

JYX



This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Jansen, Anton; Aho, Noora; Groenhof, Gerrit; Buslaev, Pavel; Hess, Berk

Title: phbuilder : A Tool for Efficiently Setting up Constant pH Molecular Dynamics Simulations in GROMACS

Year: 2024

Version: Published version

Copyright: © The Authors. Published by American Chemical Society

Rights: CC BY 4.0

Rights url: <https://creativecommons.org/licenses/by/4.0/>

Please cite the original version:

Jansen, A., Aho, N., Groenhof, G., Buslaev, P., & Hess, B. (2024). phbuilder : A Tool for Efficiently Setting up Constant pH Molecular Dynamics Simulations in GROMACS. *Journal of Chemical Information and Modeling*, 64(3), 567-574. <https://doi.org/10.1021/acs.jcim.3c01313>

phbuilder: A Tool for Efficiently Setting up Constant pH Molecular Dynamics Simulations in GROMACS

Anton Jansen,* Noora Aho, Gerrit Groenhof, Pavel Buslaev,* and Berk Hess*



Cite This: <https://doi.org/10.1021/acs.jcim.3c01313>



Read Online

ACCESS |



Metrics & More



Article Recommendations



Supporting Information

ABSTRACT: Constant pH molecular dynamics (MD) is a powerful technique that allows the protonation state of residues to change dynamically, thereby enabling the study of pH dependence in a manner that has not been possible before. Recently, a constant pH implementation was incorporated into the GROMACS MD package. Although this implementation provides good accuracy and performance, manual modification and the preparation of simulation input files are required, which can be complicated, tedious, and prone to errors. To simplify and automate the setup process, we present phbuilder, a tool that automatically prepares constant pH MD simulations for GROMACS by modifying the input structure and topology as well as generating the necessary parameter files. phbuilder can prepare constant pH simulations from both initial structures and existing simulation systems, and it also provides functionality for performing titrations and single-site parametrizations of new titratable group types. The tool is freely available at www.gitlab.com/gromacs-constantph. We anticipate that phbuilder will make constant pH simulations easier to set up, thereby making them more accessible to the GROMACS user community.



INTRODUCTION

Molecular dynamics (MD) simulations of biomolecular systems are nowadays widely applied to study a broad range of research questions, from protein folding¹ to drug design.² While parameters such as temperature and pressure are routinely taken into account in MD simulations, pH, a critical parameter influencing the structure, dynamics, and function of many biomolecules,³ cannot be controlled in standard MD simulations. To overcome this limitation and capture the effects of pH, constant pH simulation methods have been developed.^{4–28} We have recently presented an efficient implementation of constant pH for the GROMACS MD package²⁹ based on the λ -dynamics approach first introduced by Brooks and co-workers in 1996.³⁰ Similar implementations are also available in other popular MD software packages such as CHARMM²² and Amber.²⁶

While progress has been made in refining the methodological aspects of constant pH MD, the setup of such simulations remains complicated. Setup requires several steps, including preparation of the initial structure and modifying the topology and parameter files. Structure preparation requires adding dummy hydrogens to all titratable groups in the system; topology modification requires assigning charges for all protonation states of titratable groups, and modification of the parameter files includes the addition of parameters specific for constant pH MD simulations. Although these are general steps for constant pH MD,³¹ the implementation details of such setup steps can differ significantly between MD engines and often require time-consuming and error-prone manual modifications. Setting up constant pH MD also requires a deep understanding of the

system's chemistry and force field formatting, which are also specific to a particular MD engine.

Constant pH MD simulations are not the only type of simulation that is difficult to set up. Over the years, several tools have been developed to aid in the setup of MD simulations: CHARMM-GUI³² is a web environment that allows the user to generate input files of a wide range of biomolecular systems for various simulation approaches. PACKMOL³³ can generate structures of complex biomolecular systems, while BioSimSpace³⁴ is a Python framework that allows one to set up and run simulations with different software engines. Several tools were also developed to automate preparation steps for free energy simulations.^{35–40} Such tools simplify the setup process by automating the time-consuming and error-prone generation of simulation input files. However, there is currently no such tool for automatically setting up constant pH simulations for GROMACS.

To bridge this gap, we have developed phbuilder, a Python-based command line tool that automates the constant pH specific preparation steps when setting up MD simulations. phbuilder generates the required topology and parameter files, and it also adds the appropriate number of ions and neutralizing buffer particles^{27,28} to keep the simulation box

