NIH Public Access
Author Manuscript
Medchemcomm. Author manuscript; available in PMC 2012 May 1.

Published in final edited form as:
Medchemcomm. 2011 May ; 2(5): 356–370. doi:10.1039/C1MD00044F.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Computational ligand-based rational design: Role of conformational sampling and force fields in model development

Jihyun Shim and **Alexander D. MacKerell Jr.**[a]

Alexander D. MacKerell: amackere@rx.umaryland.edu

[a] HSFII, 20 Penn St. Baltimore, MD 21201, USA. Fax: 410 706 5017; Tel: 410 706 7442

Abstract

A significant number of drug discovery efforts are based on natural products or high throughput screens from which compounds showing potential therapeutic effects are identified without knowledge of the target molecule or its 3D structure. In such cases computational ligand-based drug design (LBDD) can accelerate the drug discovery processes. LBDD is a general approach to elucidate the relationship of a compound's structure and physicochemical attributes to its biological activity. The resulting structure-activity relationship (SAR) may then act as the basis for the prediction of compounds with improved biological attributes. LBDD methods range from pharmacophore models identifying essential features of ligands responsible for their activity, quantitative structure-activity relationships (QSAR) yielding quantitative estimates of activities based on physiochemical properties, and to similarity searching, which explores compounds with similar properties as well as various combinations of the above. A number of recent LBDD approaches involve the use of multiple conformations of the ligands being studied. One of the basic components to generate multiple conformations in LBDD is molecular mechanics (MM), which apply an empirical energy function to relate conformation to energies and forces. The collection of conformations for ligands is then combined with functional data using methods ranging from regression analysis to neural networks, from which the SAR is determined. Accordingly, for effective application of LBDD for SAR determinations it is important that the compounds be accurately modelled such that the appropriate range of conformations accessible to the ligands is identified. Such accurate modelling is largely based on use of the appropriate empirical force field for the molecules being investigated and the approaches used to generate the conformations. The present chapter includes a brief overview of currently used SAR methods in LBDD followed by a more detailed presentation of issues and limitations associated with empirical energy functions and conformational sampling methods.

1. Introduction

When ligands and data on the biological activities of those ligands are the only information available for drug development, computer-aided ligand based drug design (LBDD)[1-6] is an effective method to extend the knowledge of the known ligands to design compounds with improved biological activity. The importance of LBDD is emphasized by more than 50 % of current FDA-approved drugs targeting membrane proteins such as G protein coupled receptors (GPCRs), nuclear receptors, and transporters[7], for which three-dimensional (3D) structures are often not available, a necessary prerequisite for target based drug design approaches[8-10]. Considering the difficulties in determining 3D structures of membrane-

This journal is © The Royal Society of Chemistry

† Electronic Supplementary Information (ESI) available: See DOI: 10.1039/b000000x/

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

associated proteins, LBDD methodologies are anticipated to continue to have a significant impact on drug development into the foreseeable future[11, 12].

Drugs typically exert their pharmacological effects by specific interactions with their target proteins. Such unique interactions have been understood as "lock-and-key"[13], "induced-fit"[14], "conformational selection" or "population shift" hypotheses,[15-19] which are based on the inherent chemical structure of molecules, their dynamic conformational properties and how those two influence the receptor. Therefore, identifying any causation or correlation between structures and activities, referred to as a structure-activity relationship (SAR)[20-22], can be of utility for ligand design. LBDD based SAR identifies similarities and/or differences in structural or physicochemical properties of compounds and relates them to activity, including efficacy (e.g. activation or stimulation of receptors, V_{max} of enzymes), affinity (e.g. K_i), selectivity (e.g. $K_{i,\ isoform1}/K_{i,\ isoform\ 2}$), pharmacokinetics (ADME)[23, 24], drug-drug interactions, or any biological properties of interest. Various descriptors of the ligands are related to biological activities through various statistical methods, for instance, regression, classification, dimension reduction, variable selection, *etc.* from which important features of the ligands responsible for activity are identified and used to develop new leads or to optimize known ligands.

Three major categories of LBDD are quantitative structure activity relationship (QSAR)[25-27], pharmacophore modeling[28-32], and similarity searching[33-36]. Over several decades, statistics, computational algorithms, and descriptors comprising the three categories and their pipelining have led to significant improvements both in efficiency and accuracy. Programs can deal with 100~1000s of molecules to build models or search molecular properties against databases of millions compounds in a short period of time. Overall improvements have been achieved by sophisticated data mining techniques and by more accurate mathematical descriptions of molecules through molecular mechanics (MM)[37] and quantum mechanics (QM) methods[38].

Recent advances in statistical, algorithmic and chemoinformatics in relation to LBDD have been discussed elsewhere in depth[39-42]. This chapter will briefly overview LBDD followed by a detailed presentation of new developments in the areas of conformational sampling and force fields (FF) with respect to LBDD.

2. Basic components of computer-aided LBDD

2.1. Representation of molecules

Molecules may be described in different ways ranging from one- to three dimensional (3D) and higher methods. For simple counting of molecular constitutions or fragments in 1D, one can use line notation such as SMILES (Simplified molecular input line entry specification)[43] and SLN (SYBYL line notation)[44] or chemical fingerprints, such as the MACCS representation[45]. 1D representations are used for fast lookup and comparison and in some cases do not yield a unique description of the molecules, as in the MACCS fingerprints. When molecules are represented as a graph[46], atoms are nodes and bonds are edges connecting the nodes, yielding a 2D description of a molecule. Information on bond or atom types, atom size, or stereochemistry and so on can be stored in the form of matrix and readily accessed. Along with the graph representation, a simple connection table may be used to calculate 2D molecular properties, for example, molecular weight, molar refractivity, number of rotatable bonds, branching, number of hydrogen bond acceptors and donors, and sum of atomic polarizabilities. Such descriptors are widely used in QSAR analysis. For a more detailed and realistic representation of molecules, a 3D representation, typically in the form of atomic Cartesian coordinates, is required. Such 3D descriptions also allow for the calculations of various descriptors and, more importantly, can represent the bio-active

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

conformations of a molecule. This is particularly important when comparing compounds with different chemical structures that may show similar biological properties by having similar 3D placement of biologically important functional groups. 3D descriptors, such as the spatial relationship between functional groups, may be calculated using semi-empirical or *ab initio* QM methods for small size ligands (number of atoms ≤ 100) and MM for most ligands. With growing computational power, the use of QM & MM is significantly increasing in the field of LBDD[1, 38]. Beyond 3D methods are 4D and higher representations. For example, the different possible conformations of the 3D structure of a molecule may be considered a 4D representation. The remainder of section 2 will present different nD representations of molecules and their utilization in LBDD.

2.2. 2D-QSAR

A large number of 1D or 2D molecular descriptors have been developed[47, 48]. Most software packages that include a QSAR module calculate a range of descriptors such as physicochemical, electronic, topological, and shape properties. Lipinski's rule of five[49] is a classical example of a straightforward application of QSAR where bioavailability is related to descriptors including octanol/water partition coefficient (logP), molecular weight, number of hydrogen bond donors and acceptors, and number of rotatable bonds.

Development of SAR models often requires pre-processing of the descriptors prior to model development. Values are often normalized such that coefficients obtained from fitted models represent the significance of the individual descriptors. Importantly, given the large number of possible 1D/2D descriptors it is necessary that the number of descriptors used during model development be limited by selection methodologies[50, 51]. In simple terms highly correlated descriptors are typically removed from model development. Descriptor selection is then often linked to model development itself. Those descriptors most predictive of a target property are selected by iterative analysis (stepwise multiple linear regression (MLR)[52], replacement method[53]) or by learning algorithms (Genetic algorithm[54], adaptive fuzzy partition algorithm[55], Gaussian processes[56], or Genetic function approximation (GFA)[57]). Correlated or redundant descriptors may also be eliminated by partial least square (PLS)[58-61] or principle component analysis (PCA)[62, 63].

Given a suitable training set (i.e. set of compounds with known biological activities) and descriptors, one applies statistical methods according to the characteristics of the data set. When linearity is present MLR and PLS are good choices. Nonlinearity[64, 65] occurs when one handles a large number of non-homologous data sets, for example in pharmacokinetic (PK) studies where multiple biological phenomena such as absorption and metabolism can impact the biological data, or when activities are influenced by many factors such as receptor dimerization, existence of receptor isoforms, and conformational changes. Selection (or design) of the experimental data for model development can minimize nonlinearity, thereby reducing possible ambiguities in the developed QSAR but users need to keep in mind the inherent complexity of biological phenomena. Selection of the appropriate statistical methods also depends on whether the goal of the study is interpretation or predictability. Classical methods such as MLR produce explicit physical meaning but predictability is not as good as using modern statistical tools that improve predictability. However, interpretability is often compromised with improved predictability. While the MLR method, which maximizes interpretability, was the most used method in 2008, it was followed by PLS & support vector machine (SVM) approaches that yield improved predictability. Newly developed statistical methods[39] include Gene Expression Programming (GEP) [66], Project Pursuit Regression (PPR)[67], Local Lazy Regression (LLR)[68] while recent variations in QSAR approaches are hologram-QSAR[69], auto-QSAR[70], and inverse-QSAR[71] among others. Hologram-QSAR[69] partitions molecules into smaller fragments and uses size, length, and as well as additional information on those fragments as

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

descriptors. It is usually combined with PLS to derive a QSAR. Auto-QSAR[70] is an automated QSAR where the best descriptors, the best statistical methods, and validations are chosen for given set of molecules and updated as the number of molecules in training set increases. In Inverse-QSAR[71] after a QSAR model is built, distributions of descriptors yielding optimal activity are estimated and structures are generated or searched that match those distributions.

As a last step in QSAR model development, the models require validation[72, 73]. Approaches to do this include cross-validation, y-randomization, or external test set. It is generally perceived that leave-one-out or leave-n-out cross-validations do not necessarily indicate predictability directly and external validation using compounds not included in model development is recommended to verify predictability of the developed models. An overview of statistical methods used in QSAR analysis is given in Table S1 of the supporting information.

