, is obtained from dihedral angles in the structure. Improper torsion terms

The last two terms represent non-bonded interactions and are calculated pairwise between atoms *i* and *j*. The fourth term in Equation (1) is the Van der Waals potential, which is typically represented by a Lennard-Jones 6-12 potential. The r^{-6} term is the attractive component while the r^{12} term approximates Pauli repulsion. Parameters A_{ij} and B_{ij} are atom specific, while r_{ij} is a variable representing the distance between atoms i and j. The final term represents the electrostatic potential between atoms, and is modeled as a Coulomb potential. Parameters q_i and q_j are (fixed) charges on atoms i and j, while the constant ε_0 is the permittivity of free space Electrostatic interactions dominate over van der Waals forces for long-range inter-molecular interactions and they play a significant role in non-chemical binding. Each atom in a structure is assigned a partial charge from *ab initio* simulations.

Here we present a summary of the most commonly used force fields in MD simulations, namely AMBER, CHARMM, OPLS and GROMOS.

The functional form from which most of the AMBER (Assisted Model Building with Energy Refinement) force fields are derived is the one developed by Cornell and co-workers (Equation (1)) [17]. The most frequently used Amber fixed-charges force field version for proteins and nucleic acids is ff99SB [41], developed as a modification of the old ff99 force field [42]; this version allows good results to be obtained [43], [44]. A newer version for proteins studies, ff12SB [45], introduced better secondary structure prediction compared with ff99SB. Comparisons between the two versions found that ff12SB performed better [46]. Another AMBER force field used extensively is ff03 [47], [48], which introduced some changes to ff99 related to charge calculations and changes in Ψ and Φ backbone torsions for proteins. A united atom version of ff03, the ff03ua, is also available [49]. Furthermore, a general AMBER force field (GAFF) [50] was developed to include parameters for small molecules. Therefore, it is possible to use this force field to perform MD simulations of receptor-ligand complexes.

The CHARMM (Chemistry at Harvard Macromolecular Mechanics) force fields are the second most frequently used set of force fields for MD simulations. The CHARMM force fields use classical (empirical or semi-empirical) and quantum mechanical (semi-empirical or *ab initio*) energy functions for different types of molecular systems. They include parameters for proteins, nucleic acids, lipids and carbohydrates, allowing simulations on all commonly-encountered biomolecules. The initial version of CHARMM used an atom force field with no explicit hydrogens [51]. Later, the CHARMM19 parameters were developed, in which hydrogen atoms bonded to nitrogen and oxygen were explicitly represented; hydrogens bonded to carbon or sulfur are still treated as extended atoms [52]. The idea behind the CHARMM19 parameters was to obtain a balanced interaction between solute-water and water-water energies. Although this force field was tested primarily on gas-phase simulations, it is now used for peptide and protein simulation with implicit

solvent models. In CHARM22, the atomic partial charges were derived from quantum chemical calculations of the interactions between model compounds and water [53]. CHARMM22 is parameterized for the TIP3P explicit water model, although it is frequently used with implicit solvents. A corrected version of CHARMM22 with dihedral potential corrected was released as CHARMM22/CMAP [54]. CHARMM27 parameters were developed for nucleic acids (RNA, DNA) and lipid simulations [2]. Therefore, CHARMM22 and CHARM27 can be combined for the simulation of ligands or proteins binding to nucleic acids.

For the OPLS (Optimized Potentials for Liquid Simulations) force fields, the potential energy function was originally designed [51] to simulate the properties of the liquid states of water and organic liquids, and its performance was shown to be better than other force fields [52]. For proteins, a united atoms version was followed by an all-atoms version (OPLS-AA) [55]. The OPLS-AA force field uses the same parameters as the Amber force fields for bond stretching and angles. The torsional parameters were obtained by using data from *ab initio* molecular orbital calculations for 50 organic molecules and ions [56]. Several improvements and re-parameterizations were proposed later [55], [57], including for simulations of phospholipid molecules [58].

The GROMOS (Groningen Molecular Simulation) force fields were developed in conjunction with the software package of the same name to facilitate bio-molecular simulations in a university environment [58]. The initial GROMOS force field was developed for applications to aqueous or apolar solutions of proteins, nucleotides and sugars. However, a gas phase version for the simulation of isolated molecules is also available [58]. The major versions of the GROMOS force fields are GROMOS 43A1 [59], GROMOS 45A3 [60], GROMOS 53A5 and 53A6 [61], and GROMOS 54A7, 54B7 and 54A8 [62], [63]. The GROMOS force fields are united atom force fields, i.e. without explicit aliphatic (non-polar) hydrogens. These force fields are widely used for the simulation of protein folding, computational drug design, and other types of MD.

The calculation of **solvation free energies** is still one of the more challenging problems in MD simulations. Determining solvation free energy is especially difficult in aqueous bio-systems since they are relatively large [64]. Solvation free energy, ΔG_{solv}, is defined as the net energy change upon transferring a molecule from the gas phase into a solvent with which it equilibrates [65]. Solvation effects can change the physical and chemical properties of biomolecules including charge distribution, geometry, vibrational frequencies, electronic transition energies, NMR constants and chemical reactivity. Several methods for modeling solvation can be selected depending on the required accuracy and computational cost. Ordered from the highest accuracy (and most computational cost) to the lowest accuracy (and least computational cost), the types of methods are: polarizable explicit solvent, fixed charge explicit solvent, simple explicit solvent, nonlinear Poisson–Boltzmann, linear Poisson–Boltzmann, and generalized Born.

The simplest method is to treat the physical effects (e.g. electrostatic interactions, cavitation, dispersion attraction and exchange repulsion) of the solute implicitly. This method represents the solvent as a continuum environment. The most important factors to be considered in an implicit solvent model are electrostatic interactions and cavitation. The electrostatic component of the solvation free energy is the work needed to charge the solute in solution minus this work in vacuum [64], [65]. Cavitation refers to the size and shape of a cavity that the solute can occupy. Several different implicit solvent models are briefly discussed below.

The solvation free energy of a molecule, ΔG_{solv}, can be divided into two parts: electrostatic and non-electrostatic. The electrostatic energy is the energy needed to remove the charges in vacuum and in the continuum solvent environment charge the molecule again. The origin of the non-electrostatic energy is a combination of the van der Waals interactions between the solute and solvent molecules and the breaking of the water structure in the presence of the solute molecules. The generalized Born (GB) model is based on the Born approximation of point charges in a spherical cavity for each of the solute atoms. The cavity dielectric continuum represents the polarization effects of the solvent. Numerical methods are used to determine the point charges on the cavity surface that make the same electrostatic potential in vacuum as it appears from the solute's charge distribution.

The polarization energy is calculated by making approximations in the Poisson-Boltzmann (PB) equation,

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

$$\Delta G_{el} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{int}}} - \frac{1}{\varepsilon_{\text{ext}}}\right)\sum_{ij}\frac{q_i q_j}{\left[r_{ij}^2 + \alpha_i\alpha_j \exp\left(\left.-r_{ij}^2\right/4\alpha_i\alpha_j\right)\right]^{1/2}}, \quad (2)$$

where α_i is effective Born radius of particle i, r_{ij} is the distance between atoms i and j, ε_{int} and ε_{ext} the internal and external dielectric constants, respectively, and q_i is the electrostatic charge on particle i. The non-electrostatic contribution is calculated by empirical methods as a function of the solvent accessible area. This is added to the electrostatic part to yield the solvation free energy [64]–[69].

Poisson–Boltzmann model: The electrostatic force can be calculated from the PB equation (2) for solvents containing ions. Solving Poisson's equation gives the function describing the electrostatic environment that is modeled with a dielectric continuum model,

$$\nabla \cdot \varepsilon(r)\nabla\varphi(r) = -4\pi\rho(r), \quad (3)$$

where $\varphi(r)$ is the electrostatic potential, $\varepsilon(r)$ is the dielectric constant and $\rho(r)$ is the charge density. Poisson's equation has to be solved using computers and adopting numerical methods, since there is no known analytic solution to this equation except in very simple situations. The Boltzmann part, along with the assumptions of the Debye–Hückel theory, helps us treat the charge density due to ions in solution. The result can be written as

$$\nabla \cdot \varepsilon(r)\nabla\varphi(r) - \kappa \sinh[\varphi(r)] = -4\pi\rho(r). \quad (4)$$

This represents the nonlinear Poisson–Boltzmann equation, with κ denoting the Debye–Hückel parameter. Here, the charge density on the right represents the partial charges in the cavity. When the ionic strength of the solution or the potential is not very high, the equation can be linearized by expanding the second term on the left into a Taylor series and retaining only the first term to obtain

$$\nabla \cdot \varepsilon(r)\nabla\varphi(r) - \kappa\varphi(r) = -4\pi\rho(r) \quad (5)$$

The PB equation is computationally expensive to calculate without approximations [65], [67], [70].

