








SOFTWARE

Open Access



eTOX ALLIES: an automated pipeLine for linear interaction energy-based simulations

Luigi Capoferri¹, Marc van Dijk^{1†} , Ariën S. Rustenburg^{1,2†} , Tsjerk A. Wassenaar^{1,3†} , Derk P. Kooi¹ ,
Eko A. Rifai¹ , Nico P. E. Vermeulen¹  and Daan P. Geerke^{1*} 

Abstract

Background: Computational methods to predict binding affinities of small ligands toward relevant biological (off-) targets are helpful in prioritizing the screening and synthesis of new drug candidates, thereby speeding up the drug discovery process. However, use of ligand-based approaches can lead to erroneous predictions when structural and dynamic features of the target substantially affect ligand binding. Free energy methods for affinity computation can include steric and electrostatic protein–ligand interactions, solvent effects, and thermal fluctuations, but often they are computationally demanding and require a high level of supervision. As a result their application is typically limited to the screening of small sets of compounds by experts in molecular modeling.

Results: We have developed *eTOX ALLIES*, an open source framework that allows the automated prediction of ligand-binding free energies requiring the ligand structure as only input. *eTOX ALLIES* is based on the linear interaction energy approach, an efficient end-point free energy method derived from Free Energy Perturbation theory. Upon submission of a ligand or dataset of compounds, the tool performs the multiple steps required for binding free-energy prediction (docking, ligand topology creation, molecular dynamics simulations, data analysis), making use of external open source software where necessary. Moreover, functionalities are also available to enable and assist the creation and calibration of new models. In addition, a web graphical user interface has been developed to allow use of free-energy based models to users that are not an expert in molecular modeling.

Conclusions: Because of the user-friendliness, efficiency and free-software licensing, *eTOX ALLIES* represents a novel extension of the toolbox for computational chemists, pharmaceutical scientists and toxicologists, who are interested in fast affinity predictions of small molecules toward biological (off-)targets for which protein flexibility, solvent and binding site interactions directly affect the strength of ligand-protein binding.

Keywords: Binding affinity prediction, Free energy calculation, Linear interaction energy, Drug design, Computational toxicology

Background

Interactions between ligands and proteins represent an important step in many life-regulating signal-transmission processes. Upon molecular recognition a ligand can modulate protein function, thereby enhancing, inhibiting

or modulating its activity. The magnitude of the effect will depend on the strength or *affinity* of ligand-protein binding. Furthermore, in a complex biological system in which multiple interacting partners are present, the effect exerted by a ligand also depends on binding *selectivity*, i.e., the relative binding affinity toward a target when compared to other proteins. Therefore, a common goal in drug design, discovery and safety pharmacology is to obtain a compound with both high affinity for the protein of interest (target) and with high selectivity against other proteins for which activity modulation could lead to unwanted and possibly toxic events (off-targets) [1,

*Correspondence: d.p.geerke@vu.nl

†Marc van Dijk, Ariën S. Rustenburg and Tsjerk A. Wassenaar contributed equally to this work

¹ AIMMS Division of Molecular Toxicology, Department of Chemistry and Pharmaceutical Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1108, 1081 HZ Amsterdam, The Netherlands

Full list of author information is available at the end of the article

2]. In this light, computational approaches that are able to accurately predict the affinity of potential drug candidates toward targets and off-targets can help in identifying and optimizing new biologically active compounds [3, 4]. Such computational methods can be divided in ligand-based and protein-structure based approaches. In the first group, which mainly comprises quantitative structure-activity relationships (QSAR) models, statistical methods are applied to identify quantitative patterns between the structure of a chemical compound (represented as a series of molecular descriptors) and a specific biological property [5]. The fundamental assumption in QSAR is that compounds with similar structure share analogue biological properties, therefore structural information about the interacting partner is usually neglected. Furthermore, measures of similarity among structures can vary a lot depending of the metric in which it is estimated [6, 7]. Protein-structure based methods combine structural features of both the ligand and the interacting biological molecule to predict the binding affinity of the compound, usually quantified as the free energy of binding (ΔG_{bind}) [3]. In this regard, empirical scoring functions have been developed to provide a fast estimation of ΔG_{bind} during screening of large dataset of compounds [8]. However, the high efficiency of this approach comes at the expense of its accuracy, which is often low for quantitative ΔG_{bind} prediction [9]. On the other hand, more rigorous statistical-mechanics based approaches such as free-energy perturbation (FEP) [10] and thermodynamic integration (TI) [11] can provide accurate ΔG_{bind} predictions that include thermal conformational sampling by means of e.g. molecular dynamics (MD) simulation [12, 13]. However, these calculations typically require numerous and extensive simulations involving non-physical states of the system of interest, making them computationally demanding and therefore not yet suitable for screening of large sets of compounds. As an alternative, approximations to these techniques led to the development of methods in which only the physical protein-bound and unbound states of the ligand are evaluated in simulation, substantially reducing computational costs while still including thermal effects on binding [14]. Among these methods, Linear Interaction Energy (LIE) theory is an empirical approach in which binding free energies are predicted by considering only the intermolecular interactions between the ligand and its environment in both end states [15]. Although QSAR methods still represent the most commonly used approach to predict ligand-binding affinities in applied settings, structure-based models are becoming more attractive due to the increased availability of three-dimensional structures of molecular (off-)targets and of computational resources. However, extensive application of methods to

calculate ΔG_{bind} from simulation is also hindered by the high degree of supervision and technical knowledge that is typically required. To overcome these issues, we as well as others have in the past years developed extensions of the relatively efficient LIE method that open up possibilities for fully automated affinity prediction.

According to LIE, ΔG_{bind} can be computed from the difference in the ensemble Lennard–Jones ($\langle V_{lig-surr}^{LJ} \rangle$) and electrostatic ($\langle V_{lig-surr}^{Coul} \rangle$) interaction energies between the ligand and its environment as obtained from MD simulations of the ligand in complex with the protein (*protein*) and free in solvent (*water*) [15, 16]:

$$\begin{aligned} \Delta G_{bind} &= \alpha \left(\langle V_{lig-surr}^{LJ} \rangle_{protein} - \langle V_{lig-surr}^{LJ} \rangle_{water} \right) \\ &\quad + \beta \left(\langle V_{lig-surr}^{Coul} \rangle_{protein} - \langle V_{lig-surr}^{Coul} \rangle_{water} \right) \\ &= \alpha \Delta V^{LJ} + \beta \Delta V^{Coul} \end{aligned} \quad (1)$$

α and β in Eq. (1) are empirical parameters that can be fitted using a training set of ligands with known binding affinity toward a specific protein. After calibration, the LIE model can be used to predict ΔG_{bind} for query compounds with unknown affinity [16].

