# User guide for APPFEL.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 [1], 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 [2] 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 $\Delta G^o_{bind}$ [3]:

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

where $R$ is the gas constant and $C^o$ is the standard concentration of 1 M. In the APPFEL program, the calculation of $\Delta G^o_{bind}$ 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 $\Delta G^o_b$ 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:

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

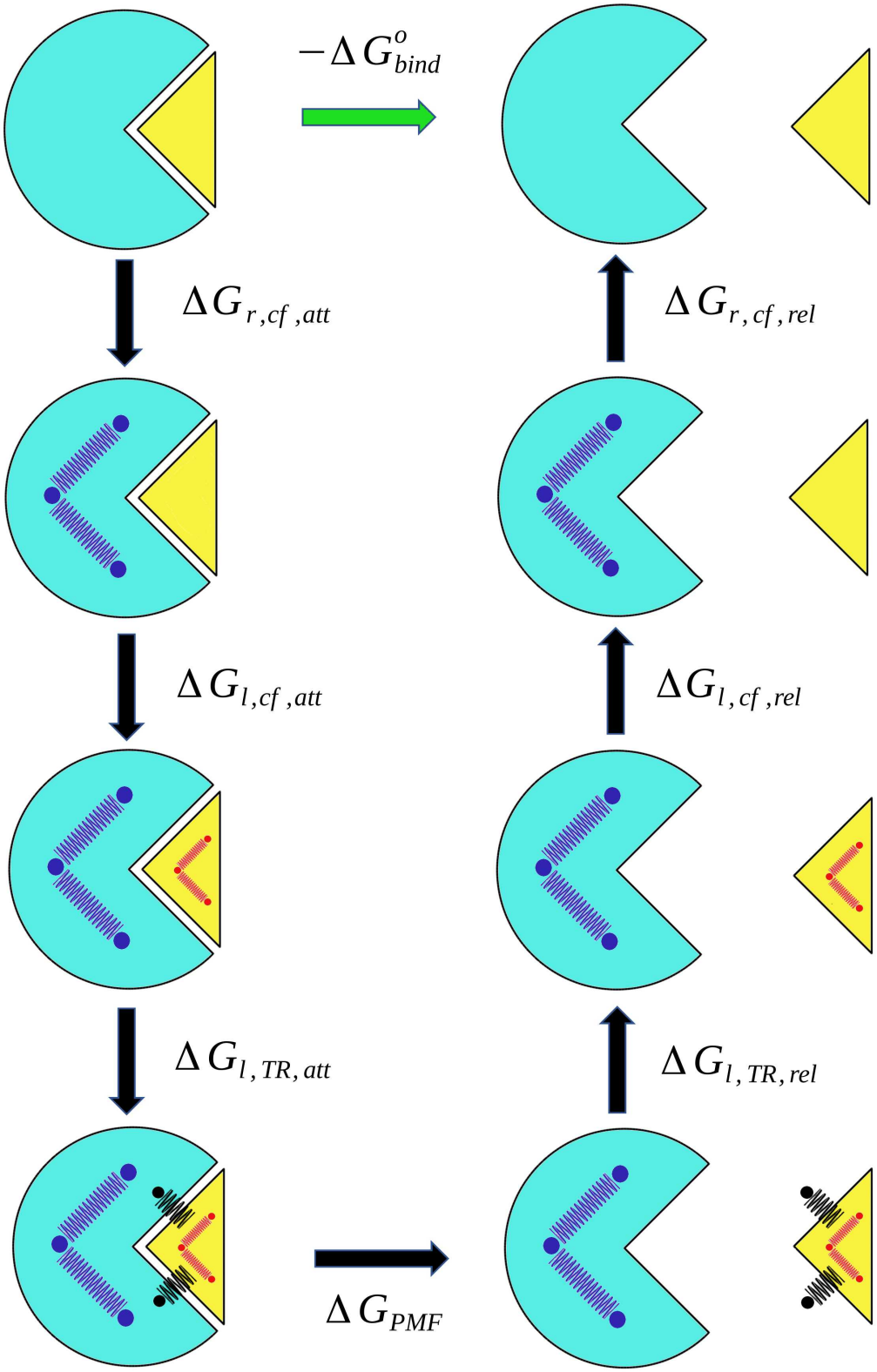


**Figure 1:** Thermodynamic cycle showing all the steps in the binding free energy calculation between the receptor (blue) and the ligand (yellow). The conformational (*cf*) restraints applied to the receptor and the ligand are shown as the blue and red springs, respectively, and the black springs denote 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 $\Delta G_{PMF}$ 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.