Received: August 16, 2023

Revised: December 5, 2023

Accepted: December 6, 2023

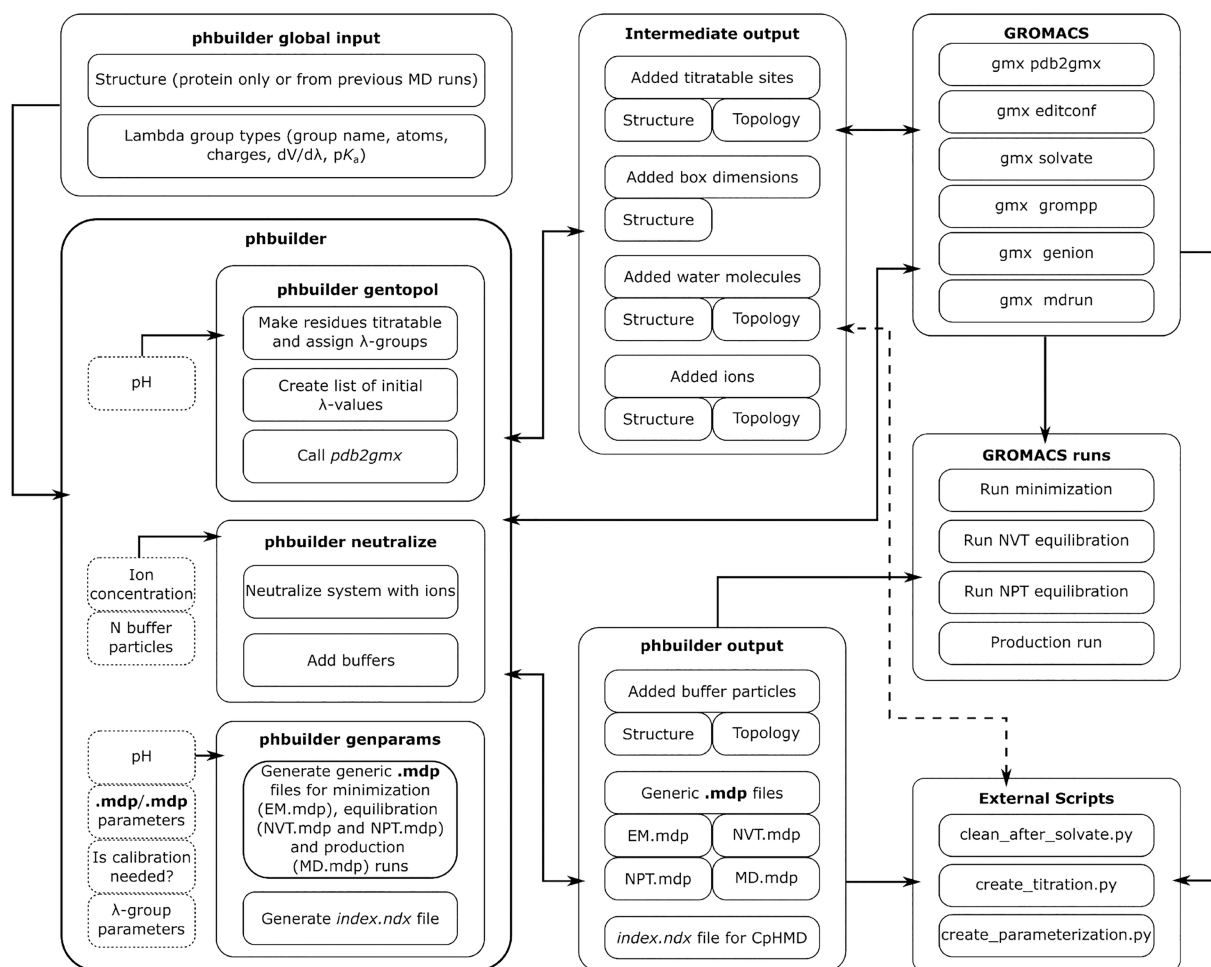


Figure 1. Schematic of the internal structure of phbuilder, as well as the movement of various input and output files and the external GROMACS programs and scripts utilized. phbuilder consists of three tools: **gentopol**, **neutralize**, and **genparams**. **gentopol** prepares the structure and topology files. **neutralize** ensures the appropriate number of ions and buffer particles for a net-neutral system. **genparams** generates the .mdp and .ndx parameter files required for running constant pH MD in GROMACS.

neutral during the constant pH MD run. To demonstrate the tool, we used it to prepare constant pH MD simulations starting from both separate structure files and existing simulation systems. Additionally, phbuilder provides functionality through the inclusion of stand-alone Python scripts for setting up titrations and parameterizations for new titratable group types. Currently, only parametrization of single-site titratable groups is supported. Multisite parametrization will be addressed in a future work and merged into phbuilder when ready. We demonstrate the parametrization feature by parametrizing a single-site model of arginine.

METHODS

phbuilder Implementation. phbuilder is a Python-based command line tool that assists in setting up and performing constant pH MD simulations with GROMACS.^{27–29} Both the tool and a user manual (including tutorials) are available at www.gitlab.com/gromacs-constantph. In simple cases such as a globular protein in a box of water, the only required input is a structure file (.pdb or .gro) and the desired simulation pH. The input can come from either an initial structure file or a prepared simulation system. During the setup process, the user can optionally provide additional parameters, such as the desired ion concentration and the number of buffer particles to

add. When the structure file is processed, all residue types should be available in the force field. If all residue types are not available, then the user will be prompted to manually provide the missing topologies. Currently, the GROMACS constant pH implementation only supports a modified version of the CHARMM36m force field,⁴¹ which includes dihedral modifications for improved sampling of the dihedral angles of the titratable groups.²⁸ In addition to system-specific input, phbuilder reads from a global parameter file, `lambda-grouptypes.dat`, which stores the topology information on the titratable group types. The details of the `lambda-grouptypes.dat` organization and modification are discussed in the following sections. With the aforementioned input, phbuilder then generates the topology, index, and generic .mdp files required for running constant pH simulations. Scripts for the automation of titration and parametrization runs are also provided.

phbuilder consists of three tools: **gentopol**, **neutralize**, and **genparams**, as presented in Figure 1. First, **gentopol** makes the residues selected by the user titratable and (re)generates the topology accordingly using `gmx pdb2gmx`. **gentopol** also generates a list of initial λ -values for each titratable group (based on the pH value set by the user) and saves the correspondence between λ -coordinates and actual residues in

the system. If necessary, the structure resulting from **gentopol** can be solvated using `gmx editconf` and `gmx solvate` tools.

