AutoSolvate: A Toolkit for Automating Quantum Chemistry Design and Discovery of Solvated Molecules

Eugen Hruska, Ariel Gale, Xiao Huang, Fang Liu*

Department of Chemistry, Emory University, Atlanta, Georgia, 30322

Abstract: The availability of large, high-quality data sets is crucial for artificial intelligence design and discovery in chemistry. Despite the essential roles of solvents in chemistry, the rapid computational data set generation of solution-phase molecular properties at the quantum mechanical level of theory was previously hampered by the complicated simulation procedure. Software toolkits that can automate the procedure to set up high-throughput explicit-solvent quantum chemistry (QC) calculations for arbitrary solutes and solvents in an open-source framework are still lacking. We developed AutoSolvate, an open-source toolkit to streamline the workflow for QC calculation of explicitly solvated molecules. It automates the solvated-structure generation, force field fitting, configuration sampling, and the final extraction of microsolvated cluster structures that QC packages can readily use to predict molecular properties of interest. AutoSolvate is available through both a command line interface and a graphical user interface, making it accessible to the broader scientific community. To improve the quality of the initial structures generated by AutoSolvate, we investigated the dependence of solute-solvent closeness on solute/solvent identities and trained a machine learning model to predict the closeness and guide initial structure generation. Finally, we tested the capability of AutoSolvate for rapid data set curation by calculating the outer-sphere reorganization energy of a large data set of 166 redox

* Electronic mail: fang.liu@emory.edu

couples, which demonstrated the promise of the AutoSolvate package for chemical discovery efforts.

I. INTRODUCTION

The availability of large, high-quality data sets is crucial for artificial intelligence (AI) design and discovery in chemistry. Due to the scarcity of large, cleaned experimental data sets for various molecular properties, computational data sets are crucial for training machine learning (ML) models used in AI design and discovery. However, most computational molecular property data sets, especially the large data sets widely used in ML benchmarking such as QM7[1, 2] and QM9,[3, 4] focus on gas phase molecular properties, such as atomization energy, ionization potential, electron affinity, and frontier orbital energies. The reason behind this is the ease of rapid generation of these data sets with quantum chemistry (QC) calculation, which is by default in the gas phase. Various toolkits have been developed to automate QC calculation and molecular property data set curation. Some are general QC workflows aiming at proving QC software interoperability (QCEngine,[5] QCArchive[6]), some focus on specific groups of molecules, such as organic molecules (e.g., RDKit[7, 8], with some support for organometallics), transition metal complexes (e.g., molSimplify[9]), and materials (e.g., pymatgen,[10] AiiDA[11]), and others focus on specific types of QC calculation, such as transition state search (e.g., AARON,[12] QChASM[13]). These workflows enable automated high-quality initial structure generation, QC input file preparation, job preparation and execution, and data collection, which lead to rapid generation of high-quality computational molecular property data sets.

However, many chemical processes related to real chemistry applications are in the solution phase. Solvent environments play an essential role in chemistry by impacting molecular properties and modifying reaction rates.[14-17] Rapid generation of large computational data sets of solution-phase molecular properties or reaction rates is crucial for developing ML models and enabling AI design in the solution phase. Nevertheless, setting up an automated workflow for accurate QC calculation of solvated molecules is more challenging than the gas phase counterpart. Although implicit solvent models, such as the polarizable continuum models,[18-24] are now widely available in QC packages and the corresponding molecular property data set curation can be easily automated with the aforementioned QC workflows, the accuracy of these calculations cannot always meet our needs in design and discovery. Accurate prediction of many molecular properties requires QC calculation in explicit solvents due to the existence of proton transfer,[25, 26] hydrogen bonds,[27-29] or other strong solute-solvent interactions. Setting up QC calculations of explicitly solvated molecules usually involves multiple steps, including building the structure of the solute molecule in a solvent box, generating a customized force field for solute/solvent, running molecular dynamics (MD) simulations at molecular mechanical (MM) and hybrid quantum mechanical and molecular mechanical (QM/MM)[30-34] level of theories to obtain equilibrated solvation configurations, and the final QC calculation of molecular properties for the microsolvated molecule extracted from the solvent box. Software toolkits that can automate the procedure to set up high-throughput explicit-solvent QC calculations for arbitrary solute molecules in an open-source framework are still scarce. Representative works in this area include the systematic microsolvation approach developed by Reiher and coworkers[35] based on stochastic structure generation and geometry optimization, and the ABCluster global optimization algorithm by Zhang and coworkers.[36, 37]

In this article, we introduce an efficient and flexible approach to the rapid generation of QC molecular property data sets for explicitly solvated molecules in our open-source AutoSolvate toolkit.[38] This streamlined procedure enables automated structure and force field parameter generation, MD simulation input file preparation, and job execution. In sections II and III, we provide a description of the code layout and the detailed routines involved in (i) explicitly solvated structure generation and MD parameter preparation, (ii) MD simulation automation, and (iii) post-processing and QC calculation. We then present benchmarking results of our approach over a 166-molecule test set. Finally, we provide the conclusions and outlook for our software toolkit.

II. CODE OVERVIEW

The developed software toolkit (AutoSolvate[38]) is an open-source workflow that facilitates high-throughput QC calculation of explicitly solvated molecular systems. It incorporates solvated structure and force field generation, automated set up and execution of MD equilibration and QM/MM solvation configuration sampling, and post-processing (automated extraction of microsolvated structure for QC calculations) (Fig. 1). The software is designed to support the calculation of arbitrary organic solute molecules in a variety of commonly used solvents, with the flexibility to accept user-provided solvents with customized force field parameters. Although currently focused on organic solute molecules, it can be easily extended to treat organometallic compounds in the near future. Our structure generation tools are designed to generate high-quality initial solvated structure by using automatically recommended solute-solvent closeness aware of the chemical features of different solute-solvent combinations. In addition to a command-line version, AutoSolvate is available through a graphical user interface (GUI) to make the code accessible to the broader scientific community.

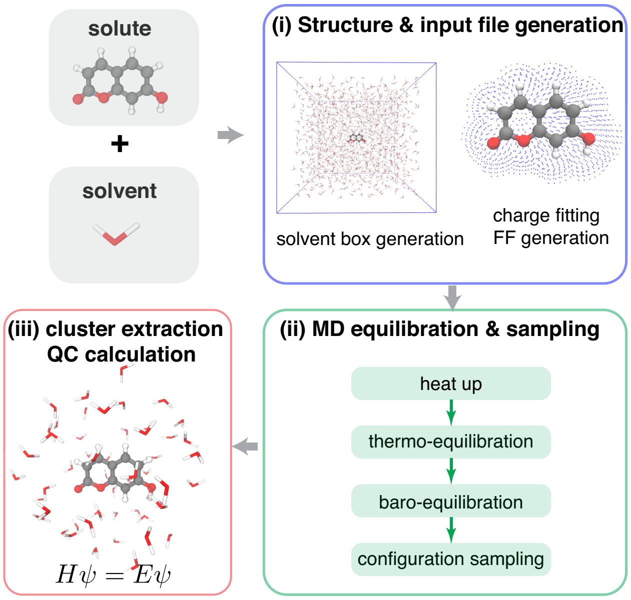


FIG. 1. Overview of AutoSolvate workflow.

III. CODE ARCHITECTURE

The AutoSolvate[38] code is written in Python 3.8, a modern high-level programming language that is widely used in the computational chemistry community. The AutoSolvate program uses a few open-source toolboxes for structure and force field parameter generation, including OpenBabel[39] to convert between chemical formats, Packmol[40] to pack solvents around the solute molecule, and AmberTools 20[40] to generate force field parameters for organic solute molecules. Portability of the code is maintained by including a conda environment[41] YAML file[42] in the releases, which ensures all dependent packages get automatically installed regardless of the operating system. Users may install the AutoSolvate toolkit on any Linux-based or Windows platform where the conda environment is available.

The user interacts with AutoSolvate in one of three ways: import as a Python Application Programming Interface (API), run through the command-line interface (CLI), or a graphical user interface (GUI). The GUI has been developed using the Python package Tkinter,[43, 44] and the

Python bindings for Tcl/Tk.[45] The GUI enables a user-friendly interface for building solvated molecular systems in a manner intended to be intuitive for the broader scientific community.

The toolkit consists of three main modules (Fig. 2) described in greater detail in the rest of this article:

(1) The structure and force field generation module.

(2) The MD simulation preparation and execution module.

(3) The extraction of microsolvated clusters for QC calculations.

Although the modules are designed to work together, each module can be called independently by the user to satisfy their special needs.

