User guide for AFFPEL.py – v1.0

1. Introduction

The Automated Protein-Protein Free Energy tooL (APPFEL.py) is an automated tool designed to computationally determine the affinity between two polypeptide chains. Examples of this type of system are the complex between two large proteins, or a protein-peptide complex. Starting only from the coordinates of the bound system, APPFEL performs all the necessary steps needed for an absolute binding free energy (ABFE) calculation combined with all-atom molecular dynamics (MD): assigning the needed parameters, building and equilibrating the simulation boxes, and performing/analyzing each of the the free energy components. The MD simulations are performed using the NAMD software, which combines high performance with a set of collective variables that is suitable for large molecules. For ABFE calculations on smaller systems, such as protein-ligand or host-guest complexes, the user is invited to try APPFEL's cousin programs BAT.py and GHOAT.py, which are freely available at https://github.com/GHeinzelmann/BAT.py and https://github.com/Gheinzelmann/GHOAT.py.

In this user guide we will first describe the theory and the methods behind the APPFEL implementation, in which the binding free energy is determined by pulling the two molecules apart in the presence of restraints. We then go through the practical aspects of the program, explaining how the equilibration and free energy stages are carried out, and detailing each of the parameters to be used in the APPFEL.py input file. Finally, we show how to add a new system to the automated workflow, allowing the calculations to be extended to several other protein complexes with minimal effort.

2. Theory and methods

2.1 Absolute binding free energy

We can relate the value of the dissociation constant K_d, between a protein receptor and a single ligand, to their absolute (or standard) binding free energy ΔG^o_{bind}:

$$\Delta G^o_{bind} = RT \ln\left(\frac{K_d}{C^o}\right) \tag{1}$$

where R is the gas constant and C^o is the standard concentration of 1 M. In the APPFEL program, the calculation of ΔG^o_{bind} is done through a series of MD simulations along an artificial path that connects the bound and unbound states (Fig. 1). This path starts with the application of a set of restraints to the two bound molecules, followed by separating them along a physical path until they do not interact anymore, and finally removing the applied restraints. By calculating the free energy variation at every step, we can obtain the value of ΔG^o_b that is valid for the spontaneous process as well (Eq. 1), since G is a state function and thus is path-independent.

Following the cycle from Fig. 1, the value of the calculated binding free energy will be written as a sum of seven components:

$$-\Delta G^o_{bind} = \Delta G_{r,cf,att} + \Delta G_{l,cf,att} + \Delta G_{l,TR,att} + \Delta G_{PMF} + \Delta G_{l,TR,rel} + \Delta G_{l,cf,rel} + \Delta G_{r,cf,rel} \tag{2}$$