2.3. nD-QSAR ($3 \le n \le 7$)

QSAR methods based on 3D descriptors (Table 1) such as molecular volume, surface area, $\Delta G_{solvation}$, dipole moments, HOMO, and LUMO, depend on the chemical and spatial features of molecules. Alternatively, 3D-QSAR[74-76] may be based on the molecular interaction field mapped onto a 3D grid surrounding the molecules of interest. The descriptors are the magnitudes of the fields at the grid points, an approach used in CoMFA (comparative molecular field analysis)[77] and CoMSIA (comparative molecular similarity indices)[78]. For example, in CoMFA polar or hydrophobic probes are placed on grid points and non-bonded interaction energies with the ligands are calculated, with the resulting values used as descriptors for each molecule, such that each molecule has, in essence, a number of descriptors that correspond to the number of grid points. The grid point values are subjected to statistical methods such as PLS or PCA and related to biological activities. Notably, these approaches require alignment of the different ligands being studied. When flexibility is added to the shape information by using multiple conformations, it is classified as 4D-QSAR[79, 80]. Although 3D-QSAR can use multiple conformers it means multiple model evaluations, with the input into each model being one static conformer for each ligand in the training set. 4D-QSAR, in one embodiment, overcomes this limitation by using grid cell occupancy descriptors calculated based on multiple conformers. 5D-QSAR[81, 82] attempts to construct a pseudo-receptor based on ligand information in combination with a GA to vary the grid point locations to produce a favorable induced-fit state. Beyond this, 6D-QSAR[83] incorporates solvation energy terms. Finally, the inclusion of 3D structure of the target from X-ray crystallography or NMR in the models represents the highest dimension, 7D, applied to date, though the approach is no longer formally LBDD.

Of the methods in Table 1, CoMFA and CoMSIA are the most widely used. Their main limitations are their dependency and sensitivity to conformations and alignments of the molecules under study. Different occupancies by different conformations or changes in molecular alignments can cause different interaction fields yielding different QSAR models. To overcome the limitations, alignment-independent 3D-QSAR was developed by transforming 3D-grid data into 2D descriptors such as GRIND (Grid independent descriptors)[84, 85] and VolSurf[84]. The approaches are listed and summarized in Table 1.

2.4. Pharmacophore modelling

The IUPAC definition[87] of a pharmacophore is “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interaction with a specific biological target structure and to trigger (or block) its biological response.” Pharmacophore modelling is closely related to 3D or 4D-QSAR and commonly used pharmacophoric keys

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

include hydrogen bond donors and acceptors, ionizable groups, aromatic rings, aliphatic hydrophobic groups, among others. In many cases superimposed compounds based on pharmacophore models are the starting point for 3D-QSAR analysis. Alternatively, pharmacophore features may be identified by building 3D-QSAR iteratively with different conformations or alignment attempted to increase activity prediction[28]. As with 3D-QSAR methods, critical steps in pharmacophore modelling are selection of bio-active conformations and structure alignments. Conformations can be pre-calculated and saved in a database or be generated on-the-fly during alignments. Various sampling methods used to generate multiple conformations will be discussed below. Using a collection of pharmacophoric keys or points (i.e. functional groups that may contribute to biological activity) and a conformational ensemble which is expected to include the active conformation, molecular alignment (or superposition)[1, 28, 31, 32, 88] is carried out by mapping fragments of the compounds with a target function being minimized. In the clique detection algorithm[89] distances between features are stored in a matrix and the differences between two matrices become the target function. Another way is using GA,[90] where a chromosome describes a molecule and genes encoded in it represent the ligand fragments to be matched to the features of the reference compound. The target function or fitness function in GA can include a range of information such as the similarities between features and volumes of aligned structures. The GA is often extended for on-the-fly conformational sampling by including geometric information of a molecule into the chromosomes. FlexS algorithm[91] is also used in pharmacophore feature mapping. It decomposes a molecule and grows it on top of the reference compound beginning from an anchor fragment. In all cases, it is important to select a reference structure that has high experimental activity, a known 3D binding conformation or a favorable docking score to facilitate interpretation of the obtained model. Details of methods used by the major pharmacophore modelling programs and recent research on alignment methods were reviewed in Leach et al.[31] and Lemmen and Lengauer[88]. Such alignment methods are also of general utility for pharmacophore-based similarity searching.

2.5. Ligand based similarity searching

Similarity searching is an effective, computationally accessible method to identify compounds with qualities similar to that of an active, lead compound. For example, if the number of ligands for which biological activity is known are too few to build a QSAR model, similarity searching may be effective. Similarity searching can be based on a number of features including chemical fingerprints, physiochemical properties as well as 2D and 3D features selected by QSAR or pharmacophore models. If compounds structurally similar to active compounds are desired, searches based on chemical fingerprints are appropriate. However, if the goal of the study is the identification of ligands that have a new scaffold/ chemical structure but maintain the desired biological activity (i.e. scaffold hopping[92]), searching based on physiochemical properties may be of utility. The most widely used descriptors in similarity searching are chemical fingerprints or large numbers of physiochemical properties. Fingerprints can have diverse properties and combinatorial or compressed fingerprints are emerging and efforts are being made to improve the fingerprint representations. To quantify the extent of similarity between compounds, different similarity measures are used alone or in combinations; these include the Tanimoto, Cosine, Hamming, Russel-Rao, and Forbes indices. Finally, it should be noted that it is useful to perform successive searches using the nearest neighbors of a query compound. A number of software packages, including MOE (Molecular Operating Environment, Chemical Computing Group[93]) and Discovery Studio (Acclerys Inc[94]) allow for similarity searching. Public small molecule databases such as PubChem[95] ZINC[96] and ChEMBL[97] or open source software[98] using those databases provide similarity searching and clustering tools. Notable are clustering techniques[99] where output structures from a similarity search are further grouped

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

into subsets to reduce redundancy and to check diversity in compounds selected, for example, for a target-based database screen[100]. Results of clustering vary based on classification algorithms, descriptors, and similarity measures[101-103] and there is no gold standard to obtain the best clustering. Therefore it is desirable to perform clustering with a combination of methods, descriptors, and similarity coefficients followed by manual evaluation of the results to achieve the desired outcome.

3. Conformational sampling

For the 3D and higher order methods it is essential that the appropriate, biologically relevant conformations be identified. Considering that drug-like molecules can have 10 or more rotatable bonds and each such bond may have 3 accessible rotamers a compound may have 3^{10} conformations that must be considered. The need to access a large number of these conformations is further emphasized by studies showing that bioactive conformations of compounds in X-ray crystal structures of ligand-protein complexes can have energies 15~20 kcal/mol higher than the global minimum[104-107]. While the value of 3^{10} is likely an overestimation due to steric clashes, it is evident that a major concern in modern LBDD methods is securing bioactive conformations given that model assessment includes multiple conformations in a large number of studies[105, 106, 108-112]. Furthermore, taking all conformations into account during model development can account for the dynamic nature of molecules; 4D-QSAR[79, 80] and the Conformationally sampled pharmacophore (CSP)[113-115] are representative methods that use this information.

Various methods are employed to generate multiple conformations of a ligand. Systematic search approaches[116-118] formally perform an exhaustive sampling of conformational space, thereby covering the whole energy surface. However, as the degrees of freedom increases the number of possible conformations becomes enormous and often includes non-physical structures. Systematic search procedures therefore often limit the number of conformations accessed and select conformations within a user-defined energy difference range. For example, MOE supports a maximum of up to 10000 conformers and each one is subjected to trial model buildup. An alternative to systematic sampling is Monte Carlo based approaches[119-121] where random changes in structures (i.e. trial moves) are attempted by rotation about a dihedral angle or other geometric change with the new conformations associated with the trial moves accepted or rejected according to the Metropolis criterion[120, 121]. In the Metropolis method a conformation with an energy, ΔE, lower than that of the previous conformation is accepted while conformations with higher, less favorable energies are kept based on the acceptance probability, $p=e^{-\Delta E/kT}$ where k is the Boltzmann constant, T is the temperature and p is compared to a random number. If p is greater than the random number the conformation is accepted, allowing higher energy conformations to be sampled. In this method energy barriers are easily overcome by increasing the effective temperature but random elements still exist during sampling leading to inefficiencies due to similar conformations being accessed. The Poling method[122] adds a penalty function (poling function) to the energy during the conformational search that is inversely proportional to the root mean square distance between conformations so sampling of similar conformation is avoided while accessing new conformation is maximized. With the same goal as the Poling method, Tabu search[123] keeps a record of previous sampled states thereby maximizing the exploration of previously unsampled conformations. When GA[124, 125] is applied for conformational sampling, each chromosome is a conformer and contains genes corresponding to structural degrees of freedom in the molecule. Chromosomes undergo mutations and crossover resulting in the sampling of diverse conformations in the descendant conformations. Molecular dynamics (MD) simulations sample conformations deterministically according to Newton's equations of motion and overcome trapping in local minima due to the inclusion of kinetic energy in the system, as

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

described below. All of the methods mentioned in this paragraph have advantages and disadvantages. For small molecules systematic search algorithms in combination with an accurate force field, as discussed below, can assure that all relevant conformations are taken into account. With larger molecules, exhaustive sampling of all accessible conformations is not feasible, MC or MD methods allow for extensive sampling of accessible conformations, though care must be taken to assure that all the relevant conformations are being sampled. One outstanding advantage of both MD and MC methods is that a variety of methods that allow for detailed representation of the biological environment of a molecule by, for example, the explicit treatment of waters and ions have been developed for these approaches. Given the wide use of MD methods for conformational sampling as well as for studies of the dynamics of molecules ranging from small ligands to large macromolecular complexes containing 1 million or more atoms, the remainder of the discussion of conformational sampling will focus on MD based approaches[126-130].

MD simulations [127, 131, 132] are based on Newton's equations of motions. The second law F=ma states that from position $r_i(t)$, velocity $v_i(t)$, and mass m_i for an atom i at time t, force $F_i(t)$ can be calculated. In MD simulations, forces are usually obtained from analytical derivatives of the potential energy function and integration methods are used to obtain new positions $r_i(t+\delta t)$ and velocities $v_i(t+\delta t)$ from the previous states, $r_i(t)$, $v_i(t)$, and $a_i(t)$. For example, the original Verlet integrator[133] uses a Taylor series expansion of position. Summing equation 1a and 1b yields 1c which determines the new position. This integration minimizes memory requirements as it is not necessary to store velocities, although they can readily be calculated using equation 1d if required.

$$r(t+\delta t)=r(t)+v(t)\delta t+\frac{1}{2}a(t)\delta t^2 \qquad \text{Equation 1a}$$

$$r(t-\delta t)=r(t)-v(t)\delta t+\frac{1}{2}a(t)\delta t^2 \qquad \text{Equation 1b}$$

$$r(t+\delta t)=2r(t)-r(t-\delta t)+a(t)\delta t^2 \qquad \text{Equation 1c}$$

$$v(t)=\frac{r(t+\delta t)-r(t-\delta t)}{2\delta t} \qquad \text{Equation 1d}$$

Leapfrog Verlet integration[134, 135] uses an expansion of positions and velocities to the second order and an interval 1/2δt instead of δt. Subtraction of equation 2b from 2a yields 2c. New positions may then be obtained by substituting v(t) from rearrangement of 2a into v(t) of 1a and truncating the expansion at the velocity term yielding equation 2d.

$$v(t+\frac{1}{2}\delta t)=v(t)+\frac{1}{2}a(t)\delta t \qquad \text{Equation 2a}$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$v(t-\frac{1}{2}\delta t)=v(t)-\frac{1}{2}a(t)\delta t \quad \text{Equation 2b}$$

$$v(t+\frac{1}{2}\delta t)=v(t-\frac{1}{2}\delta t)+a(t)\delta t \quad \text{Equation 2c}$$

$$r(t+\delta t)=r(t)+v(t+\frac{1}{2}\delta t)\delta t \quad \text{Equation 2d}$$

Extended integration methods have been developed to enhance accuracy and provide special features for the simulation system of interest[136]. For example, for simulations of aqueous solutions it may be desirable to reproduce constant pressure, temperature, or volume in accord with the specific ensemble being targeted. For example, simulations in the constant pressure, temperature and number of particles ensemble (NPT) may be used to calculate Gibbs free energies while a constant volume, temperature and number of particles ensemble (NVT) yields Helmholtz free energies. However, such simulations require the appropriate boundary conditions, such as periodic boundaries[136]. While these are necessary for sampling the conformations of ligands in the presence of explicit solvent, for LBDD MD simulations are typically performed in the absence of explicit solvent. Such simulations may be performed in the “gas phase” or using implicit solvent models to treat the solvent environment, as detailed below.