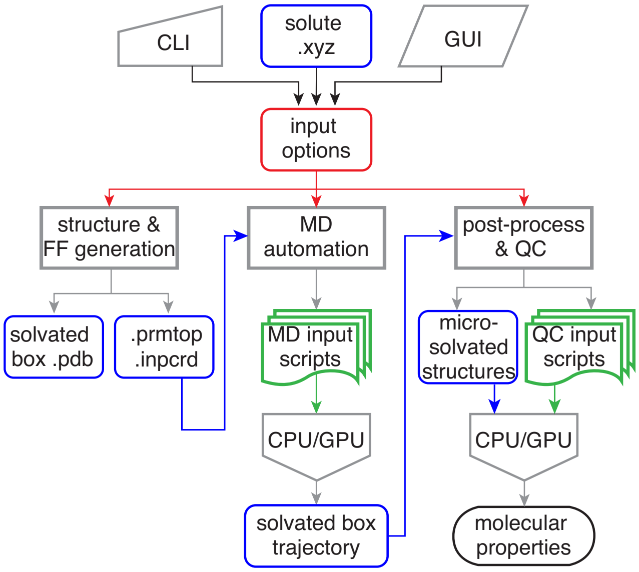


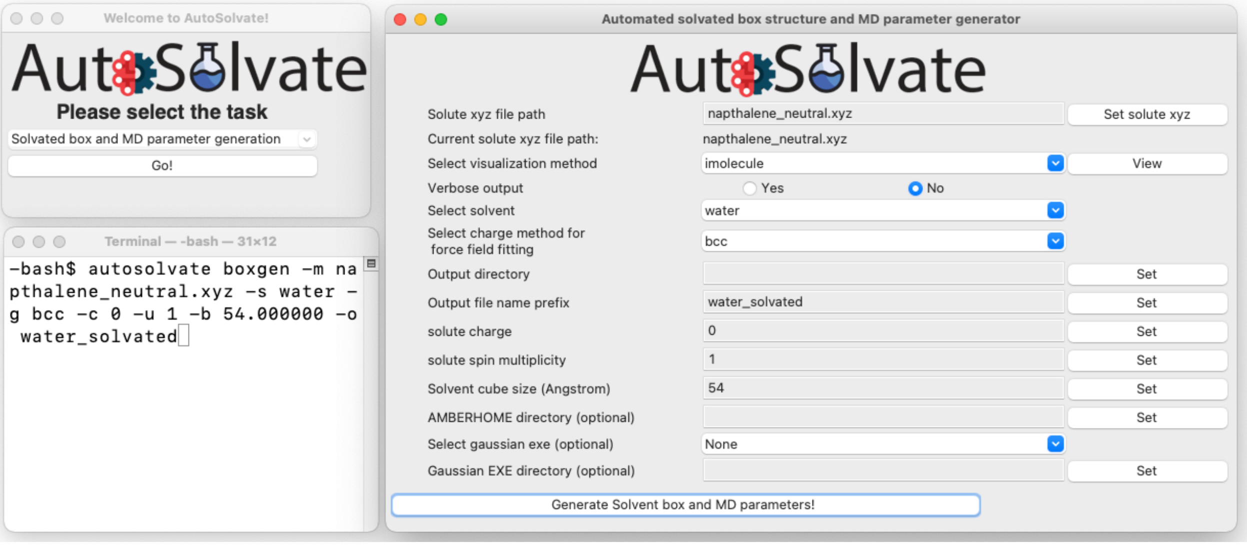
FIG. 2. Flowchart for AutoSolvate code.

A. Solvated structure and force field generation

The structure and force field generation module is the central core of AutoSolvate. This module generates the essential explicit solvent model files, the foundation for subsequent MD simulation, microsolvated cluster extraction, and QC calculations. Files generated by this module include the protein data bank (PDB)[46] file of a given solute molecule solvated in a large solvent box, the corresponding Amber parameter-topology (.prmtop) file, and the input coordinate (.inpcrd) file.

1. General approach

In the general solvated structure and force field generation approach, the user only needs to specify (i) a solute molecule xyz coordinate file, (ii) solute molecule charge and spin-multiplicity, and (iii) the solvent name as the minimal input. Using this user input (example CLI and GUI input shown in Fig. 3), the code then automatically generates the PDB file of the solvent box and the corresponding Amber parameter-topology file (.prmtop) and input coordinate files (.inpcrd). The user can specify additional input options to customize the generation results further. Detailed mechanisms and customizations are explained as follows.



Solvated box and MD parameter generation

```
-bash$ autosolvate boxgen -m na
pthalene_neutral.xyz -s water -
g bcc -c 0 -u 1 -b 54.000000 -o
 water_solvated
```

Solute xyz file path: napthalene_neutral.xyz
Current solute xyz file path: napthalene_neutral.xyz
Select visualization method: imolecule
Verbose output: ○ Yes ● No
Select solvent: water
Select charge method for force field fitting: bcc
Output directory:
Output file name prefix: water_solvated
solute charge: 0
solute spin multiplicity: 1
Solvent cube size (Angstrom): 54
AMBERHOME directory (optional):
Select gaussian exe (optional): None
Gaussian EXE directory (optional):

Fig. 3. Example solvent box structure and force field generation in GUI (upper left, right) or CLI (lower left). When using the GUI, the user enters from the main window (upper left), selects the task, and gets prompted to the structure generation window (right).

2. Solute force field generation

The user-provided solute molecules can be any organic molecule of any size, generally a non-standard residue for Amber force fields.[47] Hence, generating customized force fields of the solute is an indispensable step to generate the parameter-topology file for the solvated structure.

Specifically, the atomic charges are a set of essential parameters not readily available in the generalized Amber force field (GAFF)[48] and must be provided by the user. To simplify the procedure of selecting options and executing different modules of AmberTools (antechamber,[49] parmchk) to generate solute force field parameters, AutoSolvate automatically makes decisions for the user based on the nature of the solute (Fig. 4). If the solute molecule is open-shell (spin-multiplicity > 1), atomic charge fitting is not supported by the semi-empirical charge methods contained in antechamber and must be calculated with restrained electrostatic potential (RESP)[50, 51] charge fitting using external QC packages. AutoSolvate will automatically generate the needed input files and execute external QC packages (Gaussian 16[52] by default) to generate the electrostatic potential file, which is then fed to antechamber for RESP charge fitting. If the molecule is closed-shell (spin-multiplicity =1), AutoSolvate will directly call antechamber to do the semi-empirical AM1-BCC[53] charge fitting. The generated atomic charges, together with the input solute structure information, will be stored in Tripos mol2 file format,[54] which is sent to parmchk to check for missing force field parameters in the GAFF force field, and to generate the force field modification files (.frcmod) needed for the final prmtop file generation of the solvated structure.

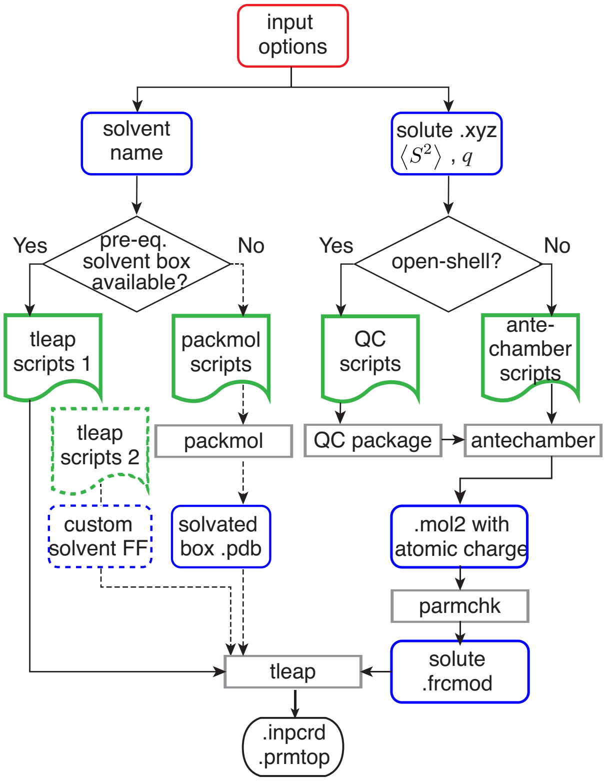


Fig. 4. Flowchart for automated solvated structure and force field generation. (left) Solvent force field parameter preparation steps for the cases with (solid path) and without (dashed path) pre-equilibrated solvent box. (right) Solute force field parameter preparation steps.

3. Solvent box generation

Currently, AutoSolvate supports two ways to generate a solvent box surrounding the solute based on the availability of pre-equilibrated solvent boxes and corresponding force fields in AmberTools (Fig. 4). For common solvents readily available in AmberTools/tleap [TIP3P water, methanol, chloroform, and N-methylacetamide (NMA)], the code prioritizes direct usage of pre-equilibrated solvent boxes. The code automatically generates the tleap commands that build the structure of the solute immersed in the pre-equilibrated solvent box. Subsequent execution of tleap simultaneously generates the solvent box PDB structure, and its parameter-topology (.prmtop) and input coordinates files (.inpcrd) needed for MD simulation (Fig 4, path indicated by solid line).

