Recent Advancements in Computing Reliable Binding Free Energies in Drug Discovery Projects

by

N. Arul Murugan, Vasanthanathan Poongavanam, U Deva Priyakumar

in

Structural Bioinformatics:

Report No: IIIT/TR/2019/-1

Centre for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2019

Recent Advancements in Computing Reliable Binding Free Energies in Drug Discovery Projects

N. Arul Murugan, Vasanthanathan Poongavanam and U. Deva Priyakumar

Abstract In recent times, our healthcare system is being challenged by many drug-resistant microorganisms and ageing-associated diseases for which we do not have any drugs or drugs with poor therapeutic profile. With pharmaceutical technological advancements, increasing computational power and growth of related biomedical fields, there have been dramatic increase in the number of drugs approved in general, but still way behind in drug discovery for certain class of diseases. Now, we have access to bigger genomics database, better biophysical methods, and knowledge about chemical space with which we should be able to easily explore and predict synthetically feasible compounds for the lead optimization process. In this chapter, we discuss the limitations and highlights of currently available computational methods used for protein–ligand binding affinities estimation and this includes force-field, ab initio electronic structure theory and machine learning approaches. Since the electronic structure-based approach cannot be applied to systems of larger length scale, the free energy methods based on this employ certain approximations, and these have been discussed in detail in this chapter. Recently, the methods based on electronic structure theory and machine learning approaches also are successfully being used to compute protein–ligand binding affinities and other pharmacokinetic and pharmacodynamic properties and so have greater potential to take forward computer-aided drug discovery to newer heights.

N. A. Murugan (✉)
Department of Theoretical Chemistry and Biology, School of Engineering
Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology,
Stockholm, Sweden
e-mail: murugan@kth.se

V. Poongavanam
Department of Physics, Chemistry, Pharmacy, University of Southern Denmark,
Campusvej 55, DK-5230 Odense M, Denmark

U. D. Priyakumar
CCNSB, International Institute of Information Technology,
Gachibowli 500032, Hyderabad, India

C. G. Mohan (ed.), *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*, Challenges and Advances in Computational Chemistry
and Physics 27, https://doi.org/10.1007/978-3-030-05282-9_7

Keywords Computational drug discovery · Free energy of binding
Hybrid QM/MM · QM fragmentation · Binding affinity · Pharmacokinetic
(PK) properties · Machine learning approach

Abbreviations

FMO	Fragment molecular orbital
MAO-B	Monoamine oxidase B
MM-GBSA	Molecular mechanics–Generalized Born Surface Area
MM-PBSA	Molecular mechanics–Poisson–Boltzmann Surface Area
PD	Pharmacodynamic
PK	Pharmacokinetic
QM/MM	Quantum mechanics/molecular mechanics

1 Introduction: Drugs and Targets

Disease can be defined as an abnormal condition that alters the function or behaviour of an organism and this can be caused by different factors, i.e. internally e.g. due to the presence of disease-causing genes or due to external factors. Externally, disease may be caused due to malnutrition or subjecting a human to severe external conditions such as exposing to radiation or pollution or microbial infections or severe physiological conditions which leads to damage or malfunctioning of body machineries. Thanks to genomic analysis of normal and diseased persons, we know that the protein profile appears quite different in these two cases and by targeting the biomacromolecules expressed in the diseased state, and we can develop methods to arrest the progress of the disease. By comparative protein profiling of normal and diseased persons or by comparing the genomes of human and pathogenic micro-organisms, [1–3] we already know the information about the potential targets, but then the problem lies in identifying whether the aberrant expression of a certain biomacromolecule is the main cause of disease or may be a side product of another key process. Once the key target (protein or enzyme) is identified, primay task is to design small molecules that can modulate the target (this can be either of inhibitor, substrate, inducer). Subsequently, the active compound (also called hit molecule) is further optimized to pre-clinical candidate. The aim of lead optimization is not only to increase the potency, but also to reduce any off-target binding. In this chapter, we will discuss how to use computational approaches not only to identify small molecules that can inhibit or modulate the catalytic process of a key enzyme that is connected with disease, but also to understand the fundamental process of biomolecular recognition which assists in the lead optimization process in the drug discovery and development projects. The properties of the ligand to be optimized are binding affinity and specificity towards a key target biomacromolecule. These target molecules can be located within microorganism or within the host organism depending upon whether the disease belongs to infectious or autoimmune category disease, respectively.

2 Optimization of Drug-Likeness

In addition to binding affinity and specificity, there are certain other properties which are to be optimized for an effective drug i.e. low toxic with improved potency and orally bioavailable for conventional dosage forms. These properties are absorption (A), distribution (D), metabolism (M), excretion (E) and toxicity (T), and they collectively are called ADMET or pharmacokinetic (PK) properties. The properties in general refer to kinetic behaviour of drugs within body and give information about the timescale required for the drug to reach the potential target and lifetime within host organism before removal through excretion (this can be shortly described as “what the body does to a drug?”). The optimization of potency (binding affinity) and then the subsequent optimization of pharmacokinetic behaviour have been the major contributing factors for the failures at the phase II and phase III clinical trials [4–6]. So, it is necessary simultaneously to optimize the potency along with ADMET properties [7]. There are also other properties that are essential for oral bioavailability such as solubility and transport properties like membrane permeability (both cellular and across blood-brain barrier).

Overall, it is apparent that drug design is challenging as we need to optimize several properties at the same time [3, 6]. In certain cases, optimizing one property may lead to unexpected changes in another property making this optimization very complex , and in those cases we need to compromise on certain properties and try to balance different properties for better PD and PK profile. For example, if the potency of a drug is superior/outstanding but then if it has very poor PK and PD properties, then one can use suitable drug delivery systems such that the drug is delivered to its target biomacromolecule. Given that drug design is an optimization process, it is inevitable to avoid the use of computers as they can be used effectively to speed up the overall process. But the only requirement is that we need to have accurately enough methods that can be written in a numerically solvable form and can reliably describe these processes involved in the drug association with a biological target and its pharmacokinetics [8]. In general, quantum mechanics is the fundamental theory which can be used to describe any atomic and molecular systems and their association process and their response to any external variables like heat, pressure and fields and to any change in physiological conditions such as pH and ionic strength. However, the complexity of the mathematics involved in solving the Schrödinger equation grows with powers of n which is number of basis functions used to describe one electron orbital used to build wave function of the molecule that describes its energy and all other properties. For example, the computational demand is at the power of three in the case density functional theory and can go to power of 5–8 for the theories which can treat the electron correlation more accurately. When the system size is comparably larger than the wavelength of light and when we are not interested in processes where the matter interacts with light or laser field, it is pragmatic to use classical mechanics to describe the molecular systems which involves relatively simple mathematics, i.e. solving Newton's equation of motion to describe the interaction within system and their association with other systems and to model their time evolution and their response to external thermodynamic variables like temperature and pressure. As per classical mechanics, once we have the force-field information for a system, its entire future and past can be predicted by solving equation of motion. Force-fields can be developed by using various available structure databases and thermodynamics data. In this chapter, We briefly cover available force-field methods for computing the binding affinity in order to rank protein–ligand complexes in drug discovery and design. In addition, briefly discuss their limitations and also present the recent advancement in computational modelling approaches based on quantum mechanical theory and machine learning algorithm in a way suitable for drug discovery applications.

3 Free Energies Relevant to Describe Potency and Pharmacokinetics

Free energy is the key variable that dictates the structure of biomacromolecular complexes (protein–ligand protein, membrane–ligand, DNA–ligand etc.) and controls various molecular association and ligand transport processes. When there are many structures possible, the one with least free energy is the most stable one. Moreover, biomacromolecular or molecular association processes such as drug binding to receptor, protein–protein binding and drug transport involve the minimization of Gibbs free energy (ΔG). Any process that involves lowering of Gibbs free energy can proceed spontaneously. By calculating the free energy change, we can predict whether an association process is feasible or not. In the case of a drug, the most relevant aspect is to understand its binding affinity or potency towards a target biomacromolecule and its association with transport proteins like albumins [9] and metabolizing enzymes such as cytochrome P450s (CYP) [10–12] and with glycoproteins responsible for absorption. The ligand binding to target biomacromolecule, transport protein and metabolizing enzymes is dictated by the change in free energy of the ligand bound to these targets when compared to that in aqueous solution. Further, it is also necessary to understand whether the compounds will pass through certain cell membranes and also how well it will dissociate once it is taken through oral dose which is dependent on the physicochemical properties like lipophilicity [13] and aqueous solubility [14]. Schematic representation of various PK computational modeling is shown in Fig. 1, and free energy of relevance is provided in Table 1.