Another type of solvation model is a probabilistic method known as the 3D reference interaction site model (3D-RISM) [64], [71]–[77]. This molecular theory of solvation simulates the solvent distributions rather than the individual solvent molecules. However, the solvation structure and the associated thermodynamics are obtained from the first principles of statistical mechanics. In this method, the 3D site density distributions of the solvent account for different chemical properties of the solvent and solute. These properties include hydrogen bonding, hydrophobic forces, and solvation thermodynamics, such as the partial molar compressibility and volume. The solvation free energy is calculated from the RISM equation as well as the closure relation [78]–[82]. Several additional advances have been made in formulating improved versions of the 3D-RISM theory including the hypernetted chain (HNC) closure approximation [71], [72]. Another derivation came from the molecular Ornstein–Zernike integral equation [82] for the solute-solvent correlation functions [74], [74], [75], [83]. Sometimes the calculated solvation free energy for ionic and polar macromolecules involves large errors due to the loss of long-range asymptotics of the correlation functions. Analytical corrections of the electrostatic long-range asymptotics for the 3D site direct correlation functions as well as the total correlation functions have been obtained [76], [77], [83]. Other developments include the closure approximation, 3D-KH closure, for solid-liquid interfaces, as well as poly-ionic macromolecules [75], [83]. 3D-RISM has been coupled with MD by making use of a multiple time step (MTS) algorithm [84], [85] or by the contraction of the solvent degrees of freedom and the extrapolation of the solvent-induced forces. This resulted in faster calculations, which is useful for larger systems [66]. A multi-scale method of multiple time steps molecular dynamics (MTS-MD) is referred to as MTS-MD/OIN/ASFE/3D-RISM-KH [66] and it converges the 3D-RISM-KH equations at large outer time steps. Converging the 3D-RISM-KH integral was obtained by using solvation force-coordinate extrapolation (SFCE) in the subspace of previous 3D-RISM-KH solutions [86]. Another developed model is the 3D-RISM-KH-NgB [87]. In this model the non-polar component of the hydration free energy obtained from 3D-RISM-KH is corrected using a modified Ng bridge function [88]. Improved performance of 3D-RISM calculations was obtained them on graphical processing units (GPUs) with a modification of the Anderson Method [89] that accelerates convergence [90].

Explicit solvation is characterized by modeling individual water molecules around a solute. There are several explicit water models available in the Amber, NAMD and Gromacs MD simulation packages, including: SPC [91], SPC/E [91], POL3 [92], TIP3P [93], TIP3P/F [94], TIP4P [93], [95], TIP4P/Ew [96] and TIP5P [97]. Examples of explicit water models are the simple point charge (SPC) model and the extended simple point charge (SPC/E) model [91]. In both of these models the water molecules are rigid. A derivative of SPC with flexible water molecules has been developed [98]. Another simple explicit model is the POL3 water model, which is a polarizable model [92]. More complex explicit water models include the transferable intermolecular potential 3 point (TIP3P) model, and its 4 and 5 point derivatives (TIP4P and TIP5P). The numbers in these models represent the number of interaction sites in each model, with just the basic oxygen atom and two hydrogen atoms modeled in the case of TIP3P [93]. A re-parameterized model of TIP3P is the TIP3P-PME/LRC, also referred to as TIP3P-F [94], which calculates electrostatic contributions by particle mesh Ewald (PME) summation and includes a long-range van der Waals correction (LRC). TIP4P [93], [95] introduced a fourth dummy atom bonded to the oxygen, improving the electron distribution of the water molecule. This model has been re-parameterized for use with Ewald sums: TIP4P/Ew [93], [95]. The five interaction points in the TIP5P [97] model include two dummy atoms, which further improves the charge distribution around the water molecule.

3. Docking Methods

Virtual screening has attracted much attention in the pharmaceutical industry [99], [100]. It provides a more economical way to screen diverse chemicals as drug candidates compared with wet-lab approaches. It consists of the creation of a chemical library of ligand structures, followed by searching for optimal ligand-receptor binding modes through docking algorithms, and finally the evaluation of binding affinities. There are three criteria that are required to successfully identify drug candidates. First, the chemical library needs to be large and contain diverse chemical structures. Second, conformational searching algorithms need to be able to search many possible binding modes within a reasonable time. Third, an appropriate scoring function needs to be utilized to correctly evaluate the binding affinity of the ligand-receptor interaction so that the ligands can be ranked. In the framework of information theory, the first and third criteria represent the fundamental information required in virtual screening process. The second criterion then can be treated as an information processing guideline. The efficiency and accuracy of this step will depend on the methods of information processing.

Genetic algorithms, which borrow from the concept of genomic evolution processes to search conformations of complex targets and chemical structures, are commonly used in docking protocols, such as AutoDock [101]. Chang et al. have offered a better alternative, MEDock [102]. Although MEDock did not completely exploit entropic-based inductive inference for searching, it does utilize the maximum entropy principle as a guideline to make decisions during this process. The fundamental question asked in MEDock is "What is the probability of finding the deepest energy valley in a ligand-target interaction energy landscape?". Maximum entropy provides a direction to update the initial guess of binding modes (described by an almost uniform distribution) to the optimal mode (a localized distribution around the global energy minimum).

Other popular docking software packages are listed below**.**

- DOCK (http://dock.compbio.ucsf.edu/) considers the biggest cluster as the active one.
- DockIt (www.metaphorics.com/products/dockit.html)
- GOLD (http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx) is a program for protein-ligand docking.
- Haddock (http://www.nmr.chem.uu.nl/haddock/)
- Fred (http://www.eyesopen.com/products/applications/fred.html)
- Flipdock (http://flipdock.scripps.edu/)

Although these algorithms also contain pocket searching and pocket-ligand matching algorithms, they are not fast enough. The program named Pocket can find the pockets and mouths of a protein much faster than DOCK. Instead of the probe sphere, it describes the pockets and mouths by the residues surrounding them. Additional references that can assist the reader in studying virtual screening processes include [103]–[114].

4. Drug Design

Quantitative structure-activity relationships (QSAR) represent an attempt to correlate structural or property descriptors of compounds with their

chemical and biological activities. These physicochemical descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects, are determined empirically or, more recently, by computational methods. Activities used in QSAR include chemical measurements and biological assays. QSAR have been applied in many disciplines, with many applications pertaining to drug design and environmental risk assessment. QSAR dates back to the 19th century when A.F.A. Cros observed that toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased **[115]**. In the 1890's, H.H. Meyer and C. E. Overton independently noticed that the toxicity of organic compounds depended on their lipophilicity [115], [116] . Later L. Hammett correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity. Hammett observed that adding substituents to the aromatic ring of benzoic acid had an orderly and quantitative effect on the dissociation constant. These relationships are termed linear free energy relationships. That is, the free energy is proportional to the logarithm of the equilibrium constant. Although they can be stated in terms of thermodynamic parameters, no thermodynamic principle states that the relationships should be true. Tables of values for numerous substituents have been published [117], [118]. A QSAR was developed based on the values of the substituents. QSAR based on Hammett's relationship utilize electronic properties as the descriptors of structures. Difficulties were encountered when applying Hammett-type relationships to biological systems, indicating that other structural descriptors were necessary. However, Hansch recognized the importance of the lipophilicity, expressed as the octanol-water partition coefficient, on biological activity [119]. This parameter provides a measure of the bioavailability of compounds, which determines, in part, the amount of the compound that arrives at the target site. Relationships were developed to correlate a structural parameter (i.e., lipophilicity) with activity, in some cases univariate relationships correlating structure and activity were found but in other cases parabolic relationships between biological response and hydrophobicity could be fit by including a $(\log P)^2$ term in the QSAR. QSAR are now developed using a variety of parameters as descriptors of the structural properties of molecules including quantum mechanically derived electronic parameters. Other descriptors to account for the shape, size, lipophilicity, polarizability, and other structural properties have also been devised. A QSAR database at Pomona College summarizes over 6000 datasets of biological and chemical QSAR. With the advent of high performance computing this field has subsequently evolved into what is now termed rational drug design or computer-assisted drug design.