For solvents not readily available in AmberTools [e.g. acetonitrile (MeCN)] and have no existing pre-equilibrated solvent box library (.off) file, the code first calls packmol[55] to pack the solute and solvent molecules in a box of the desired size, yielding a PDB file. Then the code generates the tleap commands that use the PDB file and the custom solvent force field parameters stored within AutoSolvate package to generate prmtop and inpcrd files (Fig 4, path indicated by dashed line). Furthermore, the users can also provide custom solvent library file (.off) and force field modification file (.frcmod) to enable the generation of structures solvated in other solvents not included in AutoSolvate. Custom solvent parameter files can be downloaded from databases such as the Amber Parameter Database,[56] or be found from literature.

4. Automated recommendation of solvent-solute closeness

Despite the different software (tleap vs. packmol) used in the two approaches to packing solvent boxes around the solute, both request an input parameter to control the distance, $d_{a,b}$, between a solute atom, a, and a solvent atom, b. For packmol, the parameter is a tolerance, d_{tol}, where

$$d_{a,b} \geq d_{\mathrm{tol}}. \tag{1}$$

We can see that d_{tol} is the minimum allowed distance between any solute atom and any solvent atom, which we refer to as the solvent-solute closeness hereafter. For tleap, the parameter is a scaling coefficient, s, where

$$d_{a,b} \geq \left(R_{\mathrm{VDW}}^{a} + R_{\mathrm{VDW}}^{b}\right) \cdot s, \tag{2}$$

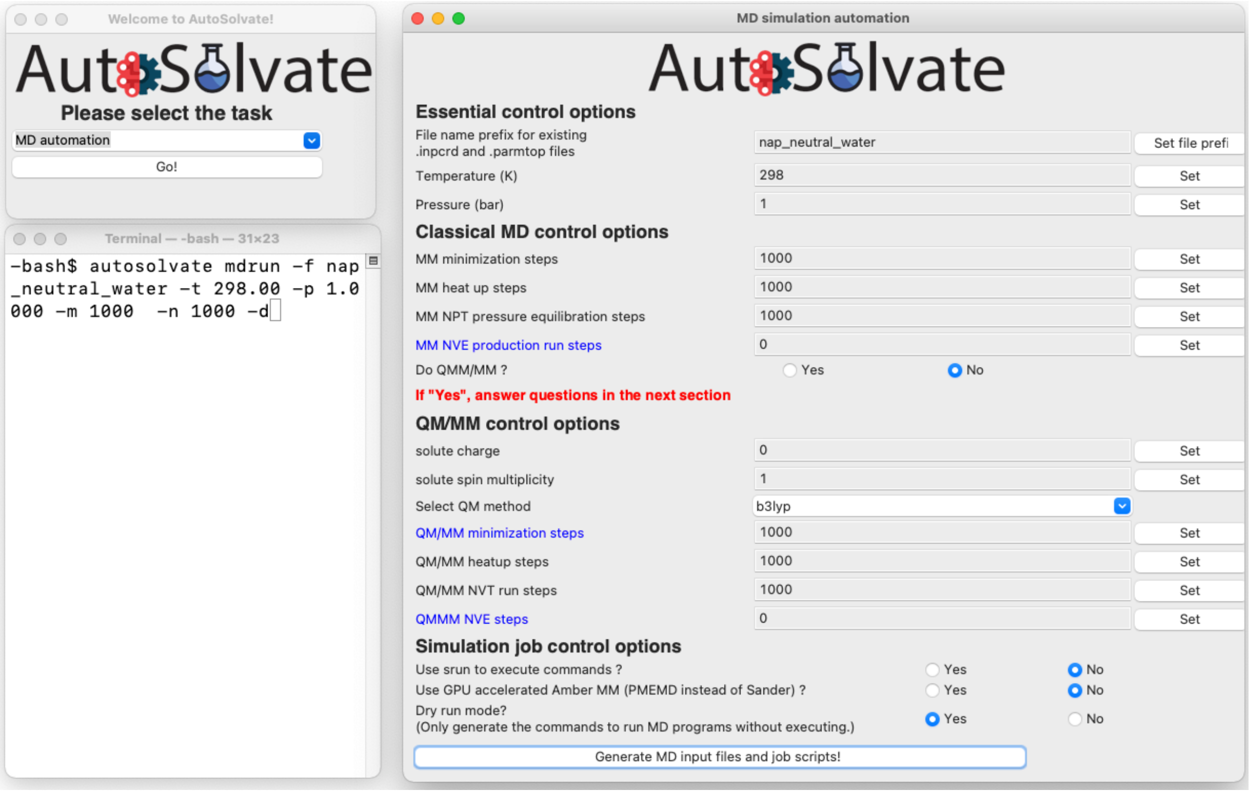
and R_{VDW}^{a} represents the van der Waals radius of atom a. Obviously, s and d_{tol} are closely related and can be converted to each other. The solvent-solute closeness (d_{tol}) impacts the generated initial structure's quality and thus influences the efficiency of the subsequent MD equilibration process. However, d_{tol} is not intuitive for users without prior experience simulating the specific solvent-solute system. To avoid bias and errors in setting an arbitrary default value of solvent-solute closeness, we developed an automated approach to estimate d_{tol} based on existing simulations in our database. This automatically estimated d_{tol} can generate initial structures with solvent-solute closeness more similar to that determined from equilibrated trajectories. Detailed implementation mechanisms are shown in Section V.

B. Simulation automation

Once the parameter-topology file and coordinates of the solvated box are generated, MD simulations are needed to further equilibrate the structure to reach desired experimental environment (temperature and pressure condition) where the molecular properties of interest are measured. After equilibration, additional MD simulations at MM or QM/MM level are performed for configuration sampling, generating production trajectories used for final solvated cluster extraction and QM property prediction. In this subsection, we describe how AutoSolvate automates different stages of the simulations.

1. Classical MD simulation automation

The classical MD simulation of the solvated system in the periodic box is composed of the following stages: energy minimization, heating the system from 0 K to the target temperature with a Langevin thermostat, pressure equilibration to the target pressure with a Berendsen barostat, and an optional stage for MM NVE production dynamics. For each of these simulation steps, AutoSolvate generates the input file for Amber, together with the Linux command line needed to execute Amber to perform the simulation. As a minimal input requirement, the user only needs to provide the charge, multiplicity, and file name of the prmtop and inpcrd files. AutoSolvate has set default values for simulation temperature, pressure, and the number of steps for each stage of the classical MD simulation. The users can also customize the simulation details for each stage from the CLI or GUI. In the GUI, all entries are pre-populated with default settings, which guides beginner users through different simulation stages in an intuitive way (Fig 5).



Please select the task: MD automation

```
-bash$ autosolvate mdrun -f nap
_neutral_water -t 298.00 -p 1.0
000 -m 1000  -n 1000 -d
```

Essential control options

File name prefix for existing .inpcrd and .parmtop files	nap_neutral_water	Set file prefi
Temperature (K)	298	Set
Pressure (bar)	1	Set

Classical MD control options

MM minimization steps	1000	Set
MM heat up steps	1000	Set
MM NPT pressure equilibration steps	1000	Set
MM NVE production run steps	0	Set
Do QMM/MM ?	○ Yes	● No

If "Yes", answer questions in the next section

QM/MM control options

solute charge	0	Set
solute spin multiplicity	1	Set
Select QM method	b3lyp	
QM/MM minimization steps	1000	Set
QM/MM heatup steps	1000	Set
QM/MM NVT run steps	1000	Set
QMMM NVE steps	0	Set

Simulation job control options

Use srun to execute commands ?	○ Yes	● No
Use GPU accelerated Amber MM (PMEMD instead of Sander) ?	○ Yes	● No
Dry run mode? (Only generate the commands to run MD programs without executing.)	● Yes	○ No

Generate MD input files and job scripts!

Fig. 5. Example MD automation in GUI (upper left, right) or CLI (lower left). When using the GUI, the user enters from the main window (upper left), selects the task, and gets prompted to the MD automation window (right).

2. QM/MM simulation automation

The QM/MM simulation automation is currently set up for the Amber/TeraChem interface. It comprises the following simulation stages in the order of execution: QM/MM energy minimization, heating the system to the target temperature with a Langevin thermostat, NVT constant-temperature equilibration, and NVE production dynamics. The QM region only includes the solute molecule by default. The users can customize the simulation details (e.g., number of MD steps, temperature, pressure) for each stage from the CLI or GUI (Fig 5). Any stage can also be opted out if the user sets the number of steps to be 0, allowing flexible control of the workflow. The entire QM/MM simulation can also be skipped if the user has no access to QM/MM calculation packages or prefers to do MM dynamics only. For each of these simulation stages, AutoSolvate generates the input file for Amber/TeraChem, together with the Linux command line needed to trigger the Amber/TeraChem interface to perform the simulation. As a minimal input requirement, the user only needs to specify the charge, spin multiplicity, and the quantum mechanical (QM) method needed for the QM region, and AutoSolvate has set default values for simulation temperature, pressure, and the number of steps for each stage of the QM/MM simulation.