Computing the free energy of binding of the drug with a biological target and other targets (such as glycoproteins, albumins, cell membranes) that mediate the drug transport across the body to relevant target area is the main goal of any computational approaches. All these drug association-related processes and

Table 1 Computing various PK and PD properties and potency as a difference in free energy of the ligand in different environments

Property	Initial medium	Final medium	Free energyof relevance
Inhibition constant/binding affinity	Water	Target enzyme	$G_{\text{enzyme}} - G_{\text{water}}$
Absorption/distribution	Water	Glycoproteins/ albumin	$G_{\text{albumin}} - G_{\text{water}}$
Metabolism	Water	Cytochromes P450	$G_{\text{P450}} - G_{\text{water}}$
Permeability	Water	Membrane	$G_{\text{membrane}} - G_{\text{water}}$
Solubility	Crystalline	Water	$G_{\text{water}} - G_{\text{crystal}}$
Off-target binding	Water	Off-target (e.g. hERG)	$G_{\text{offtarget}} - G_{\text{water}}$

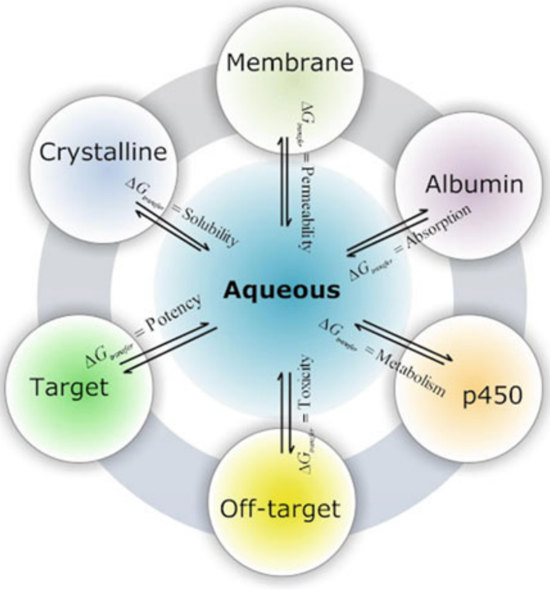

Membrane
Crystalline
Albumin
Aqueous
Target
p450
Off-target
$\Delta G_{transfer}$ = Permeability
$\Delta G_{transfer}$ = Solubility
$\Delta G_{transfer}$ = Absorption
$\Delta G_{transfer}$ = Potency
$\Delta G_{transfer}$ = Metabolism
$\Delta G_{transfer}$ = Toxicity

Fig. 1 Potency, pharmacodynamic property (such as solubility, permeability) and a few pharmacokinetic properties (such as drug absorption, distribution, metabolism and toxicity) are related to ΔG transfer, which is a free energy difference needed for driving a ligand from one environment to another

transport processes involve optimization (minimization in particular) of free energies, and the free energies associated with potency, bioavailability, drug absorption, distribution, metabolism, toxicity, solubility are listed in Table 1.

If we can calculate the free energy change for the drug to transfer from one medium to another, then we can predict how spontaneously this process will occur. As a conclusion, it can be deduced that the drug design involves calculations of free energy changes in two different media and the currently available methods are based on either force-fields or semi-empirical methods or electronic structure theory or combination of these. In this chapter, we will provide a brief outline of various computational methods available for computing free energy of binding of a ligand to a receptor which in turn can be used to optimize drug potency, absorption, distribution and metabolism. These methods are so general and also can be used to compute absolute free energies of a ligand or drug in crystalline or in different solvent environments making it feasible to predict other PD and PK properties like bioavailability, permeability and solubility [15, 16]. Very recently, machine learning approaches are also contributing to the computation of interaction energies of protein–ligand and protein–protein complexes and the progress in the use of such approaches for drug discovery projects will be discussed at the end.

4 Force-Field-Based Free Energies of Drug Target Binding

A force-field describes how the atomic and molecular systems behave at finite temperature, pressure and in a specific physiological condition or under any external fields. Force-fields have potential energy functions to explain the interactions of intermolecular and intramolecular degrees of freedom within a system. In particular, the former dictates the packing, relative orientation of molecules, while the internal geometry is dictated by the latter potentials. The currently available potential energy functions in different force-fields were parameterized using many experimental thermodynamic data and structural data [17]. For example, crystal structure database (CSD) can be easily utilized to get the information about the characteristic equilibrium radii of atoms which then dictate the size and overall structure of the materials. Similarly, the heat of vaporization can be used to get information about the well depths of the interaction potential which then gives information about transition temperatures from one phase to another. For convenience, interaction energies were modelled as the sum over pair potential where the pair potential itself is described using sum of Lennard-Jones (LJ) and electrostatic potential (refer to terms 7 and 8 of Eq. 1, respectively). As mentioned above, the two parameters of LJ potential, sigma and epsilon can be parameterized by using available structure information and thermodynamic data.

$$\begin{aligned} U_{\text{total}} = {} & \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{UB}} K_{\text{UB}}(S-S_0)^2 \\ & + \sum_{\text{dihedral}} K_x(1+\cos(cX-\delta)) + \sum_{\text{impropers}} K_{\text{impr}}(\varphi-\varphi_0)^2 \\ & + \sum_{\text{LJ}_{i\neq j}} \varepsilon_{ij}\left[\left(\frac{R_{\text{min}_{ij}}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\text{min}_{ij}}}{r_{ij}}\right)^{6}\right] + \sum_{\text{coulomb}} \frac{q_i q_j}{\varepsilon_i r'_{ij}} \end{aligned} \tag{1}$$

The LJ and electrostatic potentials describe only the intermolecular interactions, while it is not accounting for the structural changes in the molecule in the vicinities of other molecules. To describe such structural changes, we need to have as well the potential associated with changes in intramolecular structure. Usually, such a potential has terms to describe variation in bond length (bond length potential), angle (bond angle potential), dihedral angle (improper and proper dihedral angle potential) (refer to first five terms in Eq. 1). Since the classical force-field cannot describe the bond-breaking processes, a harmonic potential is in general used to describe structural changes associated with bond lengths and bond angles. However, note that the dihedral angle motion is not local and can describe conformational changes in molecules and in case of peptides these contribute to changes in the secondary structure. The total potential including both intra- and intermolecular interactions for a biomacromolecule alone or in solution or in combination with other molecules can be described by the Eq. 1.

The force constants and equilibrium values for bond length and bond angle are obtained from spectroscopic data and from the structural data, respectively. Also, electronic structure theory-based calculations can be employed to get these parameters.

In general, the binding of a ligand to a protein can be described as the equilibrium between the protein–ligand complex and the protein and the ligand (Eq. 2). The change in free energy/the binding free energy (ΔG_{Bind}) can thus be calculated as the difference between the free energies of the ligand and protein in free and bound form (Eq. 3) which is then compared to experimental binding affinity (inhibition constant or IC_{50}).

$$P_{\mathrm{Aq}} + L_{\mathrm{Aq}} \leftrightarrow PL_{\mathrm{Aq}} \tag{2}$$

$$\Delta G_{\mathrm{Bind}} = G(PL)_{\mathrm{Aq}} - G(P)_{\mathrm{Aq}} - G(L)_{\mathrm{Aq}} \tag{3}$$

All these free energies can be computed using explicit solvent models, namely SPC, TIP3P, TIP4P, TIP5P, but it is computationally very demanding. Alternatively, one can use the implicit solvent models, and then the free energies of the three systems, namely complex, receptor and ligand, can be computed with less computational effort [18]. This involves calculation of solvation free energy of a subsystem in a solvent media described with a dielectric constant which is a macroscopic parameter specific to solvent and describes its ability to polarize the solute [18]. The electrostatic interaction between the solute and the solvent is solved using generalized born (as in the MM-GBSA) or Poisson–Boltzmann (as in the MM-PBSA) to get the polar part of the solvation free energies [19]. The non-polar part of the solvation free energies is computed from the solvent-accessible surface area of the solute. In the implicit solvent model, the only used solvent parameter is dielectric constant, usually the solvent coordinates are removed from the molecular dynamics or Monte Carlo trajectories, and only the protein–ligand coordinates are used.

Force-field methods for calculating free energies (e.g. MM-GBSA or MM-PBSA) with implicit solvent models forsolvation part achieved considerable success in explaining the drug binding to a number of receptors or biomacromolecular targets [20–22]. In particular, MM-PBSA method was successfully used to predict the binding affinities of many antibacterial, antiviral benchmark datasets [22, 23]. Further, the MM-PBSA method has been extensively used to understand the interaction of various substrates with the targets that are relevant in the treatment of various neurodegenerative diseases. A detailed account of this can be found in the reference [22].

