

User guide for GHOAT.py – v1.0

1. Introduction

The Guest-HOst Affinity Tool (GHOAT.py) is a python tool designed to fully automate absolute binding free energy calculations on guest-host systems, starting only from an initial structure. It takes advantage of the large performance increases brought by the *pmemd.cuda* software from AMBER20, which can be run on common Graphics Processing Units (GPUs). In addition to their role as catalysts, guest-host systems are important since they provide small test systems for binding free energy calculations, which can be employed for parameter evaluation and optimization.

In this user guide we will first describe the theory and methodology of GHOAT.py, with the simultaneous decoupling and recoupling (SDR) approach combined with the application/removal of restraints on the guest and host. We will then explain then how the equilibrium simulations and free energy calculations are performed, and how they are analyzed in order to obtain the quantities of interest. All the parameters needed for the program input file, and how they apply to the various calculation steps, will also be described in detail. Finally, we will explain how to add a new host to our automated protocol, in addition to the ones provided by the GHOAT.py distribution.

2. Theory and method

The expression for the calculated binding free energy is defined as follows [ref]:

$$-\Delta G_{bind}^o = \Delta G_{h,att} + \Delta G_{g,conf,att} + \Delta G_{g,TR,att} + \Delta G_{sdr} + \Delta G_{g,TR,rel} + \Delta G_{g,conf,rel} + \Delta G_{h,rel} \quad (1)$$

In the equation above, the *att* index denotes attachment of restraints in the bound state, and *rel* indicates release of restraints with the guest and host separated, both in bulk solvent. The *h* and *g* indexes are for host and guest, respectively, *conf* is for conformational restraints and *TR* is for translational/rotational restraints. The ΔG_{sdr} term is the free energy of transferring the ligand from the receptor binding site to bulk with all restraints applied, using the SDR method. Each of these free energy components will be calculated using a series of simulations, as explained below.

2.1 Restraint setup

As shown above, the applied restraints can either be conformational (*conf*), meaning that they are applied in atoms belonging to the same molecule (host or guest), or translational/rotational (*TR*), which are restraints on the guest relative to the host.

The conformational restraints for both the guest and the host use the same procedure as in reference [ref], with harmonic restraints applied on all non-hydrogen dihedrals of a given molecule. These restraints are applied/released in the two ends of the calculation, and are designed to limit the conformational freedom of the host and guest during the SDR process. As in Ref [ref], their contribution to the final binding free energy is calculated using a number of simulation windows with intermediate values of the harmonic spring constants, and the result is processed using the Multistate Bennett Acceptance Ratio (MBAR) method [ref]. For the attaching process, the procedure is applied to the guest-host complex, and the release is performed with the two molecules in separate boxes.

The *TR* restraints of the guest relative to the host use three anchor atoms in the guest and three

in the host, being applied to one distance, two angles and three dihedrals formed between them (Figure). They are first applied on the bound system with the conformational restraints present on the two molecules, using a series of windows and MBAR to retrieve the *TR* attach free energy, as done in the conformational case. For their release, the following analytical expression is used [refs]:

$$\Delta G_{g,TR,rel} = k_B T \ln \left(\frac{C^o}{8\pi^2} \right) = k_B T \ln \int_0^\infty \int_0^\pi \int_0^{2\pi} \exp[-\beta(u_r + u_\theta + u_\phi)] r^2 \sin \theta d\theta d\phi dr$$

$$+ k_B T \ln \int_0^\infty \int_0^\pi \int_0^{2\pi} \exp[-\beta(u_\Theta + u_\Phi + u_\Psi)] \sin \Theta d\Theta d\Phi d\Psi dr \quad (2)$$

Here C^o is the standard concentration, $1 \text{ M} = 1/1661 \text{ \AA}^3$, and r , θ and ϕ are the distance between the H1 and G1 anchors (H1-G1), angle H2-H1-G1, and H3-H2-H1-G1 dihedral, respectively. In the last term on the right, which integrates over guest orientation relative to the host, Θ is the angle H1-G1-G2, Φ is the dihedral H2-H1-G1-G2, and Ψ is the dihedral H1-G1-G2-G3. The u terms are the potential energies from the harmonic restraints, defined as $u = k(x - x_0)^2$, with x being a given coordinate with its reference value x_0 , and k the spring constant. The $1/2$ term is omitted following the AMBER definition of harmonic restraints between single atoms.