C. Microsolvated cluster extraction and QC calculation

The number of explicit solvent molecules included in the MM or QM/MM configuration sampling is usually very large (thousands or even more), and the direct inclusion of all explicit solvents into the final QC calculations is computationally infeasible. Instead, a commonly used approach is to perform QC calculations on a microsolvated cluster with a small number of explicit solvent molecules closely interacting with the solute. AutoSolvate simplifies this post-processing step by providing a simple interface that extracts microsolvated clusters of given solvent shell size and saves them into the xyz file format widely recognized by QC packages. As a minimal input, AutoSolvate only asks for the file names of the prmtop file name, the MD trajectory (.netcdf) file name, and the desired thickness of the solvent shell. It is worth noting there are two widely-used approaches for microsolvated structure extraction: the "solvent sphere" extraction based on a cutoff for the minimum distance between the center of the solute to any atom in the solvent, and the "solvent shell" extraction based on a cutoff for the minimum distance between any atom of the solute to any atom in the solvent (Fig. 6). We use the latter "solvent shell" extraction because, for a general solute molecule whose shape deviates from a sphere, the former "solvent sphere" extraction can result in uneven solvation of different regions of the solute (Fig. 6). Users can also specify the ID of the first frame to extract (with some initial equilibration MD steps discarded) and the extraction interval (number of MD steps between two extractions) from the CLI or GUI (SI Fig. S1).

For the final QC calculations, the users can either directly run a specific QC package with the generated xyz files or load the xyz files into QCEngine[5] to enable sophisticated QC workflow at different levels of theory across many different QC packages.

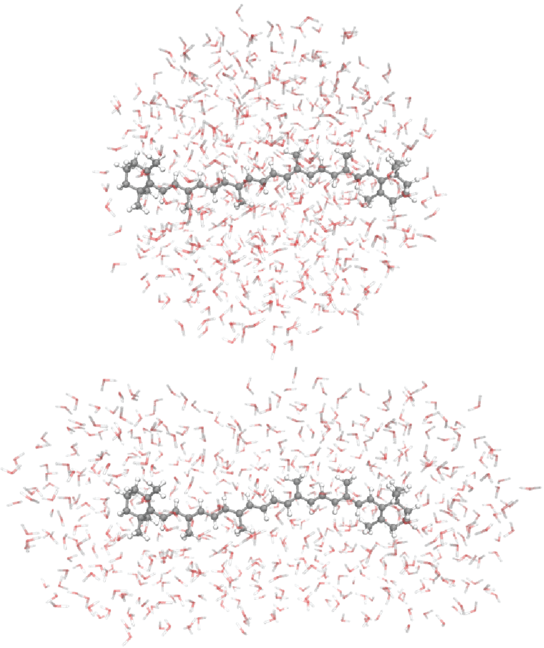


Fig. 6. Comparison of the solvent sphere (upper) and solvent shell (lower) extraction approaches for a representative non-spherical solute, β-carotene. In both approaches, 559 explicit water molecules are extracted, forming a sphere of 16.000 Å radius around the solute center of mass in the sphere extraction (left) and a solvent shell of 9.365 Å thickness around the solute (right).

IV. COMPUTATIONAL DETAILS

We use 166 organic redox couples from the ROP313[57] data (structures available in SI) set to demonstrate the use of AutoSolvate[38] for automating QC calculations of explicitly solvated molecules, and to analyze the dependence of solute-solvent closeness on solvent and solute identities. Throughout this paper the oxidized charge states of the ROP313 data set are used. We use AutoSolvate to generate the solvated structures of the benchmark set, together with input files needed for classical MD and QM/MM simulations. The solvated boxes are first minimized for maximum 2000 steps at the MM level of theory. The systems are then slowly heated to the desired temperature (default 300 K) over 20 ps with a Langevin thermostat with a collision frequency of 2 ps^{-1} and a nonbonded cutoff of 8 Å. Finally, the systems are pressure equilibrated under NPT conditions to the desired pressure (default 1 bar) over 600 ps with a Berendsen barostat with a pressure relaxation time of 1 ps. This long pressure equilibration is chosen to improve the density on the interface between the solute and solvent, which is critical for the explicit redox potential calculation (SI Fig. S2). The resulting pressure-equilibrated system is the initial structure for the QM/MM simulation with Amber/TeraChem, where the QM region is then treated with B3LYP-D3[58]/6-31G*, and the MM region is the explicit solvent. The QM/MM simulation involves an initial energy minimization (250 steps) to remove potential unexpected bond-breaking of the solute due to the abrupt switch from MM to QM description, followed by 0.5 ps of temperature equilibration with Langevin thermostat at 298.15 K with a collision frequency of 5 ps^{-1}. The following 5 ps of QM/MM trajectories under NVT conditions are used for production configuration sampling.

To demonstrate the applicability of this open-source tool, we investigate the outer-sphere contribution to the reorganization energy. The protocol to calculate reorganization energy is available in our previous publication.[59] Explicit solvent configuration sampling is performed for the reduced and oxidized charge states of the aforementioned 166 benchmark systems with AutoSolvate, and 200 snapshots of optimal solvent shell size of 4 Å are extracted from the equilibrated 5 ps of QM/MM trajectory. The C-PCM implicit solvent as implemented in TeraChem was then applied around these microsolvated clusters to estimate the energy gap between the two charge states with B3LYP-D3[58]/6-31G*. The use of C-PCM around the microsolvated cluster ensured converged energy gap at a small solvent shell size and saved computational time, as shown in our previous work.[59] Reorganization energy was successfully obtained for 151 out of the 166 benchmark systems, excluding 15 unconverged systems (listed in SI Text S1).

V. RESULTS AND DISCUSSION

In this section, we will utilize AutoSolvate to investigate the relationship between solvent/solute identities and solvent-solute closeness (d_{tol}), generating a machine learning model that predicts d_{tol} and guides explicitly solvated calculations. We will first explore two distribution functions to quantify d_{tol} given the QM/MM trajectory of an explicitly solvated system and compare their performance in identifying different solvent layers. Using the optimal minimum distance distribution function, we will analyze the trajectories of our AutoSolvate benchmark set with 166 unique organic solutes immersed in different solvents. We will then investigate the dependence of d_{tol} on solute and solvent identities, generating a machine learning (ML) model that predicts d_{tol} in MeCN solution based on molecule structure. Finally, we will also demonstrate that Autosolvate enables high-throughput calculation of complex properties such as the outer-sphere contribution to the reorganization energy of redox reactions from QM/MM trajectories.

A. Analysis on solvent-solute closeness.

As discussed in Section IIIA4, the solvent-solute closeness (d_{tol}) is an important parameter to control the generation of the explicitly solvated structure. Here we calculate two different distribution functions based on the MD trajectory of an explicitly solvated molecule and compare their performance in characterizing solvent layers and d_{tol}.

The radial distribution functions (RDFs) are the most fundamental distribution functions used to investigate the interactions between the components of a liquid. In the simple case where the solute structure is approximately rotationally symmetric, the location of solvent layers can be determined from the RDF of solvent molecules around the solute center. However, the RDFs have several obvious limitations when characterizing d_{tol} due to its ignorance of solute shape and size. First, most explicit-solvent generation algorithms (see Section IIIA4) focus on the minimum distance between any solvent atom and any solute atom. This distance is not directly comparable the distance defined in RDF, which describes the distance from the solute center of mass to solvent. Second, the solvent layer positions determined from the RDF, $G(r)$, are not directly comparable between systems with differently sized solutes, because the RDF peaks can shift with the solute size. A simple fix to these two problems is to introduce a shifted RDF,

$$G_{\text{shifted}}(d) = G(r) = \frac{c}{4\pi r^2}\frac{\Delta n}{\Delta r}, \tag{3}$$

where $d = r - r_{\text{solute}}$ estimates the distance to the solute surface, r_{solute} is the approximate radius of the solute (detail available in SI Text S2), Δn is the number of atoms in the spherical shell with a radius r to $r + \Delta r$ measured from the solute center of mass, and the scaling factor c ensures $G_{\text{shifted}}(\text{bulk}) = 1$ (Fig. 7). However, $G_{\text{shifted}}(d)$ still does not take care of the solute structure complexity, which is often non-spherical.