Similar to other free energy methods, predicted values may well depend on the conformation of the ligand-protein complex that is chosen by the user to start MD from [17]. As a remedy, Stjerschantz and Oostenbrink [18] proposed an extension to the LIE method in which the contributes obtained in simulations starting from different conformations of the ligand-protein complex could be included within a single model. ΔG_{bind} of a ligand can be expressed as averaged sum over the independent simulations i ,

$$\Delta G_{bind} = \alpha \sum_i W_i \Delta V^{LJ} + \beta \sum_i W_i \Delta V^{Coul} \quad (2)$$

where the relative contribute W_i of i can be derived from [19]

$$W_i = \frac{e^{-\frac{\Delta G_{bind,i}}{k_B T}}}{\sum_i e^{-\frac{\Delta G_{bind,i}}{k_B T}}} \quad (3)$$

with k_B Boltzmann's constant and T the temperature.

Considering that the W_i 's depend on α and β (and vice versa), Stjerschantz and Oostenbrink proposed a scheme in which α , β and W_i 's could be obtained in an iterative way during model fitting [18]. Beside leading to more accurate models due to inclusion of multiple (separate) parts of conformational space of the ligand-protein complex [18, 20], this scheme made LIE models independent from the *a priori* selection of ligand-binding

poses, laying the basis for fully unsupervised LIE free energy predictions.

Due to their complexity, fully automated predictions require the implementation and/or integration of generic procedures for (1) generation of ligand-protein conformations to start MD from, (2) preparation of the force-field topology of the system, (3) running and (4) analysis of the MD simulations, and finally (5) the actual LIE-based ΔG_{bind} estimation [21].

Partially automated workflows have been developed previously to facilitate the set up and execution of LIE-based binding free energy calculations for protein–ligand complexes [21–23]. However, they still require manual intervention (e.g. for the preparation of topologies or ligand-protein complex coordinates). Furthermore the use of commercial propriety software or textual user interfaces can limit their usage to a restricted group of users.

To overcome these issues, we present here *eTOX ALLIES* (Automated pipeLine for Linear Interaction Energy-based Simulations), which allows for the calibration and (in-house) use of LIE models for ΔG_{bind} estimations in a fully automated way. Requiring the chemical structure of ligand(s) as input (submitted as *Structure Data Format (SDF)* file), *in silico* screening of compounds can be performed: docking into the (off-)target is performed for each ligand and relevant binding poses are selected from a statistical-geometric analysis of the docking results; subsequently, the interaction energies of representative binding poses are evaluated during MD, and results are collected and employed to compute ΔG_{bind} 's. In addition, an (automated) protocol to estimate the reliability of each prediction has been implemented, according to our recently proposed approach [24]. The pipeline employs open source third-party software only, making no restriction for its use in both academic and private environment. Furthermore, a web graphical user interface (GUI) has been developed that allows for use and creation of new models in a user-friendly manner, making the tool accessible also to users that are not an expert in modeling.

Overall, *eTOX ALLIES* represents a new computational tool for pharmaceutical scientists, toxicologist and modelers, both from academia and industry, who are interested in predicting binding affinities toward (off-)targets for which structural features of the binding site and/or thermal conformational effects can significantly affect the ligand-binding process.

Implementation

Software architecture

eTOX ALLIES has been designed to provide a complete pipeline for creation and use of LIE-based models for ΔG_{bind} calculation of small molecules toward (off-)

targets with known three-dimensional structure. The code has been written in *Python 2.7* [25] and makes use of external open source softwares. Molecular (ligand) structures are handled by *Open Babel* libraries [26, 27], while *SciPy* [28] and *scikit-learn* modules are employed for statistical analysis [29]. *ParaDockS* [30] and *GROMACS* [31] are adopted as docking and MD engines, respectively. A web interface for easy handling of job and model submission and management has been built adopting the python framework *Flask* [32], and using *Open Babel* [27] and *matplotlib* [33] modules for structure and plot representations.

The software is organized in two main parts (Fig. 1): the *Job Manager* performs the steps required to obtain the descriptors used in the model for each ligand (from ligand-structure preparation and optimization, to gathering of interaction energies from MD simulations), while the *Application Programming Interface (API)* allows for easy submission of new-model calibrations or screening of query compounds.

The Job Manager

The *Job Manager* runs in background and is responsible for obtaining ligand–protein interaction energies ($\langle V_{lig-surr} \rangle$ terms in Eq. (1)). For each ligand, the process includes: (1) ligand preparation, (2) ligand topology creation and structure optimization, (3) identification of representative ligand-protein complex conformations from docking and clustering, (4) MD simulations, and (5) post-processing and gathering of interaction energies. External open source software is employed for execution of some of these steps (i.e., ligand-topology creation, docking, MD simulations) and is executed within the main framework using the *subprocess* module as interface.