4.1 Molecular Docking

Molecular docking is the most simplistic method available for computing the protein–ligand binding affinities and for finding the most stable binding mode *(pose)* for a ligand within the binding site of the protein. The scoring functions are used to decide on the binders from non-binders and their least energy binding mode and pose which can be knowledge-based, empirical and force-field-based [24–26]. In the force-field-based scoring function, the free energy of binding dictates the drug potency. The interaction energies are calculated as a sum of polar and non-polar interactions such as van der Waals and electrostatic. The change in intramolecular energies of the ligand due to conformational change is also added to the total energies just to make sure ligand conformations with unusually high energies are avoided in the search. The entropic contributions due to conformational degree of freedom are included in a simple mean; i.e. each flexible bond is associated with 0.3 kcal/mol.

The working equation to compute the interaction energy between the protein and the ligand is as given below which is a sum of van der Waals (E_{vdw}), electrostatic (E_{elec}), hydrogen bonding (E_{Hbond}) and internal energies (E_{int}). The last term refers to the change in intramolecular energy of the ligand due to binding to receptor. In the gas phase, the ligand adopts geometry where the internal energy is assumed to be zero. But when it binds to a receptor, it undergoes certain structural changes (or conformational changes) and this increase in energy is contributing to internal energy. Such contributions are usually positive to the total binding energies; however, the other contributions are dominantly negative in magnitude making the protein–ligand association to happen instead of destabilization due to increase in internal energies.

$$E_{\mathrm{dock}} = E_{\mathrm{vdw}} + E_{\mathrm{H-bond}} + E_{\mathrm{electrostatic}} + E_{\mathrm{internal}} \tag{4}$$

$$\begin{aligned} &= \sum_{\text{Protein}} \sum_{\text{ligand}} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} \right) + \sum_{\text{Protein}} \sum_{\text{ligand}} E(t) \\ &\quad \times \left(\frac{c_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}} \right) + \sum_{\text{Protein}} \sum_{\text{ligand}} 332.0 \frac{q_i q_j}{\varepsilon(d_{ij}) d_{ij}} \\ &\quad + \left\{ \sum_{\text{ligand}} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} \right) + \sum_{ligand} E(t) \left(\frac{c_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}} \right) + \sum_{\text{ligand}} 332.0 \frac{q_i q_j}{4 d_{ij} d_{ij}} \right\} \end{aligned} \tag{5}$$

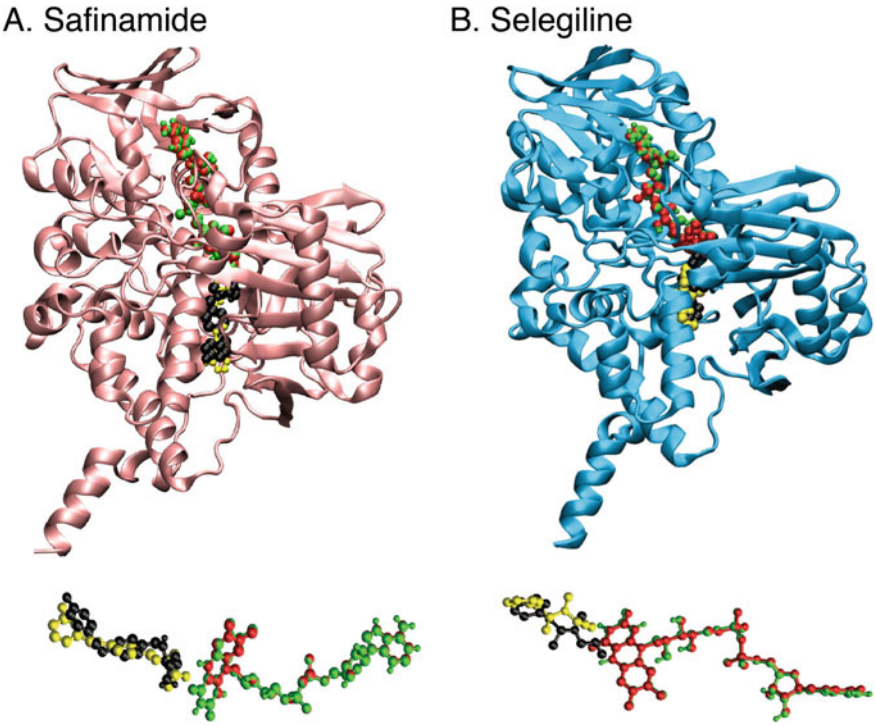


Fig. 2 Overlap of binding mode obtained from molecular docking with the experimental crystal structure

Due to the ease in doing calculations and lower computational demand, molecular docking methods are routinely used to rank the compounds according to their binding affinity or using other scoring. The pharmaceutical companies use this approach very efficiently to screen the chemical database containing millions of compounds against a potential target in the early virtual screening process before they can be synthezised as lead series. An elaborate list on use of molecular docking-based screening to design candidate drug molecules for various targets, namely G protein-coupled receptors, enzymes, ion channels, can be found in this reference [27].

Further, the binding mode and pose for number of ligands in their biological targets were predicted successfully using molecular docking tool. For example, there was a good overlap between the binding modes predicted from molecular docking and experimental crystal structure in the case of safinamide, a reversible inhibitor in monoamine oxidase B (MAO-B) (refer to Fig. 2A) [28, 29]. Interestingly, even in the case of a irreversible inhibitor such as selegiline

(L-deprenyl), where there is formation of covalent bond between the inhibitor and FAD cofactor, the predicted binding mode from molecular docking has reasonable overlap with the one from crystal structure (refer to Fig. 2B). Figure 2 shows the overlap of binding mode obtained from molecular docking with the experimental crystal structure. The target is monoamine oxidase B, and two inhibitors were considered, namely safinamide and selegiline. The former one is reversible MAO-B inhibitor, while the latter one is irreversible inhibitor which covalently bonding to the FAD cofactor.

4.2 Success Stories of Force-Field-Based Methods in Drug Discovery Projects

Drug discovery for a new disease is a complex project which requires knowledge from different domains, namely protein profiling (genomics), bioinformatics (for doing comparative genomics for target discovery), structural biology (for structure elucidation of the target), cheminformatics (chemical space), synthetic medicinal chemistry (design and synthesis of molecules), toxicology, pharmacology, pharmacokinetic property estimation, binding assay experiments, clinical studies. Computational approaches can be employed to speed up many of the intermediate steps involved in the drug discovery such as target discovery (computational comparative genomics), structure elucidation for a target (homology modelling) and lead compound prediction (using cheminformatics, virtual screening and de nova design) and ADMET property prediction (for screening the lead compounds with appropriate pharmacokinetic properties) and toxicity prediction (by studying the interaction of ligands with potential known off-targets). The chemical space consists of billions of small molecules [30], and huge genomics database suggests that there are thousands of targets and off-targets for studying the drug potency and its toxicity which makes the computational approaches as irreplaceable workhorses for the drug discovery projects. Thanks to such approaches, there are many drugs which are in the clinical trial phase as well as some of them are approved by FDA [31]. The lists include various drug compounds, namely Captopril, Dorzolamide, Saquinavir, Zanamivir, Oseltamivir, Aliskiren, Boceprevir, Nolatrexed, TMI-005, LY-517717, Rupintrivir and NVP-AUY922. In particular, the compounds Captopril and Aliskiren are used for treating heart disease, hypertension, and Saquinavir, Zanamivir, Oseltamivir, Rupintrivir are potential antiviral compounds (for HIV type I and type II, influenza virus and human rhinovirus).

4.3 Limitations of Force-Field Methods and Need for an Alternative Approach

In many occasions, force-field-based approaches were successful in explaining the ligand binding to receptors, in predicting the relative binding affinities of structurally similar ligands and in predicting the binding affinities towards various mutants of same receptors. However, many failures of these methods go unnoticed as these are not reported in general. We have noticed that the MM-GBSA and MM-PBSA methods cannot explain the relative binding affinities of indole-Substituted benzothiazoles and benzoxazoles compounds towards monoamine oxidase B and their binding specificity towards MOA-B when compared to MOA-A [32]. We have also reported that in the case of thiabendazole-based compounds the correlation between the experimental and computed binding affinities towards amyloid beta fibril using molecular docking and MM-GBSA approach when compared to quantum mechanics-based cluster model was not impressive [33].