Computer-Assisted Design (CADD), also called computer-assisted molecular design (CAMD), represents sophisticated applications of computers as tools in the drug design process. CADD attempts to find a ligand that will interact favorably with a receptor that represents the target site. Binding of ligands to the receptor may include hydrophobic, electrostatic, and hydrogen-bonding interactions. In addition, solvation energies of the ligand and receptor site are important because partial to complete desolvation must occur prior to binding. This approach to CADD optimizes the fit of a ligand in a receptor site. However, optimum fit in a target site does not guarantee that the desired activity of the drug will be enhanced or that undesired side effects will be diminished. Moreover, this approach does not consider the pharmacokinetic propertiess of the drug. Ideally, one would have 3D structural information for the receptor and the ligand-receptor complex from X-ray diffraction or NMR. In the opposite extreme, one may have no experimental data to assist in building models of the ligand and receptor, in which case computational methods must be applied without the information that the experimental data provide. Based on the information available, one can apply either ligand-based or receptor-based molecular design methods. The ligand-based approach is applicable when the structure of the receptor site is unknown, but when a series of compounds have been identified that exhibit the activity of interest. Ideally, one should have structurally similar compounds with high activity, with no activity, and with a range of intermediate activities. In recognition site mapping, an attempt is made to identify a pharmacophore, which is a template derived from the structures of these compounds. It is represented as a collection of functional groups in three-dimensional space that is complementary to the geometry of the receptor site. In applying this approach, conformational analysis is required, the extent of which is dependent on the flexibility of the compounds investigated. One strategy is to find the lowest energy conformers of the most rigid compounds and superimpose them. Conformational searching on the more flexible compounds is then done while applying distance constraints derived from the structures of the more rigid compounds. Ultimately, all of the structures are superimposed to generate the so-called pharmacophore. This template may then be used to develop new compounds with functional groups in the desired positions. This strategy assumes that the

minimum energy conformers bind most favorably in the receptor site. In fact, there is no *a priori* reason to exclude higher energy conformers as the source of activity. The receptor-based approach to CADD applies when a reliable model of the receptor site is available, as from X-ray diffraction, NMR, or homology modeling. With the availability of the receptor site, the problem is to design ligands that will interact favorably at the site, which is a docking problem.

Receptor-based drug design incorporates a number of molecular modeling techniques including docking. More recent versions of DOCK allow scoring based on force fields, which include both van der Waals and electrostatic interactions [120]. These results obtained with DOCK illustrate the potential for searching objectively for ligands complementary to receptor sites. Once potential drugs have been identified by the methods described above, other molecular modeling techniques may then be applied. For example, geometry optimization may be used to "relax" the structures and to identify low energy orientations of drugs in receptor sites. MD may assist in exploring the energy landscape, and free energy simulations can be used to compute the relative binding free energies of a series of putative drugs.

Free-energy perturbation (FEP) is considered the most accurate computational method for calculating relative solvation and binding free-energy differences. An important factor limiting the use of FEP in pharmaceutical research is its low throughput, which is due in part to the dependence on accurate molecular mechanics (MM) force field parameters for individual drug candidates and the long time required to complete the process. A novel efficient method that uses quantum mechanics (QM) for treating the solute, MM for treating the solute surroundings, and the FEP method for computing free-energy differences has been developed by Reddy et al. [121]. While considerably more CPU demanding than conventional FEP methods, this method (QM/MM-based FEP) alleviates the need for development of molecule-specific MM force field parameters and therefore may lead to future automation of FEP-based calculations.

5. Pocket prediction algorithms

Numerous software packages and web-sites can be found that assist in the process of binding pocket identification on a molecular target. A comprehensive summary of this effort can be found at:
https://bioinformatictools.wordpress.com/tag/pocket-finder/
The following is a collection of the most popular pocket prediction algorithms that are publicly available:

1. SURFNET (http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html)
2. LIGSITE^{csc} (http://scoppi.biotec.tu-dresden.de/pocket/) is a extension of LIGSITE, which uses the amino acid conservation to predict the location of pockets.
3. Pocket-Finder (http://www.bioinformatics.leeds.ac.uk/pocketfinder)
4. CASTp (http://sts-fw.bioengr.uic.edu/castp/)
5. VOIDOO (http://xray.bmc.uu.se/usf/voidoo.html)
6. PocketPicker (http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/index.html)
7. APROPOS [122]
8. PASS (http://www.ccl.net/cca/software/UNIX/pass/overview.shtml)

Pocket finding methods are also reviewed by Guo et al. [119] Pocket-Finder, VOIDOO, CASTp and LIGSITEcsc, Pocket-Picker, PASS and SURFNET are all based on geometry except Qsite-Finder, which ranks the results based on their energy values. Huang and Schroeder [120] find that different pockets binding the same ligand show greater variation in their shapes than can be accounted for by the conformational variability of the ligand. This suggests that geometrical complementarity in general is not sufficient to drive molecular recognition. Nevertheless, when considering only shape and size, a significant proportion of the recognition power of a binding pocket for its ligand resides in its shape. Laurie and Jackson [121] use a geometry and energy based method to predict the location of binding pockets. They rank the results not by volumes which most of programs use but according to the sum of the interaction energy for sites within each cluster. LIGSITE^{csc}'s authors use a geometry-based method to predict pocket and re-rank the results according to a conversion function, and find a better result. The idea is to use not only geometry but also energy or conserved properties to improve the results. In [122] the authors present a binding-sites database (SitesBase) which lists known protein-ligand binding sites. They use a geometry hashing method to do an all-against-all structure comparison and stored the results in SitesBase.
It is not easy to determine which residues belong to a pocket. Distance criteria provide a simple and

popular way to define the pocket according to the ligand's position. Typically, one defines the atom on the pocket if its distance to the ligand is less than a certain value (8 Å or 4 Å have been used in some papers).

It is also important to list key properties of binding pockets, namely:

1) Depth, for example the so-called gamma depth or the L1-depths are calculated. The gamma-depth is an index of an atom that tells us how big a sphere can touch this atom from outside of the protein. L1-depth is a statistical depth function and can measure how deep the pocket is located inside a protein.

2) Hydrophobic value. There are many indexes to measure hydrophobicity. We choose Meek's values to measure it.

3) H-bond acceptor and donor numbers on the pocket's surface.

4) Solvent-accessible surface area

5) Volume

6) Conservation: In general, the functional sites of protein are more conserved. This feature is calculated by the conservation database or blastp.

Essential information regarding protein function is generally dependent on the protein's tertiary structure. This includes the enzymatic function of a protein, and also the binding of ligands, such as small molecule inhibitors [123]. Methods developed for predicting an enzymatic function of a protein by identifying catalytic residues include: finding local characteristics of functional residues [124], [125], applying known templates of active sites [126], [127] or identifying the surface shape of active sites [128]–[132].

In order to predict ligand binding we first need to determine a 3D structure of the protein in question, which can be done using several experimental or computational methods [133], [134]. Structure-based pocket prediction employs geometrical algorithms or probe mapping/docking algorithms [135]. Comparing these two kinds of methods, it can be said that the geometrical algorithms have low computational costs in contrast to the mapping/docking and scoring of molecular fragments, but the latter algorithms have a greater physical meaning. Geometrical algorithms analyze protein surfaces, and once a structure has been determined a number of algorithms may be used to predict binding pockets on the protein surface [120], [121], [136]–[139]. One such example, SURFNET [136], fits spheres into the spaces between protein atoms and finds gap regions. The results obtained this way correspond to the cavities and keys of a given protein. An algorithm based on geometric hashing called VISGRID [140] uses the visibility of constituent atoms to identify cavities. "Active site points" are identified by PASS [139]. In this method the protein surface is coated with a layer of spherical probes, then those that clash with the protein or which are not sufficiently buried are filtered out. The active site points are identified from the final probes. Another method is LIGSITE [120], [141], which is an improvement of the POCKET algorithm [142]. This algorithm puts protein-occupied space in a grid and identifies clefts by scanning areas that are enclosed on both sides by the protein's atoms. An alpha-shape algorithm is used by CAST [137] and APROPOS [138]. DRUGSITE [135] and POCKET-FINDER [143], in addition to the protein's shape, also consider physicochemical properties for identification of ligand binding pockets. Further geometrical algorithms are TRAVEL DEPTH [144], VOIDOO [145], and CAVITY SEARCH [146]. QSITEFINDER [121] uses interaction energy computation between the protein and a van der Waals probe to find favourable binding sites. Some methods using mapping/docking and scoring of molecular fragment concepts are given by Dennis et al. [147], Kortvelyesi et al. [148], Ruppert et al. [149], and Verdonk et al. [150]. There are also several docking based methods that use ligands to probe the proteins for binding sites [151]–[154].