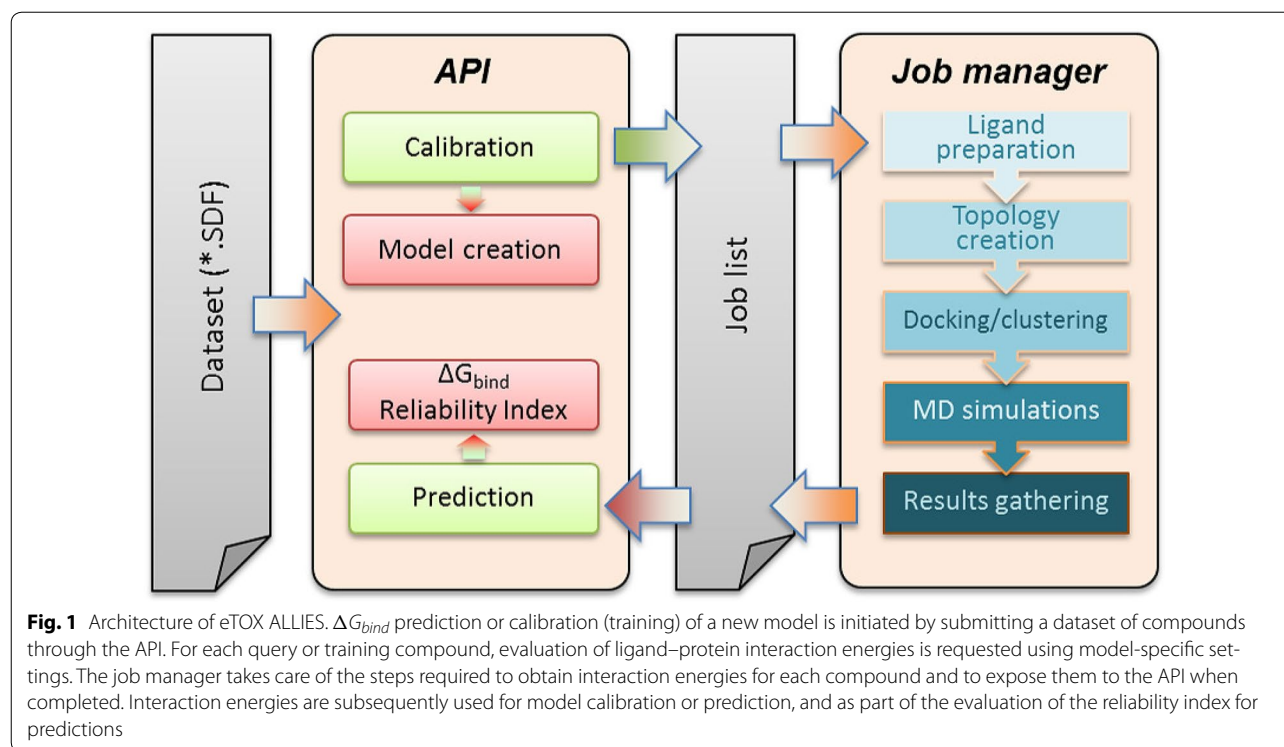
1. Ligand preparation

An *SDF* file is used to submit query or training compounds. Structural information contained in these files can be incomplete or not appropriate (e.g. improper protonation state, two-dimensional coordinates, etc.), therefore a preliminary structure optimization is performed.

During this first step, preprocessing of the ligand is performed using *Open Babel* [27], consisting of generation of 3D coordinates (in case 2D coordinates are provided only) and neutralization or protonation according to a pH of 7.4 (depending on the model settings).

2. Ligand topology creation and structure optimization

The structure of the ligand(s) must be processed to provide suitable 3D coordinates as input for molecular docking, and to generate the force-field potential parameters that will describe bonded and nonbonded interactions involving the ligand during MD (referred to as topology).



After preprocessing, the optimized structure of and atomic charges for the ligand are obtained according to the AM1-BCC method [34] and the topology of the ligand is generated according to the General Amber Force Field (GAFF) [35]. These tasks are performed by the *sqm* and *antechamber* packages provided in *Amber-Tools15* [36]. The optimized geometry of the ligand is employed as input for subsequent docking, while the topology obtained is converted to *GROMACS* [31] format using *ACPYPE* [37].

3. Identification of representative ligand-protein complex conformations

The representative ligand-protein complex conformations that will be used as starting structures for the MD simulations are obtained through clustering of binding poses obtained during molecular docking:

The optimized ligand structure is initially rotated by ± 90 degree in the x , y or z direction [38]. The ligand is subsequently docked into the protein binding site (using settings that can be defined using the API during model calibration), and maximally 50 poses with mutual RMSD of 2.0 Å are retained for each of the six rotated configurations. A principal component analysis (PCA) of the docked poses (represented as heavy-atom coordinates) is performed to reduce the number of variables, cf. [24]. The components explaining more than 5% of the initial variance are retained, and the corresponding

scores are used in subsequent k -means clustering [39]. An increasing number of clusters is considered in case it would explain at least 5% more of the variance in the score space. The medoids of the clusters obtained are considered as representative binding poses and are used as starting configurations for the MD simulations of the ligand-protein complex.

4. MD simulations

MD simulations allow the inclusion of solvent and thermal fluctuations during the evaluation of ligand–protein interaction energies.

Simulations are carried out using the *GROMACS 4.5* package [31]. An optimized version of the bash script used on the WeNMR GRID web portal [40] is adopted here to facilitate the process. Each protein-ligand complex is solvated in a dodecahedral box filled with TIP3P water [41], and Na^+ or Cl^- ions are added to neutralize the charge of the protein. After energy minimization, the system is gradually heated to 300 K in (protein and ligand) heavy-atom restrained *NVT* simulations of 10 ps simulations (at 100, 200 and 300 K, with restraining force constants of 10,000, 5000 and 50 $\text{kJ mol}^{-1} \text{nm}^{-2}$, respectively). After an additional (unrestrained) 10 ps equilibration *NVT* run, unrestrained *NpT* simulations at 1.01325 bar and 300 K are performed of few nanoseconds from which interaction energies between the ligand and its environment are obtained. The length of these

NpT production simulations can be specified by the user during model calibration. In typical cases, 2–5 different docking poses are obtained per ligand. Starting from a given protein–ligand complex conformation, two simulations are run that start from different atomic velocities, typically leading to 4–10 protein-bound simulations per ligand when using a single protein template structure. Interaction energies over the two simulations run per starting pose are averaged [20, 21]. Because of the parallelizable nature of the setup this allows to obtain free energy predictions within few hours even on a small state-of-the-art CPU cluster (ca. 10 nodes). The ligand is also simulated in explicit solvent in absence of protein and counterions in order to evaluate ligand interaction energies for the unbound state. Full details on the employed MD settings are provided elsewhere [24].

5. Postprocessing and gathering of the interaction energies

Lennard–Jones and electrostatic interaction energies between the ligand and its environment are gathered during the MD simulations. Furthermore, MD postprocessing is performed to decompose the nonbonded energy contributes of the ligand with the residues that line the binding site of the protein (for the purpose of applicability domain and reliability estimation, see below). The energies obtained are stored and used during ΔG_{bind} prediction or model calibration.

The API

An API has been created to handle the calibration of new models or the *in silico* screening of a dataset of compounds.

Upon submission of a new calculation through the API, the interaction-energies computation task is initiated through the *Job Manager*. When this process is completed, model calibration or ΔG_{bind} prediction is executed. The API provides also access to ancillary tasks: calibration of a new model involves the definition of model parameters (including e.g. definitions of the protein binding pocket, ligand-protonation states and MD simulation time) and the preparation of specific files such as for the protein topology and the formatting of the protein structure(s) in accordance with the MD and docking packages. A set of procedures is available to facilitate these preliminary processes. Additionally, every functionality is directly accessible to the user via a web GUI, which allows a user-friendly monitoring and submission of tasks.