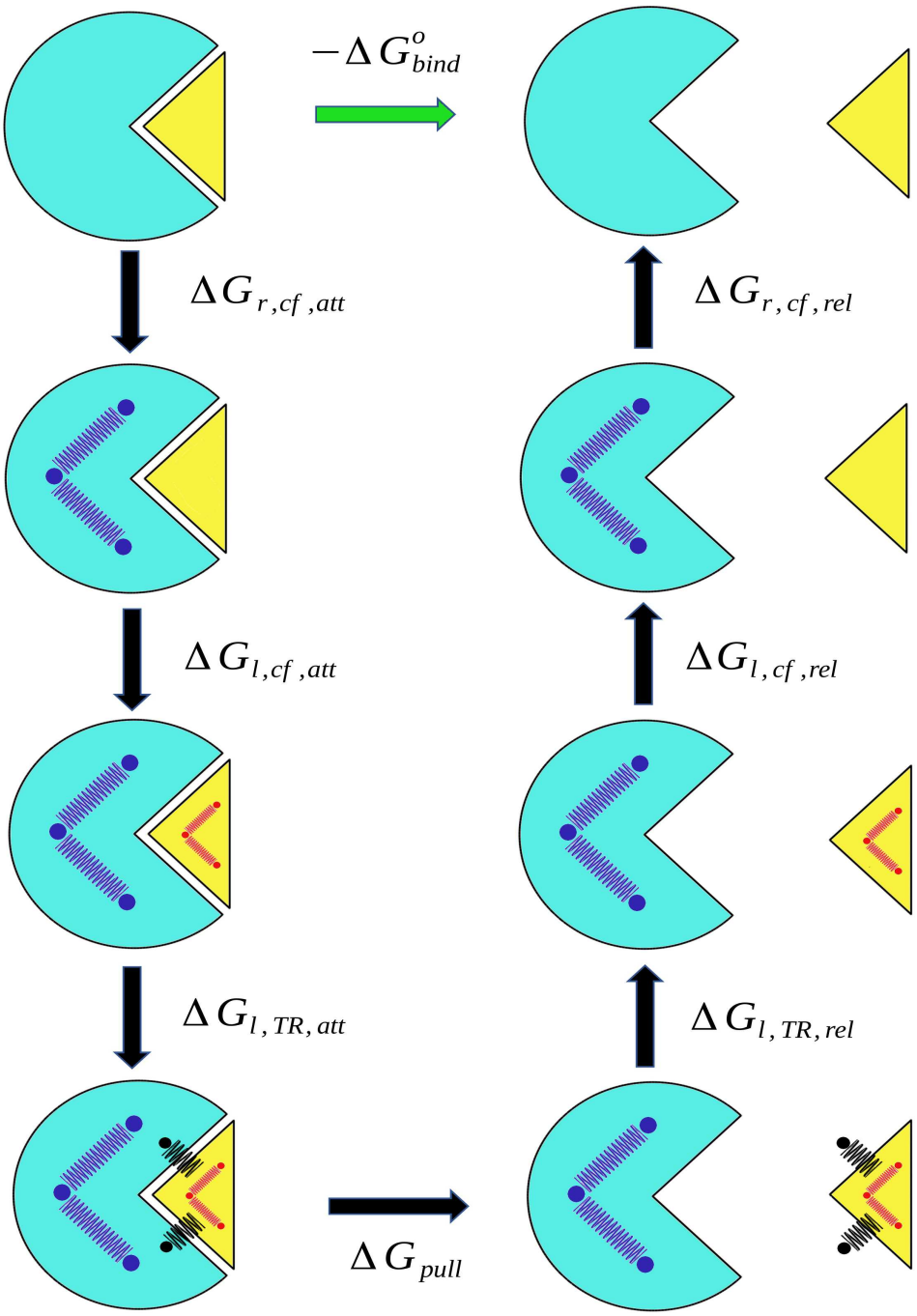

Figure 1: Thermodynamic cycle showing all the steps in the binding free energy calculation between the receptor (blue) and the ligand (yellow). The conformational (*cf*) restraints applied to the receptor and the ligand are shown as the blue and red springs, respectively, and the black springs denote the ligand translational/rotational (*TR*) restraints.

The first three terms on the right side of Eq. 2 are the free energy contributions of attaching (index *att*) restraints to the receptor (index *r*) and the ligand (index *l*), when the system is in the bound state. The nature of the restraints can be either conformational (index *cf*), or translational/rotational (index *TR*), with the former restricting the internal degrees of freedom of the molecule, and the latter used to maintain its position and overall orientation.

The ΔG_{PMF} term is the free energy obtained from the Potential of Mean Force (PMF) of bringing the ligand from the receptor binding site to a point in which they do not interact anymore, with all restraints applied to both. Once the two species are separated and each considered free in bulk solvent, the last three free energy terms on the right side of the Eq. 2 are calculated (index *rel*), by releasing each of the restraining potentials used in the pulling step.

2.2 Restraint setup

The restraint setup employed here makes use of the collective variables module from NAMD, which allows the user to apply harmonic potentials to several groups of atoms during the simulation. As noted in the previous subsection, the restraints applied to the ligand and receptor are divided into conformational (*cf*) and translational/rotational (*TR*) components.

The conformational restraints use the root mean square displacement (RMSD) of a group of *n* atoms throughout the simulation, calculated relative to a reference set of *n* atom coordinates. The restraining potential applied to this RMSD collective variable has the expression:

$$u_c = \frac{k_c}{2n} \sum_{i=1}^{n} (\vec{x}_i - \vec{x}_{0i})^2 \tag{3}$$

with k_c being the chosen force constant, $\vec{x}_i$ the position of atom *i* at a given MD-generated state and $\vec{x}_{0i}$ its position in the reference structure. The $(\vec{x}_i - \vec{x}_{0i})$ distances are computed after the set of coordinates $\vec{x}_i$ has its overall position and orientation aligned relative to $\vec{x}_{0i}$, by first centering their centers of geometry and then applying the rotation that best superimposes the two structures.

The translational/rotational restraints are present in both the receptor and the ligand during the pulling stage. Like the conformational restraints, the TR restraints are applied to a group of atoms and determined relative to a reference structure. Here, the collective variables are the distance between the centers of mass of the MD-generated and reference atom coordinates, and the relative rotation between the current and reference atom groups. For the receptor, these restraints only maintain the position and orientation of this molecule relative to the simulation box reference frame, and are not computed in the calculated free energies.