Because `gmx solvate` can accidentally place water molecules inside the protein's hydrophobic core, we provide a Python script, `clean_after_solvate.py`, which can remove those misplaced water molecules with one of two algorithms implemented:

1. Number of water molecules around other waters. This algorithm is the default. For each water molecule within the specified cutoff R_c of non-solvent molecules (default value for R_c is 5 Å), the number of water atoms $N_{\text{atoms}}^{\text{HOH}}$ within a sphere of radius R (default value is 10 Å) is calculated. For water molecules outside of the protein, this number is large (around 400), while for water molecules inside the protein, this number is smaller than $N_{\text{cutoff}} = 120$ (Figure S1). The latter water molecules are thus removed. The default values for R_c , R , and N_{cutoff} might not be optimal for every system. In case of a severely curved protein shape, it might be useful to increase R_c to get more statistics on water molecules that are far from the protein. In case of smaller simulation system sizes, reducing R is meaningful. But in general, the default values for R and R_c are reasonable. As for N_{cutoff} the default value is based on the distribution of the number of water molecules around selected waters. As demonstrated in Figure S1, $N_{\text{cutoff}} = 120$ clearly separates the majority of water molecules, which are surrounded by many water molecules, from a few waters, which have only a few other water molecules around. We recommend that users test optimal values for the system of interest themselves.
2. Number of water molecules among nearest neighbors. For each water molecule within the specified cutoff R_c of non-solvent molecules, the $N_{\text{neighbors}}$ nearest residues are detected first (default value of $N_{\text{neighbors}}$ is 20). Then, the number of water molecules around those neighbors is calculated. For water molecules far from the protein, almost all the neighbors are other water molecules, while buried residues will have only a few neighbor water molecules (Figure S1C,D). By default, water is removed if the number of water molecules in its neighbor list is less than 3. Similar to the first method, the default R_c and $N_{\text{neighbors}}$ values should be reasonable for most systems. The default cutoff value was selected, as it clearly separates water molecules that are surrounded by other water molecules from those surrounded mainly by the protein (Figure S1C). We recommend that users test optimal parameters for the system of interest themselves.

The user decides which of these two methods is used for removing misplaced water molecules.

The **neutralize** tool ensures a net-neutral charge in the simulation system. First, it adds ions to make the system neutral for the configured initial state. It can also add ions at a given concentration if specified by the user. Second, **neutralize** adds buffer particles, which will compensate for charge fluctuations during constant pH MD simulations.^{27,28,42} By default, the **neutralize** tool enforces a minimum placement distance of 0.6 nm from non-solvent molecules when replacing the solvent with ions and buffer particles. If it is observed that the ions or buffer particles are (re)placed in pockets inside a protein, this may be remedied by manually increasing the

minimum placement distance through the `-rmin` flag. By default, **neutralize** will add as many buffer particles as the total number of titratable sites N_{sites} in the system. However, the number of buffers may also be manually set by the user. The general recommendation is that the λ -value of the buffer should be close to 0.5 (which corresponds to zero charge for the buffers) during the entire simulation.²⁸ While this can be achieved by increasing the total number of buffers to up to twice the number of titratable sites, we nevertheless recommended using smaller buffer concentrations in order to not significantly alter the solvent properties. Thus, the optimal buffer concentration can be selected in the range from $2\sqrt{\pi N_{\text{sites}}}$, the average charge fluctuation,⁴³ to $2N_{\text{sites}}$, the maximal possible charge fluctuation, divided by the maximal buffer charge (0.5).

The **genparams** tool generates the `.mdp` files required for running constant pH simulations with GROMACS. Included in the `.mdp` files are general constant pH parameters, a description of the titratable group types used in the simulation, and a mapping between atoms and titratable group types (see also the original GROMACS constant pH implementation publication²⁷ and the online manual).⁴⁴ General constant pH parameters can be provided by the user together with standard `.mdp` parameters. Entries for titratable group type descriptions are based on the `lambdagrouptypes.dat` file, and the mapping between atoms and titratable group types is generated based on the topology and `lambdagrouptypes.dat`. In addition to the `.mdp` entries, the indices of atoms that are part of titratable groups are added to an `index.ndx` file. **gentopol** also sets up the initial λ -values for all λ -coordinates in the `.mdp` file based on the pH set by the user, and it provides the user the possibility to set up the barrier of the biasing potential²⁷ for each λ -coordinate. The `lambdagrouptypes.dat` file defines the parameters of titratable groups, the path to the GROMACS constant pH installation, and the force field and water model to use. The entries for a single-site titratable group are organized as follows:

```
[ ASPT ]
incl = ASP ASP1 ASPH ASPP ASH
atoms = CB CG OD1 OD2 HD2
qqA = -0.21 0.75 -0.55 -0.61 0.44
pKa_1 = 3.65
qqB_1 = -0.28 0.62 -0.76 -0.76 0.00
dvd_1 = -54.078 -144.280 278.550
-146.030 -554.270 44.621
```

First, the group name is provided inside the square brackets, followed by alternative names, which will be used to search for residues that `phbuilder` should make titratable. Next, the names of atoms that have different charges in different protonation states of the residue are given. The atomic charges of two protonation states should be provided as `qqA` and `qqB_1` entries. The order of the charges should correspond to the order of the atom names. Then, the pK_a value of the titratable group in water as well as the coefficients of the polynomial fit of the correction potential V^{MM} should be provided. For more details on V^{MM} , please check the GROMACS constant pH publication and repository.²⁷

For titratable groups described with the multisite representation (such as histidine with three protonation states), the details of each protonation state should be listed:

```
[ HSPT ]
```