In contrast, the minimum distance distribution function (MDDF)[60-62] calculates the closest distance between a solvent molecule and the solute. In this work, we introduce an estimated form of MDDF to simplify the analysis, given by

$$\text{MDDF}(d) = \frac{c}{A(d)}\frac{\Delta N}{\Delta d} \approx \frac{c}{4\pi(d + r_{\text{solute}})^2}\frac{\Delta N}{\Delta d}. \tag{4}$$

Here, $A(d)$ is the surface area of the irregular-shaped cavity where d is the distance to the solute surface, ΔN is the number of atoms in the non-spherical shell with a distance between d to $d + \Delta d$ measured from the solute surface, and the scaling factor c ensures that MDDF(bulk) equals to 1. Compared to the original MDDF definition, we substitute the surface area $A(d)$ with the approximated area $4\pi(d + r_{\text{solute}})^2$, but the results are almost identical (SI Fig. S3). The differences in the definitions of G_{shifted} and MDDF are illustrated in in Fig. 7. To compare the performance of G_{shifted} and MDDF, we use both methods to analyze the 5 ps long equilibrated QM/MM trajectory of system 11 (Fig. 7). Both G_{shifted} and MDDF show that the closest solvent molecules start to present at about 2 Å. The first solvent layer peaks at $d = 4$ Å for G_{shifted} and 3 Å for MDDF. The different locations of the first peak can be explained by the more accurate counting of atoms with a given distance to the solvent surface with MDDF. The second and third solvent layer peaks at around 6 Å and 9 Å, respectively, for MDDF, but the G_{shifted} plot is too noisy to reliably confirm the location of the two solvent layers. This result shows that MDDF is more accurate then G_{shifted} in analyzing solvent layers, and we will use MDDF in the next sections to calculate the d_{tol} of our benchmark data set generated with AutoSolvate.

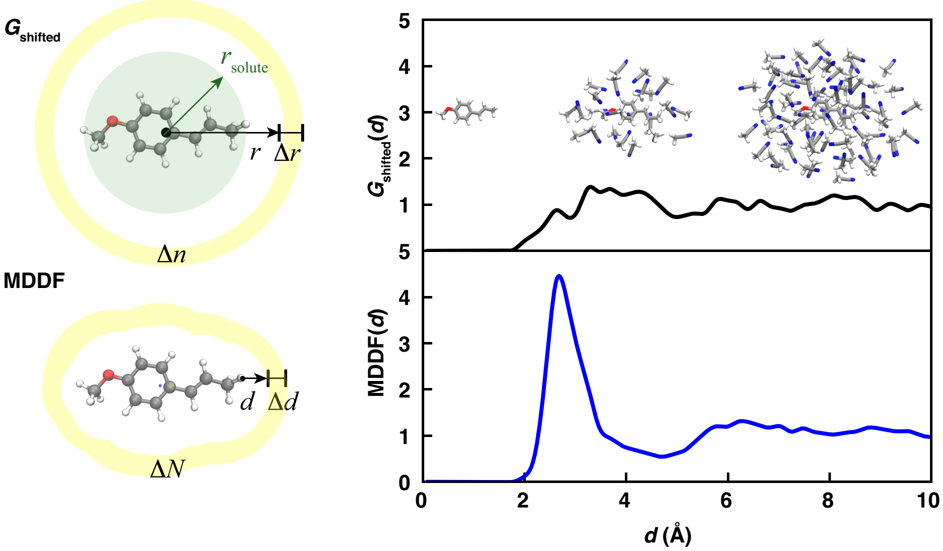


Fig. 7. (left) Illustration of the definitions of the shifted radial distribution function (G_{shifted}) v.s. the minimum distance distribution function (MDDF). Here, r is the distance from the solute center of mass, r_{solute} is the approximated solute radius, d is the distance from a solvent to the closest solute atom, Δn is the number of atoms in the spherical shell with a radius r to $r+ \Delta r$, and ΔN is the number of atoms in the non-spherical shell with a distance d to $d+ \Delta d$ measured from the solute surface. (right) The solvent layer positions for ROP313 system 11 characterized by G_{shifted} and MDDF. The inset shows the microsolvated structures with increasing solvent layer size of 0, 4 and 8 Å.

B. Machine learning predicted solvent-solute closeness.

As discussed in Section IIIA4, the solvent-solute closeness (d_{tol}), is an important parameter for initial solvation structure generation. In general, d_{tol} is system-dependent and needs to be determined from the MD trajectory of the equilibrated system. For practical initial structure generation, d_{tol} is often heuristically estimated because of the lack of an existing MD trajectory of the specific solute-solvent pair. The deviation from the equilibrated d_{tol} needs to be corrected through a longer MD equilibration process. Here we seek to solve this problem with ML. We reveal the correlation between solute/solvent identities and d_{tol} based on the large data set of equilibrated QM/MM trajectories of solvated molecules generated with AutoSolvate and build a ML model to predict d_{tol} for any solvent-solute pair. In this section, we reveal the dependence of d_{tol} on solute identity assuming the solvent is unchanged. In Section VC, we will further reveal the dependence of d_{tol} on the solvent.

We curated a d_{tol} data set for the oxidized charge states of 166 benchmark systems (see section IV) solvated in MeCN, with d_{tol} estimated as the lowest distance to the solute surface where MDDF(d_{tol}) ≥ 1. The data set is split with a 50:50 train-test ratio, and a Gradient Boost[63] ML model is trained to predict d_{tol} based on solute features including the SOAP[64] descriptor generated by DScribe[65] and the solute net charge. A low mean absolute error (MAE) of 0.09 Å is achieved for the 83 test set systems (Fig. 8). Statistics of the d_{tol} values of the test set also reveal a chemically intuitive correlation between d_{tol} and solute structure (Fig. 8). The d_{tol} value varies from 1.6 to 2.6 Å for the test set solutes, where the solutes with -OH or -SH functional groups have a noticeably smaller d_{tol} than the rest of the solutes (Fig. 8). This observation can be explained by the formation of hydrogen bonds between the -OH or -SH groups of the solute with the polar solvent, MeCN. Details about the d_{tol} data set, input features, ML model hyperparameters, and the trained ML model are available in Supporting Information (Text S3 and SI.zip). We also investigated how d_{tol}

changes as a function of simple solute properties like number of atoms, but the correlation is much weaker compared to the ML model results (SI Fig. S4 and S5).

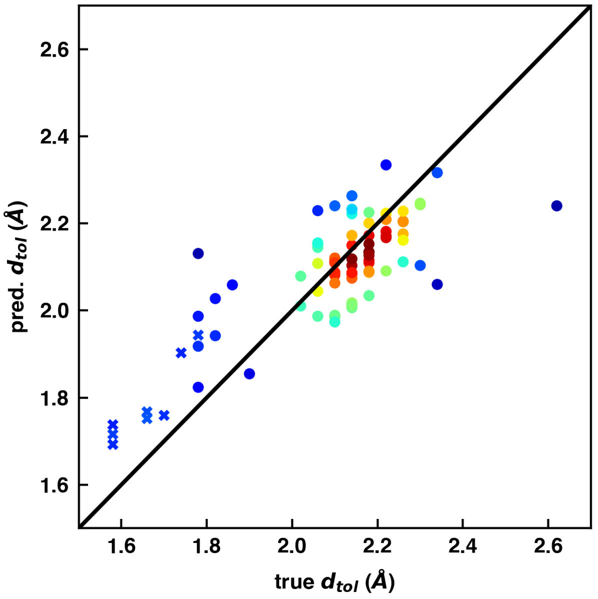


Fig. 8. Predicted ML learned d_{tol} compared with the d_{tol} obtained from the QM/MM trajectories for 83 systems in the test set. Here, d_{tol} is the minimum distance between any atom of the solute and any atom of the solvent. The x markers indicate systems with -OH or -SH functional groups, otherwise the systems are shown as circle markers. The red marker colors indicate higher Kernel density estimation for the data points.

C. Dependence of solvent-solute closeness on solvent

In this section, we reveal the dependence of d_{tol} on the solvent. We curated a d_{tol} data set for 16 benchmark solute molecules (see section IV) solvated in five solvents with different dielectric constants, ε. Statistics on this data set show that d_{tol} significantly decreases as the solvent ε increases (Fig. 9), in agreement with our intuition that polar solvents (with higher ε) interact more strongly with the solute. It is worth noting that this trend applies to all solutes in the set, as all data points shifted to lower d_{tol} values when the solvent changes from chloroform to water (Fig. 9). However, relative d_{tol} values of solvents with similar ε (methanol, MeCN, and NMA) cannot be well predicted by this simple d_{tol}-ε relation and need more sophisticated ML models trained on larger data sets with diverse solvent species. We have implemented automated d_{tol}

recommendation based on the simple d_{tol}-ε relation and will further improve this functionality with advanced ML models when more training data become available.

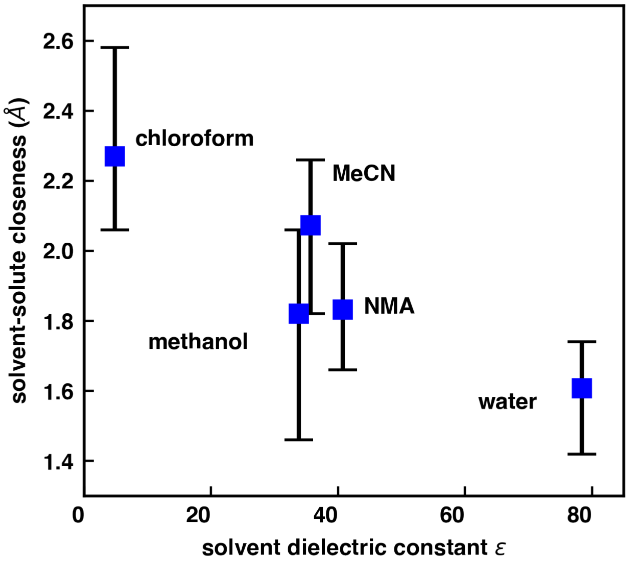


Fig. 9. The dependence of solvent-solute closeness (d_{tol}) on the solvent identity for a data set of 16 solutes selected from ROP313 data set. For each of the 5 investigated solvents, the average, minimum and maximum d_{tol} values are illustrated by the box plot.