For the ligand, the applied TR potentials will have the following expressions:

$$u_t = \frac{k_t}{2}(\vec{r} - \vec{r}_0) = \frac{k_t}{2}[x^2 + y^2 + (z - z_0)^2] \tag{4}$$

$$u_o = \frac{k_o}{2}\Omega^2 \tag{5}$$

Eq. 4 corresponds to the translational component, with $\vec{r}$ being the current position of the chosen ligand atoms center of mass, $\vec{r}_0$ the reference position, and k_t the translational spring constant

(Fig. 2). We place the origin so that $\vec{r}_0=(0,0,z_0)$, and thus we can write this equation in terms of the *x*, *y*, and *z* coordinates, as well as the value of z_0. Eq. 5 corresponds to the rotational component, with k_o as the rotational spring constant and $\Omega=\cos^{-1}(\vec{q}\cdot\vec{q}_r)$. The vectors **q** and $\mathbf{q}_r$ are the MD-generated and the reference quaternions, respectively, each made up of four components $\mathbf{q} = (q_0, q_1, q_2, q_3)$. Quaternions can represent any rotation of a rigid body in three dimensions without singularities, being an elegant alternative to the more common Euler rotation angles.

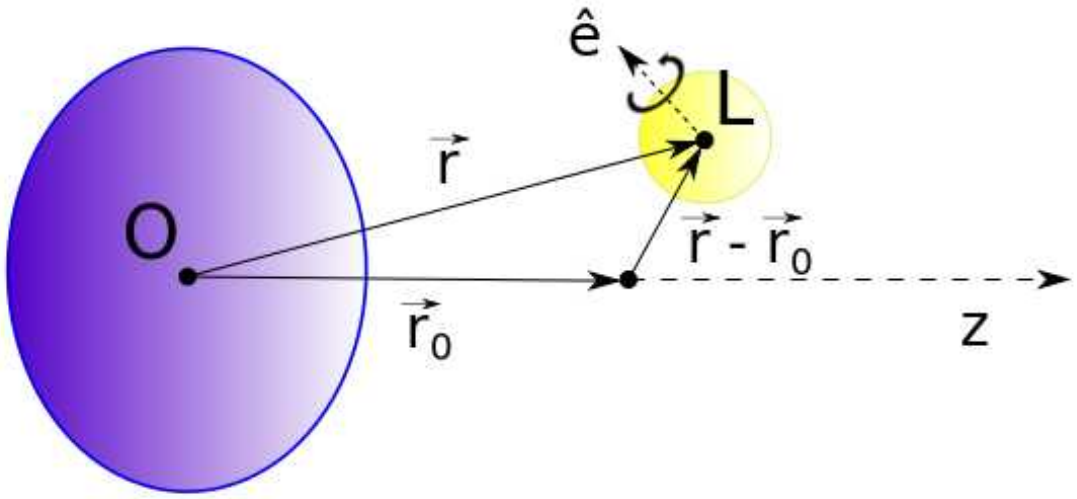

Figure 2: Scheme showing the applied restraints during the binding free energy calculations, with the receptor in blue and the ligand in yellow. Points O and L represent the origin and the center of mass of the chosen ligand atoms, respectively. The **r** vector is the position of the ligand atoms center of mass relative to the origin, and the $\mathbf{r}_0$ vector the reference position. The ê unit vector is the Euler axis of rotation, according to the quaternion representation and Euler's rotation theorem.

To obtain the free energy contributions from the application and removal of the conformational and TR restraints $\Delta G_{r,cf,att}$, $\Delta G_{l,cf,att}$, $\Delta G_{l,TR,att}$, $\Delta G_{l,cf,rel}$ and $\Delta G_{r,cf,rel}$, a set of simulation windows with intermediate values of the associated force constants is used, ranging between 0 and the final chosen value (k_c, k_t or k_o.) The potential energy output of these windows are then combined using the Multistate Bennett Acceptance Ratio (MBAR), providing the free energy difference of the process. The exception is the $\Delta G_{l,TR,rel}$ term, which is computed analytically using to the expression below:

$$\Delta G_{l,TR,rel}=k_B T \ln\left[C^o\left(\frac{2\pi k_B T}{k_t}\right)^{3/2}+\frac{1}{8\pi^2}\left(\frac{8\pi k_B T}{k_o}\right)^{3/2}\right] , \qquad (6)$$

where k_B is the Boltzmann constant and *T* is the temperature. A detailed demonstration of Eq. 6 can be found in the Appendix.