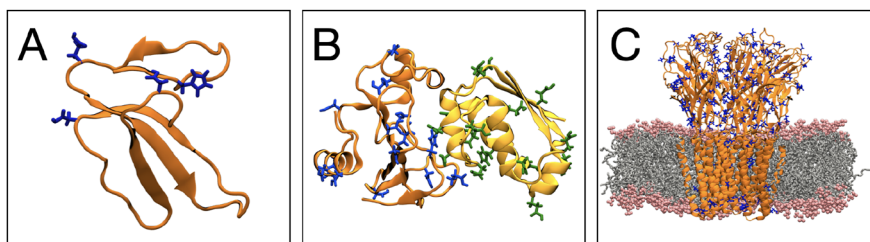



Figure 2. phbuilder was used to set up constant pH MD simulations for a number of different systems: (A) cardiotoxin V (PDB ID: 1CVO),⁴⁸ (B) the Barnase–Barstar protein complex (PDB ID: 1BRS),⁴⁹ and (C) the GLIC ion channel (PDB ID: 6ZGD).⁵⁰ The titratable residues (Asp, Glu, and His residues) are shown as blue sticks in cardiotoxin, GLIC, and Barnase–Barstar chain A and as green sticks in Barnase–Barstar chain B.

```
incl = HIS HIS1 HISA HISB HISH HISD HISE
HISP HSD HSE HSP
atoms = CB CD2 HD2 CG NE2 HE2 ND1 HD1 CE1 HE1
qqA = 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00
pKa_1 = 0.0
qqB_1 = -0.05 0.19 0.13 0.19 -0.51 0.44
-0.51 0.44 0.32 0.18
dvd1_1 = -3015.4 -13444.1 -34779. ...
pKa_2 = 6.53
qqB_2 = -0.08 -0.05 0.09 0.22 -0.36 0.32
-0.7 0.0 0.25 0.13
dvd1_2 = -1695.49 -8107.68 -22614.3 ...
pKa_3 = 6.92
qqB_3 = -0.09 0.22 0.1 -0.05 -0.7 0.0
-0.36 0.32 0.25 0.13
dvd1_3 = -111.044 -768.566 -9092.12 ...
```

The multisite representation allows for the simulation of chemically coupled titratable sites.^{27,28} In general, if a group can be found in N different protonation states, the multisite representation will include $N + 1$ states: one artificial common state and N physical states of the group. The charges for the common state (which are usually zeros) should be provided as qqA . For each physical state, i , the corresponding pK_a values, charges, and polynomial coefficients of the correction potential should be provided as pKa_i , qqB_i , and $dvd1_i$, respectively. The full details on how to select the correct pK_a values and polynomial coefficients can be found in our previous GROMACS constant pH publications.^{27,28}

After the structure and parameters for the constant pH MD simulation are automatically generated by phbuilder (and modified by the user if needed), the system needs to be minimized and equilibrated, after which the constant pH MD simulation can be started. In addition to the tools described previously, three external scripts are provided: (1) `create_parameterization.py`, to set up a parametrization of novel single-site compounds; (2) `fit_parameterization.py`; to compute the polynomial coefficients of the V^{MM} fit from the parametrization runs; and (3) `create_titration.py`, to set up titration runs of the equilibrated systems. Both the titration and parametrization require multiple constant pH MD simulations of the same system with different input conditions (pH in case of titration, fixed λ -values in case of parametrization). The scripts `create_parameterization.py` and `create_titration.py` prepare directories and input `.mdp` files for those runs automatically. `fit_parameterization.py` calculates the optimal V^{MM} fit and provides new entries for `lambdagrouptypes.dat` and `.mdp` files for this group.

Details of the procedure for obtaining V^{MM} are provided in the next section.

Parameterization with phbuilder. The results of constant pH MD simulations depend on the accuracy of the correction potential V^{MM} , which compensates for the quantum mechanical contributions to the proton affinity that are missing from the force field description. These correction potentials are specific for each titratable group and are obtained as polynomial fits to the deprotonation free energy profiles, as we discussed in detail in our previous work.^{27,28} Originally, the derivation of the correction potential V^{MM} was based on thermodynamic integration (TI).⁴⁵ However, TI can be inefficient and can lead to an inaccurate fit of V^{MM} .^{46,47} To avoid these issues, we propose a new two-step scheme for the parametrization of titratable groups: first, we run fast, but not very accurate, TI to obtain an approximated fit of V^{MM} ; then, we reweigh V^{MM} based on long sampling simulations of the group simulated at $pK_a = \text{pH}$ with the biasing barrier set to 0. Under such conditions, the distribution of the λ -coordinate should be flat if V^{MM} is fitted accurately.^{27,28} However, after a fast TI step, this might not be the case. If the potential derived from the fast TI step is insufficient, the correction to the V^{MM} can be calculated using Boltzmann inversion:

$$\Delta V^{MM} = -RT \ln p(\lambda) \quad (1)$$

where $p(\lambda)$ is the probability distribution of λ -coordinates obtained from the sampling runs. After the reweighting step and performing the sampling simulations of arginine at $pK_a = \text{pH}$ with the biasing barrier set to 0, the obtained distributions of the λ -coordinates are flat (Figure S2). phbuilder provides the functionality to perform both steps and outputs the appropriate entries for the `lambdagrouptypes.dat` and `.mdp` files, thereby making the parametrization of new molecules more straightforward. However, we would like to emphasize that parametrization is currently only supported for single-site titratable groups. We are currently developing a version for the parametrization of multisite titratable groups, which we will separately report and include in phbuilder when it is ready.

Simulation Details. To demonstrate the tool's versatility, we set up constant pH MD simulations for the following systems: (1) cardiotoxin V (PDB ID: 1CVO),⁴⁸ (2) the Barnase–Barstar complex (PDB ID: 1BRS),⁴⁹ and (3) the proton-gated ion channel GLIC (PDB ID: 6ZGD).⁵⁰ The structures of these proteins are shown in Figure 2. For cardiotoxin V, we used phbuilder to set up the system starting from a `.pdb` file from the Protein Data Bank⁵¹ as well as from the prepared simulation system. For the Barnase–Barstar complex, phbuilder was used to demonstrate the preparation of a system with multiple chains from a `.pdb` file. We also used