D. Application on reorganization energy calculation

To demonstrate the promise of AutoSolvate for chemical discovery efforts, especially the rapid data set curation of complex solution-phase molecular properties, we calculated the outer-sphere contribution to the reorganization energy[66] for the 166 benchmark redox couples based on QM/MM trajectories generated by AutoSolvate. Based on the Marcus theory of electron transfer, the reorganization energy, $\lambda,$ is defined as the energy cost required to rearrange the molecules and the dielectric environment upon this change of charge.[66]

$$\lambda = \lambda_i + \lambda_o \tag{5}$$

where λ_{i} is the inner-sphere contribution due to the solute nuclear rearrangement alone, and λ_{o} is the outer-sphere contribution due to the surrounding solvent. There are various approaches[67, 68] to calculate λ, λ_{i}, and λ_{o} from MM,[69] QM,[70, 71] or QM/MM[72, 73] calculations in implicit[70, 71, 74, 75] or explicit[72, 73] solvent models, or from experimental spectrum data.[76, 77] Specifically, λ_{o} is usually thought to be less easily calculated.[74, 75, 78]

We calculated λ within the limit of linear response approximation, where λ is estimated in terms of the thermal fluctuations of energy gaps between the two charge states:

$$\lambda^{\mathrm{var}} = \frac{\sigma_{\mathrm{ox.}}^2 + \sigma_{\mathrm{red.}}^2}{4k_{\mathrm{b}}T}. \tag{6}$$

Here, k_{b} is the Boltzmann constant, T is the system temperature, and σ is the variance of the vertical energy gap calculated from the oxidized state QM/MM trajectory ($\sigma_{\mathrm{ox.}}$) or the reduced state trajectory ($\sigma_{\mathrm{red.}}$). The superscript "var" indicates that λ is calculated based on variance. We then calculated the inner-sphere contribution, $\lambda_{\mathrm{i}}^{\mathrm{var}}$, using the same approach based on the same set of QM/MM trajectories, but the vertical energy gap was calculated on the solute structures only. The outer-sphere contribution, λ_{o}, is thus calculated as

$$\lambda_{\mathrm{o}} = \lambda^{\mathrm{var}} - \lambda_{\mathrm{i}}^{\mathrm{var}}. \tag{7}$$

The outer-sphere contribution to the reorganization energy, $\lambda_{\mathrm{o}}/\lambda$, is reported to span a wide range. The $\lambda_{\mathrm{o}}/\lambda$ values of 9 organic and organometallic dye molecules in MeCN solution calculated with implicit solvent model by Vaissier and coworkers range from 64% to 89%.[70] Another implicit solvent calculation study by Buda showed that values of 35%-100% can be observed for smaller organic molecules in various organic or aqueous solvents.[79] In contrast, the $\lambda_{\mathrm{o}}/\lambda$ values determined by experimental spectroscopy for 1st-row transition metal ions in aqueous solution are in a significantly lower range of 11%-39%.[77]

Despite the wide range of $\lambda_{\mathrm{o}}/\lambda$ values reported, most computational studies of λ_{o} focus on a small set of systems, or use implicit solvent calculations based on geometry-optimized structures without configuration sampling, due to computational challenge in QM or QM/MM configuration sampling for a large number of explicitly solvated systems. In this work we efficiently calculated the λ_o/λ ratio for a larger, diverse benchmark data set of 166 explicitly solvated systems and obtained converged results for 151 (SI Text S1). The obtained λ_o/λ values range from 0% to 40%, with an average of 14% (Fig. 10). This range is on the lower end of the abovementioned computational literature results, but the solutes investigated in our data set are different, and our calculations are based on explicit solvent configuration sampling instead of implicit solvent-based protocol.[70, 79] In addition, the outer-sphere contributions in Fig. 10 could be overestimated due to using non-polarizable solvent force field, which often underestimates the impact of solvent polarization.[67] The lowest λ_o/λ ratio is found in iodobenzene, and the highest contribution in naphthalene (Fig. 10). The higher outer-sphere contribution for naphthalene can be explained by the high rigidity of this solute, which limits the inner sphere contribution.

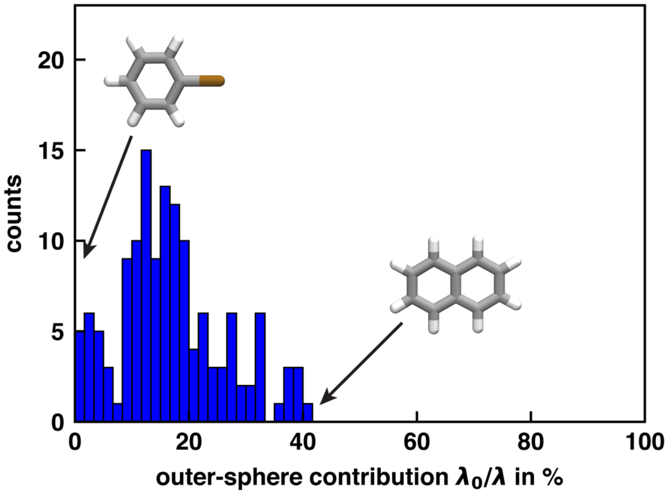


Fig. 10. Histogram of the ratio of outer-sphere contribution to the reorganization energy, λ_o/λ, for 151 converged OROP benchmark solute molecules in MeCN solvent calculated based on QM/MM trajectories. Representative solute molecules with maximum and minimum λ_o/λ values are shown in the inset. Systems with unphysical values below 0% are not shown (listed in SI Text S1). Structures of two systems with the lowest (iodobenzene) and highest (naphthalene) λ_o/λ ratios are shown in the inset.

E. Limitations and future directions

The AutoSolvate workflow currently depends on three computational molecular science packages: AmberTools for classical MD, Gaussian for RESP charge fitting needed for solute force field generation, and TeraChem for QM/MM configuration sampling. In the future, we will improve AutoSolvate's interoperability with different MD and QC packages, and therefore extend its functionality and applicability. First, the RESP charge fitting for solute force field generation will be extended to use various QC packages, especially open-source packages, such as GAMESS[80] and PSI4[81]. An alternative approach is to interface with the PyRED[82-84] program, which will allow automated RESP charge fitting for metal-containing molecules and consideration of molecular orientation. Second, the MD simulation automation can be extended to support the GPU-accelerated open-source MD package OpenMM[85], which supports many different force fields formats, including Amber, CHARMM[86], and GROMACS[87]. Finally, we aim at extending the QM/MM automation to support more QM/MM interfaces, especially the ones with polarizable force fields. This extension will improve the accuracy of the AutoSolvate workflow in predicting many important molecular properties, such as redox potential and UV/Vis absorption/fluorescence spectrum.

VI. CONCLUSIONS

Rapid curation of computational molecular property data sets is crucial for AI design and discovery for chemistry in the solutions phase. However, high-throughput explicit-solvent QC calculations were previously hampered by the complicated preparation steps involving solvated-structure generation, force field fitting, configuration sampling, and microsolvated cluster extraction. To overcome these obstacles, we developed AutoSolvate, an open-source toolkit that streamlines the workflow and enables the seamless generation of microsolvated cluster configurations that can be readily used for QC calculations. Specifically, we aimed at automated estimation of the solute-solvent closeness, a crucial parameter to control the initial structure generation of explicitly solvated systems. We investigated the dependence of solute-solvent closeness on solute identity and trained a ML model to predict solute-solvent closeness for different solutes in MeCN solvent. We also found that solvent-solute minimum distance decreases as the solvent dielectric constant increases, matching the intuition that polar solvents interact more strongly with solutes. Finally, we tested the capability of AutoSolvate for rapid data set curation by calculating the outer-sphere reorganization energy of a large data set of redox couples, which demonstrated the promise of the AutoSolvate package for chemical discovery efforts. In future versions of this package, we will further improve the interoperability of AutoSolvate workflow with more MD and QC packages, making it a helpful tool for the molecular AI design and discovery community to investigate chemistry in the solution phase.

VII. EXTERNAL MATERIAL

AutoSolvate software repository is available at https://github.com/Liu-group/AutoSolvate, and the documentation is available at https://autosolvate.readthedocs.io/. The program remains in active development. Production computations are underway using many features of the software, and test suites are expected to pass. However, users are encouraged to contact the developers as they venture afield of the verified tests.

SUPPLEMENTARY MATERIAL

See the supplementary material for calculation of solute average radius, cluster extraction in GUI and CLI, solvent-solute center distance and closeness vs. system size, NPT MM density equilibration, MDDF approximation validation, unconverged benchmark systems and systems with unphysical λ_o/λ excluded from the λ_o/λ histogram, ML hyperparameters, and data and machine learning model file description (PDF).