Like the RMSD restraints, the translational/rotational restraints are present in both the receptor and the ligand during the pulling stage. They are also 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:

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

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

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.

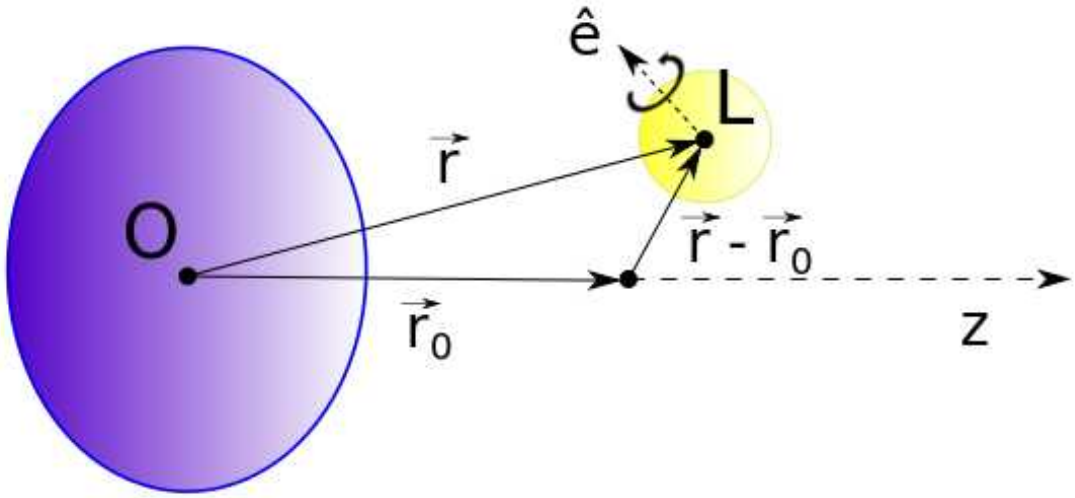


**Figure 2:** Scheme showing the applied restraints during the binding free energy calculations, with the receptor in blue and the ligand in yellow. Points O and L represent the origin and the center of mass of the ligand backbone atoms, respectively. The **r** vector is the MD-generated distance between O and L, and the $\mathbf{r}_0$ vector the restraint 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) [4], 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 Figs. 1 and 2. Here we change only the value of $z_0$ from Eq. 4, between the bound state and a state in which the receptor and the ligand are far away from each other. 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 first aligned to a reference structure, and then 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 $z_0$ 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 I. 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 | $\Delta G_{PMF}$ |
| Release of ligand TR restraints | **b** | | Ligand only | Analytical | $\Delta G_{l,TR,rel}$ |
| Release of ligand conformational restraints | **c** | **n** | Ligand only† | MBAR | $\Delta G_{l,cf,rel}$ |
| Release of receptor 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 (sections 5 and 6). 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.Å².

`rec_orient_force:` Final spring constant for the receptor orientational restraints using quaternions, as explained in section 2.2. Use units of kcal/mol.quat². 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 and section 6.3. Use units of kcal/mol.Å².

`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 for an initial equilibration of 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 (see section 6.3 for more information).

`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 (other force fields still need to be implemented).

## 6. Configuring APPFEL.py for a new system

Adding a new systems to the APPFEL workflow is a simple procedure and, once a new receptor is set, the binding free energy of any ligand can be calculated with little, if any, further adjustments. This is demonstrated in the subsections below.

### 6.1 Adding a new receptor

To apply APPFEL to a new receptor, the first step is to create a reference structure file, so that the complex is correctly oriented in space for the PMF calculations. As mentioned in the previous sections, the pulling of the ligand towards bulk solvent is done in the $+z$ direction, so this molecule needs to have free access to the solvent along this path. Here we will use the Abl-SH3 domain as an example, but the same procedure can be applied to other receptors.

Using a structure visualization/editing tool such as VMD [5], open a structure of the receptor bound to a ligand that is representative of additional ligands that will be tested for this receptor. Here we will use chains A and B from the 1bbz crystal structure [6], which has Abl-SH3 bound to a high affinity peptide. As shown in the left of Figure 3, the two chains should be reoriented in space so that the ligand can be pulled along the $+z$ direction without steric clashes, which can compromise the convergence of the calculations. Once that is done, the correctly oriented receptor structure, *without the ligand*, should be saved in the ./APPFEL/build_files/ folder as reference.pdb.

