This is the title : v0.1

Rommie E. Amaro[1*], John D. Chodera[2*], David L. Mobley[3*], Antonia S. J. S. Mey[4*], Julien Michel[4*]

[1]Institution 1; [2]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York NY 10065; [3]Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine; [4]EaStCHEM School of Chemistry, David Brewster road, Joseph Black Building, The King's Buildings, Edinburgh, EH9 3FJ, UK

This LiveCoMS document is maintained online on GitHub at https://github.com/michellab/BioMolSetupPaper; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.

This version dated November 30, 2017

Abstract Simulations of biomolecular systems are complex and often demanding but can provide great insights and valuable quantitative predictions. However, these simulations require considerable human expertise to set up well, and key aspects of system preparation and setup are often overlooked for a variety of reasons. Thus, simulations of the "same" biomolecular system prepared by different individuals or different software tools may yield substantially different results due to different choices made in system preparation, such as different handling of crystallographic waters, different treatment of missing residues or loops in a protein, or different choices for a ligand or receptor's protonation state. In some cases, human error can also play a role, with a user perhaps forgetting to check for disulfide bonds. Here, we identify the issues we believe are most important for users and/or software tools to consider when preparing biomolecular systems for simulation and present a checklist for users to follow to ensure they have handled the most crucial or troublesome aspects of system preparation, with recommendations for how to handle each step in the process when best practices have been established by the literature. We also present a detailed explanation of the items in our checklist, including why we make the recommendations we do. Additionally, because a useful checklist must necessarily be brief, we provide a discussion of other important factors to consider which did not make our major checklist. We believe this work provides a starting point to help reduce the number of simulations run which are flawed due to setup failures, and we hope the community will contribute to updating and improving this document to that end.

***For correspondence:**
email1@example.com (REA); john.chodera@choderalab.org (JDC); dmobley@mobleylab.org (DLM); antonia.mey@ed.ac.uk (ASJSM); info@julienmichel.net (JM)

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Present address: [§]Department, Institute, Country; [¶]Department, Institute, Country

1 Goals

This work focuses on providing a brief checklist of the most critical and most frequently overlooked items in the preparation of biomolecular simulations, so that that practitioners can follow this checklist, and reviewers can potentially use it in evaluating manuscripts and assessing whether method de-

scriptions are complete. We also focus on highlighting issues where further clarity may be needed, either via additional study, or from work of which we are unaware. Additionally, we value input from the community on how best to handle the issues raised, and others we may have overlooked, as an additional goal of this work is to develop a set of *community* best practices. In addition to the provided checklist, we provide an explanation of why the checklist includes the items it does and why we make the recommendations we do, in order to further facilitate training and community input.

This work can perhaps also be useful for authors who are writing papers in order to help determine what details to report in their Methods section – specifically, such sections should at minimum address the details highlighted in this document.

Our goal in this work is to *inform* about what we believe are best practices for preparation in this area and provide an aid to authors and reviewers; we have no intention of providing a set of standards which should be *enforced*. Indeed, there may be good reasons from deviating from recommended practices in certain cases; in our view, authors should simply note any deviations and explain why they made the choices they did.

2 Scope

Our focus in this work is on what is often the earliest stage of preparation of biomolecular simulations – deciding what system to simulate and which molecules and atoms are in it in which positions. We deliberately avoid the issue of assigning (and potentially developing) force field parameters, as this involves many additional complexities that warrant their own "best practices" document. We also, at least for now, focus on protein or protein-ligand systems, avoiding nucleic acids such as DNA or RNA as we lack the requisite expertise (though this work may be extended later to cover such cases). Additionally, we avoid membrane proteins for similar reasons. Thus our overall focus is on preparation for *relatively simple biomolecular simulations of soluble proteins that may include non-covalent small molecule ligands*, and especially on the force field agnostic aspects of preparation. We expect to eventually include handling of cofactors, though this is not within our initial scope.

Thus, our scope *does not include* assignment of parameters from force field libraries, and especially generation of parameters (such as for unusual nonstandard residues where literature parameters may not be available, or for covalent cofactors, complex ligands or factors, etc.). This is both because these may be research questions, and because the particular approaches applied will often be coupled to the choice of force field to a significant extent. Additionally, we do not include any aspects of choice of simulation protocol such as minimization, cutoffs, and other factors as, in our view, these all come significantly downstream of system preparation.

3 Checklist

3.1 Step 0: Know what you want to simulate

A key first step in setting up a biomolecular system is being sure you know what you intend to simulate and have planned the right simulation(s), including such aspects as:

1. What is the goal of your simulation(s)? Can you hope to achieve those goals with this simulation?
2. What do you know (or expect) about the potential timescales in your system, and can you hope to capture those with this simulation?
3. Will you be able to gather good enough statistics from this study, based on what you know about those timescales?
4. What will be the costs of your simulation (computer time, storage, memory, wallclock, network bandwidth, data throughput) and can you afford those costs?
5. What are the experimental conditions (including salt conditions, buffers, cofactors, etc.)? Do you want to match those conditions in your simulation? If not, why not?