The main reason behind is that force-field methods cannot account for the changes in the electronic structure of ligands when they are bound to the target. Usually, the charges for ligands are the same for the ligand in water as well as in the binding site of target. This is not true, the electronic structure and molecular dipole moment of the ligand can vary significantly depending upon the microenvironment [34, 35], and such polarization due to environment should be accounted for in the free energy calculations. Such a requirement automatically leads to the need for the description of the ligand using a quantum mechanical theory where the electronic degrees of freedom are treated explicitly and so the environment-specific changes in electronic structure and molecular structure can be accounted accurately [36, 37]. However, electronic structure theory is not suitable to describe protein–ligand complex systems as the number of electronic degrees of freedom is too many. So, many approximations are employed to treat the interactions between the protein–ligands in a quantum mechanical way.

5 Ab Initio Methods in Free Energy Calculations

It should be possible to calculate binding free energies using ab initio methods; however, calculation of the free energy is difficult and even intractable for large systems and an approximation is often invoked where only the energy is calculated (Eq. 6) and the temperature is assumed to be 0 K.

$$\Delta E = E_{\mathrm{Complex}} - E_{\mathrm{protein}} - E_{\mathrm{ligand}} \tag{6}$$

In this section, we briefly describe some of the known and recent developments in QM-based approaches which have been used for free energy-based drug development projects.

5.1 QM Cluster Model

In this model, the ligand and binding site residues are extracted and treated using electronic structure theory. Since the binding/catalytic site residues are mostly dictating the binding energies with ligand and the rest of the residues only play supportive role and are contributing to retain the structure of the enzyme, in particular the binding site conformation, this is reasonable approximation. Since not all the amino acids of enzyme are included in the calculation, certain approximations need to be applied. To avoid spurious charge accumulation in dangling bonds which might alter the energetics of the whole protein–ligand systems, the cut bonds are capped with hydrogens. Since the rigidity of the binding site was mostly stabilized by the rest of the protein, the free optimization of cluster might lead to unrealistic distortions in the binding site geometry. So, certain terminal residues are fixed in the space, only partial optimizations are performed, and the energies are computed for these geometries. In certain cases, the QM cluster is placed in continuum solvent to mimic the protein-like environment and the dielectric constant for the medium is chosen to be 4 [38]. There was the use of more than one quantum mechanical theory in some cluster calculations. For example as in the case of 8-Cl *TIBO* bound to human immunodeficiency virus reverse transcriptase, authors employed two-layer and three-layer ONIOM (in particular [MP2/6-31G(d), B3LYP/6-31G(d,p) and PM3]) approach to estimate the interaction energy. The residues closer to ligand are described using the high-level theory (like MP2) as these contribute to total interaction energy dominantly, while the residues far away from the ligand can be described using low-level theory as these contributions will not be very significant [39]. There are not many studies which employ this methodology to compute ligand binding energies or interaction energies with receptor [38, 40–42]. However, for modelling a number of enzymatic reactions, this method has been used successfully. In particular, the study on the enzymatic reaction of acetylene hydratase to produce vinyl alcohol using two different approaches, namely QM cluster model and QM/MM model, is worth recalling [43].

5.2 Hybrid QM/MM Approach

This approach combines the best of the two worlds, namely force-fields and electronic structure theory-based approach. Even though the receptor–ligand complex system is too large in length scale, most of the time the region of relevance to us is the ligand and certain residues that are in direct contact with the ligand. So, it is a smarter idea to split the system into two regions and use the more accurate level of theory (here, it is electronic structure theory) to describe the region of relevance and to use a relatively less accurate but cheaper (here, it is force-field) approach to describe the rest of the region. However, the harder part is the description of the interaction between these two regions or subsystems. If there is no charge transfer between these two subsystems, then one can add electrostatic and van der Waals terms to account for such interaction, and in this way the polarization of the quantum mechanical subsystem due to the system described using force-field is accounted for. The implementation is straightforward when the ligand alone is described by quantum mechanics and receptor and solvents are described using the force-fields. However, when certain residues of the receptors are to be included in the QM region, then the description of the chemical bonds between the receptor parts in QM region and MM region is a bit challenging. Methods such as hydrogen capping are developed to describe such regions, and it has become routine to use QM/MM methods for computing the protein–ligand interaction energies and free energies. Another main problem is due to the over-polarization of the terminal bonds in QM region due to atomic charges in the immediate MM region. Usually, the properties of certain atoms or groups closer to the interfacial region are moved further away into MM region so that such over-polarization is not a problem any more. Other option is to use damping function in the calculation of electrostatic contribution to deal with such effect in a mathematical way. It is also worth mentioning that the QM/MM methods in addition to energetics can be used to model the enzymatic catalytic reaction and can also be used to model the optical (linear and nonlinear) and magnetic properties of ligands when they are bound to receptors.

5.3 *Fragment Molecular Orbital*

A computationally viable strategy to evaluate the energy of an entire protein or protein–ligand complex is the fragment molecular orbital (FMO) method [44]. In the FMO method, the entire system is divided into several fragments and their energy is evaluated in the presence of all other fragments. This is known as the one-body FMO (FMO1) method. Usually, a single fragment consists of a single residue. To further enhance the quality of the calculation and include important QM effects, all pairs of fragments are evaluated in the presence of the rest of the fragments. This is known as the two-body FMO (FMO2) method. The total energy for an FMO2 calculation is given as in Eq. 7

$$E = \sum_{I}^{N} E_I + \sum_{I>J}^{N} \Delta E_{IJ} \tag{7}$$

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \tag{8}$$

where E_I is the energy of a monomer in the electrostatic potential (ESP) of all other monomers. ΔE_{IJ} is the interaction energy of fragment I and J evaluated as in Eq. 8.

5.3.1 Case study: HIV-1 RT RNase H Inhibition Screening

Many drugs or inhibitors potentially bind with metal ions in the catalytic site of enzyme or receptors in order to exhibit the therapeutic effect, e.g. magnesium ions containing enzymes such as HIV-1 integrase and RNase H [45, 46]. Thus, a good scoring function should be able to accurately calculate the metal–inhibitor interaction which impacts an overall binding affinity of individual compounds. Although the metal-binding term in the docking-scoring function is included (e.g. glide score), it considers only the anionic or highly polar interactions; therefore, ranking of actives is not appropriately achieved. On the other hand, it has been reported previously that magnesium ions in the HIV-1 reverse transcriptase-associated ribonuclease H (RNase H or RNH) play an essential role in binding and positioning of RNA–DNA duplex (natural substrate) during digestion in the viral genome reverse transcription process. Inhibition of this enzyme by chelation of magnesium ions (active site binder) is provided as an attractive approach in anti-HIV RT inhibition based drug design and discovery projects.

It is well known that the active site binder mechanism of inhibition is primarily through chelation with magnesium ions; thus, binding affinity prediction model was improved through the use of QM-based calculations by primarily considering the chelation mechanism of inhibitors with the catalytically active magnesium ions. This could be useful as a high-throughput filter in the virtual screening process.

The simplest possible model (**scenario 1**) to describe the binding of the ligand is to only describe the chelation process between the magnesium ions and the ligand in solution yielding the following approximation to Eq. 8. To further refine scenario 1, we consider in **scenario 2** geometry optimization of the protein–ligand complexes using the Qsite module (version 5.0) of the Schrödinger suite. Here, the magnesium ions and inhibitors were considered in the QM region (optimized with B3LYP and the 6-31G(d) basis set). The rest of the protein and water molecules were considered in the MM region (evaluated using the OPLS 2005 force-field) and kept frozen.

In general, docking methods could also be used for ranking compounds; however, the correlation between scoring functions and experimental values for binding free energies is rather poor in this case, and one reason is the lack of protein flexibility in the majority of the docking experiments. The correlation between molecular docking using glide score and experimental activity was quite low (n=7; $R^2 = 0.098$). However, when the atomic coordinates were used for chelation energy calculations using the DFT-B3LYP method (scenario 2; n=7; $R^2 = 0.93$) and FMO methods($R^2 = 0.80\text{-}0.94$). [47].