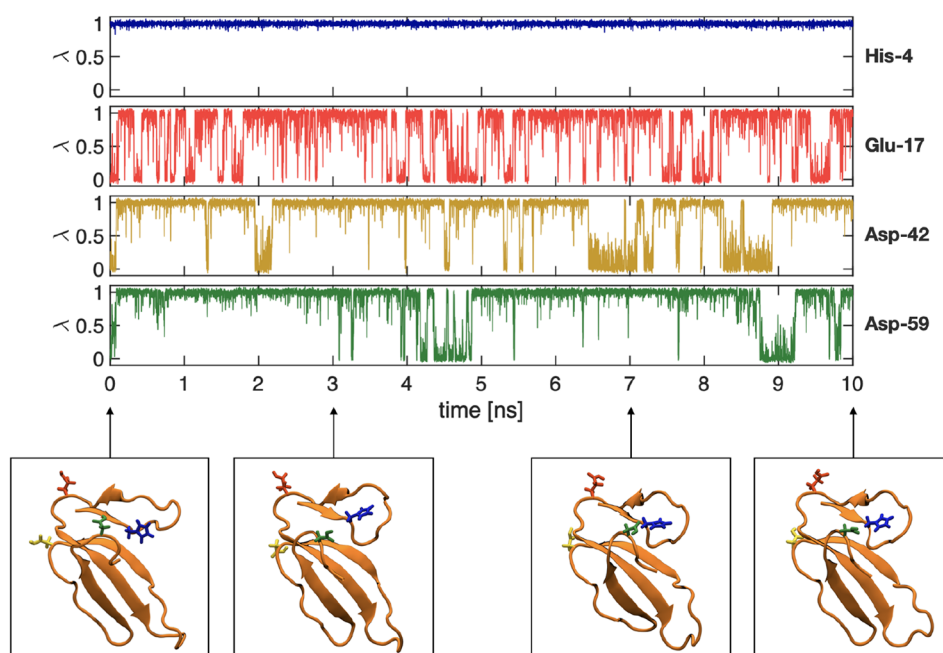


Figure 3. Evolution of the λ -coordinates over time, along with snapshots of the trajectory for a 10 ns constant pH MD simulation of cardiotoxin V (PDB ID: 1CVO)⁴⁸ at pH = 4. The upper panel presents the λ -coordinates as a function of the simulation time for His-4, Glu-17, Asp-42, and Asp-59, and the lower panels show snapshots at 0, 3, 7, and 10 ns. In the snapshots, the titratable groups are shown with the stick representation, and the colors of the groups match those of the λ -coordinate plots.

the tool to set up a constant pH MD simulation for an existing GLIC membrane system.

Additionally, `phbuilder` was utilized to parametrize the coefficients of the V^{MM} polynomial fit for arginine. Arginine has three possible deprotonated states⁵² and ideally requires a multisite description. However, here we only considered a single site of N_{η} , as we want to demonstrate setting up parametrizations for single-site groups, which are the most common. For all cases (starting from a `.pdb` file, starting from a prepared system, and parametrization), we used the same force field and simulation parameters, which are described next.

For all simulation systems, the interactions were modeled using the CHARMM36 all-atom force field⁴¹ with dihedral modifications for improved sampling of the amino acid side-chain rotamers.²⁸ This modified force field already contains the entries for the titratable glutamic acid (Glu), aspartic acid (Asp), and histidine (His) residues as well as the buffer particles.^{27,28}

To prepare the system for cardiotoxin V and the Barnase–Barstar complex (either prior to or while applying `phbuilder`), the following procedure was used: (1) the corresponding `.pdb` file was obtained from the Protein Data Bank.⁵¹ (2) The protein was placed in a cubic box with a 2.0 nm distance between the protein and the box. (3) The system was solvated with ~ 17000 CHARMM TIP3P^{53,54} waters for cardiotoxin V and ~ 31000 for Barnase–Barstar. Finally, (4) a 0.1 M concentration of Na^+ and Cl^- ions were added to neutralize the system. The system for GLIC (prior to applying `phbuilder`) was prepared using CHARMM-GUI,³² and it contained ~ 62000 CHARMM TIP3P water molecules,^{53,54} 489 POPC (1-palmitoyl-2-oleoylphosphatidylcholine) lipids, and a 0.15 M concentration of Na^+ and Cl^- ions. Arginine was simulated as an Ala–Arg–Ala tripeptide with capped termini. The box size was $\sim 5 \times 5 \times 5 \text{ nm}^3$, and the system contained ~ 5000

CHARMM TIP3P water molecules^{53,54} and a 0.15 M concentration of Na^+ and Cl^- ions. For all of the simulated systems, the input configurations, together with the topology and run parameters, are provided (see the [Data and Software Availability](#) section).

In all of the simulations, the same simulation parameters that were recommended for the CHARMM force field were used. The Coulomb interactions were computed with the particle mesh Ewald (PME) method^{55,56} with a real-space cutoff of 1.2 nm and a grid spacing of 0.14 nm. The van der Waals interactions were modeled with the Lennard-Jones potential and smoothly switched to zero in the range from 1.0 to 1.2 nm. A constant temperature was maintained at 300 K using the *v*-rescale thermostat⁵⁷ with a time constant of 0.5 ps^{-1} , and a constant pressure was maintained at 1 bar using the *c*-rescale barostat⁵⁸ with a relaxation time of 5.0 ps. For GLIC, a semi-isotropic barostat was used. In all simulations, the leapfrog integrator and a time step of 2 fs was used. The bond lengths to hydrogen atoms were constrained with the LINCS algorithm,⁵⁹ and the internal degrees of freedom of the water molecules were constrained with the SETTLE algorithm.⁶⁰

Next, details of the constant pH MD simulations constructed and performed by applying `phbuilder` are given. For the protein systems, the aspartic acid (Asp), glutamic acid (Glu), and histidine (His) residues were considered as titratable groups. The reference pK_a values were 3.65 and 4.25 for Asp and Glu, respectively, and 6.53 and 6.92 for His HSD (singly protonated at the N_{δ}) and His HSE (singly protonated at the N_{ϵ}), respectively. For the λ -particles, the mass was set to 5 atomic units, and the temperature was maintained at 300 K by using a separate *v*-rescale thermostat with a time constant of 2.0 ps^{-1} . The barrier height of the biasing potential barrier was set at 7.5 kJ/mol.