2.3 Potential of Mean Force (PMF) calculations

The PMF calculation is performed with all restraints attached, as demonstrated in Fig. 1, by varying only the value of z_0 (Eq. 4) between the bound state and a state in which the receptor and the ligand are far away from each other (Figs. 1 and 2). The free energy of this process is calculated by using a series of windows with different values of z_0, applying MBAR to combine the data from the windows and extract the free energy of the process. This technique is also known as umbrella sampling, since it uses harmonic potentials along a chosen reaction coordinate (z_0) to obtain the potential of mean force along this path.

3. Equilibrium and Steered Molecular Dynamics (SMD) simulations

The APPFEL.py workflow starts with an initial equilibration procedure, followed by a steered molecular dynamics simulation (SMD), both designed to provide the initial states needed for the full ABFE procedure. The user input parameters needed for these steps, such as temperature, ion concentration, box size and number of simulation steps, are listed in section 5.

3.1 Equilibration

The equilibration stage starts from the initial complex structure provided by the user (more details in section 6), which is solvated in a box with added ions for neutralization/ionization. An initial minimization of the system is performed, followed by a gradual heating from 0 K to the final chosen temperature. The simulation box is then coupled to a pressure reservoir at 1.0 atm and a longer simulation is performed, so that the complex will hopefully settle in a nearby free energy minimum.

3.2 SMD

The SMD process starts from the last state of the equilibration procedure above, which will also be used as the reference for the applied restraints. With all restraints attached to the receptor and ligand, the SMD step pulls the ligand from the binding site to a position in bulk in which the two species do not interact anymore. This pulling process is done along the *z* coordinate, with the final pulling distance being the distance chosen for the furthest (or last) PMF window in the APPFEL input file.

The windows from the PMF calculation will use the states from the SMD simulation, collected so that they coincide with the corresponding distances between the ligand and receptor chosen for each PMF window. This ensures a smooth transition between the non-equilibrium process of pulling the ligand, and the equilibrium sampling along a chosen number of umbrella windows.

4. Free Energy Components

Each free energy component from Eq. 2 is identified by a letter, as shown in Table 1. Components **a**, **l** and **t** are calculated using the receptor-ligand complex, and they correspond to the attachment of restraints in the bound state before the PMF calculation. The PMF, or umbrella sampling component, is denoted by the letter **u**. Components **b**, **c** and **r** correspond to the release of restraints when the two species do not interact anymore. APPFEL also brings the possibility of merging multiple attaching/releasing components into a single set of windows, which can significantly reduce the simulation time needed for a full calculation. The merged components are called **m** and **n**

for attachment and release, respectively. The first one is applied to the bound complex, and the second to the receptor and ligand simultaneously when they are separated in the box.

Table I: Binding free energy components, with the associated system, free energy method and contribution.

Description	Letter		System	Free Energy Method	Free energy term
Attachment of receptor conformational restraints	**a**	**m**	Complex	MBAR	$\Delta G_{r,cf,att}$
Attachment of ligand conformational restraints	**l**				$\Delta G_{l,cf,att}$
Attachment of ligand TR restraints	**t**				$\Delta G_{l,TR,att}$
Separation between ligand and receptor (PMF)	**u**		Complex*	MBAR	ΔG_{PMF}
Release of guest TR restraints	**b**		Ligand only	Analytical	$\Delta G_{l,TR,rel}$
Release of guest conformational restraints	**c**	**n**	Ligand only†	MBAR	$\Delta G_{l,cf,rel}$
Release of host conformational restraints	**r**		Receptor only†		$\Delta G_{r,cf,rel}$

* The **u** component connects the bound complex to the state in which the two species separated and not interacting with each other.
† For the **n** component, the receptor and ligand are in the same box, but separated and not interacting.

When the calculations are set up, the windows from each free energy component will be in folders named according to their corresponding letter followed by the window number, starting at *00*. The number of windows and their properties can be defined in the input file (section 5). The letters also identify the free energy output files, which are stored in the ./data folder of each component, after the binding free energy analysis is performed.

5. Input file

In this section we list the various parameters to be chosen in the APPFEL.py input file, which are used in the equilibrium, SMD and free energy steps listed in the previous sections:

`system`: The name of the system used for the calculations, which has to match the naming of the initial complex structure. For example, for the system variable “1bbz”, the initial pdb structure should be called 1bbz.pdb.

`rec_chain`: The chain identifier for the receptor chain in the complex pdb file. For example, if the receptor is chain A, choose “A” for this variable.

`lig_chain`: The chain identifier for the ligand chain in the complex pdb file, defined the same way as above.