High temperature MD has long been used to facilitate the crossing of high energy barriers to assure a broad sampling of conformational space. In high temperature MD, the probability of particles having the necessary velocity (or kinetic energy) to cross energy barriers is increased over room temperature simulations. While sampling in MD is driven by information on the molecular forces thereby guiding conformational sampling to physically meaningful regions, unwanted sampling may occur in high temperature MD. This is due to the high temperature leading to sampling of conformations that are inaccessible at room temperature thereby causing inefficient use of computational effort. However, care to avoid excessively high temperatures can minimize this problem and a number of protocols, referred to as simulated annealing[137], perform high temperature MD followed by room temperature MD to assure that conformations relevant to the latter are being sampled.

A simple way to improve sampling via MD simulations is to perform multiple simulations of the system starting with different initial random number seeds to assign the velocities to the particles in the system. Typically, a Gaussian distribution of velocities are randomly generated and assigned to each particle with those initial velocities satisfying a Maxwell-Boltzmann distribution defining a selected temperature. While the overall velocity distribution is approximately reproduced with the different random number seeds yielding the same macroscopic temperature, the individual atoms have different velocities, thereby directing the molecule to sample different conformations. However, this approach does not always avoid kinetic trapping for larger molecules due to large barriers often associated with large conformational changes.

Methods that go beyond the use of high temperature and multiple MD runs are referred to as generalized ensemble (GE) algorithms.[138-140] These include replica exchange MD (REXMD)[141], meta-dynamics[142, 143], accelerated MD (AMD)[144] and λ-dynamics[145] among

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

others. In GE algorithms energy barriers are overcome by adding an external biasing potential(s) to the system. This may be performed by accessing additional conformations from additional simulations, as in REXMD, or by approaches that directly modify the free energy landscape of the system. Many GE MD simulation approaches sample the free energy landscape efficiently and may be used to calculate accurate free energy differences. The free energies are often calculated by thermodynamic integration (TI) [146] or the weighted histogram analysis method (WHAM).[147]

Standard REXMD involves parallel independent simulations (replicas) at a range of temperatures and exchanges conformations between replicas according to an exchange probability (Equation 3).

$$P(exchange)=\begin{cases} e^{-\Delta}, for\ \Delta>0 \\ 1, for\ \Delta \le 0 \end{cases}$$

Equation 3a

$$\text{where } \Delta=\left(\frac{1}{kT_i}-\frac{1}{kT_j}\right)\left(E(q_j)-E(q_i)\right)$$
$$\text{in REXMD}$$

Equation 3b

$$\text{or } \Delta=\frac{1}{kT}\left[\left(E^j(q_j)-E^j(q_i)\right)-\left(E^i(q_j)-E^i(q_i)\right)\right]$$
$$\text{in HREXMD}$$

Equation 3c

In equation 3 T_i indicates the temperature of replica i, q_i is the configuration of replica i at the point of exchange, and E^i represents Hamiltonian energy of replica i. The main idea behind REXMD is that one MD trajectory in a local minima can take conformational information from another replica which may be found in another region of conformational space (e.g. across an energy barrier) but have similar energies. The probability of exchange between replicas is such that it enforces sampling of a Boltzmann distribution of conformations, thereby satisfying a proper thermodynamic ensemble as defined by the simulation conditions. Implementation of REXMD is not straightforward, with issues including how to set up the proper temperature spacing, the number of replicas, and the exchange frequency. For large systems with explicit solvent, REXMD requires a large number of replicas and small temperature difference between adjacent replicas to achieve an acceptable exchange ratio. To overcome this implicit solvent models, as discussed below, may be used thereby allowing for a significant increase in the difference in temperature between adjacent replicas. For example, in the presence of explicit solvent replicas may have temperature differences of 10 K while when implicit solvent is used 30 K may be used. In addition, hybrid methods have been developed[148]. Concerning, exchange frequency, higher exchange frequencies typically lead to enhanced sampling[149]. However, care must be taken as although high temperature enhances barrier crossing, it may shift the equilibrium between two states and make high temperature states more favorable throughout all replicas.

Hamiltonian replica exchange molecular dynamics (HREXMD)[150-152] overcomes drawbacks of REXMD by scaling the potential energy function (i.e. Hamiltonian) rather than the temperature (Equation 3c). Perturbation of the Hamiltonian can involve almost any term in a force field such as the peptide backbone conformational energies, dielectric constant or ligand-solvent interactions. HREXMD needs a lower number of replicas than

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

REXMD since the perturbation is applied locally on selected components of the system. Generally the perturbation is expressed as a function of an order parameter λ in

$$H(\lambda_i)=\lambda_i H_1+(1-\lambda_\mathrm{i})H_0 \quad \text{Equation 4}$$

where $0 \leq \lambda_\mathrm{i} \leq 1$, such that the Hamiltonian is the reference state H_0 (i.e. the ground state) when λ is zero and target state H_1 (i.e. fully perturbed) when λ is one. In HREXMD, simulations are carried on each replica with $H(\lambda_i)$ and conformations exchanged between adjacent H(λ_i) replicas thereby facilitating the crossing of energy barriers.

There are single MD approaches using dynamic variations of λ. Lambda-dynamics[145, 153] implements λ as an artificial particle that is propagated during the MD simulations thereby sampling various λ values as dictated by the free energies of the system without the need of using pre-defined λ values. The Hamiltonian is expressed by adding two more terms for dynamic λ variables to H(λ_i), as shown below, and {λ_i} is used since multiple λs can be used to perturb different components of the Hamiltonian such as electrostatic or dihedral energies. In the approach the extended Hamiltonian, comprised of the standard, ground state Hamiltonian and the perturbations associated with the λ terms is defined as

$$H_{extended}(\{\lambda_i\})=H(\{\lambda_i\})+\sum_{i=1}^{n}\frac{1}{2}m_i\lambda_\mathrm{i}^2+U^*(\{\lambda_i\}) \quad \text{Equation 5}$$

where m_i and λ_i are dynamic variables which overcome the limitation of using discrete λ values and U*({λ_i}) is a λ_i-dependent biasing potential which can take various forms to sample as many states as possible.

Another single dynamics GE method, Metadynamics[142, 143, 154] uses a history-dependent biasing potential to force selected degrees of freedom (e.g. collective variables, CV) of the system being sampled away from conformations visited frequently. This is performed by “lifting” low energy regions with a biasing potential as those regions are being sampled, thereby facilitating conformational changes away from the low energy regions. The biasing potential is the sum of Gaussian functions, V_G, that are used to fill valleys of the free energy surface as defined as follows

$$V_G(U(q),t)=h\sum_{t}e^{-\left(\frac{(U(q)-CV(t))^2}{2w^2}\right)} \quad \text{Equation 6}$$

where U(q) is the potential energy of coordinates q at time t, h is Gaussian height, w is Gaussian width, and CV(t) is the value of the CV at time t. The simulation remembers information about the added biasing potentials and the final V_G is a negative image of the free energy surface thereby allowing reconstruction of the original free energy surface. Metadynamics is able to run with multiple CVs such as distance between two atoms, angles, or torsion angles; the choice of CVs, and optimal h and w for each CV are user selected and optimization of these parameters for the system of interest is often required.

Taking a similar strategy, the orthogonal space random walk (OSRW) algorithm[155, 156] is another efficient way of conformational sampling. This strategy simultaneously perturbs the order parameter space (general term for λ or CV above) and generalized free energy space to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

overcome not only local minima trapping but also lagging of changes in the environment surrounding the CV or λ required for conformational changes to occur. OSRW uses 2D Gaussian-shaped repulsive potentials to flatten the free energy surface and avoid often-visited states. After searching the whole conformational space it is possible to select accessible conformations by order parameters associated with the conformational change.

Another approach is accelerated MD (AMD).[144] AMD is a simple but efficient sampling method that has shown good performance for biomolecules. Compared to metadynamics and OSRW, it uses a simpler form of the biasing potential (equation 7). When applying the method, the boost energy E and α, which is a tuning factor for the biasing potential's well-depth, need to be pre-defined and the biasing potential is applied when the potential energy is less than E.

$$V^{*}(r)=\begin{cases} V(r), & V(r)\geq E \\ V(r)+\frac{(E-V(r))^{2}}{\alpha+(E-V(r))}, & V(r)<E \end{cases} \qquad \text{Equation 7}$$

REXMD and HREXMD have been successfully used for conformational sampling of flexible ligands[114, 157-159], while the other GE algorithms have been used mainly in biomolecules to date. However, considering their success in conformational sampling, problems involved in flexible protein loops and ligand passage in receptors[160, 161], it is anticipated that they will be of utility to conformational sampling in LBDD.