Input features of 166 benchmark solutes, d_{tol} of the 166 benchmark solutes in MeCN, d_{tol} of a subset of 16 solutes in 5 solvents, reorganization energies for the 151 converged systems, ML model pkl file, and example python script to use ML model (ZIP).

ACKNOWLEDGMENTS

This work was supported by start-up funds provided by Emory University. This work used the Extreme Science and Engineering Discovery Environment[88] (XSEDE) Bridges-2 at Pittsburgh Supercomputing Center through allocations CHE200099 and CHE210036, which is supported by National Science Foundation grant number ACI-1548562.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning", Phys. Rev. Lett. **108**, 058301 (2012).

[2] L. C. Blum, and J.-L. Reymond, "970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13", J. Am. Chem. Soc. **131**, 8732 (2009).

[3] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17", J. Chem. Inf. Model. **52**, 2864 (2012).

[4] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules", Sci. Data **1**, 140022 (2014).

[5] D. G. A. Smith, A. T. Lolinco, Z. L. Glick, J. Lee, A. Alenaizan, T. A. Barnes, C. H. Borca, R. Di Remigio, D. L. Dotson, S. Ehlert, A. G. Heide, M. F. Herbst, J. Hermann, C. B. Hicks, J. T. Horton, A. G. Hurtado, P. Kraus, H. Kruse, S. J. R. Lee, J. P. Misiewicz, L. N. Naden, F. Ramezanghorbani, M. Scheurer, J. B. Schriber, A. C. Simmonett, J. Steinmetzer, J. R. Wagner, L. Ward, M. Welborn, D. Altarawy, J. Anwar, J. D. Chodera, A. Dreuw, H. J. Kulik, F. Liu, T. J. Martínez, D. A. Matthews, H. F. Schaefer, J. Šponer, J. M. Turney, L.-P. Wang, N. De Silva, R. A. King, J. F. Stanton, M. S. Gordon, T. L. Windus, C. D. Sherrill, and L. A. Burns, "Quantum Chemistry Common Driver and Databases (QCDB) and Quantum Chemistry Engine (QCEngine): Automation and interoperability among computational chemistry programs", J. Chem. Phys. **155**, 204801 (2021).

[6] D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard, and T. D. Crawford, "The MolSSI QCArchive project: An open-source platform to

compute, organize, and share quantum chemistry data", Wiley Interdiscip. Rev. Comput. Mol. Sci. **11**, e1491 (2021).

[7] G. Landrum, RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling; accessed 01/10/2022

[8] M. Lovrić, J. M. Molero, and R. Kern, "PySpark and RDKit: Moving towards Big Data in Cheminformatics", Mol. Inform. **38**, 1800082 (2019).

[9] E. I. Ioannidis, T. Z. H. Gani, and H. J. Kulik, "molSimplify: A toolkit for automating discovery in inorganic chemistry", J. Comput. Chem. **37**, 2106 (2016).

[10] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis", Comp. Mater. Sci. **68**, 314 (2013).

[11] S. P. Huber, E. Bosoni, M. Bercx, J. Bröder, A. Degomme, V. Dikan, K. Eimre, E. Flage-Larsen, A. Garcia, L. Genovese, D. Gresch, C. Johnston, G. Petretto, S. Poncé, G.-M. Rignanese, C. J. Sewell, B. Smit, V. Tseplyaev, M. Uhrin, D. Wortmann, A. V. Yakutovich, A. Zadoks, P. Zarabadi-Poor, B. Zhu, N. Marzari, and G. Pizzi, "Common workflows for computing material properties using different quantum engines", npj Comput. Mater. **7** (2021).

[12] Y. Guan, V. M. Ingman, B. J. Rooks, and S. E. Wheeler, "AARON: an automated reaction optimizer for new catalysts", J. Chem. Theory Comput. **14**, 5249 (2018).

[13] V. M. Ingman, A. J. Schaefer, L. R. Andreola, and S. E. Wheeler, "QChASM: Quantum chemistry automation and structure manipulation", Wiley Interdiscip. Rev. Comput. Mol. Sci. **11**, e1510 (2021).

[14] J. Tomasi, and M. Persico, "Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent", Chem. Rev. **94**, 2027 (1994).

[15] C. J. Cramer, and D. G. Truhlar, "Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics", Chem. Rev. **99**, 2161 (1999).

[16] M. Orozco, and F. J. Luque, "Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems", Chem. Rev. **100**, 4187 (2000).

[17] J. Tomasi, B. Mennucci, and R. Cammi, "Quantum Mechanical Continuum Solvation Models", Chem. Rev. **105**, 2999 (2005).

[18] S. Miertuš, E. Scrocco, and J. Tomasi, "Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects", Chem. Phys. **55**, 117 (1981).

[19] A. Klamt, and G. Schüürmann, "COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient", J. Chem. Soc., Perkin Trans. 2, 799 (1993).

[20] V. Barone, and M. Cossi, "Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model", J. Phys. Chem. A **102**, 1995 (1998).

[21] T. N. Truong, and E. V. Stefanovich, "A new method for incorporating solvent effect into the classical, ab initio molecular orbital and density functional theory frameworks for arbitrary shape cavity", Chem. Phys. Lett. **240**, 253 (1995).

[22] B. Mennucci, E. Cancès, and J. Tomasi, "Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications", J. Phys. Chem. B **101**, 10506 (1997).

[23] E. Cances, B. Mennucci, and J. Tomasi, “A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics”, J. Chem. Phys. **107**, 3032 (1997).

[24] J. Tomasi, B. Mennucci, and E. Cancès, “The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level”, J. Mol. Struct.: THEOCHEM **464**, 211 (1999).

[25] J. L. Pérez-Lustres, F. Rodriguez-Prieto, M. Mosquera, T. A. Senyushkina, N. P. Ernsting, and S. A. Kovalenko, “Ultrafast Proton Transfer to Solvent: Molecularity and Intermediates from Solvation- and Diffusion-Controlled Regimes”, J. Am. Chem. Soc. **129**, 5408 (2007).

[26] U. Raucci, M. G. Chiariello, and N. Rega, “Modeling Excited-State Proton Transfer to Solvent: A Dynamics Study of a Super Photoacid with a Hybrid Implicit/Explicit Solvent Model”, J. Chem. Theory Comput. **16**, 7033 (2020).

[27] J. M. Boereboom, P. Fleurat-Lessard, and R. E. Bulo, “Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition”, J. Chem. Theory Comput. **14**, 1841 (2018).

[28] M. M. Pinney, A. Natarajan, F. Yabukarski, D. M. Sanchez, F. Liu, R. Liang, T. Doukov, J. P. Schwans, T. J. Martinez, and D. Herschlag, “Structural Coupling Throughout the Active Site Hydrogen Bond Networks of Ketosteroid Isomerase and Photoactive Yellow Protein”, J. Am. Chem. Soc. **140**, 9827 (2018).

[29] P. A. Sigala, E. A. Ruben, C. W. Liu, P. M. B. Piccoli, E. G. Hohenstein, T. J. Martínez, A. J. Schultz, and D. Herschlag, “Determination of Hydrogen Bond Structure in Water versus Aprotic Environments To Test the Relationship Between Length and Stability”, J. Am. Chem. Soc. **137**, 5730 (2015).

[30] A. W. Götz, M. A. Clark, and R. C. Walker, “An extensible interface for QM/MM molecular dynamics simulations with AMBER”, J. Comput. Chem. **35**, 95 (2014).

[31] A. Warshel, and M. Levitt, “Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme”, J. Mol. Biol. **103**, 227 (1976).

[32] U. C. Singh, and P. A. Kollman, “A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH3Cl+ Cl− exchange reaction and gas phase protonation of polyethers”, J. Comput. Chem. **7**, 718 (1986).

[33] M. J. Field, P. A. Bash, and M. Karplus, “A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations”, J. Comput. Chem. **11**, 700 (1990).

[34] D. Bakowies, and W. Thiel, “Hybrid models for combined quantum mechanical and molecular mechanical approaches”, The Journal of Physical Chemistry **100**, 10580 (1996).

[35] G. N. Simm, P. L. Türtscher, and M. Reiher, “Systematic microsolvation approach with a cluster-continuum scheme and conformational sampling”, J. Comput. Chem. **41**, 1144 (2020).

[36] J. Zhang, and M. Dolg, “ABCluster: the artificial bee colony algorithm for cluster global optimization”, Phys. Chem. Chem. Phys. **17**, 24173 (2015).

[37] J. Zhang, and M. Dolg, “Global optimization of clusters of rigid molecules using the artificial bee colony algorithm”, Phys. Chem. Chem. Phys. **18**, 3003 (2016).

[38] E. Hruska, A. Gale, X. Huang, and F. Liu, AutoSolvate https://github.com/Liu-group/AutoSolvate; accessed 01/10/2022

[39] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, “Open Babel: An open chemical toolbox”, J. Cheminf. **3**, 1 (2011).