CADD methodology often applies protein–ligand docking methods, most commonly structure-based methods. These methods provide support to the rational design and optimization of novel drug candidates [155]. Many structure-based protein–ligand docking methods have been reported in the literature [156]–[161]. These methods generally rely on first identifying a ligand-binding pocket in the protein structure.

Pocket-ligand matching methods involve some popular algorithms to match the pocket and ligand by their complementarity of shape. Two methods can be implemented relatively easily: distance distribution and spherical harmonic-based.

1. Distance distribution: compute the distances between all atoms in the pocket or in the compound, then compare the probability distribution of the distances. If the distribution is similar, their shape is similar, too. This measure is independent of the translation and rotation shifts. So computing this index is easy and fast.

2. Spherical harmonics. The theory of spherical harmonics says that any spherical function can be decomposed as the sum of its harmonics. The pockets and the compounds can be represented by a series of spherical harmonics. We can use the expansion coefficients for spherical

harmonics to represent the shape. They are independent of rotation-invariant.
These two methods mentioned above are both global shape matching algorithms.

In view of biomedicine, the effectiveness of a drug in treating a disease hinges on the fact that the drug compound matches a functional pocket of the proteins that cause this disease. There are many algorithms for predicting the pockets, and all of these algorithms use the simplification that the largest pocket is determined as the functional pocket of a given protein. Pockets, geometric cavities and depressions in protein surfaces and structures have been identified as features of many functionally important sites on proteins. Many algorithms to find pockets, given a protein structure, have been developed. For simplicity, we assume that although each protein may have many pockets, one and only one is of functional importance. Thus, determining the functional pocket from a set of identified pockets is an important step. Certainly, the simplification that the largest pocket is the functional pocket is popular with existing algorithms, and an observed rate of correct prediction for this simple rule is 75% (LIGSITE^{csc}), 67% (LIGSITE^{cs}) 65% (LIGSITE), 54% (PASS), and 42% (SURFNET). These rates do not seem too bad; implying that indices of the pocket size (volume, surface area, etc.) are generally very important. However, by extending the set of traits used the determination accuracy can likely be improved. In general, these scores are lower than what might be obtained from a smaller, less general benchmark set. For example, SURFNET scores 83% on its original benchmark of 67 proteins. It is clear that these may not be representative of proteins in general and that a bigger, representative benchmark is needed to confidently validate accuracies. Since a shape related property, i.e. size, is known to be a factor for the match between protein and compound, we use a geometrical-based method. After this large benchmark set is built, we study the chemical and physical properties of each pocket, with the intent of finding the most important traits determining the functional pocket.
There are two databases used most frequently in the computational drug design and discovery research. The Protein Data Bank (PDB) [5] provides 3D protein structures for input into the pocket finding algorithms. The PDBbind database contains the structure of complexes from PDB. This dataset is used to estimate whether the pockets we identify are functional pockets, and as a training set to improve the prediction.
PDBbind: To validate the usefulness of the public prediction algorithms for functional pockets, a common database with a large size is needed. For example, the validation of DOCK in [162] was based on a set of 114 complexes, while the validation of LIGSITEcsc in [120] was based on a set of 48 pairs of bound/unbound structures. Among existing datasets, the "refined set" of PDBbind [163] is a good choice for a benchmark dataset due to its large size and its manually verification. The PDBbind database may be found at: http://www.pdbbind.org. The current version of the PDBbind database (version 2007) has 3124 protein-ligand complexes. 1300 of these have been manually selected to form the "refined set", with focus on the quality of structures and binding data. Further reduced from this is the "core set" of 70 triplets (210 complexes) of related proteins. The PDBbind-refined and PDBbind-core was used as the training sets to select the set of traits of pockets. The previously discussed pocket prediction algorithms, namely: SURFNET [136], LIGSITE^{csc}[120], Q-SiteFinder and Pocket-Finder [121], CAST [137], VOIDOO [145], PocketPicker [113], APROPOS [138], PASS [139] and ProShape were divided into two classes: geometry based and physico-chemical based methods. The geometry based methods can be sub-divided into four types of algorithms as follows:
POCKET, LIGSITE, and LIGSITE^{csc}. These three algorithms scan a grid for protein-solvent and surface-solvent features. POCKET employs 3 directions, LIGSITE and LIGSITE^{csc}, 7 directions; also, POCKET and LIGSITE generate the grid data directly from atom coordinates while LIGSITE^{csc} first generates the Connolly surface. SURFNET places a sphere, not containing any atoms, between two atoms. The spheres with maximal volume define the largest pocket. CAST triangulates the surface atoms and clusters triangles by merging small triangles with neighboring large triangles. PASS coats the protein with probe spheres, selects probes with many atom contacts, and then repeats coating until no new probes are kept. The pockets, or active site points, are the probes with large number of atom contacts. "POCKET" (a subprogram of ProShape software) may also be used to obtain the pockets. It uses an alpha-shape-based method similar to CAST. Benefits to using this method are as follows:

1. "POCKET" is a very efficient program. It is very fast in detecting pockets based on a fast alpha shape algorithm. Calculation of a PDB entry with 1878 atoms only takes 0.2 sec.
2. "POCKET" offers four outputs: the pockets (atoms lining these pocket are listed), the mouths (atoms of these mouths are listed) and the surface areas (including both the total areas of pockets and mouths

and the fractional contribution from each atom).

Determination of the functional pocket is essentially a classification problem. There are many classification algorithms in machine learning, such as SVM, etc. A number of methods were tested using Weka (Waikato Environment for Knowledge Analysis), [164] with the random decision forest method identified as the best. Classification trees (or decision) trees are widely used in classification and prediction. Moving upwards from the trunk of the tree, one encounters a series of forks were each branch represents particular combinations of features and traits in the data. Ultimately, at the end of each branch is a single "leaf" or classification. The random decision forest method grows many classification trees each trained using a different random subset of the training data. In use, each tree in the forest analyzes an input vector and giving a classification, and the tree "votes" for that classification. The classification having the most votes (over all the trees in the forest) is then chosen. The advantages of the random decision forest are: (a) Relatively low error (perhaps the lowest of any method), (b) No over-fitting
(c) Elegant handling of missing values, (d) Only partially black-box (e.g., results include variable importance, outlier detection), (e) Can be used for supervised and unsupervised learning problems.

6. Conclusions

This review paper 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 bio-molecular target characterization for drug action. We have also given practical information regarding identification of binding pockets for putative inhibitors of proteins and enzymes. Lists of databases, websites and publicly available software packages used in all stages of computational drug design have been presented in this paper 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] T. Zhu, H. Lee, H. Lei, C. Jones, K. Patel, M. E. Johnson, and K. E. Hevener, "Fragment-based drug discovery using a multidomain, parallel MD-MM/PBSA screening protocol," *J. Chem. Inf. Model.*, vol. 53, no. 3, pp. 560–572, Mar. 2013.

[2] M. Fujihashi, T. Ishida, S. Kuroda, L. P. Kotra, E. F. Pai, and K. Miki, "Substrate distortion contributes to the catalysis of orotidine 5'-monophosphate decarboxylase," *J. Am. Chem. Soc.*, vol. 135, no. 46, pp. 17432–17443, Nov. 2013.