2.2 Anchor atom assignment

The host anchors H1, H2 and H3 are pre-determined for a particular host, and have to be included in the GHOAT input file. Instructions on how to assign the anchors for a new host, as well as other host parameters, are shown in section 5.1. The guest anchors G1, G2 and G3 are automatically determined in the beginning of the equilibrium stage, as well as in the beginning of the free energy calculations, the latter starting from the equilibrated structure.

The choice of the guest anchor atoms is based on the position of the host anchors, using a simple procedure, similar to the one described previously [ref]. The first guest anchor G1 is the one closest to the center of mass of three host anchors H1, H2 and H3, and G2 is the one which provides the H1-G1-G2 angle closest to 90 degrees. Similarly, the G3 guest anchor is the one that provides the G1-G2-G3 angle closest to 90 degrees (figure). This is done to avoid gimbal locks in the dihedral restraints, which could cause crashes in the simulation due to large applied forces.

2.3 SDR procedure

The SDR method has been described previously in the BAT.py code [ref], so here we go over its general aspects, as well as the changes made for the GHOAT code. The simultaneous decoupling and recoupling process decouples a restrained guest from the rest of the system when the guest is bound to the host, and at the same time recouples a restrained guest in the same system, but away from the complex and considered in bulk solvent (figure). This simultaneous process allow for the computation of the binding free energy of ligands that carry a net charge, without the need for analytical corrections, which would be required if the decoupling/recoupling happened in separate boxes [refs].

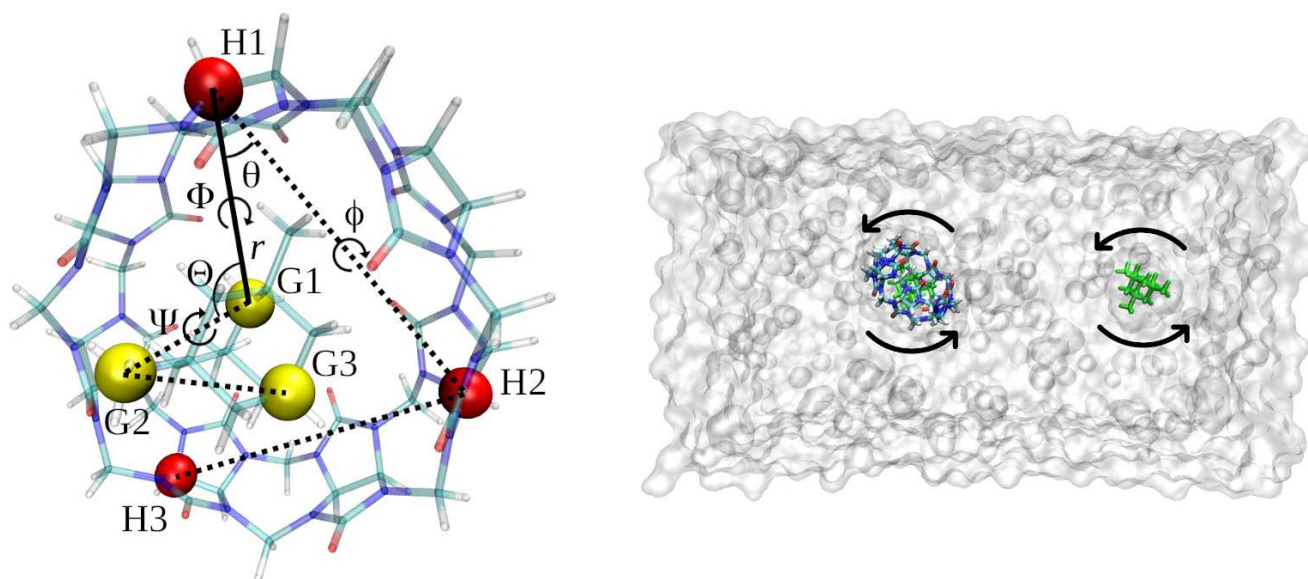
In order to keep the guest-host complex and the bulk guest away from each other during the SDR calculations, center of mass (COM) restraints are applied to all non-hydrogen atoms of the host molecule from the complex. The same way, COM restraints are applied to all non-hydrogen atoms of the guest molecule that is located in bulk solvent (figure). This ensures that the conformational space of both molecules are not affected by the COM restraints, even though this space is already limited by the *conf* restraints. Note that both molecules are still allowed to rotate around their center of mass, which

does not affect the calculations, as long as they don't get too close together. The distance between them can be optimally chosen in the input file, as explained in Section 4.

