Supplementary Material for:

MBX: A Many-Body Energy and Force Calculator for Data-Driven Many-Body Simulations

Marc Riera,[1, a)] Christopher Knight,[2, b)] Ethan F. Bull-Vulpe,[1] Xuanyu Zhu,[1] Henry Agnew,[1] Daniel G. A. Smith,[3] Andrew C. Simmonett,[4] and Francesco Paesani[1, 5, 6, 7, c)]

[1)]*Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California 92093, USA*

[2)]*Argonne National Laboratory, Computational Science Division, Lemont, IL 60439, United States*

[3)]*Molecular Sciences Software Institute, Blacksburg, Virginia 24060, USA*

[4)]*Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

[5)]*Materials Science and Engineering, University of California San Diego, La Jolla, California 92093, USA*

[6)]*Halicioğlu Data Science Institute, University of California San Diego, La Jolla, California 92093, United States*

[7)]*San Diego Supercomputer Center, University of California San Diego, La Jolla, California 92093, USA*

[a)]Electronic mail: mrierari@ucsd.edu

[b)]Electronic mail: knightc@anl.gov

[c)]Electronic mail: fpaesani@ucsd.edu

S1. THE NRG FILE

When used as a standalone code or interfaced with external software such as LAMMPS[1] and i-PI,[2] MBX requires a special input NRG file containing information the `System` class needs to construct the monomers. The NRG file can contain one or more systems, each of which is a group of molecules. Each molecule can contain one or more monomers, but in the current version of MBX, each molecule is composed of a single monomer. Each system is delimited by the `SYSTEM` and `ENDSYS` keywords to denote the beginning and end of a system, respectively. Each system can be composed by one or more molecules, and they are delimited by the `MOLECULE` and `ENDMOL` keywords. Each molecule can be composed by one or multiple monomers, delimited by the `MONOMER monomer_id` and `ENDMON` keywords, where `monomer_id` is the name of the monomer. The monomers available in the current version of MBX and their corresponding names are listed in Table S1. Within each pair of `MONOMER monomer_id` and `ENDMON` tags, the atoms and

TABLE S1. Monomer type and name in MBX

Molecule	MBX name	Atom order
H_2O	h2o	OHH
Li ion	li+	Li
Na ion	na+	Na
K ion	k+	K
Rb ion	rb+	Rb
Cs ion	cs+	Cs
F ion	f-	F
Cl ion	cl-	Cl
Br ion	br-	Br
I ion	i-	I
Ar	ar	Ar
CO_2	co2	COO
CH_4	ch4	CHHHH
N_2O_5	n2o5	ONNOOOO
NH_3	nh3	NHHH

their coordinates are written in the same format as in a standard *xyz* file. Each line contains the atom name followed by the corresponding *x*, *y*, and *z* coordinates.

Shown below is an example of an NRG file for two systems. The first system contains a CO_2 molecule, and the second system contains a H_2O molecule and a Na^+ ion. When performing energy and gradient calculations during molecular dynamics, there should only be one system in the NRG file, representing the current geometry of the atoms in the simulation box. For other applications, NRG files with multiple systems may be passed, and MBX will iteratively perform calculations on each of them.

```
SYSTEM One CO2
MOLECULE
MONOMER co2
C  0.0 0.0 0.0
O  1.1 0.0 0.0
O -1.1 0.0 0.0
ENDMON
ENDMOL
ENDSYS

SYSTEM Na+ H2O
MOLECULE
MONOMER na+
Na 0.0 0.0 0.0
ENMON
ENDMOL
MOLECULE
MONOMER h2o
O 2.0  0.0 0.0
H 2.5  0.5 0.0
H 2.5 -0.5 0.0
ENDMON
ENDMOL
ENDSYS
```