[3] G. Tiwari and D. Mohanty, "An in silico analysis of the binding modes and binding affinities of small molecule modulators of PDZ-peptide interactions," *PloS One*, vol. 8, no. 8, p. e71340, 2013.

[4] Y. Fukunishi and H. Nakamura, "Improved estimation of protein-ligand binding free energy by using the ligand-entropy and mobility of water molecules," *Pharm. Basel Switz.*, vol. 6, no. 5, pp. 604–622, 2013.

[5] D. Zahn, "Molecular dynamics simulation of ionic conductors: perspectives and limitations," *J. Mol. Model.*, vol. 17, no. 7, pp. 1531–1535, Jul. 2011.

[6] N. T. Wood, E. Fadda, R. Davis, O. C. Grant, J. C. Martin, R. J. Woods, and S. A. Travers, "The influence of N-linked glycans on the molecular dynamics of the HIV-1 gp120 V3 loop," *PloS One*, vol. 8, no. 11, p. e80301, 2013.

[7] A. Chaudhuri, I. Sarkar, and S. Chakraborty, "Comparative analysis of binding sites of human meprins with hydroxamic acid derivative by molecular dynamics simulation study," *J. Biomol. Struct. Dyn.*, Nov. 2013.

[8] G. Leonis, T. Steinbrecher, and M. G. Papadopoulos, "A contribution to the drug resistance mechanism of darunavir, amprenavir, indinavir, and saquinavir complexes with HIV-1 protease due to flap mutation I50V: a systematic MM-PBSA and thermodynamic integration study," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 2141–2153, Aug. 2013.

[9] T. Yoda, Y. Sugita, and Y. Okamoto, "Hydrophobic core formation and dehydration in protein folding studied by generalized-ensemble simulations," *Biophys. J.*, vol. 99, no. 5, pp. 1637–1644, Sep. 2010.

[10] T. Yoda, Y. Sugita, and Y. Okamoto, "Salt effects on hydrophobic-core formation in folding of a helical miniprotein studied by molecular dynamics simulations," *Proteins*, Nov. 2013.

[11] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *J. Mol. Graph.*, vol. 8, no. 1, pp. 52–56, Mar. 1990.

[12] R. A. Engh and R. Huber, "Accurate bond and angle parameters for X-ray protein structure refinement," *Acta Crystallogr. A*, vol. 47, no. 4, pp. 392–400, Jul. 1991.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

[14] H. Lodish, A. Berk, S. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Molecular Cell Biology," in *Molecular Cell Biology*, Fourth., W. H. Freeman, p. Section 15.4.

[15] 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, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, May 1995.

[16] T. Arakawa, N. Kamiya, H. Nakamura, and I. Fukuda, "Molecular dynamics simulations of double-stranded DNA in an explicit solvent model with the zero-dipole summation method," *PloS One*, vol. 8, no. 10, p. e76606, 2013.

[17] Y. Liu, E. Haddadian, T. R. Sosnick, K. F. Freed, and H. Gong, "A novel implicit solvent model for simulating the molecular dynamics of RNA," *Biophys. J.*, vol. 105, no. 5, pp. 1248–1257, Sep. 2013.

[18] R. Elber, A. P. Ruymgaart, and B. Hess, "SHAKE parallelization," *Eur. Phys. J. Spec. Top.*, vol. 200, no. 1, pp. 211–223, Nov. 2011.

[19] X. Ye, Q. Cai, W. Yang, and R. Luo, "Roles of boundary conditions in DNA simulations: analysis of ion distributions with the finite-difference Poisson-Boltzmann method," *Biophys. J.*, vol. 97, no. 2, pp. 554–562, Jul. 2009.

[20] T. Darden, L. Perera, L. Li, and L. Pedersen, "New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations," *Struct. Lond. Engl. 1993*, vol. 7, no. 3, pp. R55–60, Mar. 1999.

[21] M. C. Patra, S. N. Rath, S. K. Pradhan, J. Maharana, and S. De, "Molecular dynamics simulation of human serum paraoxonase 1 in DPPC bilayer reveals a critical role of transmembrane helix H1 for HDL association," *Eur. Biophys. J. EBJ*, vol. 43, no. 1, pp. 35–51, Jan. 2014.

[22] R. C. Harris, A. H. Boschitsch, and M. O. Fenley, "Influence of grid spacing in Poisson-Boltzmann equation binding energy estimation," *J. Chem. Theory Comput.*, vol. 9, no. 8, pp. 3677–3685, Aug. 2013.

[23] M. R. Reddy, C. R. Reddy, R. S. Rathore, M. D. Erion, P. Aparoy, R. N. Reddy, and P. Reddanna, "Free Energy Calculations to Estimate Ligand-Binding Affinities in Structure-Based Drug Design," *Curr. Pharm. Des.*, Aug. 2013.

[24] M. Adler and P. Beroza, "Improved ligand binding energies derived from molecular dynamics: replicate sampling enhances the search of conformational space," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 2065–2072, Aug. 2013.
[25] P. Kar, R. Lipowsky, and V. Knecht, "Importance of polar solvation and configurational entropy for design of antiretroviral drugs targeting HIV-1 protease," *J. Phys. Chem. B*, vol. 117, no. 19, pp. 5793–5805, May 2013.
[26] J. M. Sanders, M. E. Wampole, M. L. Thakur, and E. Wickstrom, "Molecular determinants of epidermal growth factor binding: a molecular dynamics study," *PloS One*, vol. 8, no. 1, p. e54136, 2013.
[27] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 20, pp. 12562–12566, Oct. 2002.
[28] F. Osterberg and J. Aqvist, "Exploring blocker binding to a homology model of the open hERG K+ channel using docking and molecular dynamics methods," *FEBS Lett.*, vol. 579, no. 13, pp. 2939–2944, May 2005.
[29] E. Jenwitheesuk and R. Samudrala, "Virtual screening of HIV-1 protease inhibitors against human cytomegalovirus protease using docking and molecular dynamics," *AIDS Lond. Engl.*, vol. 19, no. 5, pp. 529–531, Mar. 2005.
[30] R. Tatsumi, Y. Fukunishi, and H. Nakamura, "A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor," *J. Comput. Chem.*, vol. 25, no. 16, pp. 1995–2005, Dec. 2004.
[31] S. Hu, H. Yu, Y. Liu, T. Xue, and H. Zhang, "Insight into the binding model of new antagonists of kappa receptor using docking and molecular dynamics simulation," *J. Mol. Model.*, vol. 19, no. 8, pp. 3087–3094, Aug. 2013.
[32] W. Hu, S. Deng, J. Huang, Y. Lu, X. Le, and W. Zheng, "Intercalative interaction of asymmetric copper(II) complex with DNA: experimental, molecular docking, molecular dynamics and TDDFT studies," *J. Inorg. Biochem.*, vol. 127, pp. 90–98, Oct. 2013.
[33] X. Huang, G. Zheng, and C.-G. Zhan, "Microscopic binding of M5 muscarinic acetylcholine receptor with antagonists by homology modeling, molecular docking, and molecular dynamics simulation," *J. Phys. Chem. B*, vol. 116, no. 1, pp. 532–541, Jan. 2012.
[34] N. Moitessier, C. Henry, B. Maigret, and Y. Chapleur, "Combining pharmacophore search, automated docking, and molecular dynamics simulations as a novel strategy for flexible docking. Proof of concept: docking of arginine−glycine−aspartic acid-like compounds into the αvβ3 binding site," *J. Med. Chem.*, vol. 47, no. 17, pp. 4178–4187, Aug. 2004.
[35] A. Tarcsay, G. Paragi, M. Vass, B. Jójárt, F. Bogár, and G. M. Keserű, "The impact of molecular dynamics sampling on the performance of virtual screening against GPCRs," *J. Chem. Inf. Model.*, vol. 53, no. 11, pp. 2990–2999, Nov. 2013.
[36] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters.," *Proteins*, vol. 65, no. 3, pp. 712–25, Nov. 2006.
[37] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *J. Comput. Chem.*, vol. 21, no. 12, pp. 1049–1074, Sep. 2000.
[38] L. Wickstrom, A. Okur, and C. Simmerling, "Evaluating the performance of the ff99SB force field based on NMR scalar coupling data," *Biophys. J.*, vol. 97, no. 3, pp. 853–856, Aug. 2009.
[39] E. A. Cino, W.-Y. Choy, and M. Karttunen, "Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations," *J. Chem. Theory Comput.*, vol. 8, no. 8, pp. 2725–2740, Aug. 2012.
[40] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, T. . Cerutti, I. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Anderson, I. Kolossváry, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman, "AMBER 14." University of California, San Francisco., 2014.
[41] Y. Zhang and C. Sagui, "The gp41(659-671) HIV-1 antibody epitope: a structurally challenging small peptide," *J. Phys. Chem. B*, vol. 118, no. 1, pp. 69–80, Jan. 2014.
[42] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *J. Comput. Chem.*, vol. 24, no. 16, pp. 1999–2012, Dec. 2003.
[43] M. C. Lee and Y. Duan, "Distinguish protein decoys by Using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model," *Proteins Struct. Funct. Bioinforma.*, vol. 55, no. 3, pp. 620–634, May 2004.