The SDR calculation is performed separately for the electrostatic and Lennard-Jones components of the guest-host interactions, using the expression:

$$\Delta G_{sdr} = \Delta G_{elec} + \Delta G_{LJ} \quad , \quad (3)$$

where the subscript *LJ* denotes Lennard-Jones interactions, and *elec* the electrostatic interactions. The calculation of these free energy contributions takes place through a series of simulation windows that are ran independently, with the final free energy value being computed using either Thermodynamic Integration with Gaussian Quadrature (TI-GQ), or the MBAR method as with the restraint calculations.



3. Equilibrium simulations

Starting from an initial structure of the guest-host complex, GHOAT prepares the system for the equilibrium simulations, so that the free energy calculations will (hopefully) start from a free energy minimum. The necessary parameters are also generated at this stage, if they are not already provided by the user. GHOAT is able to use the GAFF or GAFF2 parameters for the bonded and LJ interactions, and the AM1-BCC charge model for the partial charges of the guest and host.

The equilibrium stage consists of a series of simulations, in which the restraints on the guest relative to the host are gradually released, followed by a final unrestrained simulation. The final state of the complex after this procedure will be the reference state for all the free energy calculations from the next section.

3. Free Energy Components

As

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 host conformational restraints	a	Complex	MBAR	$\Delta G_{h,att}$
Attachment of guest conformational restraints	l	Complex	MBAR	$\Delta G_{g,conf,att}$
Attachment of guest TR restraints	t	Complex	MBAR	$\Delta G_{g,TR,att}$
Simultaneous dec/recoupling of ligand charge interactions	e	Complex + bulk guest	MBAR/TI	ΔG_{elect}
Simultaneous dec/recoupling of ligand LJ interactions	v	Complex + bulk guest	MBAR/TI	ΔG_{LJ}
Release of guest TR restraints	b	Guest only	Analytical	$\Delta G_{g,TR,rel}$
Release of guest conformational restraints	c	Guest only	MBAR	$\Delta G_{g,conf,rel}$
Release of host conformational restraints	r	Host only	MBAR	$\Delta G_{h,rel}$

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 0. The number of windows and their properties can be defined in the input file. The letters also identify the free energy output files, which are stored in the ./data folder of each component, after the analysis is performed. More information on the nature of each of the restraints, and the free energy methods we use, can be found in Refs. [7,8].

4. Input file

Various options concerning the creation of the systems, simulations and analysis, can be chosen in the input file:

calc_type: Accepts the options “dock” or “crystal”, for a receptor ligand pair of pdb files, or a complex co-crystal structure, respectively. The system initial pdb files should be located in the ./all-poses folder.

celpp_receptor: Sets the name of the receptor, followed by _docked. For example, choose “LMCSS-5uf0_5uez” for a receptor file called LMCSS-5uf0_5uez_docked.pdb. The naming is based on the CELPP challenge procedure. For a crystal structure, put the four letter identifier for the structure, for example “5uf0” for the 5uf0.pdb crystal structure.

poses_list: The list of poses that will be used for the calculations. The list should be placed in brackets and separated by commas. Ex: “[0,1,2,3,4]”. The docked poses files in the ./all-poses folder must be listed accordingly as pose0.pdb, pose1.pdb, pose2.pdb, etc. This parameter is not used for crystal structure calculations.

ligand_name: The ligand residue name in the docked poses above, if **calc_type** is set to “dock” (Ex: “LIG”), or in the initial crystal structure, if **calc_type** is set to “crystal” (Ex: “89J” for the 5uf0 ligand.).

P1, P2 and P3: These define the anchor atoms of the receptor, which have to be determined beforehand. The original protein sequence numbering should be used here, using AMBER masks to define each atom. Ex: “:403@CA” for the CA atom of residue 403.

fe_type: Type of binding free energy calculation. If double decoupling with restraints will be performed, choose “dd-rest”. For only the DD components without computing the free energy of attaching/releasing restraints, choose “dd”. The same goes for “sdr-rest” or “sdr”, when using the SDR method, and “pmf-rest” or “pmf” when using APR. One can also choose the option “custom”, for a chosen set of components.