[40] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, “PACKMOL: a package for building initial configurations for molecular dynamics simulations”, J. Comput. Chem. **30**, 2157 (2009).
[41] Anaconda Software Distribution https://docs.anaconda.com/; accessed 01/10/2022
[42] O. Ben-Kiki, C. Evans, and B. Ingerson, “Yaml ain't markup language (yaml™) version 1.1”, Working Draft 2008-05 **11** (2009).
[43] J. E. Grayson, *Python and Tkinter programming* (Manning Publications Co. Greenwich, (2000),
[44] J. W. Shipman, “Tkinter 8.4 reference: a GUI for Python”, New Mexico Tech Computer Center **54** (2013).
[45] B. B. Welch, K. Jones, and J. Hobbs, *Practical Programming in Tcl/Tk* (Prentice Hall Professional, (2003),
[46] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola, “Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules”, Acta Crystallogr. D. **54**, 1078 (1998).
[47] J. W. Ponder, and D. A. Case, “Force fields for protein simulations”, Adv. Protein Chem. **66**, 27 (2003).
[48] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field”, J. Comput. Chem. **25**, 1157 (2004).
[49] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, “Automatic atom type and bond type perception in molecular mechanical calculations”, J. Mol. Graphics Modell. **25**, 247 (2006).
[50] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model”, J. Phys. Chem. **97**, 10269 (1993).
[51] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman, “Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation”, J. Am. Chem. Soc. **115**, 9620 (2002).
[52] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 16 Rev. C.01; accessed 01/10/2022
[53] A. Jakalian, D. B. Jack, and C. I. Bayly, “Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation”, J. Comput. Chem. **23**, 1623 (2002).
[54] Tripos, Tripos Mol2 File Format (1988)
[55] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, “PACKMOL: A package for building initial configurations for molecular dynamics simulations”, Journal of Computational Chemistry **30**, 2157 (2009).
[56] Bryce Group, Amber Parameter Database, http://amber.manchester.ac.uk/; accessed 01/10/2022
[57] H. Neugebauer, F. Bohle, M. Bursch, A. Hansen, and S. Grimme, “Benchmark Study of Electrochemical Redox Potentials Calculated with Semi-empirical and DFT Methods”, J. Phys. Chem. A 10.1021/acs.jpca.0c05052 (2020).

[58] S. Grimme, S. Ehrlich, and L. Goerigk, “Effect of the damping function in dispersion corrected density functional theory”, J. Comput. Chem. **32**, 1456 (2011).

[59] E. Hruska, A. Gale, and F. Liu, “Bridging the experiment-calculation divide: machine learning corrections to redox potential calculations in implicit and explicit solvent models”, J. Chem. Theory Comput. https://doi.org/10.1021/acs.jctc.1c01040 (2022).

[60] L. Martínez, and S. Shimizu, “Molecular Interpretation of Preferential Interactions in Protein Solvation: A Solvent-Shell Perspective by Means of Minimum-Distance Distribution Functions”, J. Chem. Theory Comput. **13**, 6358 (2017).

[61] L. Martínez, “ComplexMixtures.jl: Investigating the structure of solutions of complex-shaped molecules from a solvent-shell perspective”, J. Mol. Liq., 117945 (2021).

[62] L. Cai, W. Lv, H. Zhu, and Q. Xu, “Molecular dynamics simulation on adsorption of pyrene-polyethylene onto ultrathin single-walled carbon nanotube”, Physica E. **81**, 226 (2016).

[63] H. F. Jerome, “Greedy function approximation: A gradient boosting machine”, Ann. Stat. **29**, 1189 (2001).

[64] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, “Comparing molecules and solids across structural and alchemical space”, Phys. Chem. Chem. Phys. **18**, 13754 (2016).

[65] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, “DScribe: Library of descriptors for machine learning in materials science”, Comput. Phys. Commun. **247**, 106949 (2020).

[66] R. A. Marcus, and N. Sutin, “Electron transfers in chemistry and biology”, Bba-rev. Bioenergetics **811**, 265 (1985).

[67] J. Blumberger, “Recent Advances in the Theory and Molecular Simulation of Biological Electron Transfer Reactions”, Chem. Rev. **115**, 11191 (2015).

[68] J. Blumberger, “Free energies for biological electron transfer from QM/MM calculation: method, application and critical assessment”, Phys. Chem. Chem. Phys. **10**, 5651 (2008).

[69] B. S. Brunschwig, and N. Sutin, “Energy surfaces, reorganization energies, and coupling elements in electron transfer”, Coordin. Chem. Rev. **187**, 233 (1999).

[70] V. Vaissier, P. Barnes, J. Kirkpatrick, and J. Nelson, “Influence of polar medium on the reorganization energy of charge transfer between dyes in a dye sensitized film”, Phys. Chem. Chem. Phys. **15**, 4804 (2013).

[71] L. Eberson, R. Gonz·lez-Luque, J. Lorentzon, M. Merch·n, and B. O. Roos, “The ab initio calculation of inner sphere reorganization energies of inorganic redox couples”, J. Am. Chem. Soc. **115**, 2898 (1993).

[72] E. Falbo, and T. J. Penfold, “Redox Potentials of Polyoxometalates from an Implicit Solvent Model and QM/MM Molecular Dynamics”, J. Phys. Chem. C **124**, 15045 (2020).

[73] X. Jiang, Z. Futera, and J. Blumberger, “Ergodicity-Breaking in Thermal Biological Electron Transfer? Cytochrome C Revisited”, J. Phys. Chem. B **123**, 7588 (2019).

[74] K. A. Sharp, “Calculation of electron transfer reorganization energies using the finite difference Poisson-Boltzmann model”, Biophys. J. **74**, 1241 (1998).

[75] Y.-P. Liu, and M. D. Newton, “Reorganization Energy for Electron Transfer at Film-Modified Electrode Surfaces: A Dielectric Continuum Model”, J. Phys. Chem. **98**, 7162 (1994).

[76] S. U. M. Khan, and J. O'M. Bockris, “Contribution of inner-sphere reorganization in electron-transfer reaction in solution”, Chem. Phys. Lett. **99**, 83 (1983).

[77] S. U. Khan, and J. O. Bockris, “Relative contributions of inner-and outer-shell reorganization in electron-transfer reactions in solution”, J. Phys. Chem. **87**, 4012 (1983).

[78] E. Maggio, N. Martsinovich, and A. Troisi, “Evaluating Charge Recombination Rate in Dye-Sensitized Solar Cells from Electronic Structure Calculations”, J. Phys. Chem. C **116**, 7638 (2012).
[79] M. Buda, “On calculating reorganization energies for electrochemical reactions using density functional theory and continuum solvation models”, Electrochim. Acta **113**, 536 (2013).
[80] G. M. J. Barca, C. Bertoni, L. Carrington, D. Datta, N. De Silva, J. E. Deustua, D. G. Fedorov, J. R. Gour, A. O. Gunina, E. Guidez, T. Harville, S. Irle, J. Ivanic, K. Kowalski, S. S. Leang, H. Li, W. Li, J. J. Lutz, I. Magoulas, J. Mato, V. Mironov, H. Nakata, B. Q. Pham, P. Piecuch, D. Poole, S. R. Pruitt, A. P. Rendell, L. B. Roskop, K. Ruedenberg, T. Sattasathuchana, M. W. Schmidt, J. Shen, L. Slipchenko, M. Sosonkina, V. Sundriyal, A. Tiwari, J. L. Galvez Vallejo, B. Westheimer, M. Włoch, P. Xu, F. Zahariev, and M. S. Gordon, “Recent developments in the general atomic and molecular electronic structure system”, J. Phys. Chem. **152**, 154102 (2020).
[81] D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. Deprince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, “PSI4 1.4: Open-source software for high-throughput quantum chemistry”, J. Chem. Phys. **152**, 184108 (2020).
[82] F.-Y. Dupradeau, A. Pigache, T. Zaffran, C. Savineau, R. Lelong, N. Grivel, D. Lelong, W. Rosanski, and P. Cieplak, “The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building”, Phys. Chem. Chem. Phys. **12**, 7821 (2010).
[83] E. Vanquelef, S. Simon, G. Marquant, E. Garcia, G. Klimerak, J. C. Delepine, P. Cieplak, and F.-Y. Dupradeau, “R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments”, Nucleic Acids Res. **39**, W511 (2011).
[84] F. Wang, J.-P. Becker, P. Cieplak, and F.-Y. Dupradeau, in *Abstr. Pap. Am. Chem. S.* (American Chemical Society, 2014).
[85] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics”, Plos Comput. Biol. **13**, e1005659 (2017).
[86] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, “CHARMM: The biomolecular simulation program”, J. Comput. Chem. **30**, 1545 (2009).
[87] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”, SoftwareX **1-2**, 19 (2015).
[88] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, “XSEDE: Accelerating Scientific Discovery”, Comput. Sci. Eng. **16**, 62 (2014).