[44] L. Yang, C.-H. Tan, M.-J. Hsieh, J. Wang, Y. Duan, P. Cieplak, J. Caldwell, P. A. Kollman, and R. Luo, "New-generation amber united-atom force field," *J. Phys. Chem. B*, vol. 110, no. 26, pp. 13166–13176, Jul. 2006.
[45] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004.
[46] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, Mar. 1988.
[47] G. Kaminski and W. L. Jorgensen, "Performance of the AMBER94, MMFF94, and OPLS-AA force fields for modeling organic liquids," *J. Phys. Chem.*, vol. 100, no. 46, pp. 18010–18013, Jan. 1996.
[48] A. D. MacKerell, J. Wiórkiewicz-Kuczera, and M. Karplus, "CHARMM22 Parameter Set," *Harv. Univ. Dep. Chem. Camb. MA*, 1995.
[49] M. Feig, Alexander D. MacKerell, and C. L. Brooks, "Force field influence on the observation of π-helical protein structures in molecular dynamics simulations," *J. Phys. Chem. B*, vol. 107, no. 12, pp. 2831–2836, Mar. 2003.
[50] A. D. MacKerell, N. Banavali, and N. Foloppe, "Development and current status of the CHARMM force field for nucleic acids," *Biopolymers*, vol. 56, no. 4, pp. 257–265, Jan. 2000.
[51] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," *J. Am. Chem. Soc.*, vol. 118, no. 45, pp. 11225–11236, Jan. 1996.
[52] D. S. Maxwell, J. Tirado-Rives, and W. L. Jorgensen, "A comprehensive study of the rotational energy profiles of organic systems by ab initio MO theory, forming a basis for peptide torsional parameters," *J. Comput. Chem.*, vol. 16, no. 8, pp. 984–1010, Aug. 1995.
[53] K. Kahn and T. C. Bruice, "Parameterization of OPLS-AA force field for the conformational analysis of macrocyclic polyketides," *J. Comput. Chem.*, vol. 23, no. 10, pp. 977–996, Jul. 2002.
[54] S. W. I. Siu, K. Pluhackova, and R. A. Böckmann, "Optimization of the OPLS-AA force field for long hydrocarbons," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1459–1470, Apr. 2012.
[55] X. Daura, A. E. Mark, and W. F. Van Gunsteren, "Parametrization of aliphatic CHn united atoms of GROMOS96 force field," *J. Comput. Chem.*, vol. 19, no. 5, pp. 535–547, Apr. 1998.
[56] L. D. Schuler, X. Daura, and W. F. van Gunsteren, "An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase," *J. Comput. Chem.*, vol. 22, no. 11, pp. 1205–1218, Aug. 2001.
[57] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1656–1676, Oct. 2004.
[58] M. M. Reif, M. Winger, and C. Oostenbrink, "Testing of the GROMOS force-field parameter set 54A8: structural properties of electrolyte solutions, lipid bilayers, and proteins," *J. Chem. Theory Comput.*, vol. 9, no. 2, pp. 1247–1264, Feb. 2013.
[59] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7," *Eur. Biophys. J. EBJ*, vol. 40, no. 7, pp. 843–856, Jul. 2011.
[60] H. Freedman, "Solvation Free Energies from a Coupled Reference Interaction Site Model/simulation Approach (Thesis)," Department of Chemistry, University of Utah, 2005.
[61] D. M. Chipman, "Vertical electronic excitation with a dielectric continuum model of solvation including volume polarization. I. Theory," *J. Chem. Phys.*, vol. 131, no. 1, p. 014103, 2009.
[62] T. Luchko, S. Gusarov, D. R. Roe, C. Simmerling, D. A. Case, J. Tuszynski, and A. Kovalenko, "Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber," *J. Chem. Theory Comput.*, vol. 6, no. 3, pp. 607–624, Mar. 2010.
[63] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the Amber biomolecular simulation package," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 3, no. 2, pp. 198–210, Mar. 2013.
[64] M. Orozco and F. J. Luque, "Theoretical methods for the description of the solvent effect in biomolecular systems," *Chem. Rev.*, vol. 100, no. 11, pp. 4187–4226, Nov. 2000.
[65] B. I. Simkin, *Quantum chemical and statistical theory of solutions: a computational approach*. London ; New York: Ellis Horwood, 1995.
[66] S. Höfinger, "Solving the Poisson-Boltzmann equation with the specialized computer chip MD-GRAPE-2," *J. Comput. Chem.*, vol. 26, no. 11, pp. 1148–1154, Aug. 2005.

[67] D. Chandler, J. D. McCoy, and S. J. Singer, "Density functional theory of nonuniform polyatomic systems. I. General formulation," *J. Chem. Phys.*, vol. 85, no. 10, p. 5971, 1986.
[68] D. Chandler, J. D. McCoy, and S. J. Singer, "Density functional theory of nonuniform polyatomic systems. II. Rational closures for integral equations," *J. Chem. Phys.*, vol. 85, no. 10, p. 5977, 1986.
[69] D. Beglov and B. Roux, "An integral equation to describe the solvation of polar molecules in liquid water," *J. Phys. Chem. B*, vol. 101, no. 39, pp. 7821–7826, Sep. 1997.
[70] A. Kovalenko and F. Hirata, "Potential of mean force between two molecular ions in a polar molecular solvent: a study by the three-dimensional Reference Interaction Site Model," *J. Phys. Chem. B*, vol. 103, no. 37, pp. 7942–7957, Sep. 1999.
[71] A. Kovalenko and F. Hirata, "Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model," *J. Chem. Phys.*, vol. 110, no. 20, p. 10095, 1999.
[72] A. Kovalenko and F. Hirata, "Potentials of mean force of simple ions in ambient aqueous solution. I. Three-dimensional reference interaction site model approach," *J. Chem. Phys.*, vol. 112, no. 23, p. 10391, 2000.
[73] A. Kovalenko and F. Hirata, "Potentials of mean force of simple ions in ambient aqueous solution. II. Solvation structure from the three-dimensional reference interaction site model approach, and comparison with simulations," *J. Chem. Phys.*, vol. 112, no. 23, p. 10403, 2000.
[74] H. C. Andersen, "Optimized cluster expansions for classical fluids. I. General theory and variational formulation of the mean spherical model and hard sphere Percus-Yevick equations," *J. Chem. Phys.*, vol. 57, no. 5, p. 1918, 1972.
[75] P. J. Rossky and H. L. Friedman, "Accurate solutions to integral equations describing weakly screened ionic systems," *J. Chem. Phys.*, vol. 72, no. 10, p. 5694, 1980.
[76] F. Hirata and P. J. Rossky, "An extended rism equation for molecular polar fluids," *Chem. Phys. Lett.*, vol. 83, no. 2, pp. 329–334, Oct. 1981.
[77] F. Hirata, "Application of an extended RISM equation to dipolar and quadrupolar fluids," *J. Chem. Phys.*, vol. 77, no. 1, p. 509, 1982.
[78] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids with Applications to Soft Matter*. Burlington: Elsevier Science, 2013.
[79] F. Hirata, Ed., *Molecular theory of solvation*. Dordrecht ; Boston: Kluwer Academic Publishers, 2003.
[80] M. E. Tuckerman, B. J. Berne, and G. J. Martyna, "Molecular dynamics algorithm for multiple time scales: Systems with long range forces," *J. Chem. Phys.*, vol. 94, no. 10, p. 6811, 1991.
[81] M. Tuckerman, B. J. Berne, and G. J. Martyna, "Reversible multiple time scale molecular dynamics," *J. Chem. Phys.*, vol. 97, no. 3, p. 1990, 1992.
[82] I. Omelyan and A. Kovalenko, "Generalised canonical–isokinetic ensemble: speeding up multiscale molecular dynamics and coupling with 3D molecular theory of solvation," *Mol. Simul.*, vol. 39, no. 1, pp. 25–48, Jan. 2013.
[83] J.-F. Truchon, B. M. Pettitt, and P. Labute, "A cavity corrected 3D-RISM functional for accurate solvation free energies," *J. Chem. Theory Comput.*, vol. 10, no. 3, pp. 934–941, Mar. 2014.
[84] K.-C. Ng, "Hypernetted chain solutions for the classical one-component plasma up to Γ=7000," *J. Chem. Phys.*, vol. 61, no. 7, p. 2680, 1974.
[85] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *J. ACM*, vol. 12, no. 4, pp. 547–560, Oct. 1965.
[86] Y. Maruyama and F. Hirata, "Modified Anderson method for accelerating 3D-RISM calculations using graphics processing unit," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3015–3021, Sep. 2012.
[87] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, Nov. 1987.
[88] J. W. Caldwell and P. A. Kollman, "Structure and properties of neat liquids using nonadditive molecular dynamics: water, methanol, and N-methylacetamide," *J. Phys. Chem.*, vol. 99, no. 16, pp. 6208–6219, Apr. 1995.
[89] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, p. 926, 1983.
[90] D. J. Price and C. L. Brooks, "A modified TIP3P water potential for simulation with Ewald summation," *J. Chem. Phys.*, vol. 121, no. 20, p. 10096, 2004.
[91] W. L. Jorgensen and J. D. Madura, "Temperature and size dependence for Monte Carlo simulations of TIP4P water," *Mol. Phys.*, vol. 56, no. 6, pp. 1381–1392, Dec. 1985.
[92] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew," *J. Chem. Phys.*, vol. 120, no. 20, p. 9665, 2004.
[93] M. W. Mahoney and W. L. Jorgensen, "A five-site model for liquid water and the