Model calibration

Calibration of a new model can be performed by submitting a training set of compounds, provided that the experimental binding free energy ΔG^{obs} toward the (off-)

target is included as a common associated data field in the *SDF* file for the entire series of molecules. Calibration is performed after the computation of the ligand–protein interaction energies from the different independent simulations of the training set compounds. It involves (i) LIE model parameter fitting, and (ii) Applicability Domain (AD) definition.

(i) *LIE fitting* α and β coefficients are fitted using an adapted version of the iterative scheme proposed by Stjernschantz and Oostenbrink [18, 24], Fig. 2. An offset parameter γ (in kJ mol^{-1}) [42, 43] can optionally be included in model fitting, which is in that case added as a constant to Eq. (2). Initially, arbitrary values are assigned to the LIE coefficients and ΔG_{bind} is computed for every pose. The contribute of each pose to the total free energy of binding of a single compound is obtained according to Eq. (3). Using the weighted sums (according to the contribute of each pose) α and β are re-optimized and the new α and β coefficients are used to update the contribute of each pose to the total interaction energy for each compound, etc. [18]. This process is repeated iteratively until α and β are converged.

Models including increasing number of poses with lowest ΔG_{bind} for each compound are created and evaluated based on the standard deviation error in prediction (SDEP) obtained during internal leave-one-out cross-validation. The model with lowest SDEP is stored and an applicability domain for this model is created according to the approach proposed by Capoferri et al. [24].

(ii) *AD definition* In *eTOX ALLIES*, the space of information used by the LIE model is evaluated according to five different metrics: (1) range of ΔG^{obs} values; (2) chemical similarity of the ligand, in order to take into account possible effects of rarely occurring functional groups and, implicitly, of the force-field parametrization; (3) average ligand–protein interaction energies obtained during MD simulations: to evaluate the distribution of the variables used by the model; and finally, per-residue contributes to the (4) Lennard–Jones and (5) electrostatic interactions between the ligand and protein during MD, to topographically map the interactions of the ligand with specific regions of the protein binding site (possibly characterized by different electrostatic and hydrophobic properties). During calibration of the model, the space delimited by the training set is defined as follows (cf. [24]):

1. For the range of ΔG^{obs} values, the cutoffs are defined as minimum and maximum training set values.
2. Chemical similarity is expressed as Tanimoto scores (TSs) between pairs of molecules, represented as MACCS fingerprints [44]. Every training compound is compared with the other elements of the training

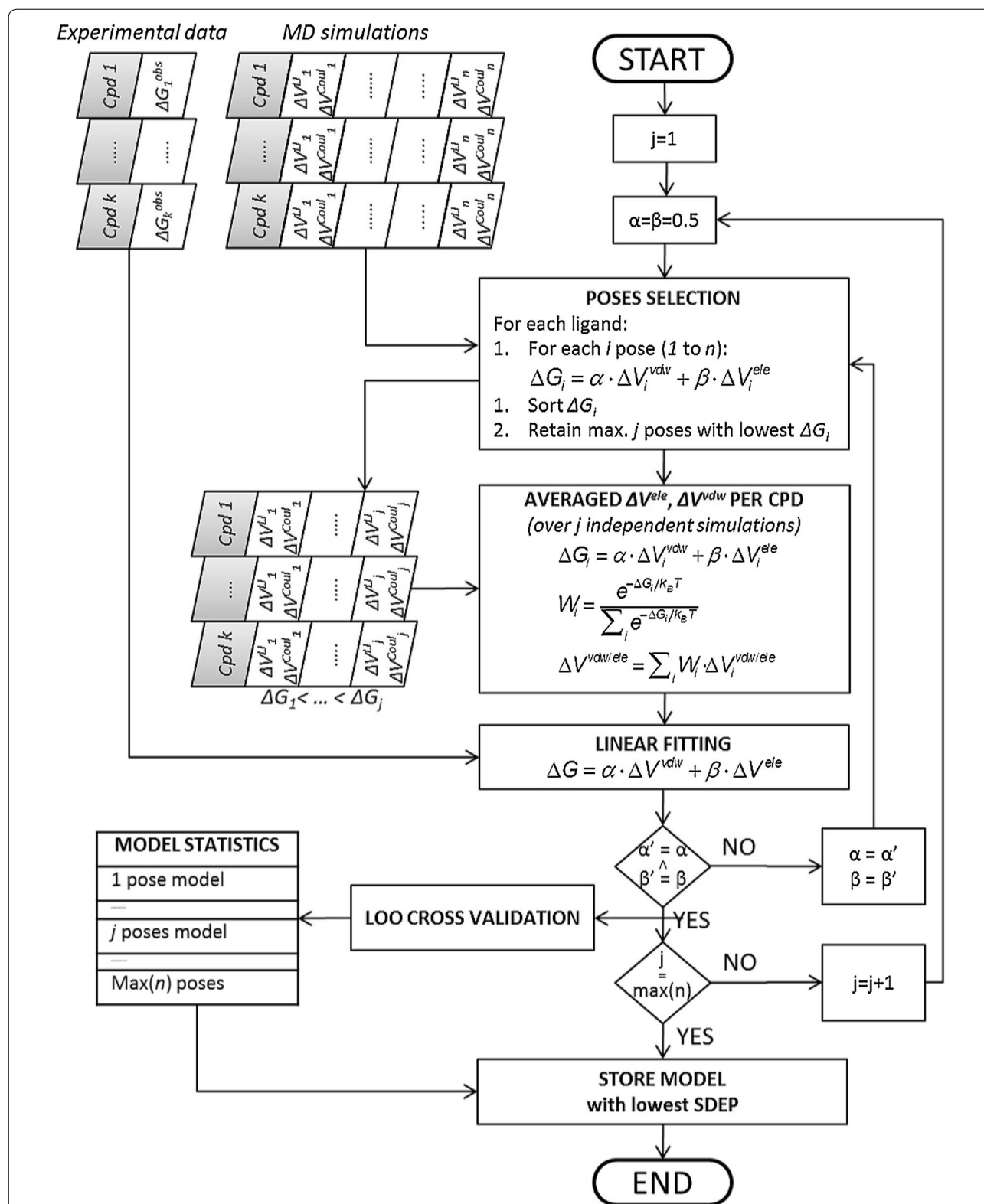


Fig. 2 Iterative LIE fitting of a model calibrated using k training compounds (Cpds) for which n simulations are run (note that n can be different per Cpd). The final model will contain simulation results for the number of poses (looped over using index j) for which the standard deviation error in prediction (SDEP) is lowest as determined in leave-one-out (LOO) cross validation. Note that ΔV 's are averaged over two MD simulations starting from different atomic starting velocities

set and the TS with the most similar compound is stored for every ligand. The lowest value is used as cutoff.