In addition to the sampling algorithms, an important consideration is the role of solvent in conformational sampling. Conformational changes and sampling are dependent on the surrounding environment of all molecules such that energetically favorable conformations in gas phase may not be favorable in solution (or a receptor binding site) and vice versa. This occurs due to water competing for favorable intramolecular interactions, for example, by disrupting intramolecular hydrogen bonds or reduced intramolecular dipole-dipole interactions. Alternatively, water can impact the orientation of hydrophobic groups, which may remain “accessible” in the gas phase, but cluster together in the presence of solvent. Therefore sampling conformations in the presence of explicit water molecules is ideal when studying biological systems but such calculations are computationally expensive due to the presence of additional particles in the system. In addition, the viscosity of the water surrounding a molecule can slow conformational sampling during MD simulations, thereby further increasing the computational costs. An effective alternative to explicit solvent are continuum solvent models that allow for distributions of conformations to be obtained that approximate an explicit solvent environment while allowing for efficient sampling by avoiding the increased number of particles and the viscosity issues with water. A large number of implicit solvent models are available and recent reviews on the topic have been presented[162-164]. Simple implicit solvent models use constant or distance-dependent dielectric constants. Another approach is using solvation parameters for each atom based on their solvent accessibility so that particles in the system have varying responses to environments. More accurate solutions to solvation effects are obtained by the Poisson-Boltzmann (PB)[165, 166] or Generalized Born (GB)[167] models. In MD, the most widely used methods are GB models due to the high efficiency of their analytical solution and comparable accuracy with respect to PB models that typically require numerical solutions. GB models calculate the electrostatic component of the solvation free energy as shown in Equation 8, with that term added to the total energy.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\Delta G_{elec} \approx -\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2+R_iR_j\exp\left(\frac{r_{ij}^2}{4R_iR_j}\right)}}$$

Equation 8

In equation 8 ε is the dielectric constant of solvent, q_i is the partial atomic charge of atom i, r_{ij} is the interatomic distance between atom i and j, and R_i, which is the most important parameter in GB models, is the effective Born radii of atom i. The effective Born radii can be understood as an atom's degree of burial within the solute or radius exposed to the solvent environment. A number of GB models have been developed such as GBMV[168] and GBSW[169] in CHARMM[170, 171] and GB^{OBC} [163] in AMBER[172, 173] which differ primarily in the way that the effective Born radii are calculated.

The final portion of this section presents a simple example of sampling the conformational space of a peptide using three MD-based sampling approaches based on the CHARMM22/CMAP protein force field[174]. Figure 1 shows the extent of sampling attained by MD in both vacuum and explicit solvent (Fig. 1a and 1b) and by two sampling methods (single MD vs. HREXMD using explicit solvent, Fig. 1b and 1c) for the flexible opioid pentapeptide, Leu-enkephalin. For the explicit solvent HREXMD, λ=0, 0.14, 0.19, 0.27, 0.37, 0.52, 0.72, and 1.0 where 1.0 represents a CHARMM phi, psi energy surface that is flat for non-glycine residues, as implemented using the CMAP tool[174]. Simulations were performed using the REPDSTR module in CHARMM for 10 ns, attempting exchanges every 0.5 ps for the HREXMD. The range of conformational sampling was measured by the 2D probability distribution of the distance between two aromatic rings (A and B) and the angle between two aromatic rings and N-terminal nitrogen (N). Additional details of the simulation methodology are included in the supporting information. Comparison of figures 1a and 1b show that the sampling of conformational space by MD differs in gas phase vs explicit solvent. In explicit solvent, structures with longer AB distances and larger ANB angles are being sampled. These represent more extended structures due to the presence of solvent; in the gas phase the peptide folds back on itself leading to more compact structures as no solvent is available to compete for intramolecular interactions. As expected, HREXMD (1c) samples a similar range of conformations as the standard explict solvent MD (1b), but the extent of sampling is more complete using the same amount of simulation time. The more complete sampling by HREXMD is due to the simulation overcoming an energy barrier in the vicinity of 10~12 Å for the AB distance and 60~100° for the ANB angle. Although this example is not a rigorous test from which better performance can be proved based on more efficient sampling, it points out the importance of the simulation method and solvent environment when performing conformational sampling.

An important consideration when performing conformational sampling is the extent of convergence; have all the accessible conformations of the molecule been sampled? When conformational sampling is done by MD, convergence of sampling may be checked by continuing the simulation time and testing if additional conformations are being sampled. If additional conformations are not being sampled the sampling may be considered converged. For conventional MD, root mean square deviations of overall structure, distance between atoms or functional groups (as in Figure 1), or torsions may be used for simple evaluation of convergence. Alternatively, differences in probability distributions between two intervals of a trajectory (e.g. the first and second half) can also indicate if the simulation has reached convergence. For GE methods, convergence of the calculated free energy surface indicates adequate sampling. However, it should be emphasized that the appearance of convergence does not necessarily mean that true convergence has been attained. There is always the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

possibility that a molecule, especially more complex molecules such as polypeptides, may have access to significantly different conformations than those accessed in the performed simulations.

4. Force fields (FF)

While the appropriate sampling approaches can assure that the required range of conformations is being accessed it is the underlying energy function that largely determines the probability of the conformations being sampled. While QM methods can supply this information their computational demand limits their utility for sampling large numbers of conformations for even small molecules. Accordingly, it is necessary to use molecular mechanics energy functions. While such functions are computationally efficient, they are based on simple terms that require a set of parameters to allow for the energy and forces on a molecule to be accurately calculated, as described below. These parameters, therefore, dictate the applicability and quality of the force field and a number of force fields are available for drug-like molecules. In the remainder of this article an overview of the force fields most commonly used for ligands will be presented, including examples of the ability of selected force fields to reproduce QM conformational energies of two example ligands.

A force field consists of a potential energy function and the associated parameters that allow the energy and forces to be calculated as a function of the molecular structure and conformation. Potential energy functions used in molecular mechanics typically include terms for bond stretching, angle bending, rotation around bonds (dihedral or torsion angles), out of plane motions (improper angles), and non-bonded interactions (electrostatic and van der Waal energies). Such force fields are referred to as Class I models. In Class II force fields crossterms to treat correlation between bonds and angles, angles and torsions, and so on are included and different variations of the nonbond terms may be used. Equations 9 and 10 show examples of potential energy functions in the two classes, which are classified by their simplicity and potential transferability. Detailed descriptions of the example functional forms are found in Brooks *et al.*[170] for the class I force field in CHARMM and in Plimpton[175, 176] for a class II FF. Class I FFs include those specialized for biomolecules (see below). The simple functional form shown in Equation 9 is computationally efficient allowing them to handle macromolecules in aqueous or other condensed phase environments. Equation 10 is a typical energy function used in a Class II FF and employs more complex form which facilitates (but doesn't necessarily dictate) transferability across a wider range of molecules.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$
\begin{aligned}
Class I{:}U(\overrightarrow{R}) = & \sum_{bonds} K_b(b-b_0)^2 \\
& + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{Urey-Bradley} K_{UB}(S-S_0)^2 \\
& + \sum_{dihedrals} K_\phi\,(1+\cos(n\phi-\delta)) \\
& + \sum_{impropers} K_\omega(\omega-\omega_0)^2 \\
& + \sum_{residues} U_{CMAP}(\phi,\varphi) \\
& + \sum_{nonbonded\ pairs} \left\{\varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12}-2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right]+\frac{q_iq_j}{4\pi\varepsilon r_{ij}}\right\}
\end{aligned}
\qquad \text{Equation 9}
$$

In equation 9 b, θ, S, n, φ, δ, and ω represent the bond length, angle, distance between 1-3 atoms, multiplicity, torsion, phase, and improper angle, respectively, and the subscript zero indicates the equilibrium geometry parameter. K_x are the associated force constants. The Lennard-Jones 6-12 equation is commonly used to model the van der Waals energy between atom i and j, where ε_{ij}, r_{ij}, and $R_{min,ij}$ are the well-depth, interatomic distance, and interaction distance at minimum of the energy between atoms i and j, respectively. In the electrostatic or Coulombic term, q_i and q_j are partial atomic charges and ε is the dielectric constant.

For equation 10 the same symbols are used as in Equation 9 and additionally b′ and θ ′ used in the crossterms represent the second bond or angle associated with the cross interaction.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$
\begin{aligned}
ClassII{:}U(\overrightarrow{R}) =& \sum_{bonds}[K_{b2}(b-b_0)^2+K_{b3}(b-b_0)^3+K_{b4}(b-b_0)^4] \\
&+\sum_{angles}[K_{\theta 2}(\theta-\theta_0)^2+K_{\theta 3}(\theta-\theta_0)^3+K_{\theta 4}(\theta-\theta_0)^4] \\
&+\sum_{dihedrals}[K_{\phi 1}(1+\cos\phi)+K_{\phi 2}(1+\cos 2\phi)+K_{\phi 3}(1+\cos 3\phi)] \\
&+\sum_{impropers}K_{\omega}(\omega)^2 \\
&+\sum_{nonbonded\ pairs}\left\{\varepsilon_{ij}\left[2\left(\frac{R_{\mathrm{min},ij}}{r_{ij}}\right)^9-3\left(\frac{R_{\mathrm{min},ij}}{r_{ij}}\right)^6\right]+\frac{q_iq_j}{4\pi\varepsilon r_{ij}}\right\} \\
&+\sum_{bonds}\sum_{bonds'}K_{bb'}(b-b_0)(b'-b'_0)+\sum_{angles}\sum_{angles'}K_{\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0) \\
&+\sum_{bonds}\sum_{angles}K_{b\theta}(b-b_0)(\theta-\theta_0) \\
&+\sum_{bonds}\sum_{dihedrals}(b-b_0)[K_{b,\phi 1}\cos\phi+K_{b,\phi 2}\cos 2\phi+K_{b,\phi 3}\cos 3\phi] \\
&+\sum_{bonds'}\sum_{dihedrals}(b'-b'_0)[K_{b',\phi 1}\cos\phi+K_{b',\phi 2}\cos 2\phi+K_{b',\phi 3}\cos 3\phi] \\
&+\sum_{angles}\sum_{dihedrals}(\theta-\theta_0)[K_{\theta,\phi 1}\cos\phi+K_{\theta,\phi 2}\cos 2\phi+K_{\theta,\phi 3}\cos 3\phi] \\
&+\sum_{angles}\sum_{angles'}\sum_{dihedrals}K_{\theta\theta'\phi}(\theta-\theta_0)(\theta'-\theta'_0)\cos\phi
\end{aligned}
\qquad \text{Equation 10}
$$

To move from a potential energy function to a FF requires determination of the values of the parameters, a process referred to a parameter optimization or parametrization. Parameters to be optimized in equations 9 and 10 include force constants, equilibrium geometries, partial atomic charges, well depth and interaction distance at minimum energy ($R_{min,ij}$) and so on. The goal of parameter optimization is reproducing a collection of quantum mechanical and/or experimental observables for the ligands of interest.

Parameters for macromolecules such as proteins, nucleic acids, lipids and carbohydrates have been paid special attention and optimized extensively in Class I FFs such as AMBER(Assisted Model Building with Energy Refinement)[172, 173], CHARMM(Chemistry at Harvard Molecular Mechanics)[170, 171], GROMACS[177] and OPLS(Optimized Potential for Liquid Simulations)[178, 179]. Subsequently, parameters for drug-like molecules were added to be compatible with the individual biomolecular FFs while maintaining the accuracy of the "parent" biomolecular FF. To achieve this, the parent and related small molecule parameters use the same form of the potential energy function and strategy used for optimization of the FF. This is necessary to provide consistent and balanced energy and force evaluations during simulations of small molecule-biomolecular complexes. For example, the CHARMM General FF (CGenFF) [180] follows the standard optimization procedure of the CHARMM additive biomolecular FF[181].