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 e v]”. When using the custom option with the **e** and **v** components, set the **dec_method** variable to either “dd” or “sdr”. This is necessary to either build the system with only the complex (DD), or with an additional ligand in bulk (SDR), as shown in Table 1.

dec_method: Choice between the DD and SDR methods to build the **e** and **v** components, choosing “dd” or “sdr” for this variable. Only needed when the “custom” option is chosen for the **components** variable, in other cases it will follow the **fe_type** choice.

sdr_dist: Distance (in Å) between the bound ligand and the copy of the ligand located in bulk solvent (measured along the z axis), as required for the SDR method. The value of this variable should be large enough that the interactions of the complex with the bulk copy of the ligand are negligible.

release_eq: The weights for the gradual release of the restraints in the equilibrium stage, going from 100 (fully restrained) to 0 (unrestrained). Each option will be a new simulation, and they are performed in sequence. Use a list of letters separated by spaces inside a bracket to define these weights. Ex: “[5.00 2.50 1.00 0.00]”. A single 0.00 inside the brackets (Ex: “[0.00]”) will run just one equilibrium simulation without any ligand restraints.

attach_rest: List of weights for the spring constant of each window during the attaching/releasing of restraints using MBAR (components **a**, **l**, **t**, **c** and **r**). 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.

`translate_apr`: Windows distances (in Å) for the umbrella sampling (pulling) procedure of APR, identified by the letter **u**. It starts from 0.00 (bound state) until the desired reference distance between the receptor and the ligand in the unbound state. The number of windows is the size of the array. Ex: “[0.00 1.00 2.00 3.00 4.00 5.00]” for a total of 6 windows, ending 5.00 Å away from the initial reference distance between N1 and L1, which is 5.00 Å by construction.

`lambdas`: Lambda values for the decoupling procedure in DD and SDR, going from 0.00 to 1.00. Ex: For a 12-point Gaussian quadrature, choose “[0.00922 0.04794 0.11505 0.20634 0.31608 0.43738 0.56262 0.68392 0.79366 0.88495 0.95206 0.99078]” for the lambda array values.

`pull_ligand`: Choice to pull the ligand from the binding site or not during preparation. This is needed for the APR method, but not needed for DD and SDR. Choose “yes” or “no” for this option.

`pull_spacing`: The interval between each ligand position during the preparation simulations, if the option above is set to yes. The final distance is the last value in the `translate_apr` array. Ex: “0.1” for a pulling interval of 0.1 Å.

`rec_distance_force`: Distance spring constant for the receptor rigid body restraints, identified as D2 in Ref. [8]. Use units of kcal/mol.Å².

`rec_angle_force`: Angle and torsion angle spring constants for the receptor rigid body restraints, identified as A3, A4, T4, T5 and T6 in Ref. [8]. Use units of kcal/mol.rad². The forces from `rec_distance_force` and `rec_angle_force` are included to keep the receptor in the laboratory reference frame.

`rec_dihcf_force`: Final spring constant for the protein conformational dihedral restraints, if this option is activated (see `rec_bb` variable). The nature of these restraints, and how they are implemented, are explained in Ref. [9]. Use units of kcal/mol.rad².

`rec_discf_force`: Final spring constant for the protein conformational distance restraints between the anchor atoms. Use units of kcal/mol.Å².

`lig_distance_force`, `lig_angle_force`, `lig_dihcf_force` and `lig_discf_force`: Final spring constants for the ligand restraints, defined the same way as the receptor above. The value of `lig_distance_force` also designates the spring constant used during the pulling simulations. The nature of the ligand conformational restraints, and how they are implemented, are explained in Ref. [9].

`water_model`: The water model used in the calculations. Supported options are “TIP3P”, “TIP4PEW” and “SPCE”.

`num_waters`: Number of waters used in the simulations of the complex and the *apo* protein.

`buffer_x` and `buffer_y`: Options for the water padding in the x and y axes of the system, remembering that the pulling is done along the z coordinate. The dependent variable is the padding in the z-axis, so make sure you have enough waters to cover the complex, and also allow the pulling of the ligand if APR is used.

`lig_buffer`: Water padding in the three Cartesian axes for the box with only the ligand in it.

`neutralize_only`: Option to add ions only to neutralize the system, or to also include an additional number of ions. Accepts options “yes” or “no”.

`cation` and `anion`: Cation and anion species to be used, accepts all ions supported by the Joung and Cheatham monovalent ion parameters [10]. Ex: “Na⁺” and “Cl⁻”.