S2. JSON CONFIGURATION FILE

The JSON file contains parameters that control the details of the MB-nrg energy/force calculation. An example of JSON file for i-PI is shown below:

```
{
   "Note" : "This is a configuration file",
   "MBX" : {
       "box" : [21.0,0.0,0.0,0.0,21.0,0.0,0.0,0.0,21.0],
       "twobody_cutoff"   : 9.0,
       "threebody_cutoff" : 7.0,
       "dipole_tolerance" : 1E-8,
       "dipole_max_it"    : 100,
       "dipole_method"     : "cg",
       "alpha_ewald_elec" : 0.6,
       "grid_density_elec" : 2.5,
       "spline_order_elec" : 6,
       "alpha_ewald_disp" : 0.6,
       "grid_density_disp" : 2.5,
       "spline_order_disp" : 6,
       "ignore_2b_poly" : [],
       "ignore_3b_poly" : []
   } ,
   "i-pi" : {
       "port" : 34567,
       "localhost" : "localhost3"
   }
}
```

When MBX is interfaced with LAMMPS, the following section related to the interface with i-PI should be removed from the JSON file:

```
{
   "i-pi" : {
       "port" : 34567,
       "localhost" : "localhost3"
   }
}
```

Below is a brief summary of each one of the configuration features:

- `box` refers to the simulation cell. It is either a 9 element list, comma-separated and enclosed by brackets with the 3 vectors of the box: $[a_x, a_y, a_z, b_x, b_y, b_z, c_x, cy, c_z]$; a 6 element list if the periodic cell is given by the lengths of the three sides and their angles: $[A, B, C, \alpha, \beta, \gamma]$; or an empty list if the energy calculation is performed in the gas phase.

- `twobody_cutoff` is the cutoff distance for the 2-mer selection process ($r^{\mathrm{2B}}_{\mathrm{cutoff}}$) as well as the distance at which the physics-based interactions are truncated in real space. For gas-phase calculations, `twobody_cutoff` should be set to a value larger than the dimensions of the system being simulated in order to ensure that all electrostatic and dispersion interactions are properly accounted for in real space. For calculations in periodic boundary conditions, `twobody_cutoff` should be set to the largest outer cutoff distance adopted by any 2-body switching function (s^{2B}) participating in a $V^{\mathrm{2B}}_{\mathrm{ML}}$ term describing the system being simulated or 9.0 Å (which is the recommended minimum value for the real-space cutoff for the physics-based terms), whichever is larger.

- `threebody_cutoff` is the cutoff distance for the 3-mer selection process ($r^{\mathrm{3B}}_{\mathrm{cutoff}}$). See Section S3 of this document for details on choosing a value of $r^{\mathrm{3B}}_{\mathrm{cutoff}}$. The user is referred to the original publications describing each MB-nrg PEF for specific details about the corresponding 3-body cutoff distances.

- `dipole_tolerance` (ε) is the convergence tolerance for the self-consistent calculation of the induced dipole moments. A dipole moment calculation achieves convergence when

$$|\mu_i^{t+1} - \mu_i^t|^2 < \varepsilon, \forall\, i$$

- `dipole_max_it` is the maximum number of iterations allowed for the calculation of the dipole moments. If the number of iterations exceeds this value, the calculation of the dipole moments is considered to have diverged.

- `dipole_method` defines the method used for the calculation of the induced dipole moments. The options available in the current version of MBX are: `iter` (iterative), `cg` (conjugate gradient), and `aspc` (always stable predictor-corrector). `aspc` is only recommended for MD simulations.

- `alpha_ewald_elec` / `alpha_ewald_disp` is the splitting parameter used by the PME algorithm to partition Coulomb/dispersion interactions into short- and long-range components.[3] It is set to 0 when the calculation/simulation is performed in the gas phase since, in this case, all electrostatic and dispersion contributions are calculated in real space.

- `grid_density_elec` / `grid_density_disp` is the number of grid points per Å in the PME mesh for the calculation of electrostatic and dispersion interactions, respectively.[3]

- `spline_order_elec` / `spline_order_disp` is the order of the splines used for interpolation within the PME mesh used to calculate electrostatic and dispersion interactions, respectively.[3]

- `ignore_2b_poly` is a list of monomer pairs for which the 2-body PIPs should not calculated. Example: "ignore_2b_poly" : [["na","h2o"]] will omit PIP contributions to the sodium-water 2-body energy.