After the construction of the systems, the potential energy of each system was minimized by using the steepest descent

algorithm, followed by 10 ps of NVT equilibration and 10 ps of NPT equilibration. During these simulations, the λ -coordinates of all titratable groups were kept constant. Then, a 10 ns constant pH simulation was performed for cardiotoxin V from a .pdb file at pH = 4 and from an existing system at pH = 4, for Barnase–Barstar at pH = 7, and for GLIC at pH = 4. For the parametrization of arginine, 1 ns simulations with λ -coordinates fixed at values ranging from -0.1 to 1.1 with 0.1 steps were first run to obtain an initial guess for the V^{MM} . Then, 10 replicas of 100 ns simulations with pH = pK_a and the biasing barrier set to 0 were run. From those sampling simulations, new coefficients for the V^{MM} polynomial fit were calculated, and another 10 replicas of 100 ns simulations with pH = pK_a and the biasing barrier set to 0 with updated V^{MM} were run.

RESULTS AND DISCUSSION

Here, we demonstrate the tool's efficacy by preparing constant pH MD simulations for initial structure files, for structures from existing MD simulation systems, and for the parametrization of a new titratable group. All systems were subsequently simulated for 10 ns to demonstrate that the system setup was correct.

Setting up Simulations from Protein Structure. Two constant pH MD simulations were set up from initial protein structures. These were cardiotoxin V (four titratable groups) and the Barnase–Barstar complex (30 titratable groups). phbuilder successfully set up both of these systems. We subsequently performed 10 ns simulation runs, and the protonation states of the titratable groups evolved together with the atomic coordinates (Figure 3 for cardiotoxin V and Figure S3 for the Barnase–Barstar complex).

Setting up Simulations for a Prepared System. As mentioned previously, phbuilder can also convert existing simulations to constant pH simulations. To demonstrate this feature, we set up constant pH MD simulations based on an existing cardiotoxin V system and from a membrane protein that was prepared using CHARMM-GUI.³² We used CHARMM-GUI as an example of a preparation workflow for membrane proteins, but phbuilder can work with any workflow that prepares structures for MD simulations, such as inflatgro⁶¹ or g_membed.⁶²

We first set up and performed a 10 ns constant pH simulation of cardiotoxin V from an existing simulation system and made a comparison to the initial cardiotoxin V simulation performed earlier. In both cases, we found that histidine was always protonated, while aspartic acids were predominantly deprotonated and underwent only a few transitions. Glutamic acid showed more than 10 transitions between protonation states (Figures 3 and S4). To show that phbuilder may also be used to set up more complex systems, we next used it to set up a constant pH simulation for the pentameric ion channel GLIC embedded in a POPC bilayer. We successfully managed to take an embedded structure and use the tool to set up the constant pH input files from there. The subsequent 10 ns simulation showed that the protonation states evolved alongside the atomic coordinates (Figure S5).

Parameterization of New Titratable Groups. Obtaining a sufficiently accurate polynomial fit can sometimes be difficult, which is why we introduced a new two-step approach for the parametrization: (i) acquire an initial polynomial fit by performing a fast, but inaccurate, TI calculation to obtain an approximate correction potential; then, (ii) systematically

improve this approximate potential by reweighting the expansion coefficients of the polynomials based on inverse Boltzmann statistics. phbuilder automates the setup of both steps, analyzes the trajectories, and calculates the optimal polynomial coefficients for V^{MM} . Here, we demonstrate this new parametrization protocol by performing a parametrization for a single-site model of arginine. The fast TI step gave somewhat skewed histograms (Figure S2), which were subsequently corrected by applying a reweighting of the approximate fit using inverse Boltzmann statistics, finally yielding consistently flat λ -distributions (Figures 4 and S2).

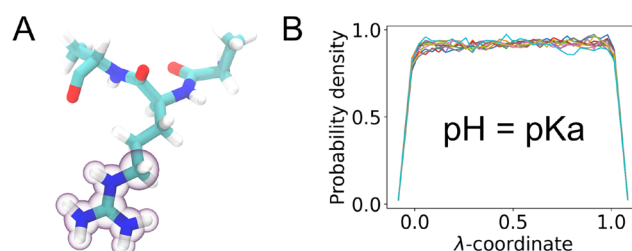


Figure 4. Final parametrization result for arginine. (a) The structure of the tripeptide (titratable atoms highlighted). (b) The distribution of the λ -coordinate for simulations performed at pH = pK_a and with the biasing barrier set to 0.

CONCLUSIONS

Constant pH MD simulations allow for more accurate modeling of pH dependence compared to traditional MD simulations. While a constant pH MD algorithm was recently implemented in the GROMACS MD package,²⁷ it has been proven difficult to efficiently prepare such simulations due to the complex and error-prone nature of their setup process. Here, we have presented phbuilder, a Python-based system builder tool that automates the setup process. It allows the user to set up systems for constant pH MD simulations starting from both initial structures and existing simulation systems. The user can either follow the default protocol, which will generate a generally suitable setup, or provide additional details if some specific tuning is required for the system. phbuilder also allows the user to automatically set up protein titrations or parametrize new titratable groups. We expect that with the help of phbuilder, constant pH MD simulations will be more straightforward to set up and will thereby become more readily accessible to the GROMACS user community.

DATA AND SOFTWARE AVAILABILITY

Both the phbuilder tool, the GROMACS constant pH MD beta version, and the modified CHARMM36m force field are available for download free of charge at www.gitlab.com/gromacs-constantph. All simulation setup files are available at <https://doi.org/10.5281/zenodo.10468944>.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.3c01313>.