`num_cations`: Number of cations to be added after neutralization, for the desired ion concentration, for simulations of the complex and the *apo* protein. The number of anions is the dependent variable, since the systems are always neutral.

`num_cations_ligbox`: Number of cations to be added after neutralization, for the desired ion concentration, for the smaller ligand box.

`hmr`: Use hydrogen mass repartitioning [11] or not. Accepts options “yes” and “no”.

`temperature`: Temperature of the simulated systems, in Kelvin (K).

`eq_steps1`: Number of steps for each simulation of the gradual release of restraints, during the equilibration procedure.

`eq_steps2`: Number of steps for the last simulation of the equilibration procedure, in which the ligand is unrestrained.

`prep_steps1`: Number of steps for the first simulation of the preparation step, in which the ligand is fully restrained in the bound state.

`prep_steps2`: Number of steps for each of the pulling simulations, during the preparation procedure, if `pull_ligand` is set to yes. The distance separation of each of these steps is defined in the `pull_spacing` option.

`[component]_steps1`: Number of steps of equilibration, for each window of 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 stage of each window of the various components of the free energy calculation, in which data is collected.

`rec_bb`: Choice to use or not protein (receptor) backbone dihedral restraints, accepting “yes” or “no”.

`bb_start` and `bb_end`: If the option above is activated, these variables define the residue range of the protein backbone dihedral restraints, using the original protein sequence numbering.

`bb_equil`: Choice to keep the protein backbone restraints fully attached during the ligand equilibration, or leave it unrestrained so it can adapt to the docked ligand. If the latter is chosen, the reference coordinates for the backbone restraints come from the final state of equilibration. Accepts “yes” or “no”.

`l1_x`: distance in the x axis between the first protein anchor atom P1 and the center of the “strike zone” for the search of the ligand first anchor L1. More details on this procedure can be found in section 6 of this guide and also in Ref. [9].

`l1_y`: Same as the previous one, but for the y axis distance.

`l1_z`: Minimum distance between the first protein anchor atom P1 and the first ligand anchor L1, in the z axis.

`l1_zm`: Maximum distance between the first protein anchor atom P1 and the first ligand anchor L1, in the z axis.

`l1_range`: Size of the “strike zone” for the first ligand anchor atom search, which is a square with sides having twice the value of this parameter ($2 \times l1_range$).

`min_adis` and `max_adis`: Minimum and maximum distance between the ligand anchors.

`dec_int`: Type of integration method for the decoupling/recoupling components of the binding free energy calculation (e, v, f and w). If “TI” is chosen, Gaussian quadrature is applied, if “MBAR” is chosen, the latter is used to calculate these components. Remember that the lambda values have to be suitable for either type of integration method.

`weights`: Weights for Gaussian quadrature calculations, in case the TI option is chosen above. These values must correspond to the values in the `lambdas` array, for the procedure to be applied correctly. In the case of a 12-point Gaussian quadrature, write “[0.02359 0.05347 0.08004 0.10158 0.11675 0.12457 0.12457 0.11675 0.10158 0.08004 0.05347 0.02359]” for this variable.

`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. This option is also used for the calculation of the uncertainties of each free energy component [8].

`ntpr`, `ntwr`, `ntwe`, `ntwx`, `cut`, `gamma_ln`, `barostat` and `dt`: Options for running the various simulations, such as output frequency, non-bonded cutoff, barostat type, time step, and others. These use the same variables as the ones from the *pmemd.cuda* simulation input file, and their definitions can be found in the AMBER user guide.

`receptor_ff`: Choice of force field for the receptor atoms, such as “protein.ff14SB” or “protein.fb15”. Also supports parameters for DNA and RNA, such as “DNA.OL15” and “RNA.OL3”.

`ligand_ff`: Choice of force field for the ligand Lennard-Jones and bonded parameters. Accepts either “gaff” or “gaff2”. The ligand partial charges are always determined by the AM1-BCC model.

`ligand_ph`: Reference pH for the protonation of the ligand, before generating parameters.