### 6.2 Adding new ligands to a receptor

Once a new receptor is configured as above, APPFEL can calculate the binding free energy of any other polypeptide ligand that binds to the same binding site. To do that, the structure of one or more receptor-ligand complexes (in pdb format) should be added to the ./APPFEL/structures/ folder. Here we will again use the 1bbz.pdb crystal structure as an example, with chains A and B being the two molecules of interest, the former being the receptor and the latter the ligand. This should be set in the APPFEL input file accordingly, using options `rec_chain` and `lig_chain`. The name of complex pdb structure (1bbz) should also be included as an input, using the `system` variable. More details on these options can be found in section 5.

### 6.3 Conformational restraints

One of the most important aspects of binding free energy calculations involving large and flexible molecules is an optimal set of conformational restraints, which can strongly affect both the convergence and accuracy of the calculations. Here we have developed a simple but very general approach to tackle this issue, combining NAMD's RMSD collective variable with an effective procedure for choosing restrained atoms, based on the well-known and accessible VMD syntax for atom selection. Detailed instructions on how to use this syntax can be found in the VMD tutorial and user guide: see for example the "Displaying different selections" section of the tutorial (https://www.ks.uiuc.edu/Training/Tutorials/vmd/tutorial-html/), and the shortcut to the relevant section of the VMD user guide https://www.ks.uiuc.edu/Research/vmd/current/ug/node90.html.

#### *6.3.1 Atom selections*

Concerning the conformational changes that could affect the calculations, the most critical residues are the receptor ones that directly interact with the ligand molecule, and vice-versa. These groups can undergo severe changes from their initial state during the SMD pulling of the ligand, and also the intermediate windows from the PMF calculation. With that in mind, we can use the VMD syntax explained above to choose all non-hydrogen atoms belonging to receptor residues that are within 4 angstroms of the ligand, typing for the for the `rec_res` option (including the quotes): *"(segname A) and (same residue as within 4 of segname B) and (noh)"*. APPFEL always defines the segment (or segname) A as the receptor and segname B as the ligand, regardless of the chain definitions for each in the initial complex structure file. Using the same idea for the ligand residues

that will have the RMSD restraints applied, one can choose for the `lig_res` variable *"(segname B) and (same residue as within 4 of segname A) and (noh)"*. The resulting chosen residues are shown in the right of Fig. 3. Additional atoms can be added to these selections if needed, such as a section of the backbone that could suffer deformation during the separation of the two molecules.

*6.3.2 RMSD spring constants*

The spring constants chosen for the RMSD restraints should take into account the number of atoms selected for the associated collective variables since, as shown in Eq. 3, the expression for the total RMSD restraining potential is divided by the number of atoms ($n$). Therefore, in order to keep the same restraining potential applied to each atom, the `rec_rmsd_force` and `lig_rsmd_force` spring constants should be increased linearly with the number of atoms chosen for the `rec_res` and `lig_res` selections, respectively. This approach will avoid applying weak RMSD restraints when they are applied to a large number of atoms.