3. To compare average ligand–protein interaction energies obtained during MD simulations, the distribution of the simulations (in terms of ΔV^{LJ} and ΔV^{Coul}) employed during LIE fitting is characterized according to its average and covariance matrix [21].
4. Per-residue contributes to the Lennard–Jones interactions between ligand and protein: for each compound, the weighted sum (according to weights W_i of the corresponding simulation energies) of the Lennard–Jones interaction energies is computed for every residue located in proximity of the binding site. A principal component analysis (PCA) of the residue contributes is performed, in which components are retained if they include more than 5% of the original variance, to summarize the principal ligand–protein interactions explored during fitting.
5. Per-residue contributes to the electrostatic interactions between ligand and protein: same approach as described for the analysis of Lennard Jones interactions (4).

The parameters obtained during calibration are coupled to a specific model version and can be used to estimate the reliability of the ΔG_{bind} prediction of query compounds for the specific target.

Prediction and reliability index

Calibrated models can be used for *in silico* screening of datasets of compounds. After gathering of the ligand–protein interaction energies, ΔG_{bind} is computed. Additionally, an index that takes into account the reliability of the prediction is provided, expressed as total number of AD metrics (0 to 5) in which the query compound is found to deviate from the training set. A query compound is considered to *not* deviate from the training set according to the different metrics reported above if [24]:

1. the predicted ΔG_{bind} value is within the range of the training set experimental values;
2. the query compound shows, for at least one training-set compound, a similarity score (as TS) that is equal or higher than the cutoff defined during calibration;
3. in terms of the average ligand–protein interaction energies, the simulations for the query compound are within 95 percentile of the training set distribution, evaluated as Mahalanobis distances from the centroid;
4. the weighted sums of the per-residue contributes to the Lennard–Jones interactions of the ligands are projected onto the principal component space of the training set and show score and orthogonal distances

that are within the 95 percentile of the training set distribution;

5. the per-residue contributes to the electrostatic interactions are similar to the training set distribution when evaluated analogously as for the Lennard–Jones interactions.

A low total number of deviations (e.g. 0 or 1) corresponds to high reliability estimations, while higher numbers indicate low reliability of the predicted ΔG_{bind} [24].

Model preparation

Before calibration, a model needs to be configured, e.g. in terms of the choice for the protonation treatment of the ligand, by preparing the protein conformation and topology, and by defining the binding site coordinates and residues.

In *eTOX ALLIES*, model settings can be defined through the GUI, and the tedious preparation of files required for MD and docking is automated upon submission of the protein 3D structure as *PDB* format. This process is described hereafter.

1. Preparation of the protein structure: tautomeric states and rotamers of the residue side chains are obtained using *reduce* (*AmberTools*) [36].
2. Topology and coordinates of the protein are generated according to the Amber 14SB force field [45] by *LEaP* (*AmberTools*) [36]. In case the protein is a Cytochrome P450, special force-field parameters for both the heme domain and its coordinating cysteine are employed [46]. After conversion of topology and coordinate files to the *GROMACS* format by *ACPYPE* [37], hydrogen atom masses are increased to 4.032 amu to allow for timesteps of 4 fs during MD [47].
3. Structure templates for molecular docking are generated according to specific software requirements.
4. Docking software requires definition of the binding site (as radius and coordinates of the center of the sphere around it). For Cytochrome P450s, the center of docking can be automatically assigned according to the position of the heme domain atoms [21], otherwise it can be defined manually.
5. Residues lining the cavity of the binding site can be defined either manually or automatically as the residues for which any of the heavy atoms is within 16 Å from the (docking) center of the binding site.

Extendibility

The pipeline has been developed in order to provide high flexibility in terms of exploitation of computational resources, software implementation, and job

management. A list of the most important features are listed hereafter.

- High-performance computing cluster (HPCC): a specific interface is implemented that allows execution of MD simulations (the most compute-intensive part of the calculations) on a HPCC platform, instead of the local machine. Connection takes place via ssh tunnelling and is based on a *paramiko* python interface [48].
- Integration with *eTOXlab*: the *eTOXlab* software constitutes a framework for the creation of QSAR models and their deployment in production environments [49]. *eTOX ALLIES* models can be connected and used through *eTOXlab* in order to provide multiple modeling techniques in a single interface.
- Docking software: interfaces have been developed to integrate use of *PLANTS* (free for academic use) [50] and *ParaDockS* (released under GPL license) [30], which are dynamically loaded during execution of the program, according to model settings. Minor modifications to these modules can provide interfaces for other docking software packages.
- Force field: Amber-based force fields are adopted here because of the availability of free-license tools for the creation of ligand and protein topologies. All the operations that are force-field related are included in a specific module that is dynamically loaded in analogy to the docking software interfaces. Similarly, support for different force fields can be implemented in a straightforward way.
- Job identification: a specific python class has been implemented to handle submission and managing of jobs between the *API* and the *Job Manager* using *JSON* objects. In case of an extensive load of work, the class can be replaced in order to make use of a database management system.

Results and discussion

The web GUI

A web GUI has been developed that allows access to and monitoring of the functions provided by the API. The interface is directly accessible from a standard web browser, thus reducing the problem of dependency from specific libraries. In this way, the software can be deployed on a virtual machine and loaded on any machine, while being accessible from the host machine through the web browser. The functions accessible through the GUI include:

- overview of the available models, in terms of settings and statistics and creation of links with *eTOXlab* (Fig. 3);

- model preparation for a new (off-)target protein (Fig. 4);
- calibration of the model (Fig. 5);
- *in silico* screening of a dataset of compounds (Fig. 5);
- overview of the running jobs, with details about compounds, status, and results of the screening (Fig. 6).