Example of the `clean_after_solvate.py` application, the evolution of the λ -coordinates of the test systems, and the reweighting step during parametrization (PDF)

AUTHOR INFORMATION

Corresponding Authors

Anton Jansen – Department of Applied Physics and Swedish e-Science Research Center, Science for Life Laboratory, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden; orcid.org/0000-0003-2835-7987; Email: ajansen@kth.se

Pavel Buslaev – Nanoscience Center and Department of Chemistry, University of Jyväskylä, 40014 Jyväskylä, Finland; orcid.org/0000-0003-2031-4691; Email: pavel.i.buslaev@jyu.fi

Berk Hess – Department of Applied Physics and Swedish e-Science Research Center, Science for Life Laboratory, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden; Email: hess@kth.se

Authors

Noora Aho – Nanoscience Center and Department of Chemistry, University of Jyväskylä, 40014 Jyväskylä, Finland; orcid.org/0009-0008-1243-0215

Gerrit Groenhof – Nanoscience Center and Department of Chemistry, University of Jyväskylä, 40014 Jyväskylä, Finland; orcid.org/0000-0001-8148-5334

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jcim.3c01313>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the Swedish Research Council (Grant 2019-04477), the Academy of Finland (Grants 311031, 332743, and 342908), and the BioExcel CoE (Grants H2020-INFRAEDI-02-2018-823830 and 101093290). We also thank Paul Bauer and M. Eric Irrgang for their support during the project.

REFERENCES

- (1) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.
- (2) Cournia, Z.; Allen, B.; Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.
- (3) Barroso daSilva, F. L.; Dias, L. G. Development of constant-pH simulation methods in implicit solvent and applications in biomolecular systems. *Biophysical reviews* **2017**, *9*, 699–728.
- (4) Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct., Funct., Bioinf.* **1997**, *27*, 523–544.
- (5) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **2002**, *117*, 4184–4200.
- (6) Bürgi, R.; Kollman, P. A.; Van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 469–480.
- (7) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *Journal of computational chemistry* **2004**, *25*, 2038–2048.
- (8) Meng, Y.; Roitberg, A. E. Constant pH replica exchange molecular dynamics in biomolecules using a discrete protonation model. *J. Chem. Theory Comput.* **2010**, *6*, 1401–1412.
- (9) Itoh, S. G.; Damjanović, A.; Brooks, B. R. pH replica-exchange method based on discrete protonation states. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3420–3436.
- (10) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.
- (11) Swails, J. M.; Roitberg, A. E. Enhancing conformation and protonation state sampling of hen egg white lysozyme using pH replica exchange molecular dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 4393–4404.
- (12) Börjesson, U.; Hünenberger, P. H. Explicit-solvent molecular dynamics simulation at constant pH: methodology and application to small amines. *J. Chem. Phys.* **2001**, *114*, 9706–9719.
- (13) Lee, M. S.; Salsbury, F. R., Jr; Brooks, C. L., III Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738–752.
- (14) Khandogin, J.; Brooks, C. L. Constant pH molecular dynamics with proton tautomerism. *Biophysical journal* **2005**, *89*, 141–157.
- (15) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. Constant pH molecular dynamics in explicit solvent with λ -dynamics. *J. Chem. Theory Comput.* **2011**, *7*, 1962–1978.
- (16) Wallace, J. A.; Shen, J. K. Continuous constant pH molecular dynamics in explicit solvent with pH-based replica exchange. *J. Chem. Theory Comput.* **2011**, *7*, 2617–2629.
- (17) Wallace, J. A.; Shen, J. K. Charge-leveling and proper treatment of long-range electrostatics in all-atom molecular dynamics at constant pH. *J. Chem. Phys.* **2012**, *137* (18), 184105.
- (18) Goh, G. B.; Knight, J. L.; Brooks, C. L. Constant pH molecular dynamics simulations of nucleic acids in explicit solvent. *J. Chem. Theory Comput.* **2012**, *8*, 36–46.
- (19) Chen, W.; Wallace, J. A.; Yue, Z.; Shen, J. K. Introducing titratable water to all-atom molecular dynamics at constant pH. *Biophysical journal* **2013**, *105*, L15–L17.
- (20) Lee, J.; Miller, B. T.; Damjanovic, A.; Brooks, B. R. Constant pH molecular dynamics in explicit solvent with enveloping distribution sampling and Hamiltonian exchange. *J. Chem. Theory Comput.* **2014**, *10*, 2738–2750.
- (21) Dobrev, P.; Donnini, S.; Groenhof, G.; Grubmüller, H. Accurate Three States Model for Amino Acids with Two Chemically Coupled Titrating Sites in Explicit Solvent Atomistic Constant pH Simulations and p K a Calculations. *J. Chem. Theory Comput.* **2017**, *13*, 147–160.
- (22) Huang, Y.; Chen, W.; Wallace, J. A.; Shen, J. All-atom continuous constant pH molecular dynamics with particle mesh Ewald and titratable water. *J. Chem. Theory Comput.* **2016**, *12*, 5411–5421.
- (23) Huang, Y.; Harris, R. C.; Shen, J. Generalized Born based continuous constant pH molecular dynamics in Amber: Implementation, benchmarking and analysis. *J. Chem. Inf. Model.* **2018**, *58*, 1372–1383.
- (24) Harris, R. C.; Shen, J. GPU-accelerated implementation of continuous constant pH molecular dynamics in amber: p K a predictions with single-pH simulations. *J. Chem. Inf. Model.* **2019**, *59*, 4821–4832.
- (25) Harris, R. C.; Liu, R.; Shen, J. Predicting reactive cysteines with implicit-solvent-based continuous constant pH molecular dynamics in amber. *J. Chem. Theory Comput.* **2020**, *16*, 3689–3698.
- (26) Harris, J. A.; Liu, R.; Martins de Oliveira, V.; Vázquez-Montelongo, E. A.; Henderson, J. A.; Shen, J. GPU-Accelerated All-atom Particle-Mesh Ewald Continuous Constant pH Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2022**, *18*, 7510–7527.
- (27) Aho, N.; Buslaev, P.; Jansen, A.; Bauer, P.; Groenhof, G.; Hess, B. Scalable constant pH molecular dynamics in GROMACS. *J. Chem. Theory Comput.* **2022**, *18*, 6148–6160.
- (28) Buslaev, P.; Aho, N.; Jansen, A.; Bauer, P.; Hess, B.; Groenhof, G. Best practices in constant pH MD simulations: accuracy and sampling. *J. Chem. Theory Comput.* **2022**, *18*, 6134–6147.
- (29) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