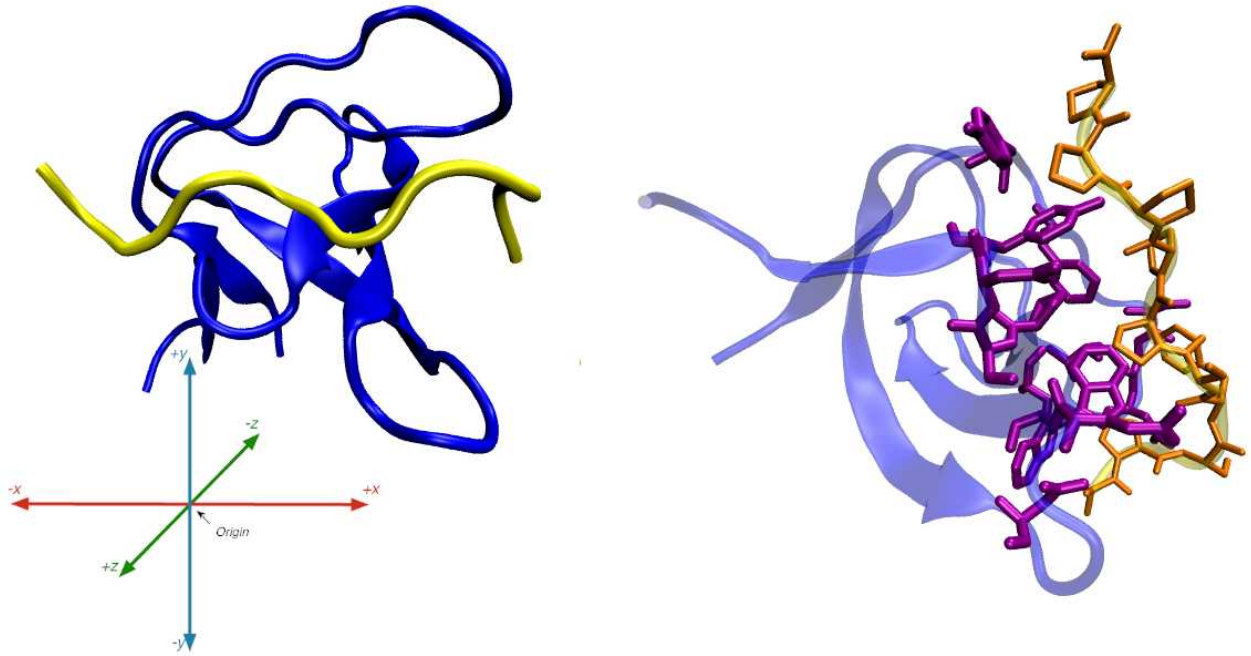


**Figure 3:** *(left)* Reoriented complex to use the receptor (blue) as the reference.pdb file, with the ligand (yellow) being pulled in the $+z$ direction. *(right)* Chosen residues from the receptor and the ligand to be conformationally restrained, with the ligand residues in orange and the receptor residues in purple.

## 6.4 Additional options

There are a few additional important options that have to be included in the APPFEL input file, in order to obtain optimal binding free energy results in a timely fashion. We discuss them briefly in the subsections below, with more information on the associated variables shown in section 5:

*6.4.1 Box size and center*

The box size in the three axes should be large enough to fit the complex and to provide sufficient dielectric screening between the various periodic boxes. Since the separation between the receptor and the ligand is performed along the $z$ axis, one might want to make the box longer along this

coordinate. There is also the choice of an arbitrary position for the box center in the *z* direction, in order to avoid periodicity problems if the receptor and ligand are pulled too far from each other. The box with only the receptor in it, used for component **r**, will have the same dimensions as the box for the complex. The box for the separate ligand, component **c**, has the same size in the three axes, also selected in the APPFEL input file.

The grid size for the Particle Mesh Ewald (PME) [7] electrostatic potential calculations will have the same values as the box sizes for each Cartesian coordinate. For speed, the grid sizes should have small integer factors such as 2, 3 and 5. Therefore, it is recommended that the box sizes in the three axes described above also follow this rule, avoiding prime numbers or numbers that have only large integer factors.

### *6.4.2 PMF and restraint windows*

A common concern when applying umbrella sampling calculations is sufficient overlap between adjacent windows, avoiding the appearance of regions that have insufficient or even no sampling along the reaction coordinate [8]. This could cause a series of problems, ranging from slow/poor convergence to meaningless or spurious results. If the RMSDs of the interacting residues are properly restrained, a good strategy is to reduce the spacing between the first umbrella windows (`pmf_dist` variable), and increase the spacing when the ligand interactions with the receptor become weaker.

The restraint windows (or weights), determined using the `rest_wgt` variable, should be closer together for values closer to 0.0, and can be further apart when the restraints become stronger. This is because more sampling is needed when the system flexibility is less hampered by the applied restraints, due to the increase in the available configurational space of the molecule.

### *6.4.3 Number of steps*

APPFEL also allows the choice of the number of equilibrium and production steps for the windows of each free energy component. Here there is a trade-off between the umbrella sampling (**u**) and the conformational (**a**, **l**, **r** and **c**) windows. Strong RMSD restraints applied to a large number of atoms will accelerate convergence of the PMF, requiring fewer steps on this component. However, more sampling will be needed to apply/remove these restraints, since the conformational space of the restrained atoms will increase with their number.