3.2 Protein

1. Sequence of the protein – is it the protein you wanted to simulate? (Assays, crystal structure, intended) And what conditions do you want to simulate (are they the structural conditions)? Oligomeric state (how many chains do we need to simulate?) (Chodera/Amaro)
2. Structure selection (Mey)
    - X-ray structural data
    - Dealing with NMR structural ensembles
    - Other sources of structural data: e.g. cryo-EM
3. Disulfide bonds depending on reducing/oxidizing environments (Amaro)
4. Post-translational modifications (Chodera)
5. Model in missing residues and loops (Chodera/Amaro tag-team)
    - Schrodinger tools model in short loops, but truncate long loops with proton caps (?)
    - UCSF Modeller terrible for loops without care; Rosetta Model works well ? (JDC says)
    - (But Amaro reports success stories with Schrodinger tools if one knows when to use them and does not try to push the tool beyond the limit of reasonable application - eg loops < 10-12 residues, otherwise need templated homology models)

6. Assign protein protonation states/tautomers via a consistent, well-documented, and (ideally) reproducible procedure
    - PROPKA may be a good automated tool [ref] though it can be sensitive to the specific snapshot chosen for analysis, so propkatraj (https://github.com/Becksteinlab/propkatraj) may be worth using to provide a more thorough analysis
    - Multi-Conformer Continuum Electrostatics (MCCE) is another well-established option [ref]
    - Constant pH MD simulations may offer an alternative as they develop further [ref Case, McCammon, Roitberg, Shen, Roux, Chodera]
    - Careful hand selection may be needed in some cases but must be carefully documented and is too onerous to be recommended as the only solution (that is, we recommend hand selection be applied only after applying an algorithmic approach)
7. Crystal waters (Amaro, Mey)
8. Metal ions (JDC)

3.3 Ligands

1. Select protonation state/tautomer (Chodera)
2. Select the correct ligand binding mode
    - Are you sure you have the right ligand?
    - If the starting point is crystallographic, does the electron density support its presence?
    - Is it present only because of crystal contacts or packing issues?
    - If you are modeling the ligand in, how will you place it? Will you use docking, shape overlays, knowledge of binding modes of similar ligands, or other factors?
    - If you have modeled it in, how will you determine whether this is the correct binding mode? Might other binding modes be possible? How would this affect your results?
    - Will you need to run simulations of the ligands to validate the putative binding mode?

3.4 Counterions/water

1. No counterions, minimal counterions, or physiological ionic strength? (Mobley)

4 Other things to think about that didn't make the checklist

Here, we focus on items which are also important part of system preparation, but which are less frequent causes of critical failures or are not as often or easily overlooked.

4.1 Proteins

Termini: Build them, or cap? Depends on length; if too long can contribute to timescales. Write some guidelines into how to know.

4.2 Ligands

While simple proteins without covalent modifications or nonstandard residues can typically be handled by library force fields for biomolecular simulations, ligands will involve parameter assignment. As noted above, parameterization is not within the scope of this work, but it is important to note that there are roughly two categories of approaches for handling of ligand parameter assignment: General-purpose small molecule force fields (e.g. GAFF, GAFF2, OPLS2, OPLS3, etc.) which provide generic parameters that can easily be assigned to the vast majority of small organic molecules, possibly with a partial charge calculation required; and more advanced force fields which require parameter development for each new molecule to be considered (e.g. AMOEBA, … ??).

Handling of covalently bound cofactors, adducts, and certain nonstandard residues (those without library parameters available) can be an even more complex issue than handling of ligand parameterization, often turning into a research topic requiring considerable care.

4.3 Counterions/water

(minor) Select simulation box size (Mey) – important, but first thing to check in analysis.

Ultimately, finalizing certain aspects of setup will require selection of water and ion models, which is often coupled (at least partially) to the choice of force field. The choice of these models is outside our scope, but is important to note. Many different water models are available, with varying degrees of quality for pure solvent properties. The choice of water model is also coupled to the choice of model for ions – for example, different water models can require different models for monatomic ions in order to prevent aggregation of ions [refs].

(minor) Add solvent then ions, or ions then solvent? Pre-equilibrate ion-water mixture? (Amaro) Determine electrostatic potential around molecule and place ions at minima (default AMBER approach does this in solvent, replacing some solvent with ions)

5 Detailed explanation of checklist items

5.1 Step 0: Know what you want to simulate

5.2 Protein

5.3 Ligands

5.4 Counterions/water

6 Conclusions

7 Acknowledgments

8 Author Contributions

(Explain the contributions of the different authors here)

For a more detailed description of author contributions, see the GitHub issue tracking and changelog at https://github.com/michellab/BioMolSetupPaper.

9 Other Contributions

(Explain the contributions of any non-author contributors here) For a more detailed description of contributions from the community and others, see the GitHub issue tracking and changelog at https://github.com/michellab/BioMolSetupPaper.

10 Potentially Conflicting Interests

Declare any potentially conflicting interests here, whether or not they pose an actual conflict in your view.

11 Funding Information

FMS acknowledges the support of NSF grant CHE-1111111.