5. Adding new ligands to a given protein

In the example provided in the BAT folder, binding free energy calculations are performed for the 5uf0 crystal structure, as well as 5 docked poses of the same ligand docked to the receptor with the 5uez structure. The protein receptor is the second bromodomain of the BRD4 protein – BRD4(2). The `./all-poses` folder in this

example contains the original 5uf0.pdb file, as well as a set of pdb files for the docked poses and receptor. The same procedure can be applied to any ligand that binds to BRD4(2), as explained below:

5.1 Docked complexes: In order to generate a set of pdb files for the docked poses and receptor, there are two options, either do a manual docking with chosen input files, or using the CELPP challenge workflow. Both options use AutoDockTools [12], Chimera [13], and AutoDock Vina [14] to prepare the files and run the docking, so you must have them in your path in order to perform this stage. For the first option, the ./docking-files/Vina-example folder has a simple docking workflow using the *dock.bash* script and the input files for the 5uf0 system, which already outputs the files in the right format for use with BAT.

The docking can also be integrated into the CELPP challenge, using the procedure from the CELPPade tutorial (<https://doc.google.com/document/d/1iJcPUktbdrRftAA8cuVa32Ri1TPr2XvZVqTccDja2OMh>), which the user needs to be familiarized with. The same scripts from the “internal_autodockvina_contestant” model from this tutorial can be used for docking, except for *internal_autodockvina_contestant_dock.py*. To output the docked structures suitable for use with BAT, you should use instead the *BAT_dock.py* script, which is included inside the ./docking-files/CELPP-docking folder. Once you run the CELPP docking, the necessary pdb structures for various different receptors will be inside the ./4-docking folder that was just created with the docking results.

5.2 Co-crystal structure: In the case of a co-crystal structure, the *calc_type*, *ligand_name* and *celpp_receptor* options in the input file have to be adjusted properly, as explained in section 4. The script *prep-crystal.tcl*, inside the ./BAT/build_files folder, is used by VMD [15] to “clean” the original file and leave only a single chain of the receptor-ligand complex. The editing of this tcl script usually not needed, but is necessary if the original pdb file has more than one chain. Keep in mind that the “MMM” identifier will be replaced by the ligand name, so it does not have to be changed beforehand.

6. Adding a new protein system

BAT.py can be extended to several protein systems, by including a reference alignment file and adjusting a few parameters in the input.in file. The ./systems-library folder contains the necessary data for a few systems, and more can be requested by contacting the author directly. The user can also set up a new protein system, so below we show a few good practices, using the second BRD4 bromodomain as an example.

6.1: Aligning the protein: The first step is to create the reference.pdb file, so that the complex can be aligned relative to it using MUSTANG, before the BAT.py simulations are performed. If using APR, the ligand pulling direction is along the z axis and towards positive values (z+), so the ligand must have free access to the solvent along this direction. This is not needed if only DD is to be employed. The reference file is created by rotating a structure of the desired protein with a ligand bound, in this example the 5uf0 crystal structure of BRD4(2), and then saving it in the ./BAT/build_files folder as reference.pdb. One way of doing this is using VMD (<http://www.ks.uiuc.edu/Training/Tutorials/vmd/vmd-tutorial.pdf>), but other programs such as Chimera can also be employed for that purpose. Figure 1 shows the 5uf0 structure before (red) and after the rotation (blue), with the ligand now having access to the solvent along the z direction. The reference.pdb file does not need to have the same sequence as the receptor input file, so the same reference created here for BRD4(2) can be extended to other bromodomain homologs that share the same fold, such as BRD4(1), CREBBP and BAZ2B.