As we have discussed above that an effective virtual screening of RNase H inhibitors from large chemical databases could be achieved using the combination of docking and QM-based refinement calculations. In order to identify a novel chemotype for RNase H inhibition and to validate previously developed computational methods, the best models were used to screen the Specs database (containing 277,325 drug-like compounds for purchase) for HIV-1 RNase H inhibition screening (Fig. 3). A set of 1205 compounds was obtained at the end of the docking-based virtual screening, and these compounds were subsequently used for

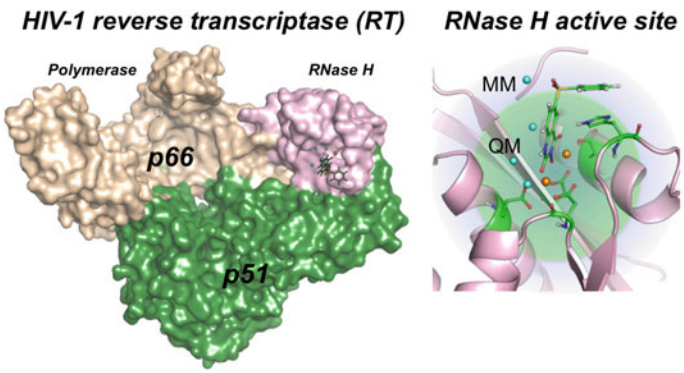


Fig. 3 Structure of HIV-1 reverse transcriptase enzyme and its active site. The components of the QM and MM region for geometry optimization are shown

QM-based refinement calculation based on density functional theory (DFT) calculations as described above (Eq. 8). The best-ranked 180 compounds from the screening were sorted for further inspection. To select a diverse set of structures for the biochemical assays, these compounds were clustered according to the structural similarity. Of the 50 structural clusters, 25 structurally diverse compounds, with the best scores, were chosen and purchased from the chemical vendor (www.specs.net) to be tested against the HIV RT-associated RNase H function in enzymatic assays. The overall workflow of the virtual screening process is shown in Fig. 4. Out of 25 compounds tested, 3 compounds inhibited the RNase H activity

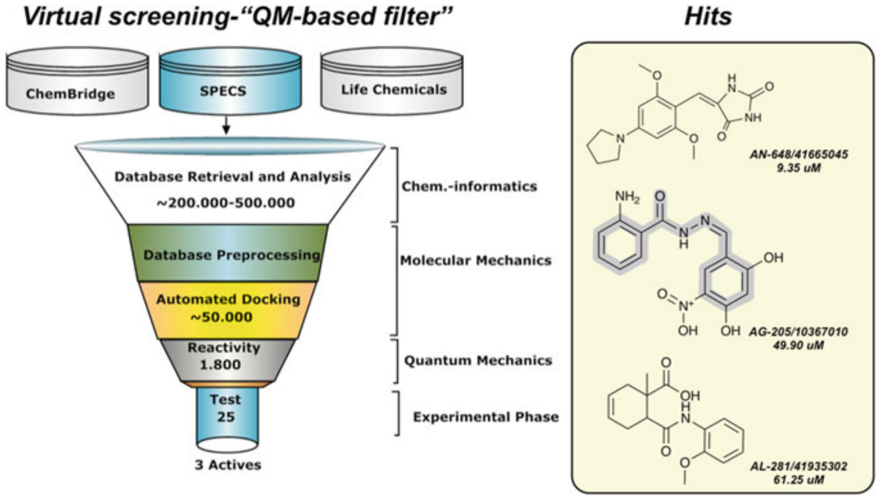


Fig. 4 Overall workflow of structure-based virtual screening strategies applied. Initial hit molecules from the first screening are provided and one of the compound is highlighted in regions where it shares a common structural pattern with known RNase H inhibitor BHMP07

below an IC_{50} value of 100 μM and compound AN-648/41665045 showed an IC_{50} value of 9.35 μM. Notably, none of these compounds has previously been reported as an inhibitor for RNase H [48]. (Fig. 4).

5.4 QM Fragmentation Approach

In this approach, whole protein is fragmented into individual amino acids, and the fragment-wise interaction energies with ligand are calculated and added together to get the total interaction energies as in the equation below. Since the whole protein is broken into individual fragments, the size of the protein is not a problem any more and a high-level electronic structure theory such as Møller–Plesset perturbation theory or coupled cluster method that accounts for electronic correlation explicitly can be used to compute the subsystem interaction energies [49].

$$\Delta E = \sum_{i=1}^{n} \Delta E_{(\mathrm{Ai-ligand})} \tag{9}$$

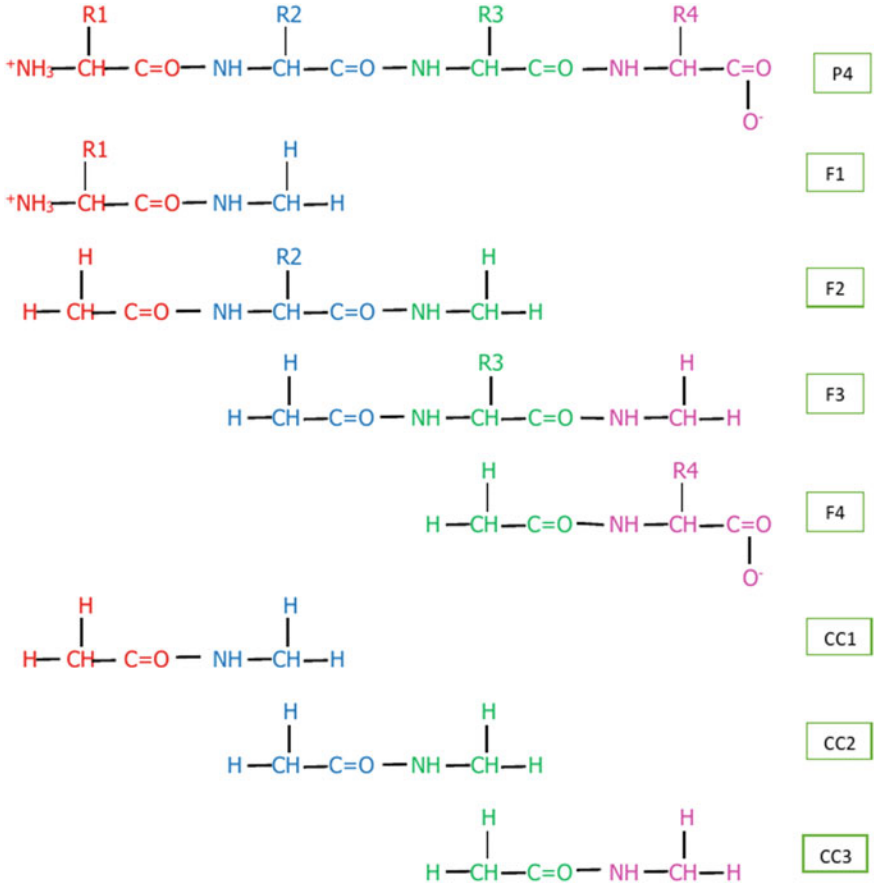
where Ai is the ith amino acid in a receptor, n is the total number of amino acids, and Ai-ligand refers to the ith residue–ligand complex. The ΔE is interaction energy between the ith residue and the ligand, which itself is computed as below:

$$\Delta E_{\mathrm{Ai-ligand}} = E_{\mathrm{Ai-ligand}} - E_{\mathrm{Ai}} - E_{\mathrm{ligand}} \tag{10}$$

The amino acids are cut along the peptide bond and capped either with hydrogens or with other functional groups to mimic protein-like environment around the residue. When the hydrogen atoms are employed as capping atom, then the above equation for the calculation of interaction energy is sufficient. However, it is appropriate to use $–NHCH_3$ and $–CO–CH_3$ as capping groups for either side of the amino acids. Moreover, the additional contributions to the interaction energies due to these capping residues should be removed as below:

$$E_{p-L} = \sum_{k=1}^{N-2} E_{F_k-L} - \sum_{k=1}^{N-3} E_{CC_k-L} - \sum_{k=1}^{N-2} E_{Fk} + \sum_{k=1}^{N-3} E_{CC_k} - E_L \tag{11}$$

In this case, the interactions are due to two molecular entities (a ligand and an amino acid) at a time and so we completely ignore the three-body contributions to the total interaction energies. In other words, this is similar to making an assumption that interaction between an amino acid and the ligand is not modulated by the presence of the neighbouring amino acids (or fragments). However, by doing additional calculations for estimating the interaction energies of dipeptide (or in units of two amino acids) and ligand at a time, such three-body contributions can be included. The expression for interaction energy is now a bit more complicated and



Scheme 1 Construction of various capped fragments for a peptide made of four amino acids (and so three peptide bonds). As can be seen, there are eventually four fragments (referred to F1, F2, F3 and F4). Each peptide bond can be capped with three pairs of –CO–CH_3 and –NH–CH_3 groups, so there are three conjugate caps (referred to CC1, CC2 and CC3), and the interactions of these with ligands should be removed as these are counted twice. (Note the positive sign for these contributions in the equation above.) It can be seen for a peptide with n amino acids there can be $n - 1$ fragments formed and $n - 2$ conjugate gaps possible if we fragment them using a scheme shown above

involves the calculation of trimer (two residues and a ligand), dimer (one residue and ligand) and monomer energies.