The opposite, at least in principle, is also true. If ones chooses to perform more sampling during the PMF calculations, there is less need for the receptor and the ligand to be rigid, since the conformational space will be spontaneously sampled during this process. The choice to prioritize the PMF or the conformational windows will depend on the system, but the results should always be consistent as long as there is sufficient sampling on all steps.

### *6.4.4 Editing the psf generator file*

It is common for proteins to have modifications in one or more residues, for example an acetylated N-terminus, disulfide bonds, protonation of residues, among others. As long as they are supported by the CHARMM36 topology and parameter files [9], they can be included by editing the scripts that create the .psf, or protein structure files. These scripts can be found inside the APPFEL/build_files folder from the APPFEL package, and are called psf.tcl, psf-lig.tcl and psf-rec.tcl. The first one is used to build the complex, the second the box with only the ligand in the it, and the third the box with only the receptor in it.

If these files are not modified, the systems will be built with the default definitions, such as regular N- and C- termini and aspartate/glutamate residues with a negative charge. In order to apply

the changes properly, the user needs to be familiarized with the protein structure file generator (psfgen), used by VMD. A complete user guide can be found here https://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/ug.pdf. The VMD-L Mailing List is also very useful to resolve possible issues and see examples.

# Appendix – derivation of TR restraints release in bulk

In order to obtain Eq. 6, used for the analytical release of the ligand TR restraints, we will use the rigid rotator approximation, separating the center of mass translational restraints from Eq. 4 and the orientational restraints from Eq. 5.

Considering the ligand to be in isotropic bulk media after the separation from the receptor, the free energy of releasing the harmonic potentials $u_t$ and $u_o$ to the standard concentration $C^o$ can be written as the expression:

$$\Delta G_{l,TR,rel} = k_B T \ln\left[\left(\frac{C^o}{8\pi^2}\right) + \int \exp(-\beta u_t)\, d\vec{r} + \int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} \exp(-\beta u_o)\sin\theta\, d\psi\, d\phi\, d\theta\right] \quad \text{(A1)}$$

The first integral on the right side is performed over all Cartesian space, and can be solved with the expression for the $u_t$ potential and an integration in the three axes.

$$\int \exp(-\beta u_t)\, d\vec{r} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(\frac{k_t x^2}{2k_B T}\right)\exp\left(\frac{k_t y^2}{2k_B T}\right)\exp\left(\frac{k_t (z-z_0)^2}{2k_B T}\right) dz\, dy\, dx = \left(\frac{2\pi k_B T}{k_t}\right)^{3/2} \quad \text{(A2)}$$

The expression above has already been derived in previous studies that use absolute coordinates for the translational restraints [10,11].

The second integral on the right side is done over the space of Euler angles, but we are using quaternions for the orientational restraints. Therefore, a connection between these two representations needs to be established to correctly obtain the analytical result for $\Delta G_{l,TR,rel}$ . We will start with the expression for the $u_o$ potential in an isotropic (bulk solvent) media using the quaternion representation:

$$u_o = \frac{k_o}{2}\Omega^2 = \frac{k_o}{2}[\cos^{-1}(\vec{q}\cdot\vec{q}_r)]^2 = \frac{k_o}{2}[\cos^{-1}(q_0)]^2 \sim \frac{k_e}{2}(\theta^2 + \phi^2 + \psi^2) \quad \text{(A3)}$$

where $\theta$, $\varphi$ and $\psi$ are the three Euler angles, **q** is the current quaternion given by the components **q** = ($q_0$ , $q_1$ , $q_2$ , $q_3$), and $\mathbf{q}_r$ the unit quaternion relative to a reference structure, given by $\mathbf{q}_r$ =(1, 0, 0, 0). The potential on the rightmost side of Eq. A3 is a valid approximation for small rotations, which considers the rotation axes of the $\theta$, $\varphi$ and $\psi$ Euler angles to be orthogonal and independent. Thus we can write $u_o$ a function of an orientational spring constant applied to Euler angles, $k_e$, and this expression can be replaced in the rotational integral from Eq. A1.