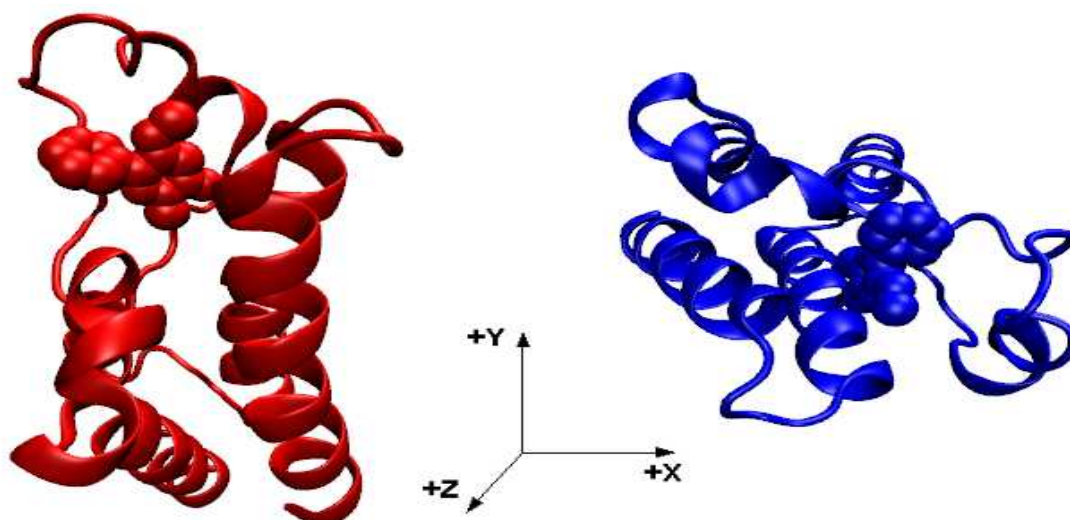


Figure 1: The 5uf0 structure before (red) and after (blue) the rotation, so that the latter has the ligand with free access to the solvent in the +z direction.

6.2: Choosing the protein anchors: Once the reference.pdb file with the desired orientation is created, it is time to choose the three protein anchor atoms. Starting from this file, choose a tentative ligand first anchor atom L1 [8], with the lowest or one of the lowest values for the z coordinate from all the ligand atoms, and with the x and y coordinates not too far from the center of the binding site (Figure 2). Even though the L1 anchor is going to be chosen automatically when you run BAT.py, an estimate of its location is needed to choose the protein anchors. The protein P1 anchor is then chosen using a few rules:

- 1 – Should be a backbone atom (CA, C or N) and part of a stable secondary structure such as an alpha-helix or a beta sheet, always avoiding loop regions due to their increased flexibility.
- 2 – There should be a distance ranging from 10 Å to 15 Å between P1 and the L1 anchor atom in the z axis, and an absolute value between 5.0 Å and 10 Å for this same distance in the xy plane.

In the example for the 5uf0 structure (Figure 2), the tentative L1 atom has a $\{x1\ y1\ z1\}$ distance vector from the CA atom from the protein 403 residue (P1) being equal to $\{-0.74\ -6.16\ 13.03\}$, with the $z1$ distance being 13.03 Å and the distance in the xy plane $\sqrt{x1^2 + y1^2}$ being 6.20 Å. Since it satisfies our criteria, we choose :403@CA for the first protein anchor P1.

The choice of the other protein anchors P2 and P3 are chosen after the first one, and also follow a few guidelines:

- 1 – Should also be backbone atoms part of stable secondary structures such as alpha-helix or beta sheet, always avoiding loop regions due to their increased flexibility.
- 2 – The N2-P1-P2 and P1-P2-P3 angles (Figure 2) should not be close to 0 or 180 degrees (better if close to 90°), and the distance between the protein anchors should be large (at least 10 Å). This is to avoid large forces during the simulations due to a gimbal lock. Like the ligand anchors, the positioning of the dummy atoms is done automatically, with an example shown in Figure 2. More information on their location and the full restraint setup can be found on Refs [8,9].