As an example of parameter optimization the approach used in the CHARMM additive force field for small molecules, with which we are intimately familiar, will be used. Typically, optimization of CGenFF parameters is performed as follows. Force constants, equilibrium bond lengths and valence angles are parametrized to reproduce experimental or QM

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

vibrational frequencies and geometries. Dihedral angle force constants, phases, and multiplicities are optimized targeting QM potential energy scans or spectroscopic data such as NMR J coupling constants. Charges are optimized by evaluating optimal distances and interaction energies of water interaction with the drug-like molecule based on QM calculations as well as dipole moments and optimization of LJ parameters is guided by the reproduction of pure solvent or crystal experimental data. To date, CGenFF includes approximately 150 atom types, 400 bond, 1200 angle, 3000 torsion parameters explicitly optimized based on 500 model compounds. When extending the force field to new chemical entities, parameters for the new molecules not already available in the CGenFF may be assigned by analogy for the bond/angle/dihedral and LJ terms, while determination of the partial atomic charges is based on a bond-charge increment algorithm extended to included angle- and dihedrals increments that have been trained to reproduce CGenFF charges for over 500 model compounds (K. Vanommeslaeghe and A.D. MacKerell, Jr., work in progress). A web-based utility in the context of the ParamChem project is available to perform these functions. An important feature of CGenFF when automatically assigning parameters is information about the quality of the assigned parameter based on a penalty score. This is important as the ability of parameters to be transferred between molecules in the context of empirical force fields is limited, as shown below, and it allows users to know which parameters require validation and further optimization to obtain the required level of accuracy. However, as with conformational sampling, even in cases where the parameters are directly transferred to a new molecule, the possibility that those parameters may not perform with adequate accuracy exists, such that the aware user is advised to perform validation tests of the transferred parameters, as previously described.[180]

Beyond CGenFF there are a number of other small molecule FFs designed to be compatible with biomolecular FFs. GAFF (General AMBER FF)[182] was developed for the simulation of pharmaceutical compounds with the AMBER biomolecular FF. It is based on QM optimization of about 3000 model compounds and geometric information from the Cambridge structure database (CSD)[183]. It has 57 atom types, 700 bond length parameters, 3000 angle parameters, and 500 dihedral angle parameters. Beyond these available parameters, the Antechamber toolkit[184] is used to assign parameters for novel molecules. SwissParam is a web-based utility used to generate CHARMM consistent parameters for ligands. It takes internal energy parameters and charges from MMFF (Merck Molecular FF)[185, 186] while cubic and quadratic terms for bond, angle, and improper energies that are present in the Class II force field are truncated as required for use with the Class I CHARMM additive FF. In addition, van der Waals energy parameters are from the CHARMM additive FF based on atom type similarity. However, it should be noted that the nonbond parameters being derived in a different manner than that of the parent biomolecular FF make them formally incompatible with the CHARMM biomolecular FF. OPLS-AA (All Atom) emphasizes parameters to reproduce the conformational energetic and condensed phase properties of small molecules for use in biological environments. Initial parameters were adopted from the OPLS-UA (united atom), AMBER, and CHARMM FFs and 50 model compounds were optimized focusing on torsion and non-bonded parameters[178]. OPLS-AA uses experimental liquid properties as target data during parameter optimization. Thus, Class I biomolecular FFs have been extended to include parameters for a range of small molecules though the extent of chemical space covered and the quality of the parameters for those molecules vary significantly.

Class II FFs were initially designed to treat a wide range of small molecules. Examples include CFF/CVFF (consistent valence FF)[187], MM2 (molecular mechanics)[188], MM3[189], MM4[190], MMFF94[185], and Tripos 5.2 FF[191]. These FFs are typically not optimized with respect to interactions with the environment, with the exception of MMFF, limiting their applicability. In general, Class II force fields were optimized to reproduce geometries,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

vibrational spectra and conformational energies in the gas phase, with the various cross and higher order terms in the energy functions (Equation 10) included to allow for both better reproduction of those properties as well as facilitate transferability of the parameters to a wider range of compounds. MMFF94 is currently one of the most widely used FFs for small molecules and is available in numerous LBDD software packages for small molecule simulations. Its goal is broad applicability and QM data for over 3000 molecules and condensed phase data for 2800 CSD compounds were used to optimize and validate the parameters. Allinger and coworkers have developed the MM1-4 FF series achieving high accuracy for organic molecules with respect to geometries, conformational properties and heats of formation. Upon going from MM1 to MM2, the MMx series shifted to a simpler form of Class II FF and MM3 lead to further improvements by including more model compounds, additional experimental data, and higher energy conformations during parameter optimization. MM4 represents a further extension of a Class II FF due to the inclusion of four-fold torsional energy terms, torsion-improper-torsion cross terms, bond-torsion-bond cross terms, two torsion-bond cross terms for central and terminal bonds each and so on. MM4 was optimized targeting thermodynamic quantities ΔH, ΔS, ΔG, and geometries from QM calculations or experimental spectroscopic data. All of these class II FFs are primarily utilized for organic compounds in the gas phase, though MMFF94 has shown limited use in macromolecular condensed phase simulations.

Use of a FF for energy evaluation, energy minimization, MD simulation or other sampling approach represents a significant, important step forward in most modern LBDD studies and the quality of the FF plays an important role in the outcomes of such studies. As emphasized in the preceding paragraph, the various FFs were optimized targeting a training set of molecules. Accordingly, each FF may be anticipated to reproduce the energies and forces of the molecules in the respective training sets with reasonable accuracy. However, the question of transferability remains such that how accurate is the treatment of a molecule not in the training set originally used to optimize the FF. While a full investigation to address this issue represents a significant challenge, two examples of the transferability of selected FFs will be given targeting QM dihedral potential energy scans of dimethyltryptamine and dimethylamino[1,4]diazepine which are analogues of serotonin and clozapine.

Figure 2 shows dihedral potential energy surface generated by MMFF, SwissParam, CGenFF, and, in 2b, CGenFF after additional parameter optimization along with QM data obtained at the MP2/6-31G* level using the Gaussian03 program[192]. Dihedral angles, shown as curved arrows in the figure, were rotated in 15° increments and the geometries were optimized at each step. MMFF, SwissParam and CGenFF parameters were input into CHARMM and geometries were minimized to an RMS force of < 10^{-6} kcal/mol/Å. With dimethyltryptamine, MMFF and SwissParam underestimated the height of energy barriers and CGenFF had a different peak shape at 45° and 330°. Since SwissParam adopted parameters from MMFF, its energy surfaces were similar to that of MMFF, though not identical. This is due to the different representations of the nonbond terms, which contribute to the energy surfaces and emphasize the problem with mixing parameters from different force fields. Overall, the shape of the surfaces for dimethyltryptamine are acceptable for all three FFs. However, results for dimethylamino[1,4]diazepine emphasize that caution needs to be taken as when transferring parameters to new compounds (Figure 2b and Table 2). The MMFF energy surface has local minima around 150° and 270° which will lead to errors in conformational sampling; similar problems are present with the parameters generated by SwissParam. When energies as a function of conformation are incorrectly represented the conformations selected from the sampling approach will typically be incorrect or the probabilities of those conformations improperly represented. Initial parameters from CGenFF showed poor agreement with the QM PES, but the FF is in significantly better agreement following optimization of selected dihedral angle parameters. The results with all

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

the tested FFs indicate the limited ability to transfer parameters to new molecules. An advantage of CGenFF is that penalty values are provided for each parameter assignment. This alerts the user to possible limitations in the FF, such as occur in dimethylamino[1,4]diazepine's conformational energy. In such cases validation of the parameters and additional optimization should be performed as required. Efforts to extend the ParamChem web server to include an automated interface for parameter validation and optimization are ongoing (K. Vanommeslaghe, S. Pamidighantam, M. Sheetz and A.D. MacKerell, Jr. Work in progress).

5. Conformationally sampled pharmacophore (CSP)

Leveraging the ability to perform extensive conformational sampling of small molecules using a properly optimized FF for the ligands of interest facilitated the development of a novel approach in our laboratory, the conformationally sampled pharmacophore (CSP)[113-115, 159, 193, 194]. CSP is a LBDD approach based on extensive sampling of conformational space, under the assumption that such sampling will lead to inclusion of the bioactive conformation being sampled despite that conformation not being known. The use of all accessible conformations in the CSP approach allows for probability distributions of different geometric features and/or physical properties to be determined, as shown in Figure 1 for Leu-Enkephalin. As 4D-QSAR uses occupancy of lattice points on a 3D grid by conformations of the ligands being studied as descriptors, CSP uses probability distributions of pharmacophoric features (e.g. distances, angles and dihedrals) as descriptors for model development. The use of all accessible conformations in CSP has allowed it to be applied successfully to highly flexible molecules such as peptidic opioids[113-115] and bile acids[159, 193, 194]. A strength of the method is the ability to connect the pharmacophore models to molecular details of the ligands being studied thereby facilitating physical interpretation of the models and applying the knowledge for ligand optimization, including rational drug design. By including all conformations, CSP can often recognize subtle differences among structurally similar compounds as well common pharmacophore features among diverse compounds. By using the quantitative overlap of pharmacophoric feature probability distributions of different ligands rather than conformations of the ligands themselves during model fitting, the molecular alignment problem is eliminated. However, once a suitable model is developed the conformational distributions from the MD simulations used in CSP model development may be used to guide possible superposition thereby identifying the biologically relevant conformations of the ligands. For example, CSP model for δ-opioid receptor ligands demonstrated how flexible peptidic opioids can be superimposed with non-peptidic opioids[114].

For proper CSP modelling, accurate FF and efficient conformation searching are needed. To date, MMFF and CGenFF have been used successfully. Conformational sampling has used extended MD simulations alone at both room and high temperatures[113-115] and temperature REXMD simulations in implicit solvent[114, 159, 194]. Figure 3 shows the general procedure used in CSP modelling. Once conformations are pre-enumerated for the training set of compounds, which may be performed using any of the above sampling methods, pharmacophore development is performed in an automated, computationally feasible fashion. As for identification of pharmacophore features, aromatic ring, ionizable groups, or hydrogen bond donors and acceptors can be identified, the associated probability distributions between the features calculated and, based on the extent of overlap of those distributions, the various combination of overlaps iteratively regressed against biological data, with those features yielding the best correlation with experimental data used for further model development. Notably, the CSP method can be readily combined with physicochemical descriptors to further facilitate model development[159, 194].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6. Conclusions

Presented is an overview of computational ligand-based drug design approaches currently in use in rational drug design. Over the last 2-3 decades, a large number of methods have been developed and many of these have been implemented in readily accessible software packages. While this convenience is important for utilization of these methods, it is essential that users understand the assumptions and limitations in those methods allowing for decision on the suitability of the methods for a given project, what kind of knowledge one can obtain through the study, and which aspect is the limiting factor with respect to producing accurate SAR models. As many LBDD approaches require extensive sampling of conformational space emphasis in this article was placed on recent FF development and the use of MD simulations and related techniques for conformational sampling.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by National Institutes of Health grants DK67530, DA13583, and DA19634, Accelrys Inc., and the University of Maryland Computer-Aided Drug Design Center for computational resources.