reproduction of the density anomaly by rigid, nonpolarizable potential functions," *J. Chem. Phys.*, vol. 112, no. 20, p. 8910, 2000.
[94] K. Toukan and A. Rahman, "Molecular-dynamics study of atomic motions in water," *Phys. Rev. B*, vol. 31, no. 5, pp. 2643–2648, Mar. 1985.
[95] T. Honma, "Recent advances in de novo design strategy for practical lead identification," *Med. Res. Rev.*, vol. 23, no. 5, pp. 606–632, Sep. 2003.
[96] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," *Nat. Rev. Drug Discov.*, vol. 4, no. 8, pp. 649–663, Aug. 2005.
[97] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009.
[98] D. T.-H. Chang, Y.-J. Oyang, and J.-H. Lin, "MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm," *Nucleic Acids Res.*, vol. 33, no. suppl 2, pp. W233–W238, Jul. 2005.
[99] S. M. Saberi Fathi, D. T. White, and J. A. Tuszynski, "Geometrical comparison of two protein structures using Wigner-D functions: Geometrical Comparison of Protein Structures," *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. 10, pp. 2756–2769, Oct. 2014.
[100] R. J. Morris, R. J. Najmanovich, A. Kahraman, and J. M. Thornton, "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons," *Bioinformatics*, vol. 21, no. 10, pp. 2347–2355, May 2005.
[101] Y. Kalidas and N. Chandra, "PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins," *J. Struct. Biol.*, vol. 161, no. 1, pp. 31–42, Jan. 2008.
[102] L. J. Latecki, V. Megalooikonomou, Q. Wang, and D. Yu, "An elastic partial shape matching technique," *Pattern Recognit.*, vol. 40, no. 11, pp. 3069–3080, Nov. 2007.
[103] A. C. Good, "Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: Adding flexibility to the search for ligand kin," *J. Mol. Graph. Model.*, vol. 26, no. 3, pp. 656–666, Oct. 2007.
[104] P. K. Agarwal, N. H. Mustafa, and Y. Wang, "Fast Molecular Shape Matching Using Contact Maps," *J. Comput. Biol.*, vol. 14, no. 2, pp. 131–143, Mar. 2007.
[105] P. Shilane and T. Funkhouser, "Distinctive regions of 3D surfaces," *ACM Trans. Graph.*, vol. 26, no. 2, p. 7–es, Jun. 2007.
[106] M. E. Bock, C. Garutti, and C. Guerra, "Discovery of Similar Regions on Protein Surfaces," *J. Comput. Biol.*, vol. 14, no. 3, pp. 285–299, Apr. 2007.
[107] A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton, "Shape variation in protein binding pockets and their ligands," *J. Mol. Biol.*, vol. 368, no. 1, pp. 283–301, Apr. 2007.
[108] L. Xie and P. E. Bourne, "A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites," *BMC Bioinformatics*, vol. 8, no. Suppl 4, p. S9, 2007.
[109] M. Weisel, E. Proschak, and G. Schneider, "PocketPicker: analysis of ligand binding-sites with shape descriptors," *Chem. Cent. J.*, vol. 1, no. 1, p. 7, 2007.
[110] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, and H. A. Carlson, "Binding MOAD, a high-quality protein ligand database," *Nucleic Acids Res.*, vol. 36, no. Database, pp. D674–D678, Dec. 2007.
[111] S. Borman, "New QSAR Techniques Eyed For Environmental Assessments: Expert system, spectroscopy method use readily available data to develop quantitative structure-activity relationships for broad compound classes," *Chem. Eng. News*, vol. 68, no. 8, pp. 20–23, Feb. 1990.
[112] R. L. Lipnick, "Charles Ernest Overton: narcosis studies and a contribution to general pharmacology," *Trends Pharmacol. Sci.*, vol. 7, pp. 161–164, Jan. 1986.
[113] C. Hansch, A. Leo, and R. W. Taft, "A survey of Hammett substituent constants and resonance and field parameters," *Chem. Rev.*, vol. 91, no. 2, pp. 165–195, Mar. 1991.
[114] C. Hansch, A. Leo, and D. H. Hoekman, Eds., *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*. Washington, DC: American Chemical Society, 1995.
[115] C. Hansch, "Quantitative approach to biochemical structure-activity relationships," *Acc. Chem. Res.*, vol. 2, no. 8, pp. 232–239, Aug. 1969.
[116] E. C. Meng, B. K. Shoichet, and I. D. Kuntz, "Automated docking with grid-based energy evaluation," *J. Comput. Chem.*, vol. 13, no. 4, pp. 505–524, May 1992.
[117] M. R. Reddy, U. C. Singh, and M. D. Erion, "Development of a Quantum Mechanics-Based Free-Energy Perturbation Method: Use in the Calculation of Relative Solvation Free Energies," *J. Am. Chem. Soc.*, vol. 126, no. 20, pp. 6224–6225, May 2004.
[118] K. P. Peters, J. Fauck, and C. Frömmel, "The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure

Using only Geometric Criteria," *J. Mol. Biol.*, vol. 256, no. 1, pp. 201–213, Feb. 1996.
[119] Z. Guo, B. Li, L.-T. Cheng, S. Zhou, J. A. McCammon, and J. Che, "Identification of Protein-Ligand Binding Sites by Level-Set Variational Implicit Solvent Approach," *J Chem Theory Comput*, Submitted.
[120] B. Huang and M. Schroeder, "LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation," *BMC Struct. Biol.*, vol. 6, p. 19, 2006.
[121] A. T. R. Laurie and R. M. Jackson, "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites," *Bioinforma. Oxf. Engl.*, vol. 21, no. 9, pp. 1908–1916, May 2005.
[122] N. D. Gold and R. M. Jackson, "Fold Independent Structural Comparisons of Protein–Ligand Binding Sites for Exploring Functional Relationships," *J. Mol. Biol.*, vol. 355, no. 5, pp. 1112–1124, Feb. 2006.
[123] L. Polgár, "The catalytic triad of serine peptidases," *Cell. Mol. Life Sci. CMLS*, vol. 62, no. 19–20, pp. 2161–2172, Oct. 2005.
[124] S. D. Mooney, M. H.-P. Liang, R. DeConde, and R. B. Altman, "Structural characterization of proteins using residue environments," *Proteins*, vol. 61, no. 4, pp. 741–747, Dec. 2005.
[125] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "Recognition of functional sites in protein structures," *J. Mol. Biol.*, vol. 339, no. 3, pp. 607–633, Jun. 2004.
[126] J. S. Fetrow, A. Godzik, and J. Skolnick, "Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity," *J. Mol. Biol.*, vol. 282, no. 4, pp. 703–711, Oct. 1998.
[127] A. C. Wallace, N. Borkakoti, and J. M. Thornton, "TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites," *Protein Sci. Publ. Protein Soc.*, vol. 6, no. 11, pp. 2308–2323, Nov. 1997.
[128] M. L. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.
[129] B. B. Goldman and W. T. Wipke, "QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock)," *Proteins*, vol. 38, no. 1, pp. 79–94, Jan. 2000.
[130] B. S. Duncan and A. J. Olson, "Approximation and characterization of molecular surfaces," *Biopolymers*, vol. 33, no. 2, pp. 219–229, Feb. 1993.
[131] T. E. Exner, M. Keil, and J. Brickmann, "Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory," *J. Comput. Chem.*, vol. 23, no. 12, pp. 1176–1187, Sep. 2002.
[132] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database eF-site," *Protein Sci. Publ. Protein Soc.*, vol. 12, no. 8, pp. 1589–1595, Aug. 2003.
[133] B. Rupp and J. Wang, "Predictive models for protein crystallization," *Methods San Diego Calif*, vol. 34, no. 3, pp. 390–407, Nov. 2004.
[134] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinforma. Oxf. Engl.*, vol. 22, no. 2, pp. 195–201, Jan. 2006.
[135] J. An, M. Totrov, and R. Abagyan, "Comprehensive identification of 'druggable' protein ligand binding sites," *Genome Inform. Int. Conf. Genome Inform.*, vol. 15, no. 2, pp. 31–41, 2004.
[136] R. A. Laskowski, "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions," *J. Mol. Graph.*, vol. 13, no. 5, pp. 323–330, 307–308, Oct. 1995.
[137] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design," *Protein Sci. Publ. Protein Soc.*, vol. 7, no. 9, pp. 1884–1897, Sep. 1998.
[138] K. P. Peters, J. Fauck, and C. Frömmel, "The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria," *J. Mol. Biol.*, vol. 256, no. 1, pp. 201–213, Feb. 1996.
[139] G. P. Brady Jr and P. F. Stouten, "Fast prediction and visualization of protein binding pockets with PASS," *J. Comput. Aided Mol. Des.*, vol. 14, no. 4, pp. 383–401, May 2000.
[140] B. Li, S. Turuvekere, M. Agrawal, D. La, K. Ramani, and D. Kihara, "Characterization of local geometry of protein surfaces with the visibility criterion," *Proteins*, vol. 71, no. 2, pp. 670–683, May 2008.
[141] M. Hendlich, F. Rippmann, and G. Barnickel, "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins," *J. Mol. Graph. Model.*, vol. 15, no. 6, pp. 359–363, 389, Dec. 1997.
[142] D. G. Levitt and L. J. Banaszak, "POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids," *J. Mol. Graph.*, vol. 10, no. 4, pp. 229–234, Dec. 1992.

[143] J. An, M. Totrov, and R. Abagyan, "Pocketome via comprehensive identification and classification of ligand binding envelopes," *Mol. Cell. Proteomics MCP*, vol. 4, no. 6, pp. 752–761, Jun. 2005.
[144] R. G. Coleman and K. A. Sharp, "Travel depth, a new shape descriptor for macromolecules: application to ligand binding," *J. Mol. Biol.*, vol. 362, no. 3, pp. 441–458, Sep. 2006.
[145] G. J. Kleywegt and T. A. Jones, "Detection, delineation, measurement and display of cavities in macromolecular structures," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 50, no. Pt 2, pp. 178–185, Mar. 1994.
[146] C. M. Ho and G. R. Marshall, "Cavity search: an algorithm for the isolation and display of cavity-like binding regions," *J. Comput. Aided Mol. Des.*, vol. 4, no. 4, pp. 337–354, Dec. 1990.
[147] S. Dennis, T. Kortvelyesi, and S. Vajda, "Computational mapping identifies the binding sites of organic solvents on proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 7, pp. 4290–4295, Apr. 2002.
[148] T. Kortvelyesi, M. Silberstein, S. Dennis, and S. Vajda, "Improved mapping of protein binding sites," *J. Comput. Aided Mol. Des.*, vol. 17, no. 2–4, pp. 173–186, Apr. 2003.
[149] J. Ruppert, W. Welch, and A. N. Jain, "Automatic identification and representation of protein binding sites for molecular docking," *Protein Sci. Publ. Protein Soc.*, vol. 6, no. 3, pp. 524–533, Mar. 1997.
[150] M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet, and P. Willett, "SuperStar: improved knowledge-based interaction fields for protein binding sites," *J. Mol. Biol.*, vol. 307, no. 3, pp. 841–859, Mar. 2001.
[151] A. A. Bliznyuk and J. E. Gready, "Simple method for locating possible ligand binding sites on protein surfaces," *J. Comput. Chem.*, vol. 20, no. 9, pp. 983–988, 1999.
[152] S. J. Campbell, N. D. Gold, R. M. Jackson, and D. R. Westhead, "Ligand binding: functional site location, similarity and docking," *Curr. Opin. Struct. Biol.*, vol. 13, no. 3, pp. 389–395, Jun. 2003.
[153] M. Glick, D. D. Robinson, G. H. Grant, and W. G. Richards, "Identification of ligand binding sites on proteins using a multi-scale approach," *J. Am. Chem. Soc.*, vol. 124, no. 10, pp. 2337–2344, Mar. 2002.
[154] C. Sotriffer and G. Klebe, "Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design," *Farm. Soc. Chim. Ital. 1989*, vol. 57, no. 3, pp. 243–251, Mar. 2002.
[155] A. D. Andricopulo, L. B. Salum, and D. J. Abraham, "Structure-based drug design strategies in medicinal chemistry," *Curr. Top. Med. Chem.*, vol. 9, no. 9, pp. 771–790, 2009.
[156] B. Waszkowycz, D. E. Clark, and E. Gancia, "Outstanding challenges in protein-ligand docking and structure-based virtual screening," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 1, no. 2, pp. 229–259, Mar. 2011.
[157] C. N. Cavasotto and A. J. W. Orry, "Ligand docking and structure-based virtual screening in drug discovery," *Curr. Top. Med. Chem.*, vol. 7, no. 10, pp. 1006–1014, 2007.
[158] R. L. DesJarlais, M. D. Cummings, and A. C. Gibbs, "Virtual Docking:How Are We Doing And How Can We Improve?," *Front. Drug Des. Discov. Struct.-Based Drug Des. 21st Century*, vol. 3, no. 1, pp. 81–103, 2007.
[159] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *Br. J. Pharmacol.*, vol. 153 Suppl 1, pp. S7–26, Mar. 2008.
[160] M. Kontoyianni, P. Madhav, E. Suchanek, and W. Seibel, "Theoretical and practical considerations in virtual screening: a beaten field?," *Curr. Med. Chem.*, vol. 15, no. 2, pp. 107–116, 2008.
[161] T. Tuccinardi, "Docking-based virtual screening: recent developments," *Comb. Chem. High Throughput Screen.*, vol. 12, no. 3, pp. 303–314, Mar. 2009.
[162] D. T. Moustakas, P. T. Lang, S. Pegg, E. Pettersen, I. D. Kuntz, N. Brooijmans, and R. C. Rizzo, "Development and validation of a modular, extensible docking program: DOCK 5," *J. Comput. Aided Mol. Des.*, vol. 20, no. 10–11, pp. 601–619, Dec. 2006.
[163] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, "The PDBbind Database: Methodologies and Updates," *J. Med. Chem.*, vol. 48, no. 12, pp. 4111–4119, Jun. 2005.
[164] I. H. Witten, *Data mining: practical machine learning tools and techniques*, 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman, 2005.

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html