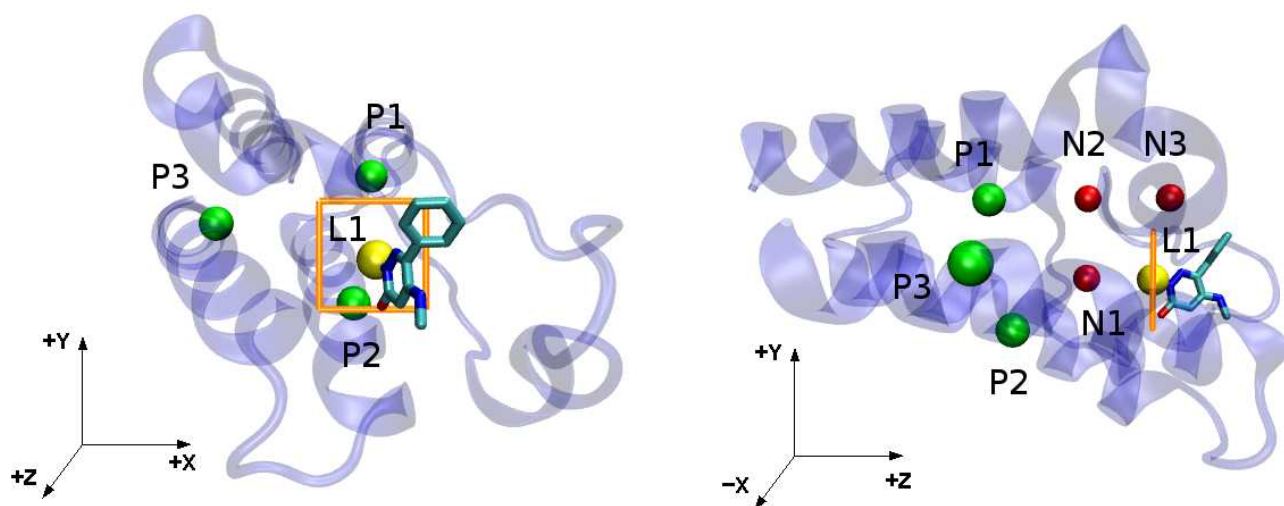


Figure 2: (left) The ligand tentative first anchor L1 (yellow) and the three protein anchors P1, P2 and P3 (green). The strike zone for L1 search is shown in orange. (right) The same structure, but showing the the yz plane. The dummy atoms N1, N2 and N3 (red), which are placed automatically when running BAT.py, are shown as an example.

6.3: Determining the input values for ligand anchor search: Once P1, P2 and P3 are chosen, a few more parameters are needed for the input file. The *x1* and *y1* coordinates determined above can be used for the *l1_x* and *l1_y* parameters. The minimum (*l1_z*) and maximum (*l1_zm*) P1-L1 distances in the z axis can have safe values of 8.50 Å and 20.0 Å, respectively. The *l1_range* parameter, which defines a “strike zone” when searching for the first ligand anchor L1, can also be safely defined as 3.0 Å (Figure 2). The *min_adis* and *max_adis* define the minimum and maximum distance between the ligand anchors, and can usually be left with values of 4.0 Å and 8.0 Å, respectively. For smaller ligands, *min_adis* might have to be reduced to 3.5 Å or even 3.0 Å, and *max_adis* could be increased in the case of larger ligands.

Even though this is a heuristic approach, if applied correctly the protein and the ligand will be properly restrained during the calculations, allowing us to obtain the absolute binding free energy of several molecules to a given protein without any further adjustments.

7. References

- [1] D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco.
- [2] J. Wang, W. Wang, P.A. Kollman, and D.A. Case.. (2006) "Automatic atom type and bond type perception in molecular mechanical calculations". *Journal of Molecular Graphics and Modelling*, **25**, 247-260.
- [3] J. Wang, R.M. Wolf, J.W. Caldwell, and P. A. Kollman, D. A. Case (2004) "Development and testing of a general AMBER force field". *Journal of Computational Chemistry*, **25**, 1157-1174.

- [4] A. Jakalian, B. L. Bush, D. B. Jack, and C.I. Bayly (2000) "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method". *Journal of Computational Chemistry*, **21**, 132-146.
- [5] A. Jakalian, D. B. Jack, and C.I. Bayly (2002) "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". *Journal of Computational Chemistry*, **16**, 1623-1641.
- [6] A. S. Konagurthu, J. Whisstock, P. J. Stuckey, and A. M. Lesk. (2006) "MUSTANG: A multiple structural alignment algorithm". *Proteins*, **64**, 559-574.
- [7] G. Heinzelmann, N. M. Henriksen, and M. K. Gilson. (2017) "Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain" *Journal of Chemical Theory and Computation*, **13**, 3260-3275.
- [8] 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.
- [9] M. R. Shirts and J. Chodera (2008) "Statistically optimal analysis of samples from multiple equilibrium states." *Journal of Chemical Physics*, **129**, 129105.
- [10] I. S. Joung and T. E. Cheatham III (2008). "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations". *The Journal of Physical Chemistry B*, **112**, 9020-9041.
- [11] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. (2015) "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning". *Journal of Chemical Theory and Computation*, **11**, 1864-1874.
- [12] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. (2009) "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility" *Journal of Computational Chemistry*, **30**, 2785-2791.
- [13] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. (2004). "UCSF Chimera - A Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry*, **25**, 1605-1612.
- [14] O. Trott and A. J. Olson. (2010) "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading," *Journal of Computational Chemistry*, **31**, 455-461.
- [15] W. Humphrey, A. Dalke and K. Schulten. (1996) "VMD - Visual Molecular Dynamics", *Journal of Molecular Graphics*, **14**, 33-38.