QM fragmentation energies can be further made sophisticated by computing the individual monomer, dimer, trimer energies with an embedding scheme which allows the interaction between these fragments with the rest of the protein through an effective Hamiltonian. This part is methodologically very similar to the above-discussed QM/MM approach where the QM system interacts with MM subsystem through electrostatic and van der Waals interaction. However, care

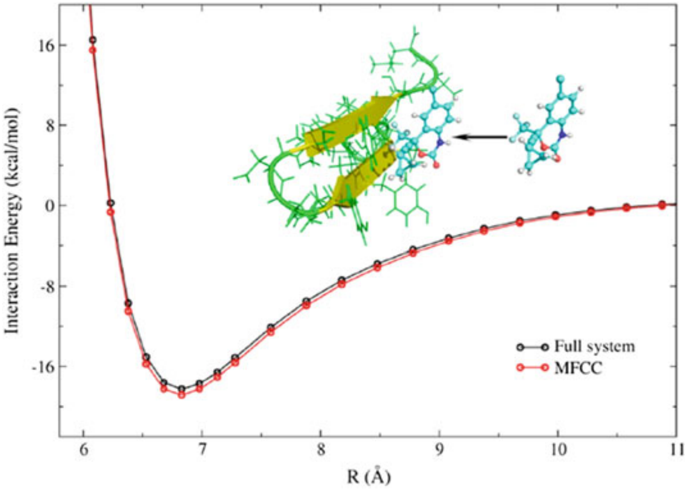


Fig. 5 Interaction energy calculated using M062X/6-311G** for Efavirenz with a fragment of HIV-1 reverse transcriptase containing residues in the range from Asn175 to Leu193 of chain A. Reprinted (adapted) with permission from (Acc. Chem. Res., **2014**, 47 (9), pp 2748–2757). Copyright (2014) American Chemical Society

should be taken to make sure that certain subsystem interactions are not double counted or in general over-counted (Scheme 1).

QM fragmentation scheme has been employed successfully to compute the interaction of ligands with various drug targets. Interestingly, certain studies showed that the computed interaction energy is comparable to that of QM cluster model. Figure 5 compares the interaction energy of Efavirenz with HIV NNRT target based on the two approaches, namely QM cluster and QM fragmentation schemes [50]. As can be seen, the interaction energy as obtained from QM fragmentation scheme agrees well with the full QM model suggesting that the former scheme is accurate as well as quite inexpensive.

Recently, it has been shown [51] that QM fragmentation method was able to correctly reproduce the relatively larger binding affinity of a tracer, FDDNP towards tau fibril when compared to amyloid beta fibril. In contrary, the MM-GBSA-based method predicted that FDDNP has a larger binding affinity towards amyloid beta fibril which is not in agreement with experimental binding affinity data. As shown in Fig. 6, it is necessary to include interaction energy of the ligand with water (within a cut-off of 15 Å) with that of its interaction with protein residues to correctly reproduce the experimental binding affinity data.

QM fragmentation scheme has been applied to compute not only protein–ligand interaction energies but also solvation energy, molecular electrostatic potential and properties such as NMR chemical shifts of the ligands in solvent and bio-environment [50]. Even the electron density of the whole biomolecule can be obtained using this approach.

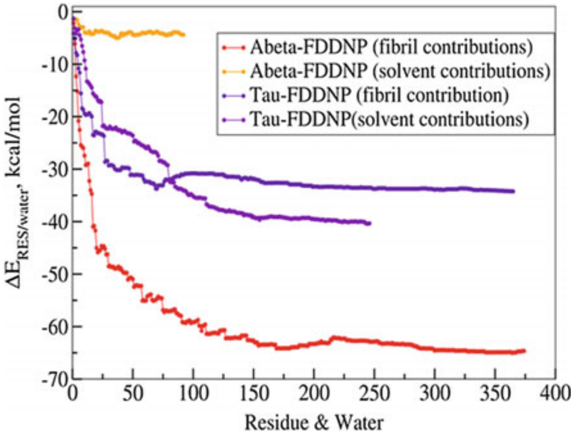


Fig. 6 Total interaction energy between the FDDNP tracer and amyloid and tau fibrils with increasing number of residues (related to increased cut-off). Also, the interaction energy of tracer with solvents located near the binding site is shown with increasing number of solvent. The residues and water solvents were first arranged with increased distance from the tracer centre of mass, and their contributions were computed and added to the total interaction energy. As can be seen with inclusion of around 125 residues, the major part of interaction energy with amyloid and tau fibril is retrieved. The figure has been reproduced with peermission from (ACS Chem. Neurosci., **2018**, 9 (7), pp 1757–1767). Copyright (2018) American Chemical Society

6 Entropic Contributions in the QM-Based Free Energy Calculations

So far in all the electronic structure theory-based approaches, we have only seen how to compute the interaction energies between a receptor and a ligand. However, the quantity of interest is the free energy of binding and not the interaction energy. For this, we also need to add the entropic contributions. The translational, rotational and vibrational contributions to the entropies are computed from the translational, rotational and vibrational partition functions as given in the reference by Yu et al. [52]. The translational and rotational contributions to the protein–ligand association are usually positive, while the vibrational contributions favour the association process. The vibrational contribution has been often reported to be much smaller in quantity when compared to the translational and rotational contributions. In some cases, we have noticed that the addition of translational and rotational contributions to total interaction energy yielded positive binding free energies. The computation of absolute free energy (including all these different entropic contributions) still remains as a challenge as there are no detailed benchmarking studies on the estimation of the translational and rotational contributions and their relative contributions to binding free energies.

7 Machine Learning (ML)-Based Approaches for Drug Discovery

Application of machine learning (ML) methods to problems in chemistry, biology, materials, etc., has taken a huge leap during the last few years. Specifically, a number of problems related to accurate intermolecular potentials, [53] drug design, [54] protein–protein interaction, [55] viable retrosynthetic pathways, [56] stability of solids, [57] potential energy surfaces [58], etc., are being addressed [59]. Advances that are being made in this space in terms of tackling problems in a way that was not thought about even few years are rapid, and the number of papers that are being published in this area is increasing exponentially. Unlike in the most research areas of science and technology, traditional ML methods such as single-layer neural networks or random forest have been applied in the area of computer-aided drug design long time ago. However, modern deep learning methods within ML are expected to make significant contributions to the area of drug design in the coming days [54, 60, 61]. Given that the last fifteen years have witnessed prolific generation of experimental data in terms of synthesizable compounds, their pharmacodynamic and pharmacokinetic properties, application of data-driven methods is likely to advance the field significantly. The following sections give a brief account of ML and some of the recent successes in application of ML methods in areas relevant to various drug design projects, including off-targets [62].

There are two fundamentally different methods in ML: supervised and unsupervised learning. Given a large data of inputs and outputs, supervised learning methods try to learn a function so that given a set of inputs, output may be predicted. Supervised learning methods such as the artificial neural networks (ANNs) are pertinent in quite a few drug discovery applications. On the other hand, unsupervised learning methods learn structure within the data when only inputs are available, which are typically applied dimensional reduction, pattern recognition, etc. Most of the ML methods are based on ANNs that connect the input and the output layers via an interconnected neural network (hidden layer(s)). The ANNs consist of a number of layers with each containing a number of neurons. Output of one of the layers is taken as input of the next layer, and the output values are calculated using an activation function. Fully connected deep neural network (DNN), recurrent neural network (RNN), convolutional neural network (CNN) and autoencoders are some of the variants of ANNs that are very successful as efficient methods for statistical modelling in a variety of fields. For a detailed account of different machine learning methods relevant to drug discovery, the readers may refer to the review by Lavecchia [63].