- (30) Kong, X.; Brooks, C. L. λ -dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105*, 2414–2423.
- (31) Henderson, J. A.; Lui, R.; Harris, J. A.; Huang, Y.; de Oliveira, V. M.; Shen, J. A Guide to the Continuous Constant pH Molecular Dynamics Methods in Amber and CHARMM [Article v1. 0]. *Living Journal of Computational Molecular Science* **2022**, *4*, 1563–1563.
- (32) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **2008**, *29*, 1859–1865.
- (33) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry* **2009**, *30*, 2157–2164.
- (34) Hedges, L.; Mey, A.; Laughton, C.; Gervasio, F.; Mulholland, A.; Woods, C.; Michel, J. BioSimSpace: An interoperable Python framework for biomolecular simulation. *Journal of Open Source Software* **2019**, *4*, 1831.
- (35) Jespers, W.; Esguerra, M.; Åqvist, J.; Gutiérrez-de Terán, H. QligFEP: an automated workflow for small molecule free energy calculations in Q. *Journal of cheminformatics* **2019**, *11*, 26.
- (36) Homeyer, N.; Gohlke, H. FEW: a workflow tool for free energy calculations of ligand binding. *Journal of computational chemistry* **2013**, *34*, 965–973.
- (37) Suruzhon, M.; Senapathi, T.; Bodnarchuk, M. S.; Viner, R.; Wall, I. D.; Barnett, C. B.; Naidoo, K. J.; Essex, J. W. ProtoCaller: robust automation of binding free energy calculations. *J. Chem. Inf. Model.* **2020**, *60*, 1917–1921.
- (38) Carvalho Martins, L.; Cino, E. A.; Ferreira, R. S. PyAutoFEP: An Automated Free Energy Perturbation Workflow for GROMACS Integrating Enhanced Sampling Methods. *J. Chem. Theory Comput.* **2021**, *17*, 4262–4273.
- (39) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: automating setup for alchemical free energy simulations. *J. Chem. Inf. Model.* **2015**, *55* (12), 2485.
- (40) Kim, S.; Oshima, H.; Zhang, H.; Kern, N. R.; Re, S.; Lee, J.; Roux, B.; Sugita, Y.; Jiang, W.; Im, W. CHARMM-GUI free energy calculator for absolute and relative ligand solvation and binding free energy simulations. *J. Chem. Theory Comput.* **2020**, *16*, 7207–7218.
- (41) Huang, J.; MacKerell, A. D., Jr CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, *34*, 2135–2145.
- (42) Hub, J. S.; de Groot, B. L.; Grubmüller, H.; Groenhof, G. Quantifying artifacts in Ewald simulations of inhomogeneous systems with a net charge. *J. Chem. Theory Comput.* **2014**, *10*, 381–390.
- (43) Donnini, S.; Ullmann, R. T.; Groenhof, G.; Grubmüller, H. Charge-neutral constant pH molecular dynamics simulations using a parsimonious proton buffer. *J. Chem. Theory Comput.* **2016**, *12*, 1040–1051.
- (44) Buslaev, P.; Aho, N.; Jansen, A.; Bauer, P.; Hess, B.; Groenhof, G. In *GROMACS constant pH manual*. <https://gitlab.com/gromacs-constantph/constantph> (last accessed 2023-09-24).
- (45) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (46) Shirts, M. R.; Pande, V. S. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **2005**, *122*, 144107.
- (47) Paliwal, H.; Shirts, M. R. A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *J. Chem. Theory Comput.* **2011**, *7*, 4115–4134.
- (48) Singhal, A. K.; Chien, K. Y.; Wu, W. G.; Rule, G. S. Solution structure of cardiotoxin V from *Naja naja atra*. *Biochemistry* **1993**, *32*, 8036–8044.
- (49) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-Å resolution. *Biochemistry* **1994**, *33*, 8878–8889.
- (50) Rovšnik, U.; Zhuang, Y.; Forsberg, B. O.; Carroni, M.; Yvonne-dotter, L.; Howard, R. J.; Lindahl, E. Dynamic closed states of a ligand-gated ion channel captured by cryo-EM and simulations. *Life Science Alliance* **2021**, *4*, e202101011.
- (51) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.
- (52) Norberg, J.; Foloppe, N.; Nilsson, L. Intrinsic relative stabilities of the neutral tautomers of arginine side-chain models. *J. Chem. Theory Comput.* **2005**, *1*, 986–993.
- (53) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-induced forces between two hydrophilic groups. *J. Phys. Chem.* **1994**, *98*, 2198–2202.
- (54) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902–1921.
- (55) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (56) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (57) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
- (58) Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *J. Chem. Phys.* **2020**, *153*, 114107.
- (59) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.
- (60) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* **1992**, *13*, 952–962.
- (61) Schmidt, T. H.; Kandt, C. LAMBADA and InflateGRO2: Efficient Membrane Alignment and Insertion of Membrane Proteins for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2012**, *52* (10), 2657–2669.
- (62) Wolf, M. G.; Hoefling, M.; Aponte-Santamaría, C.; Grubmüller, H.; Groenhof, G. *g_membed*: Efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. *J. Comput. Chem.* **2010**, *31* (11), 2169–2174.