References

1. Leach, AR.; Gillet, VJ. An Introduction to Chemoinformatics. Springer; 2007.
2. Merz, KM.; Ringe, D.; Reynolds, CH. Drug Design: Structure- and Ligand-Based Approaches. Cambridge University Press; 2010.
3. Wermuth, CG. The practice of medicinal chemistry. Elsevier/Academic Press; 2008.
4. Young, DC. Computational drug design: a guide for computational and medicinal chemists. Wiley-Interscience; 2009.
5. Andricopulo AD. Current Topics in Medicinal Chemistry. 2009; 9:754–754. [PubMed: 19754392]
6. Kapetanovic IM. Chemico-Biological Interactions. 2008; 171:165–176. [PubMed: 17229415]
7. Overington JP, Al-Lazikani B, Hopkins AL. Nat Rev Drug Discov. 2006; 5:993–996. [PubMed: 17139284]
8. Gane PJ, Dean PM. Current Opinion in Structural Biology. 2000; 10:401–404. [PubMed: 10981625]
9. Jhoti, H.; Leach, AR. Structure-based drug discovery. Springer; 2007.
10. Marrone TJ, Briggs JM, McCammon JA. Annual Review of Pharmacology and Toxicology. 1997; 37:71–90.
11. Costanzi S, Tikhonova IG, Harden TK, Jacobson KA. Journal of Computer-Aided Molecular Design. 2009; 23:747–754. [PubMed: 18483766]
12. Tropsha A, Wang SX. Ernst Schering Foundation Symposium Proceedings. 2006:49–73. [PubMed: 17703577]
13. Fischer E. Berichte der deutschen chemischen Gesellschaft. 1894; 27:2985–2993.
14. Koshland DE. Proceedings of the National Academy of Sciences of the United States of America. 1958; 44:98–104. [PubMed: 16590179]
15. Boehr DD, Nussinov R, Wright PE. Nature Chemical Biology. 2009; 5:789–796.
16. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Protein Science: A Publication of the Protein Society. 2000; 9:10–19. [PubMed: 10739242]
17. Ma B, Kumar S, Tsai CJ, Nussinov R. Protein Engineering. 1999; 12:713–720. [PubMed: 10506280]
18. Tsai CJ, Kumar S, Ma B, Nussinov R. Protein Science: A Publication of the Protein Society. 1999; 8:1181–1190. [PubMed: 10386868]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

19. Tsai CJ, Ma B, Nussinov R. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96:9970–9972. [PubMed: 10468538]

20. Free SM, Wilson JW. Journal of Medicinal Chemistry. 1964; 7:395–399. [PubMed: 14221113]

21. Hansch C, Fujita T. Journal of the American Chemical Society. 1964; 86:1616–1626.

22. Hansch, C.; Selassie, C. Quantitative structure-activity relationship-a historical perspective and the future. Elsevier; Oxford: 2007.

23. Ekins S, Boulanger B, Swaan PW, Hupcey MAZ. Molecular Diversity. 2002; 5:255–275. [PubMed: 12549676]

24. Ekins S, Ecker GF, Chiba P, Swaan PW. Xenobiotica; the Fate of Foreign Compounds in Biological Systems. 2007; 37:1152–1170.

25. Winkler, DA. Molecular analysis and genome discovery. John Wiley and Sons; 2004.

26. Gedeck P, Lewis RA. Current Opinion in Drug Discovery & Development. 2008; 11:569–575.

27. Scior T, Medina-Franco JL, Do QT, Martínez-Mayorga K, Yunes Rojas JA, Bernard P. Current Medicinal Chemistry. 2009; 16:4297–4313. [PubMed: 19754417]

28. Martin, YC.; John, BT.; David, JT. Comprehensive Medicinal Chemistry II. Elsevier; Oxford: 2007. p. 119-147.

29. Güner, OF. Pharmacophore perception, development, and use in drug design. Internat'l University Line; 1999.

30. Wermuth, CG. Pharmacophores and pharmacophore searches. Wiley-VCH; 2006.

31. Leach AR, Gillet VJ, Lewis RA, Taylor R. Journal of Medicinal Chemistry. 2010; 53:539–558. [PubMed: 19831387]

32. Wolber G, Seidel T, Bendix F, Langer T. Drug Discovery Today. 2008; 13:23–29. [PubMed: 18190860]

33. Gillet, VJ.; Willett. Compound selection using measures of similarity and dissimilarity In Comprehensive medicinal chemistry II. Elsevier; 2007.

34. Willett P. Drug Discovery Today. 2006; 11:1046–1053. [PubMed: 17129822]

35. Geppert H, Vogt M, Bajorath Jr. Journal of Chemical Information and Modeling. 50:205–216. [PubMed: 20088575]

36. Shanmugasundaram K, Rigby AC. Combinatorial Chemistry & High Throughput Screening. 2009; 12:984–999. [PubMed: 20025564]

37. Leach, AR.; John, BT.; David, JT. Comprehensive Medicinal Chemistry II. Elsevier; Oxford: 2007. p. 87-118.

38. Zhou T, Huang D, Caflisch A. Current Topics in Medicinal Chemistry. 10:33–45. [PubMed: 19929831]

39. Liu P, Long W. International Journal of Molecular Sciences. 10:1978–1998. [PubMed: 19564933]

40. Yap CW, Li H, Ji ZL, Chen YZ. Mini Reviews in Medicinal Chemistry. 2007; 7:1097–1107. [PubMed: 18045213]

41. Güner O, Clement O, Kurogi Y. Current Medicinal Chemistry. 2004; 11:2991–3005. [PubMed: 15544485]

42. Gasteiger, J.; Engel, DT. Chemoinformatics: a textbook. Wiley-VCH; 2003.

43. Weininger D. Journal of Chemical Information and Computer Sciences. 1988; 28:31–36.

44. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. Journal of Chemical Information and Computer Sciences. 1997; 37:71–79.

45. Durant JL, Leland BA, Henry DR, Nourse JG. Journal of Chemical Information and Computer Sciences. 2002; 42:1273–1280. [PubMed: 12444722]

46. Bonchev, D.; Rouvray, DH. Chemical graph theory: introduction and fundamentals. Taylor & Francis; 1991.

47. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. Wiley-VCH; 2002.

48. Helguera AM, Combes RD, González MP, Cordeiro MNDS. Current Topics in Medicinal Chemistry. 2008; 8:1628–1655. [PubMed: 19075771]

49. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Advanced Drug Delivery Reviews. 2001; 46:3–26. [PubMed: 11259830]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

50. González MP, Terán C, Saíz-Urra L, Teijeira M. Current Topics in Medicinal Chemistry. 2008; 8:1606–1627. [PubMed: 19075770]

51. Miller, AJ. Subset selection in regression. CRC Press; 2002.

52. Draper, NR.; Smith, H. Applied regression analysis. Wiley; 1981.

53. Mercader AG, Duchowicz PR, Fernández FM, Castro EA. Chemometrics and Intelligent Laboratory Systems. 2008; 92:138–144.

54. Devillers, J. Genetic algorithms in molecular modeling. Academic Press; 1996.

55. Pintore M, van de Waterbeemd H, Piclin N, Chrétien JR. European Journal of Medicinal Chemistry. 2003; 38:427–431. [PubMed: 12750031]

56. Obrezanova O, Csányi G, Gola JMR, Segall MD. Journal of Chemical Information and Modeling. 2007; 47:1847–1857. [PubMed: 17602549]

57. Rogers D, Hopfinger AJ. Journal of Chemical Information and Computer Sciences. 1994; 34:854–866.

58. Geladi P, Kowalski B. Analytica Chimica Acta. 1986; 185:1–17.

59. Rosipal R, Krämer N. Subspace, Latent Structure and Feature Selection. 2006:34–51.

60. Wold, H. Partial least squares. Wiley; New York: 1985.

61. Wold S, Ruhe A, Wold H, Dunn WJ Iii. SIAM Journal on Scientific and Statistical Computing. 1984; 5:735–743.

62. Jolliffe. Principal Component Analysis. Springer-Verlag; New York: 2002.

63. Wold S, Esbensen K, Geladi P. Chemometrics and Intelligent Laboratory Systems. 1987; 2:37–52.

64. Michielan L, Moro S. Journal of Chemical Information and Modeling. 50:961–978. [PubMed: 20527756]

65. Sakiyama Y. Expert Opinion on Drug Metabolism & Toxicology. 2009; 5:149–169. [PubMed: 19239395]

66. Si HZ, Wang T, Zhang KJ, Hu ZD, Fan BT. Bioorganic & Medicinal Chemistry. 2006; 14:4834–4841. [PubMed: 16580211]

67. Friedman JH, Stuetzle W. Journal of the American Statistical Association. 1981; 76:817–823.

68. Kulkarni AJ, Jayaraman VK, Kulkarni BD. Combinatorial Chemistry & High Throughput Screening. 2009; 12:440–450. [PubMed: 19442070]

69. Trevor, W Heritage; David, R Lowis. Rational Drug Design. American Chemical Society; 1999. p. 212-225.

70. Cartmell J, Enoch S, Krstajic D, Leahy DE. Journal of Computer-Aided Molecular Design. 2005; 19:821–833. [PubMed: 16416245]

71. Wong WW, Burkowski FJ. Journal of Cheminformatics. 2009; 1:4–4. [PubMed: 20142987]

72. Golbraikh A, Tropsha A. Journal of Molecular Graphics & Modelling. 2002; 20:269–276. [PubMed: 11858635]

73. Kohavi R. IJCAI. 1995:1137–1145.

74. Clark RD. Current Topics in Medicinal Chemistry. 2009; 9:791–810. [PubMed: 19754395]

75. Cross S, Cruciani G. Drug Discovery Today. 15:23–32. [PubMed: 19150413]

76. Verma J, Khedkar VM, Coutinho EC. Current Topics in Medicinal Chemistry. 10:95–115. [PubMed: 19929826]

77. Cramer RD, Patterson DE, Bunce JD. Journal of the American Chemical Society. 1988; 110:5959–5967.

78. Klebe G, Abraham U, Mietzner T. Journal of Medicinal Chemistry. 1994; 37:4130–4146. [PubMed: 7990113]

79. Andrade CH, Pasqualoto KFM, Ferreira EI, Hopfinger AJ. Molecules (Basel, Switzerland). 15:3281–3294.

80. Klein CD, Hopfinger AJ. Pharmaceutical Research. 1998; 15:303–311. [PubMed: 9523319]

81. Polanski J. Current Medicinal Chemistry. 2009; 16:3243–3257. [PubMed: 19548875]

82. Vedani A, Dobler M. Journal of Medicinal Chemistry. 2002; 45:2139–2149. [PubMed: 12014952]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

83. Vedani A, Dobler M, Lill MA. Journal of Medicinal Chemistry. 2005; 48:3700–3703. [PubMed: 15916421]

84. Cruciani G, Pastor M, Guba W. European Journal of Pharmaceutical Sciences: Official Journal of the European Federation for Pharmaceutical Sciences. 2000; 11 2:S29–39. [PubMed: 11033425]

85. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. Journal of Medicinal Chemistry. 2000; 43:3233–3243. [PubMed: 10966742]

86. G. D. G. Hawkins, D. J.;Lynch, G. C.;Chambers, C. C.;Rossi, I.;Storer, J. W.;Li, J.;Zhu, T.;Thompson, J. D.;Winget, P.;Rinaldi, D.;Liotard, D. A.;Cramer, C. J.;Truhlar, D. G., University of Minnesota: Minneapolis, MN, 2002.

87. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA, James AB. Academic Press. 1998:385–395.

88. Lemmen C, Lengauer T. Journal of Computer-Aided Molecular Design. 2000; 14:215–232. [PubMed: 10756477]

89. Brint AT. J Chem Inf Comput Sci. 1987; 27:152–158.

90. Jones G, Willett P, Glen RC. Journal of Computer-Aided Molecular Design. 1995; 9:532–549. [PubMed: 8789195]

91. Lemmen C, Lengauer T, Klebe G. Journal of Medicinal Chemistry. 1998; 41:4502–4520. [PubMed: 9804690]

92. Schneider G, Schneider P, Renner S. QSAR & Combinatorial Science. 2006; 25:1162–1171.

93. Chemical Computing Group. http://www.chemcomp.com

94. Acclerys. http://accelrys.com

95. PubChem. http://pubchem.ncbi.nlm.nih.gov/

96. Irwin JJ, Shoichet BK. Journal of Chemical Information and Modeling. 2005; 45:177–182. [PubMed: 15667143]

97. ChEMBL. http://www.ebi.ac.uk/chembldb

98. Villoutreix BO, Renault N, Lagorce D, Sperandio O, Montes M, Miteva MA. Current Protein & Peptide Science. 2007; 8:381–411. [PubMed: 17696871]

99. Barnard JM, Downs GM. Journal of Chemical Information and Computer Sciences. 1992; 32:644–649.

100. Cerchietti LC, Ghetu AF, Zhu X, Da Silva GF, Zhong S, Matthews M, Bunting KL, Polo JM, Farès C, Arrowsmith CH, Yang SN, Garcia M, Coop A, MacKerell AD, Privé GG, Melnick A. Cancer Cell. 17:400–411. [PubMed: 20385364]

101. Martin YC, Kofron JL, Traphagen LM. Journal of Medicinal Chemistry. 2002; 45:4350–4358. [PubMed: 12213076]

102. Matter H. Journal of Medicinal Chemistry. 1997; 40:1219–1229. [PubMed: 9111296]

103. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M. Journal of Medicinal Chemistry. 2006; 49:6802–6810. [PubMed: 17154510]

104. Perola E, Charifson PS. Journal of Medicinal Chemistry. 2004; 47:2499–2510. [PubMed: 15115393]

105. Agrafiotis DK, Gibbs AC, Zhu F, Izrailev S, Martin E. Journal of Chemical Information and Modeling. 2007; 47:1067–1086. [PubMed: 17411028]

106. Foloppe N, Chen IJ. Current Medicinal Chemistry. 2009; 16:3381–3413. [PubMed: 19515013]

107. Nicklaus MC, Wang S, Driscoll JS, Milne GW. Bioorganic & Medicinal Chemistry. 1995; 3:411–428. [PubMed: 8581425]

108. Günther S, Senger C, Michalsky E, Goede A, Preissner R. BMC Bioinformatics. 2006; 7:293–293. [PubMed: 16764718]

109. Kirchmair J, Laggner C, Wolber G, Langer T. Journal of Chemical Information and Modeling. 2005; 45:422–430. [PubMed: 15807508]

110. Kirchmair J, Wolber G, Laggner C, Langer T. Journal of Chemical Information and Modeling. 2006; 46:1848–1861. [PubMed: 16859316]

111. Loferer MJ, Kolossváry I, Aszódi A. Journal of Molecular Graphics & Modelling. 2007; 25:700–710. [PubMed: 16815716]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

112. Moock TE, Henry DR, Ozkabak AG, Alamgir M. Journal of Chemical Information and Computer Sciences. 1994; 34:184–189.

113. Bernard D, Coop A, MacKerell AD. Journal of the American Chemical Society. 2003; 125:3101–3107. [PubMed: 12617677]

114. Bernard D, Coop A, MacKerell AD. Journal of Medicinal Chemistry. 2007; 50:1799–1809. [PubMed: 17367120]

115. Bernard D, Coop A, MacKerell AD Jr. Journal of Medicinal Chemistry. 2005; 48:7773–7780. [PubMed: 16302816]

116. Beusen DD, Berkley Shands EF, Karasek SF, Marshall GR, Dammkoehler RA. Journal of Molecular Structure: THEOCHEM. 1996; 370:157–171.

117. Lipton M, Still WC. Journal of Computational Chemistry. 1988; 9:343–355.

118. Garland, R Marshall; Barry, DC.; Heinz, E Bosshard; Richard, A Dammkoehler; Deborah, A Dunn. Computer-Assisted Drug Design. AMERICAN CHEMICAL SOCIETY; 1979. p. 205-226.

119. Li Z, Scheraga HA. Proceedings of the National Academy of Sciences of the United States of America. 1987; 84:6611–6615. [PubMed: 3477791]

120. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. The Journal of Chemical Physics. 1953; 21:1087–1087.

121. Metropolis N, Ulam S. Journal of the American Statistical Association. 1949; 44:335–341. [PubMed: 18139350]

122. Smellie A, Teig SL, Towbin P. Journal of Computational Chemistry. 1995; 16:171–187.

123. Cvijovicacute D, Klinowski J. Science. 1995; 267:664–666. [PubMed: 17745843]

124. Nair N, Goodman JM. Journal of Chemical Information and Computer Sciences. 1998; 38:317–320.

125. Parrill AL. Drug Discovery Today. 1996; 1:514–521.

126. Brooks C, Case DA. Chemical Reviews. 1993; 93:2487–2502.

127. Brooks, CL.; Karplus, M.; Pettitt, BM. Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics. John Wiley and Sons; 1990.

128. Karplus M, Petsko GA. Nature. 1990; 347:631–639. [PubMed: 2215695]

129. McCammon, JA.; Harvey, SC. Dynamics of Proteins and Nucleic Acids. Cambridge University Press; 1988.

130. van Gunsteren WF, Berendsen HJC. Angewandte Chemie International Edition in English. 1990; 29:992–1023.

131. Frenkel, D.; Smit, B. Understanding molecular simulation: from algorithms to applications. Academic Press; 2002.

132. Rapaport, DC. The art of molecular dynamics simulation. Cambridge University Press; 2004.

133. Verlet L. Physical Review. 1967; 159:98–98.

134. Potter DE. Computational Physics. 1988 Books on Demand.

135. Hockney RW. Methods Comput Phys. 1970; 9:136–211.

136. Tuckerman ME, Martyna GJ. The Journal of Physical Chemistry B. 2000; 104:159–178.

137. Kirkpatrick S, Gelatt CD, Vecchi MP. Science. 1983; 220:671–680. [PubMed: 17813860]

138. Chipot, C.; Pohorille, A. Free energy calculations: theory and applications in chemistry and biology. Springer; 2007.

139. Mitsutake A, Sugita Y, Okamoto Y. Biopolymers. 2001; 60:96–123. [PubMed: 11455545]

140. Okamoto Y. Journal of Molecular Graphics & Modelling. 2004; 22:425–439. [PubMed: 15099838]

141. Sugita Y, Okamoto Y. Chemical Physics Letters. 1999; 314:141–151.

142. Laio A, Parrinello M. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:12562–12566. [PubMed: 12271136]

143. Micheletti C, Laio A, Parrinello M. Physical Review Letters. 2004; 92

144. Hamelberg D, Mongan J, McCammon JA. The Journal of Chemical Physics. 2004; 120:11919–11929. [PubMed: 15268227]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

145. Kong X, Brooks CL. The Journal of Chemical Physics. 1996; 105:2414–2414.

146. Hermans J, Yun RH, Anderson AG. Journal of Computational Chemistry. 1992; 13:429–442.

147. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Journal of Computational Chemistry. 1992; 13:1011–1021.

148. Okur A, Wickstrom L, Layten M, Geney R, Song K, Hornak V, Simmerling C. Journal of Chemical Theory and Computation. 2006; 2:420–433.

149. Sindhikara D, Meng Y, Roitberg AE. The Journal of Chemical Physics. 2008; 128:024103–024103. [PubMed: 18205439]

150. Fukunishi H, Watanabe O, Takada S. The Journal of Chemical Physics. 2002; 116:9058–9067.

151. Sugita Y, Kitao A, Okamoto Y. The Journal of Chemical Physics. 2000; 113:6051, 6042–6051, 6042.

152. Sugita Y, Okamoto Y. cond-mat/0009119. 2000

153. Knight JL, Brooks CL. Journal of Computational Chemistry. 2009; 30:1692–1700. [PubMed: 19421993]

154. Laio A, Gervasio FL. Reports on Progress in Physics. 2008; 71:126601–126601.

155. Zheng L, Chen M, Yang W. Proceedings of the National Academy of Sciences. 2008; 105:20227–20232.

156. Zheng L, Chen M, Yang W. The Journal of Chemical Physics. 2009; 130:234105–234105. [PubMed: 19548709]

157. Sanbonmatsu KY, García AE. Proteins. 2002; 46:225–234. [PubMed: 11807951]

158. Su L, Cukier RI. The Journal of Physical Chemistry B. 2007; 111:12310–12321. [PubMed: 17918879]

159. González PM, Acharya C, MacKerell AD, Polli JE. Pharmaceutical Research. 2009; 26:1665–1678. [PubMed: 19384469]

160. Lee S, Chen M, Yang W, Richards NGJ. Journal of the American Chemical Society. 132:7252–7253. [PubMed: 20446682]

161. Provasi D, Bortolato A, Filizola M. Biochemistry. 2009; 48:10020–10029. [PubMed: 19785461]

162. Feig M, Brooks CL. Current Opinion in Structural Biology. 2004; 14:217–224. [PubMed: 15093837]

163. Onufriev A, Bashford D, Case DA. Proteins. 2004; 55:383–394. [PubMed: 15048829]

164. Koehl P. Current Opinion in Structural Biology. 2006; 16:142–151. [PubMed: 16540310]

165. Honig B, Nicholls A. Science (New York, NY). 1995; 268:1144–1149.

166. Jackson, JD. Classical electrodynamics. Wiley; 1999.

167. Still C, Tempczyk A, Hawley R, Hendrickson T. Journal of the American Chemical Society. 1990; 112:6127–6129.

168. Lee MS, Feig M, Salsbury FR, Brooks CL. Journal of Computational Chemistry. 2003; 24:1348–1356. [PubMed: 12827676]

169. Im W, Lee MS, Brooks CL. Journal of Computational Chemistry. 2003; 24:1691–1702. [PubMed: 12964188]

170. Brooks BR, Brooks CL, MacKerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. Journal of Computational Chemistry. 2009; 30:1545–1614. [PubMed: 19444816]

171. MacKerell, ADJ.; Brooks, B.; Brooks, CLI.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. John Wiley & Sons; Chichester: 1998.

172. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. Journal of Computational Chemistry. 2005; 26:1668–1688. [PubMed: 16200636]

173. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. Journal of the American Chemical Society. 1995; 117:5179–5197.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

174. MacKerell AD, Feig M, Brooks CL. Journal of Computational Chemistry. 2004; 25:1400–1415. [PubMed: 15185334]

175. Plimpton S. J Comput Phys. 1995; 117:1–19.

176. LAMMPS. http://lammps.sandia.gov

177. Hess B, Kutzner C, van der Spoel D, Lindahl E. Journal of Chemical Theory and Computation. 2008; 4:435–447.

178. Jorgensen WL, Maxwell DS, Tirado-Rives J. Journal of the American Chemical Society. 1996; 118:11225–11236.

179. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. The Journal of Physical Chemistry B. 2001; 105:6474–6487.

180. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, MacKerell AD. Journal of Computational Chemistry. 2009; 31:671–690. [PubMed: 19575467]

181. MacKerell AD. Journal of Computational Chemistry. 2004; 25:1584–1604. [PubMed: 15264253]

182. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Journal of Computational Chemistry. 2004; 25:1157–1174. [PubMed: 15116359]

183. Allen FH. Acta Crystallographica Section B, Structural Science. 2002; 58:380–388.

184. Wang J, Wang W, Kollman PA, Case DA. Journal of Molecular Graphics & Modelling. 2006; 25:247–260. [PubMed: 16458552]

185. Halgren TA. Journal of Computational Chemistry. 1996; 17:616–641.

186. Halgren TA. Journal of Computational Chemistry. 1996; 17:490–519.

187. Lifson S, Hagler AT, Dauber P. Journal of the American Chemical Society. 1979; 101:5111–5121.

188. Allinger NL. Journal of the American Chemical Society. 1977; 99:8127–8134.

189. Allinger NL, Yuh YH, Lii JH. Journal of the American Chemical Society. 1989; 111:8551–8566.

190. Nevins N, Lii JH, Allinger NL. Journal of Computational Chemistry. 1996; 17:695–729.

191. Clark M, Cramer RD, Van Opdenbosch N. Journal of Computational Chemistry. 1989; 10:982–1012.

192. M. J. T. Frisch, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.;, Gaussian, Inc. Wallingford CT, 2004.

193. Rais R, Acharya C, MacKerell AD, Polli JE. Molecular Pharmaceutics. 2010; 7:2240–2254. [PubMed: 20939504]

194. Rais R, Acharya C, Tririya G, MacKerell AD, Polli JE. Journal of Medicinal Chemistry. 2010; 53:4749–4760. [PubMed: 20504026]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

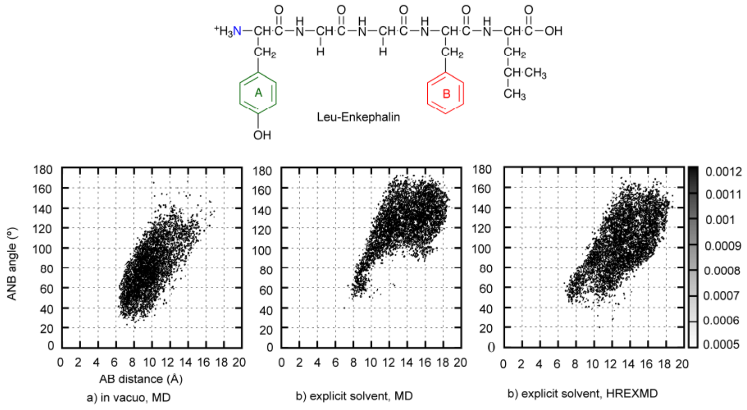

Figure 1.

2D-probability distribution of AB distance-ANB angle pair of Leu-Enkephalin. Pharmacophoric point A represents the centroid of the aromatic ring of tyrosine, B is the centroid of the aromatic ring of phenylalanine, and N is the basic nitrogen. a) through c) compares different sampling of conformational space by a) gas phase MD, b) explicit solvent MD and c) explicit solvent HREMD. Simulation details are in the supporting information.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

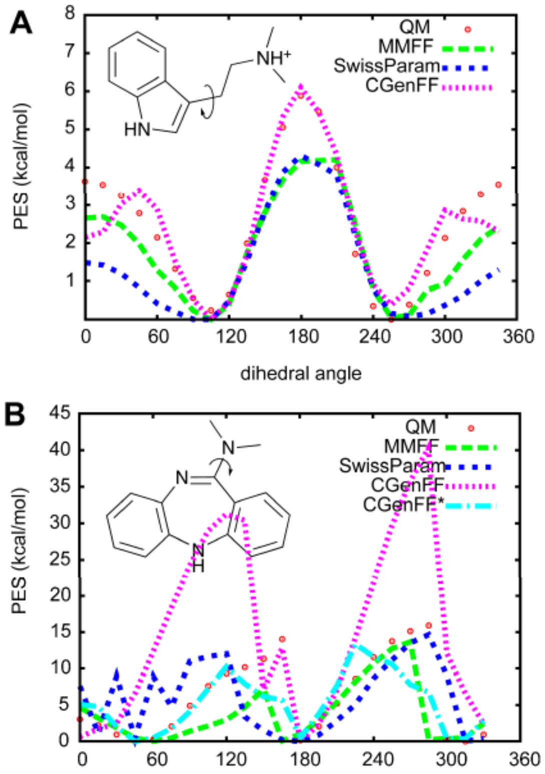

Figure 2.

Comparison of conformational energy surface. A) is potential energy surface (PES) of dimethyltryptamine and B) shows that of dimethylamino-dibenzo[1,4]diazepine.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

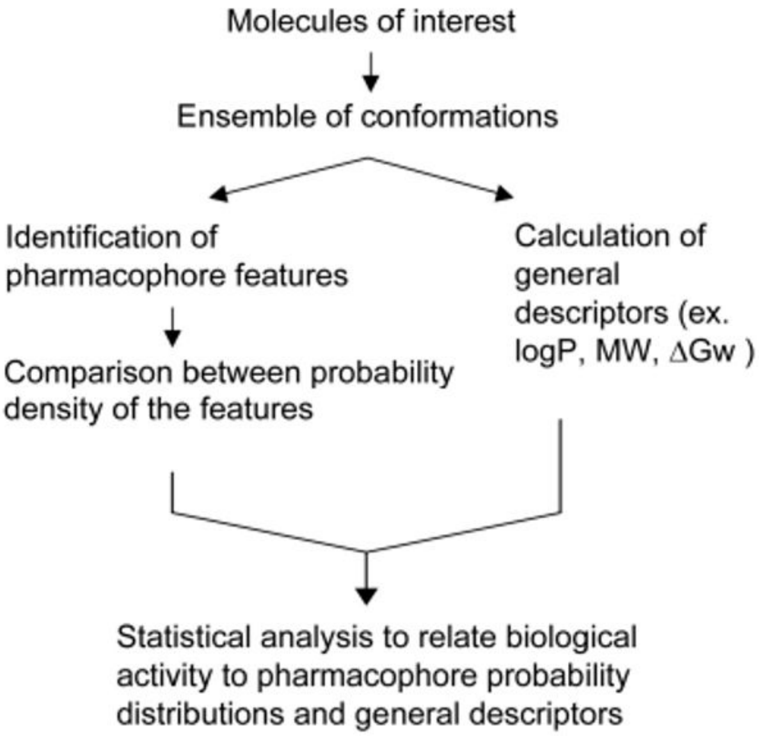
Molecules of interest
Ensemble of conformations
Identification of pharmacophore features
Comparison between probability density of the features
Calculation of general descriptors (ex. logP, MW, ΔGw)
Statistical analysis to relate biological activity to pharmacophore probability distributions and general descriptors

Figure 3. Schematic diagram of CSP procedures

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1
Overview of nD-QSAR approaches ($3 \le n \le 7$)

method	description
CoMFA	Probes placed on grid points in the 3D field around a molecule experience an interaction energy with the ligands that defines the molecular shape and electrostatic properties in the surrounding environment.
CoMSIA	It expands CoMFA by including hydrophobic and hydrogen bonding contributions and calculates how these contributions are similar between molecules.
GRIND	It eliminates alignment-dependency by using distances between 3D grid points. Highly relevant regions among a set of molecules are selected as nodes and the intensity of molecular interaction field at those nodes are used as descriptors. The program ALMOND provides tools to compute, analyze, and interpret the GRIND.
VolSurf	Information on 3D grid voxels (shape, electrostatic, volume) are compressed into 2D numerical descriptors by image analysis tools.
4D-QSAR	Multiple conformations in a grid box generate the occupancies at grid points, with those occupancies used as the descriptors.
5D-QSAR	Multiple hypothetical binding pockets are generated around ligands based on a 3D grid and the receptor models are evolved by GA with the most favorable binding pocket model evaluated by relative free energy of ligand binding.
6D-QSAR	It includes optimization of structures in aqueous solution and calculates solvation energy and charges by semi-empirical QM method, AMSOL[86]. Ligands' arrangement in pseudo-binding pocket is determined by MC simulation.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Root mean square deviation from QM potential energy surface

	RMSD	
FF	**dimethyltryptamine**	**dimethylamino-dibenzo[1,4]diazepine**
MMFF	0.87	5.04
SwissParam	1.42	5.42
CGenFF	0.62	12.84
CGenFF*	n.d.	3.59