7.1 Structure-Based ML Approaches in Drug Design

As explained in previous sections, molecular recognition is a fundamental phenomenon behind all biological processes and in drug binding. While the number of drug-like molecules that could be synthesized is estimated to be around 10^{60}, the current experimental techniques cannot possibly screen all of these within reasonable time and expense. Computational methods such as docking calculations address this to some extent; however, the accuracy of the scoring functions behind these algorithms is still not good enough to efficiently narrow the search space that can be explored by experiments. Recently, it has been shown that machine learning (ML)-based scoring functions can predict binding affinities better than the classical scoring functions that are primarily used in computer-aided drug design [64]. Wojcikowski et al. recently reported a systematic study on the performance of ML-based scoring functions and compared it to well-known established methods [65]. They proposed a scoring function based on the random forest method (RF-Score-VS) that was trained on about 15,000 active and 900,000 inactive molecules against about 100 different drug targets . However, the authors do indicate that use of better molecular representations and descriptors will further increase the success of machine learning scoring functions. In addition, Kinnings et al. also showed that the support vector machines (SVMs) can be used to improve the performance of scoring functions. They constructed two prediction models; one is a regression model to predict the IC_{50} values, and the second is a classification model that was shown to perform very well across the entire data set [66]. Ragoza et al. have proposed a CNN-based model for scoring functions that can be used in structure-based drug design [67]. The model used the existing three-dimensional structures of protein–ligand complexes to train a model that predicts the binding affinity corresponding to any protein and ligand. The model was systematically trained by including a series of structural and binding variations such as high affinity binders, low affinity binders, correct binding pose and incorrect poses. They found that the scoring function obtained based on the CNN algorithm performs significantly better than AutoDock Vina in terms of predicting both binding poses and affinities. Recently, Dror and co-workers proposed a method named Siamese Atomic Surfacelet Network (SASNet) that applies CNN to predict protein–protein binding interfaces with high accuracy compared to the previously available knowledge-based and ML-based methods [58]. Interestingly, the training was done on a biased data where binding-driven conformational changes are not part of the data set; however, the model was shown to perform very well suggesting that the model has possibly learned the inherent structural and dynamic properties of proteins in general. In addition to the traditional neural network-based algorithms, there are other deep learning methods such as reinforcement learning that have been found to be very effective in drug design. Reinforcement learning is based on two neural networks, namely the generative and predictive neural networks. Popova et al. have recently proposed, Reinforcement Learning for Structural Evolution (ReLeaSE), a de novo method based on reinforcement learning [68]. Initially, the generative and predictive networks are trained individually using one of the supervised learning methods followed by training of both models together. This allows for predicting new chemical structures with desired biological activities. They have shown test cases by generating libraries of molecules with desired melting points, hydrophobicity and biological activities. They propose that it is possible to use a similar approach for optimization of multiple properties such as biological

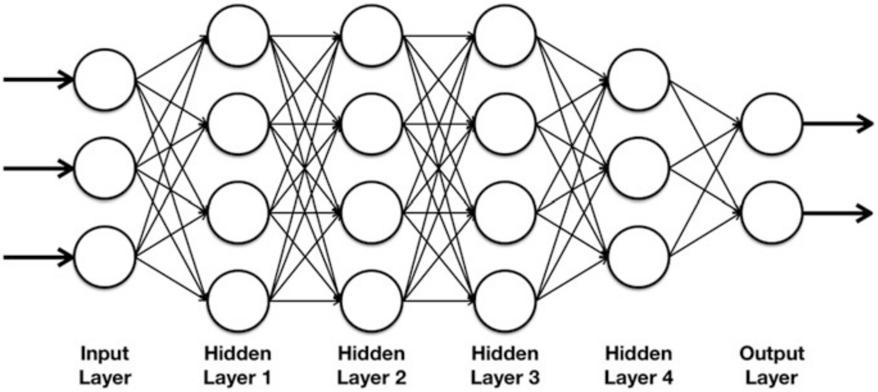


Fig. 7 Schematic representation of a multi-layer feed forward ANN

activity and different ADMET properties simultaneously to identify small molecules with desired pharmacodynamic and pharmacokinetic properties at the same time. Machine learning methods are also being used in ligand-based drug design projects . The main goal of ligand-based drug design activities is to predict how a chemical structure can be modified to achieve desired biological activity and/or ADMET properties. The aim of QSAR, one of the major methodologies in ligand-based drug design, is to generate a predictive regression model that gives a relationship between biological activity (or any other property) and a set of molecular descriptors. Such an exercise is inherently very suitable for traditional machine learning algorithms, and hence it has been adopted very early [69]. Supervised learning algorithms such as neural networks, random forest, SVMs and k-nearest neighbour have been used in QSAR [55, 60, 61]. Similarly, application of unsupervised methods such as clustering methods, principal component analysis and independent component analysis has been successful. Schematic representation of ANN workflow is shown in Fig. 7.

7.2 Future Prospects of AI-ML in Drug Design

Machine learning methods have been used in ligand-based drug design for a long time and have been reasonably successful. During the last five years, applications of deep learning algorithms have showed a lot of promise in terms of their superior performance compared to traditional ML methods used in drug design. The rate of advance of computational methodologies that are traditionally applied to drug design seems to be far lower than the advances that are being made by machine learning-based methods. The availability of high-quality data, improved biophysical experimental techniques, increasing computational resources/power and faster evolution of machine learning methods such as deep learning are further pushing the drug design efforts in right direction. Although the efforts seem to be fragmented at this point of time, further involvement of research groups with varied backgrounds and availability of clean data are expected to inspire emergence of more efficient workflows in drug design that combine traditional methods and machine learning methods.

8 Conclusions

The current drug discovery projects can be benefited a lot from advancements in the structure elucidation methods such as cryogenic electron microscopy, NMR spectroscopy, X-ray crystallograpy and from computational free energy calculation methods. This chapter presents various computational approaches available for estimating the free energies of drugs in different environments which surrogate various components of fate of drugs in the biosystem and this not only limited to drug binding to its targets but also other interactions and its relevant properties e.g. ADMET. We present various free energy calculation methods which use force-field, semi-empirical and ab initio electronic structure theory-based methods. Until a decade ago, using the electronic structure theory method for studying the structure and energetics of biomacromolecule was formidable. Thanks to the fragmentation and effective Hamiltonian approaches, it is possible to employ these methods for computing the interaction energy between ligand and biomacromolecular fragments reliably. Here, various working principles of these approaches along with key illustrative examples are presented. Even though the methods appear very promising for computing the free energy of the ligands in solvent or in biomacromolecular (such as enzyme, membrane, fibril, DNA and RNA) environment, we need to systematically study various receptor–ligand systems and test for their ability to reproduce experimental binding affinity and other pharmacokinetic parameters before employing them as lead compounds in drug discovery projects. While physics-based methods such as those mentioned above are important and unavoidable, alternative approaches based on machine learning algorithms that exploit existing experimental/computational data are emerging to be powerful tools for drug design. We expect that elegant combination of traditional physics-based methods, better computational power and more sophisticated machine learning algorithms will enable efficient and accurate quantification of protein–ligand binding affinities for improved lead identification/optimization processes in the drug design and discovery projects.

References

1. Rask-Andersen M, Almen MS, Schioth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10:579–590
2. Lenz GR, Nash HM, Jindal S (2000) Chemical ligands, genomics and drug discovery. Drug Discov Today 5(4):145–156
3. Knowles J, Gromo G (2003) Target selection in drug discovery. Nat Rev Drug Discov 2: 63–69

4. Kubinyi H (2003) Drug research: myths, hype and reality. Nat Rev Drug Discov 2(8):665–668
5. Hodgson John (2001) ADMET-turning chemicals into drugs. Nat Biotechnol 19(8):722
6. Caldwell GW (2000) Compound optimization in early-and late-phase drug discovery: acceptable pharmacokinetic properties utilizing combined physicochemical, in vitro and in vivo screens. Curr Opin Drug Discov Devel 3(1):30–41
7. Zamora I, Oprea T, Cruciani G, Pastor M, Ungell AL (2003) Surface descriptors for protein — ligand affinity prediction. J Med Chem 46(1):25–33
8. De Waterbeemd Van, Han Eric Gifford (2003) ADMET in silico modelling: towards prediction paradise. Nat Rev Drug Discov 2(3):192–204
9. Colmenarejo G (2003) Insilico prediction of drug-binding strengths to human serum albumin. Med Res Rev 23(3):275–301
10. Guengerich FP (2006) Cytochrome P450 s and other enzymes in drug metabolism and toxicity. AAPS J 8(1):E101–E111
11. Vasanthanathan P, Hritz J, Taboureau O, Olsen L, Jorgensen FS, Vermeulen NPE, Oostenbrink C (2009) Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. J Chem Inf Model 49:43–52
12. Vasanthanathan P, Olsen L, Jorgensen FS, Vermeulen NPE, Oostenbrink C (2010) Calculation of Binding Free Energy for CYP1A2 Ligands by Using Empirical Free Energy Method. Drug Metab Dispos 38:1347–1354
13. Leung SS, Mijalkovic J, Borrelli K, Jacobson MP (2012) Testing physical models of passive membrane permeation. J Chem Inf Model 52(6):1621–1636
14. Westergren J, Lindfors L, Höglund T, Lüder K, Nordholm S, Kjellander R (2007) In silico prediction of drug solubility: 1. Free energy of hydration. J Phys Chem 111(7):1872–1882
15. Rossi Sebastiano M, Doak BC, Backlund M, Poongavanam V, Over B, Ermondi G, Caron G, Matsson P, Kihlberg J (2018) Impact of dynamically exposed polarity on permeability and solubility of chameleonic drugs beyond the rule of 5. J Med Chem 61(9):4189–4202
16. Wan J, Zhang L, Yang GF, Zhan CG (2004) Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: a study of quantum chemical descriptors from density functional theory. J Chem Inf Comput Sci 44:20
17. Hopfinger AJ, Pearlstein RA (1984) Molecular mechanics force-field parameterization procedures. J Comput Chem 5(5):486–99.99–2105
18. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Acc Chem Res 33:889–897
19. Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10:449–461
20. Wang W, Donini O, Reyes CM, Kollman PA (2001) Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 30:211–243
21. Massova I, Kollman PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspect Drug Discov 18:113–135
22. Wang C, Greene DA, Xiao L, Qi R, Luo R (2018) Recent developments and applications of the MMPBSA method. Front Mol Biosci 10(4):87
23. Wang J, Morin P, Wang W, Kollman PA (2001) Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. J Am Chem Soc 123(22):5221–5230
24. Meng XY, Zhang HX, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des 7(2):146–157
25. Jain AN (2006) Scoring functions for protein-ligand docking. Curr Protein Pept Sci 7:407–420
26. Stahl M, Rarey M (2001) Detailed analysis of scoring functions for virtual screening. J Med Chem 44:1035–1042