Application studies

Here we demonstrate the advantage of our fully automated pipeline for (iterative) LIE model calibration and predictions for three pharmaceutically relevant proteins: Cytochrome P450 (CYP) isoform 1A2, nuclear receptor (NR) Farnesoid X receptor (FXR) and Janus Kinase 2 (JAK2). Many CYPs are flexible in nature and also NRs and kinases may bind ligands in different protein conformations and/or binding poses. Human CYPs metabolize a large variety of drug-like compounds. Drugs that tightly bind to CYPs can inhibit them and alter metabolic pathways of co-administered drugs, therefore leading to potential adverse reactions. Hence, affinity toward CYPs is of great relevance in safety pharmacology. A LIE model for binding affinity toward the isoform CYP 1A2 was recently published, in which a similar protocol as presented here was adopted [24], but in which ligand preparation was only semi-automated. Using a MD simulation time of 2.5 ns per pose, optimal α and β values were found to be 0.587 and 0.267, respectively. The root mean square error (RMSE) for the model was 4.1 kJ mol⁻¹ and the standard deviation error in prediction during leave-one-out cross validation (SDEP_{CV}) was found to be 4.3 kJ mol⁻¹. Using a different docking software (ParaDockS) than in [24], a new CYP 1A2 model was calibrated automatically using the same dataset of compounds and protein structure (for which the center of the binding site was defined automatically based on heme domain coordinates). A model was created in *eTOX ALLIES* based on simulations of 1 ns (replicated twice) for each relevant ligand binding pose, in which optimal α (0.594) and β (0.315) were comparable to the published model. The RMSE was 3.9 kJ mol⁻¹ and the SDEP_{CV} was 4.3 kJ mol⁻¹, and the SDEP for an external set of (15) compounds for which the number of AD deviations = 0 (SDEP_{EXT,0}) was determined at 4.9 kJ mol⁻¹.

As other examples, we used *eTOX ALLIES* to develop LIE binding affinity models for benzimidazole-like compounds binding to FXR in the context of the D3R Grand Challenge 2 for blind binding prediction [51, 52], and for phenylaminopyrimidines binding to JAK2. Crystal structures of the proteins were obtained from the Brookhaven protein database (PDB ID 3OMK for FXR [53], 5UT6 for JAK2 [54]) for which the center of docking was

eTOX ALLIES

HOME
SUBMIT
MODELS
RUNNING JOBS
THEORY
ABOUT

Available models

Model name:

1A2
MCL1

Model Version:

1

Model Details

Model	1A2
Docking software	paradocks
pH-dependent ligand protonation	False
forceField	amber
Simulation time	1.0
Protein Topology	MD_conf_1.top
Charge Protein	7.0
Residue lining Catalytic site	28,29,65,73,74,75,76,78,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,100,103,104,107,108,111,115,128,156,159,160,161,162,163,164,165,166,167,168,170,171,172,176,183,184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,204,219,220,221,223,224,225,226,227,228,230,231,250,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,339,342,343,346,347,348,349,350,351,352,353,354,355,374,375,376,377,378,379,382,415,416,417,418,419,420,422,423,424,425,426,427,428,429,430,431,432,433,434,435,438,460,461,462,463,464,465,466,467,468,470,481
Protein Positional restraints	ref_conf_1-posre.itp

Protein 1

Version Details

Creation Date	2015/12/05 20:44:07
Version	LIE vers 12/2015
num Simulations	2
Equation	$\Delta G \text{ (kJ mol}^{-1}\text{)} = 0.594 \times \Delta V_{VDW} + 0.315 \times \Delta V_{Ele} + 0.0$
RMSE	3.913 (kJ mol ⁻¹)
SDEP	4.340 (kJ mol ⁻¹)

1	<chem>c1(c(cccc1Nc1c(cccc1)C(=O)O)C)C</chem>
2	<chem>c12cccc1nc1c(e2N)CCCC1</chem>
3	<chem>c1ccc(c(c1)[N+](=O)[O-])[C@H]1C=C(NC(=C1C(=O)OC)C)C(=O)OC</chem>
4	<chem>c1cc(ccc1OCC)NC(=O)C</chem>
5	<chem>c12c3c(ccc1c(=O)cc(o2)c1ccccc1)cccc3</chem>
6	<chem>c1cc(cc(c1N/C=N/O)C)CCCC</chem>
7	<chem>c1(ccccc1Nc1cc(ccc1)C(F)(F)F)C(=O)O</chem>

Fig. 3 Web GUI: models page. A list of available models and calibrated versions are available here. Configuration parameters are shown in the Model Details section, while statistics about the calibrated model version are shown in the Version Details section

eTOX ALLIES

HOME
SUBMIT
MODELS
RUNNING JOBS
THEORY
ABOUT

Create new model

Model name:

Neutral ligand:

Is it Cyt P450?:

Simulation time (ns):

Residues around binding site:

Guess: **List:**

Protein conformation 1

PDB file: Nessun file selezionato.

Binding site Coordinates (in Å):

Guess: **X:** **Y:** **Z:** **radius (Å):**

Fig. 4 Web GUI: creating a new model. A new model can be calibrated through this page in a straightforward way. Multiple relevant protein structures can be included in a single model and uploaded in PDB format

determined as the center of mass of the co-crystallized ligand (which was removed before docking). FXR model calibration and validation data were obtained for benzimidazole agonists with direct therapeutic potential and derived from IC_{50} inhibition data reported by Richter et al. [53, 55] After splitting up these data into sets of 22 training compounds and of 8 test compounds for which all applicability domain criteria were fulfilled, the thus derived experimental binding free energies were used to obtain a LIE model (based on twice replicated 1 ns simulations) with $\alpha = 0.333$ and $\beta = 0.121$, and the additional off-set γ with a value of $-13.0 \text{ kJ mol}^{-1}$ (RMSE = 3.8 kJ mol^{-1} , $SDEP_{CV} = 4.1 \text{ kJ mol}^{-1}$, $SDEP_{EXT,0} = 5.0 \text{ kJ mol}^{-1}$) [52]. For JAK2 we used two 1 ns production simulations as well and phenylaminopyrimidine IC_{50} data from [56] for model calibration (22 training compounds; 4 test compounds with all AD criteria fulfilled). A LIE model was obtained with $\alpha = 0.497$, $\beta = 0.044$, RMSE = 4.3 kJ mol^{-1} , $SDEP_{CV} = 4.9 \text{ kJ mol}^{-1}$ and $SDEP_{EXT,0} = 3.8 \text{ kJ mol}^{-1}$.

Additional file 1: Figure S1 presents time series for ligand–environment interaction energies and protein–ligand atom-positional RMSDs, obtained from MD simulations used in model calibration and illustrating absence of large configurational changes, as needed when applying Eqs. (2) and (3) [19].

Conclusions

We have presented an open source framework for unperturbed protein–ligand binding affinity (free energy) computation using iterative linear interaction energy (LIE) theory. Functionalities are available and implemented in a web GUI to submit predictions and/or (re) calibrate LIE models in a straightforward way. Output of our MD and LIE based pipeline includes predictions as well as reliability indices. External software only comprises open source third-party softwares, and a specific interface is implemented to enable running MD simulations on high-performance computing clusters.

eTOX ALLIES

HOME
SUBMIT
MODELS
RUNNING JOBS
THEORY
ABOUT

Submit new job

Type of calculation:

Prediction
 Calibration

Model name:

1A2

Model version:

1

Load SDF file:

Sfogliala... Nessun file selezionato.