To find the relation between the quaternion force constant $k_o$ and its Euler counterpart $k_e$, we use the fact that, for small angles, the value of $q_0$ is less than, but close to 1, so that the following approximation holds:

$$[\cos^{-1}(q_0)]^2 \sim 1 - q_0^2 = q_1^2 + q_2^2 + q_3^2 \quad \text{(A4)}$$

with the last equality coming from the quaternion normality condition. The conversion from quaternions to Euler angles can be performed using the transformations [12]:

$$\phi = \tan^{-1}\left[\frac{2(q_0 q_3 + q_1 q_2)}{1-2(q_2^2+q_3^2)}\right] \qquad \theta = \sin^{-1}[2(q_0 q_2 - q_1 q_3)] \qquad \psi = \tan^{-1}\left[\frac{2(q_0 q_1 + q_2 q_3)}{1-2(q_1^2+q_2^2)}\right] \qquad \text{(A5)}$$

For small rotations, $q_0$ is close to unity, and $q_1$, $q_2$, $q_3 << 1$, so we can drop the terms multiplied twice by $q_1$, $q_2$ or $q_3$. We can also use the small angle approximation for the $\sin^{-1}$ and $\tan^{-1}$ functions, obtaining the simple relations $q_1 \sim \psi/2$, $q_2 \sim \theta/2$ and $q_3 \sim \phi/2$, with the Euler angles defined in radians. Combining this result with equations A3 and A4, we now write the $u_o$ potential in terms of these quantities:

$$u_o \sim \frac{k_o}{2}(q_1^2+q_2^2+q_3^2) \sim \frac{k_o}{8}(\theta^2+\phi^2+\psi^2) \qquad \text{(A6)}$$

Plugging this expression to the rotational integral from Eq. A1, and considering $sin\ \theta \sim 1$ for orthogonal rotation axes, we finally obtain the result from Eq. 6:

$$\int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\left(\frac{k_o\theta^2}{8k_BT}\right)\exp\left(\frac{k_o\phi^2}{8k_BT}\right)\exp\left(\frac{k_o\psi^2}{8k_BT}\right) d\psi\, d\phi\, d\theta = \left(\frac{8\pi k_B T}{k_o}\right)^{3/2} \qquad \text{(A7)}$$

For the approximations above to be valid, it is essential that the value of $k_o$ for the ligand is sufficiently large. For example, a value of $k_o$ = 1000 kcal/mol.quat², equivalent to $k_e = k_o/4$ = 250 kcal/mol.rad², is already suitable. The receptor translational and orientational spring constants should be much stronger than the ligand ones, so as to avoid the coupling between the movements of the two species.

## 7. References


[1] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, *et al.* (2020) "Scalable molecular dynamics on CPU and GPU architectures with NAMD." *Journal of Chemical Physics*, **153**, 044130.

[2] G. Heinzelmann and M. K. Gilson (2021). "Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation". *Scientific Reports*, **11**, 1116.

[3] M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon (1997). "The statistical-thermodynamic basis for computation of binding affinities: A critical review." *Biophysical Journal* **72**, 1047–1069.

[4] M. R. Shirts and J. Chodera (2008) "Statistically optimal analysis of samples from multiple equilibrium states." *Journal of Chemical Physics*, **129**, 129105.

[5] W. Humphrey, A. Dalke and K. Schulten. (1996) "VMD - Visual Molecular Dynamics", *Journal of Molecular Graphics*, **14**, 33-38.

[6] M. T. Pisabarro, L. Serrano and M. Wilmans (1998). “Crystal structure of the Abl-SH3 domain complexed with a designed high-affinity peptide ligand: implications for SH3-ligand interactions”. *Journal of Molecular Biology*, **281**, 513-521.

[7] T. Darden, D. York and L. Pedersen. (1993) “Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems”. *Journal of Chemical Physics*, **98**, 10089.

[8] P. C. Chen and S. Kuyucak (2011). “Accurate Determination of the Binding Free Energy for KcsA-Charybdotoxin Complex from the Potential of Mean Force Calculations with Restraints”. *Biophysical Journal*, **100**, 2466-2474.

[9] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, *et al.* (2010) “CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Field.” *Journal of Computational Chemistry*, **31**, 671−690.

[10] G. Heinzelmann, P. C. Chen and S. Kuyucak (2014). “Computation of Standard Binding Free Energies of Polar and Charged Ligands to the Glutamate Receptor GluA2” *Journal of Physical Chemistry B*, **118**, 1813-1824.

[11] ) J. Hermans and L. Wang. (1997) “Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme.” *Journal of the American Chemical Society*, **119**, 2707−2714.

[12] J. L. Blanco (2013). “A tutorial on SE(3) transformation parameterizations and on-manifold optimization” Technical Report nº. 012010, Universidad de Málaga.