27. Kellogg GE. (2006) In: Ekins S (ed) Computer applications in pharmaceutical research and development. Wiley, Hoboken, NJ
28. Binda C, Wang J, Pisani L, Caccia C, Carotti A, Salvati P, Edmondson DE, Mattevi A (2007) Structures of human monoamine oxidase B complexes with selective noncovalent inhibitors: safinamide and coumarinanalogs. J Med Chem 50(23):5848–5852
29. De Colibus L, Li M, Binda C, Lustig A, Edmondson DE, Mattevi A (2005) Three-dimensional structure of human monoamine oxidase A (MAO A): relation to the structures of rat MAO A and human MAO B. Proc Natl Acad Sci 102(36):12684–12689
30. Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 52 (11):2864–2875
31. Talele TT, Khedkar SA, Rigby AC (2010) Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr Top Med Chem 10(1):127–141
32. Nam MH, Park M, Park H, Kim Y, Yoon S, Sawant VS, Choi JW, Park JH, Park KD, Min SJ, Lee CJ (2017) Indole-substituted benzothiazoles and benzoxazoles as selective and reversible MAO-B inhibitors for treatment of Parkinson’s disease. ACS Chem Neurosci 8(7):1519–1529
33. Balamurugan K, Murugan NA, Ågren H (2016) Multistep modeling strategy to improve the binding affinity prediction of PET tracers to Aβ42: case study with styrylbenzoxazole derivatives. ACS Chem Neurosci 7(12):1698–1705
34. Murugan NA, Aidas K, Kongsted J, Rinkevicius Z, Agren H (2012) NMR spin-spin coupling constants in polymethine dyes as polarity indicators. Chem Eur J 18:11677–11684
35. Murugan NA, Kongsted J, Rinkevicius Z, Agren H (2012) Color modeling of protein optical probes. Phys Chem Chem Phys 14:1107–1112
36. Ryde U, Soderhjelm P (2016) Ligand-binding affinity estimates supported by quantum-mechanical methods. Chem Rev 116:5520–5566
37. Cavalli A, Carloni P, Recanatini M (2006) Target-related applications of first principles quantum chemical methods in drug design. Chem Rev 106:3497–3519
38. Nikitina E, Sulimov V, Zayets V, Zaitseva N (2004) Semiempirical calculations of binding enthalpy for protein—ligand complexes. Int J Quantum Chem 97:747–763
39. Saen-oon S, Kuno M, Hannongbua S (2005) Binding energy analysis for wild-type and Y181C mutant HIV-1 RT/8-Cl TIBO complex structures: Quantum chemical calculations based on the ONIOM method. Proteins Struct Funct Bioinf 61(4):859–869
40. Perakyla M, Pakkanen TA (1994) Quantum mechanical model assembly study on the energetics of binding of arabinose, fucose, and galactose to L-arabinose-binding protein. Proteins Struct Funct Genet 20:367–372
41. Perakyla M, Pakkanen TA (1995) Model assembly study of the ligand binding by p-hydroxybenzoate hydroxylase: correlation between the calculated binding energies and the experimental dissociation constants. Proteins Struct Funct Genet 21:22–29
42. Nikitina E, Sulimov V, Grigoriev F, Kondakova O, Luschekina S (2006) Mixed implicit/ explicit solvation models in quantum mechanical calculations of binding enthalpy for protein —ligand complexes. Int J Quantum Chem 106:1943–1963
43. Liao RZ, Thiel W (2012) Comparison of QM-only and QM/MM models for the mechanism of tungsten-dependent acetylene hydratase. J Chem Theory Comput 8(10):3793–3803
44. Fedorov DG, Kitaura K (2007) Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. J Phys Chem 111:6904–6914
45. Klumpp K, Hang JQ, Rajendran S, Yang Y, Derosier A et al (2003) Two metal ion mechanism of RNA cleavage by HIV RNase H and mechanism-based design of selective HIV RNase H inhibitors. Nucleic Acids Res 31:6852–6859
46. Budihas SR, Gorshkova I, Gaidamakov S, Wamiru A, Bona MK et al (2005) Selective inhibition of HIV-1 reverse transcriptase-associated ribonuclease H activity by hydroxylatedtropolones. Nucleic Acids Res 33:1249–1256
47. Poongavanam V, Steinmann C, Kongsted J (2014) Inhibitor ranking through QM based chelation calculations for virtual screening of HIV-1 RNase H inhibition. PLoS ONE 9(6): e98659

48. Poongavanam V, Corona A, Steinmann C, Scipione L, Grandi N, Pandolfi F, Santo RD, Esposito F, Tramontano E, Kongsted J (2018) Structure-guided approach identifies a novel class of HIV-1 ribonuclease H inhibitors: binding mode insights through magnesium complexation and site-directed mutagenesis studies. Med Chem Comm 9:562–575
49. Zhang Lei, Li Wei, Fang Tao, Li Shuhua (2017) accurate relative energies and binding energies of large ice-liquid water clusters and periodic structures. J Phys Chem 121(20):4030–4038
50. He X, Zhu T, Wang X, Liu J, Zhang JZ (2014) Fragment quantum mechanical calculation of proteins and its applications. Acc Chem Res 47(9):2748–2757
51. Murugan NA, Nordberg A, Ågren H (2018) Different positron emission tomography tau tracers bind to multiple binding sites on the tau fibril: insight from computational modeling. ACS Chem Neurosci 9 (7):1757–1767
52. Yu YB, Privalov PL, Hodges RS (2001) Contribution of translational and rotational motions to molecular association in aqueous solution. Biophys J 81(3):1632–1642
53. von Lilienfeld OA (2018) Quantum machine learning in chemical compound space. Angew Chem Int Ed 57(16):4164–4169
54. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Disco Today 23(6):1241–1250
55. Townshend RJ, Bedi R, Dror RO (2018) Generalizable protein interface prediction with end-to-end learning. arXiv preprint arXiv:1807.01297
56. Segler MH, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555(7698):604
57. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559(7715):547
58. Smith JS, Isayev O, Roitberg AE (2017) ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci 8(4):3192–3203
59. Chattopadhyay A, Zheng M, Waller MP, Priyakumar UD (2018) A probabilistic framework for constructing temporal relations in replica exchange molecular trajectories. J Chem Theory Comput 14(7):3365–3380
60. Klepsch F, Vasanthanathan P, Ecker GF (2014) Ligand and structure-based classification models for Prediction of P-glycoprotein inhibitors. J Chem Inf Model 54:218–229
61. Poongavanam V, Kongsted J (2013) Virtual screening models for prediction of HIV-1 RT associated RNase H inhibition. PLoS ONE 16(8):e73478. https://doi.org/10.1371/journal.pone.0073478
62. Vasanthanathan P, Lastdrager J, Oostenbrink C, Commandeur JNM, Vermeulen NPE, Jørgensen FS, Olsen Lars (2011) Identification of CYP1A2 ligands by structure-based and ligand-based virtual screening. Med Chem Comm 2:853–859
63. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Disco Today 20(3):318–331
64. Colwell LJ (2018) Statistical and machine learning approaches to predicting protein-ligand interactions. Curr Opin Struct Biol 49:123–128
65. Wójcikowski M, Ballester PJ, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 7:46710
66. Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, Bourne PE (2011) A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 51(2):408–419
67. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR (2017) Protein-ligand scoring with convolutional neural networks. J Chem Inf Model 57(4):942–957
68. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci Adv 4(7):eaap7885
69. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Disco Today 23(8):1538–1546.M