SDF Activity field:

Activity

Submit calculation

Fig. 5 Web GUI: submit page. This page offers the possibility to submit new screenings (i.e., prediction(s) for a compound or set of compounds listed in a SDF file to be uploaded). Calibration of a new model version (recalibration) can be performed by changing the default setting for Type of calculation

eTOX ALLIES

HOME
SUBMIT
MODELS
RUNNING JOBS
THEORY
ABOUT

	Job name	Model	Ver.	n	Date Sub	Status	Type	Del
(+)	/home/modeler/VUALIE/iLIETmp3qrssp.dsr	MCL1	1	38	2016/02/12 16:21:15	DONE	CAL	<input type="button" value="DEL"/>
(-)	/home/modeler/VUALIE/iLIETmp3HKFE.dsr	MCL1	1	4	2016/02/15 09:44:14	DONE	PRED	<input type="button" value="DEL"/>

Rep	SMILE	Job name	Status	ΔG_{EXP}	ΔG_{CALC}	CI	Del
Click Me	<chem>O=C([O-])c1sc2c(cccc2c1CCCOc1cc(C)c(Cl)c(C)c1)Cl</chem>	/home/modeler/VUALIE/iLIETmpgnF7OM.job	DONE	NONE	-37.05	0	<input type="button" value="DEL"/>
Click Me	<chem>O=C([O-])c1c(e2c(ccce2)o1)CCCOc1ccce2c1ccce2</chem>	/home/modeler/VUALIE/iLIETmpKfcKU.job	DONE	NONE	-31.52	0	<input type="button" value="DEL"/>
Click Me	<chem>O=C([O-])c1c(e2c(ccce2)[nH]1)CCCOc1ccce1)Oe1ccce1</chem>	/home/modeler/VUALIE/iLIETmpS8Yyi3.job	DONE	NONE	-37.62	1	<input type="button" value="DEL"/>
Click Me	<chem>O=C([O-])c1c(e2c(ccce2)[nH]1)CCCOc1ccce2c(c1)CC2</chem>	/home/modeler/VUALIE/iLIETmp3EBdM.job	DONE	NONE	-34.08	0	<input type="button" value="DEL"/>

Fig. 6 Web GUI: running jobs page Status and results from current screening(s) are shown here. For each dataset screening, a dropdown menu shows details about calculations for each compound

Availability and requirements

Project name: eTOX ALLIES

Project home page: <https://github.com/GeerkeLab/eTOX-ALLIES>

Operating systems: Linux OS or Mac OS X

Programming language: Python (MD runner in Bash)

Other requirements: ParaDockS or PLANTS, GROMACS 4.5.x, ACPYPE, AmberTools15

License: GPL v2

Any restrictions to use by non-academics: None.

Additional file

Additional file 1: Figure S1. Time series for a selection of MD simulations used to calibrate the CYP 1A2 (left panels), FXR (middle panels) and JAK2 LIE models (right panels): Coulomb (blue) and van der Waals (violet) interaction energies (V) between the ligand and its environment (upper three rows); and protein (backbone)-ligand atom root-mean-square positional deviations (RMSD) from the MD starting structure (bottom row, different colors refer to individual RMSD time series).

Abbreviations

AD: applicability domain; API: application programming interface; CYP: cytochrome P450; FEP: free energy perturbation; GUI: graphical user interface; HPCC: high-performance computing cluster; LIE: linear interaction energy; LOO: leave-one-out; MD: molecular dynamics; PCA: principal component analysis; QSAR: quantitative structure-activity relationship; RMSE: root mean square error; SDEP: standard deviation error in prediction; SDF: structure data file; TI: thermodynamic integration; TS: tanimoto score.

Authors' contributions

Main development and architecture: LC; design: LC, MvD, DPG; integration into eTOXlab: MvD, ASR; automation MD: TAW; implementation docking: ASR, DPK; application studies: LC, EAR; project management and supply: NPEV, DPG; manuscript writing: LC, DPG. All authors contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Author details

¹ AIMMS Division of Molecular Toxicology, Department of Chemistry and Pharmaceutical Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1108, 1081 HZ Amsterdam, The Netherlands. ² Present Address: Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. ³ Present Address: Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, 9747 AG Groningen, The Netherlands.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Archived version under the project home page, see the Availability and Requirements section for further details.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking under Grant Agreement No. 115002 (eTOX), resources of which are composed of financial contribution from the European Union Seventh Framework Programme (FP7/20072013) and EFPIA companies in kind contribution. E.A.R. acknowledges financial support from the Indonesia Endowment Fund for Education (LPDP) and D.P.G. acknowledges financial support from The Netherlands Organization for Scientific Research (NWO, VIDI Grant 723.012.105).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 12 July 2017 Accepted: 1 November 2017

Published online: 21 November 2017

References

- Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem (Int Ed)*. 41(15):2644–2676
- Whitebread S, Hamon J, Bojanic D, Urban L (2005) Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 10(21):1421–1433
- Jorgensen WL (2009) Efficient drug lead discovery and optimization. *Acc Chem Res* 42(6):724–733
- Stumpfe JG, Davis AM, Muresan S, Haeberlein M, Chen H (2013) Chemical predictive modelling to improve compound quality. *Nat Rev Drug Discov* 12(12):948–962
- Arkadiusz Z, Dudek TA, Galvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High Throughput Screen* 9(3):213–228
- Medina-Franco JL (2013) Activity cliffs: facts or artifacts? *Chem Biol Drug Des* 81(5):553–556
- Stumpfe D, Hu Y, Dimova D, Bajorath J (2014) Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J Med Chem* 57(1):18–28
- Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G et al (2015) Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* 137(7):2695–2703
- Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL (2004) Assessing scoring functions for protein-ligand interactions. *J Med Chem* 47(12):3032–3047
- Zwanzig RW (1954) High-temperature equation of state by a perturbation method. 1. Nonpolar gases. *J Chem Phys* 22(8):1420–1426
- Beveridge DL, DiCapua FM (1989) Free energy via molecular simulation: applications to chemical and biomolecular systems. *Ann Rev Biophys Chem* 18:431–492
- Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS (2011) Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol* 21(2):150–160
- Hansen N, van Gunsteren WF (2014) Practical aspects of free-energy calculations: a review. *J Chem Theory Comput* 10(7):2632–2647
- Singh N, Warshel A (2010) Absolute binding free energy calculations: on the accuracy of computational scoring of protein-ligand interactions. *Proteins* 78(7):1705–1723
- Aqvist J, Medina C, Samuelsson JE (1994) A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* 7(3):385–391
- Aqvist J, Marelus J (2001) The linear interaction energy method for predicting ligand binding free energies. *Comb Chem High Throughput Screen* 4(8):613–626
- Mobley DL, Dill KA (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* 17(4):489–498
- Stjerschantz E, Oostenbrink C (2010) Improved ligand-protein binding affinity predictions using multiple binding modes. *Biophys J* 98(11):2682–2691
- Hritz J, Oostenbrink C (2009) Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers. *J Phys Chem B* 113(38):12711–12720
- Perić-Hassler L, Stjerschantz E, Oostenbrink C, Geerke DP (2013) CYP 2D6 binding affinity predictions using multiple ligand and protein conformations. *Int J Mol Sci* 14(12):24514–24530
- Vosmeer CR, Pool R, van Stee MF, Peric-Hassler L, Vermeulen NPE, Geerke DP (2014) Towards automated binding affinity prediction using an iterative linear interaction energy approach. *Int J Mol Sci* 15(1):798–816
- Marelus J, Kolmodin K, Feierberg I, Åqvist J (1998) Q: a molecular dynamics program for free energy calculations and empirical valence bond simulations in biomolecular systems. *J Mol Graph Model* 16(4–6):213–225