`fe_type`: Type of binding free energy calculation. For a full calculation with all free energy components, choose “all”. For only the PMF calculation without computing the free energy of attaching/releasing restraints, choose “pmf”, or “rest” for restraints only. For a full calculation using

the merged **m** and **n** components, choose “express”. One can also choose the option “custom”, for a chosen set of components (see below).

`components:` If the option “custom” is set in the option above, choose the components you want to calculate, using a list of letters separated by spaces inside a bracket. Ex: “[c l u r]”.

`rest_wgt`: The weights for the attachment/release of restraints using a set of windows, going from 0 (unrestrained) to 100 (fully restrained), used for all components except **b** and **u**. The total number of windows for each of these components will be the size of the array. Ex: “[0.00 2.00 4.00 16.00 64.00 100.00]” for a total of 6 windows.

`pmf_dist:` Windows distances (in Å) for the PMF calculation, identified by the letter **u**. It starts from 0.00 (bound state) until the desired maximum distance between the receptor and the ligand in the unbound state. The total number of windows will be the size of the array. Ex: “[0.00 0.50 1.00 1.50 2.00 2.50]” for a total of 6 umbrella windows.

`blocks:` Number of blocks for block data analysis. This separates the simulation data in blocks and provides the results for each, so the temporal variation and convergence of the results can be assessed. The standard deviation across the blocks is used for the calculation of the uncertainties of each free energy component.

`num_sim:` Number of production simulations for each window after equilibration (still need to fully implement)

`rec_trans_force:` Final spring constant for the receptor center of mass translational restraints, as explained in section 2.2. Use units of kcal/mol.Å^2.

`rec_orient_force:` Final spring constant for the receptor orientational restraints using quaternions, as explained in section 2.2. Use units of kcal/mol.quat^2. The same way as radians, the quaternion unit (called quat here) is dimensionless.

`rec_rmsd_force:` Final spring constant for the receptor RMSD restraints, as explained in section 2.2. Use units of kcal/mol.Å^2. The total RMSD restraining potential (Eq. 3) is divided by the number of atoms, so a greater number of restrained atoms generally requires a larger value for this option.

`lig_trans_force, lig_orient_force, lig_rmsd_force:` Same as the receptor definitions above, but for the ligand restraint final spring constants.

`water_model`: The water model used in the calculations. Currently only “TIP3P” is supported and defined as default.

`boxsize_x,boxsize_y, boxsize_z:` Simulation box size in the x, y and z directions for the complex simulations, and simulations with only the receptor in the box (component **r**).

`box_z_center:` Center of the box in the *z* axis for the complex simulations, and simulations with only the receptor in the box (component **r**). Useful to maximize the pulling length in the *z* axis without box periodicity problems.

`boxsize_ligand:` Simulation box size in the three Cartesian axes for the simulations with only the ligand in the box.

`cation and anion:` Cation and anion species to be used, accepts all ions supported by the CHARMM force field. Ex: “SOD” and “CLA” (still need to fully implement).

`ion_conc:` Salt concentration of the chosen ions for all simulation boxes. Use units of mol/L. (Ex. “0.15”). For neutralization of the box only, without additional ions, set this option to 0.00.

`temperature:` Temperature of all simulated systems after the initial heating, in Kelvin (K).

`eq_steps`: Number of steps for the equilibrium simulations after heating, as explained in section 3.1.

`smd_steps`: Total number of steps for the SMD simulation after equilibrium, as explained in section 3.2.

`[component]_steps1`: Number of steps of equilibration, for each window from the various components of the free energy calculation, with the component letters shown in Table I. No data is collected during this simulation.

`[component]_steps2:` Number of steps for the production simulations of each window from the various components of the free energy calculation, in which data is collected.

`rec_restr:` Chosen atoms for the RMSD restraints applied to the receptor, using the VMD syntax for atom selection. For example, for non-hydrogen atoms belonging to receptor residues that are within 4 angstroms of the ligand, type for this option “(segname A) and (same residue as within 4 of segname B) and (noh)”. More details on this syntax can be found in the VMD tutorial and User Guide.

`lig_restr:` Same as the option above, but for the ligand atoms.

`Restartfreq, dcdfreq, xstfreq, outputPressure, outputEnergies, colvarsTrajFrequency, cutoff, langevinDamping, timestep:` Various options for the simulations, such as time step and output frequency of the various quantities. The variables are the same used for a regular NAMD run, and their meaning can be found in the NAMD tutorial and User Guide.

`force_field:` Force field used for all simulations, with CHARMM as default (still needs to be fully implemented).

6. Adding a new system to APPFEL.py