23. Homeyer N, Gohlke H (2013) FEW: a workflow tool for free energy calculations of ligand binding. *J Comput Chem* 34(11):965–73
24. Capoferri L, Verkade-Vreeker MCA, Buitenhuis D, Commandeur JNM, Pastor M, Vermeulen NPE et al (2015) Linear interaction energy based prediction of cytochrome P450 1A2: binding affinities with reliability estimation. *PLoS ONE* 10(11):e0142232
25. Python Software Foundation. Python Language Reference, version 2.7; 2007–2010. <http://www.python.org>
26. O'Boyle NM, Morley C, Hutchison GR (2008) Pybel: a python wrapper for the openbabel cheminformatics toolkit. *Chem Cent J* 2:5
27. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3(1):33
28. van der Walt S, Colbert C, Varoquaux C (2011) The NumPy Array: a structure for efficient numerical computation. *Comput Sci Eng* 13:22–30
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
30. Meier R, Pippel M, Brandt F, Sippl W, Baldauf C (2010) ParaDockS: a framework for molecular docking with population-based metaheuristics. *J Chem Inf Model* 50(5):879–889
31. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4(3):435–447
32. Ronacher A. Flask, version 0.10; 2013. <http://flask.pocoo.org/>
33. Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9(3):90–95
34. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J Comput Chem* 21(2):132–146
35. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25(9):1157–74
36. Case DA, Berryman JT, Betz RM, Cerutti DS, Cheatham TE III, Darden TA et al (2015) AMBER 2015. University of California San Francisco, San Francisco
37. Sousa da Silva AW, Vranken WF (2012) ACPYPE-AnteChamber Python Parser interface. *BMC Res Notes* 5(1):367
38. Hritz J, Santos R, Oostenbrink C (2008) Impact of plasticity and flexibility on docking results for Cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J Med Chem* 51:7469–7477
39. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1: Statistics. The Regents of the University of California, pp 281–297. <http://projecteuclid.org/euclid.bsm/1200512992>
40. van Dijk M, Wassenaar TA, Bonvin AMJJ (2012) A flexible, grid-enabled web portal for GROMACS molecular dynamics simulations. *J Chem Theory Comput* 8(10):3463–3472
41. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935
42. Aqvist J, Luzhkov VB, Brandsdal BO (2002) Ligand binding affinities from MD simulations. *Acc Chem Res* 35(6):358–365
43. Carlson HA, Jorgensen WL (1995) An extended Linear Response method for determining free energies of hydration. *J Phys Chem* 99(26):10667–10673
44. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 42(6):1273–1280
45. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser K, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* 11(8):3696–3713
46. Shahrokh K, Orendt A, Yost GS, Cheatham TE (2012) Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J Comput Chem* 33(2):119–133
47. Feenstra KA, Hess B, Berendsen HJC (1999) Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* 20(8):786–798
48. Available from: www.paramiko.org
49. Carrió P, López O, Sanz F, Pastor M (2015) eTOXlab, an open source modeling framework for implementing predictive models in production environments. *J Cheminform* 7(1):8
50. Korb O, Stützel T, Exner TE (2009) Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* 49(1):84–96
51. See: <https://drugdesigndata.org/about/grand-challenge-2>
52. Rifai EA, van Dijk M, Vermeulen NPE, Geerke DP (2017) Binding free energy predictions of Farnesoid X receptor (FXR) agonists using a linear interaction energy (LIE) approach with reliability estimation: application to the D3R grand challenge 2. *J Comput Aided Mol Des*. <https://doi.org/10.1007/s10822-017-0055-0>
53. Richter HGF, Benson GM, Bleicher KH, Blum D, Chaput E, Cleemann N, Feng S, Gardes C, Grether U, Hartman P, Kuhn B, Martin RE, Plancher J-M, Rudolph MG, Schuler F, Taylor S (2011) Optimization of a novel class of benzimidazole-based farnesoid X receptor (FXR) agonists to improve physicochemical and ADME properties. *Bioorg Med Chem Lett* 21(4):1134–1140
54. Newton AS, Deiana L, Puleo DE, Cisneros JA, Cutrona KJ, Schlessinger J, Jorgensen WL (2017) JAK2 JH2 fluorescence polarization assay and crystal structures for complexes with three small molecules. *ACS Med Chem Lett* 8(6):614–617
55. Richter HGF, Benson GM, Blum D, Chaput E, Feng S, Gardes C, Grether U, Hartman P, Kuhn B, Martin RE, Plancher J-M, Rudolph MG, Schuler F, Taylor S, Bleicher KH (2011) Discovery of novel and orally active FXR agonists for the potential treatment of dyslipidemia and diabetes. *Bioorg Med Chem Lett* 21(1):191–194
56. Burns CJ, Bourke DG, Andrau L, Bu X, Charman SA, Donohue AC, Fantino E, Farrugia M, Feutrill JT, Joffe M, Kling MR, Kurek M, Nero TL, Nguyen T, Palmer JT, Phillips I, Shackelford DM, Sikanyika H, Styles M, Su S, Treutlein H, Zeng J, Wilks AF (2009) Phenylaminopyrimidines as inhibitors of Janus kinases (JAKs). *Bioorg Med Chem Lett* 19(20):5887–5892

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