- `ignore_3b_poly` is a list of monomer triples for which the 3-body PIPs should not be calculated. Example: "ignore_3b_poly" : [["na","h2o","h2o"]] will omit PIP contributions to the sodium-water-water 3-body energy.

- `port` (only used when MBX is interfaced with i-PI) specifies the internet/unix port that is used when communicating with i-PI. It is recommended to set the port number to a value larger than 34500.

- `localhost` (only used when MBX is interfaced with i-PI) is the name of the socket. It must match the name in the XML input file of i-PI.

S3. SWITCHING FUNCTIONS AND CUTOFF DISTANCES

In each of the n-body machine-learned terms ($V_{\mathrm{ML}}^{n\mathrm{B}}$), a switching function ($s^{n\mathrm{B}}$) is used to switch off the corresponding n-body PIP ($V_{\mathrm{PIP}}^{n\mathrm{B}}$) as the 1-mers in the n-mer are separated:

$$V_{\mathrm{ML}}^{n\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\ldots,\mathrm{M}_l|K\right) = s^{n\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\ldots,\mathrm{M}_l|K\right) V_{\mathrm{PIP}}^{n\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\ldots,\mathrm{M}_l|K\right) \tag{S1}$$

where $\mathrm{M}_i,\mathrm{M}_j,\ldots,\mathrm{M}_l$ are monomers that belong to an n-mer of type K.

MBX allows the $V_{\mathrm{ML}}^{n\mathrm{B}}$ for each n-mer type to adopt a different form of the switching function. A common form for the 2-body switching function adopted by the MB-nrg PEFs available in MBX is:

$$s^{2\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j|K\right) = s^{2\mathrm{B}}\left(r_{ij}|r_{\mathrm{in}}^K, r_{\mathrm{out}}^K\right) \tag{S2}$$

where r_{in}^K and r_{out}^K are the inner and outer cutoffs for 2-mer type K, r_{ij} is the distance between the first atom in the ith and jth 1-mer, and $s^{2\mathrm{B}}(r|r_{\mathrm{in}},r_{\mathrm{out}})$ is a 2-body switching function defined as

$$s^{2\mathrm{B}}(r|r_{\mathrm{in}},r_{\mathrm{out}}) = \begin{cases} 1 & \text{if } r < r_{\mathrm{in}} \\ 0.5\left(1+\cos(\pi x)\right) & \text{if } r_{\mathrm{in}} \le r < r_{\mathrm{out}} \\ 0 & \text{if } r_{\mathrm{out}} \le r \end{cases} \tag{S3}$$

with

$$x = \left(\frac{r - r_{\mathrm{in}}}{r_{\mathrm{out}} - r_{\mathrm{in}}}\right) \tag{S4}$$

in which, by design, $s^{2B}(r|r_{\mathrm{in}},r_{\mathrm{out}})$ smoothly goes from 1 to 0 as r moves from the inner cutoff r_{in} to the outer cutoff r_{out}.

Possible forms of the 3-body and 4-body switching functions are:

$$s^{3\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\mathrm{M}_k|r_{\mathrm{in}}=0\right) = \frac{1}{3}\left(s_{ij}s_{ik} + s_{ij}s_{jk} + s_{ik}s_{jk}\right) \tag{S5}$$

$$s^{3\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\mathrm{M}_k\right) = s_{ij}s_{ik} + s_{ij}s_{jk} + s_{ik}s_{jk} - 2s_{ij}s_{ik}s_{jk} \tag{S6}$$

$$\begin{aligned} s^{4\mathrm{B}}\left(\mathrm{M}_i,\mathrm{M}_j,\mathrm{M}_k,\mathrm{M}_l\right) = {} & 3\, s_{ij}s_{ik}s_{il}s_{jk}s_{jl}s_{kl} \\ & - s_{ik}s_{il}s_{jk}s_{jl}s_{kl} - s_{ij}s_{il}s_{jk}s_{jl}s_{kl} - s_{ij}s_{ik}s_{jk}s_{jl}s_{kl} \\ & - s_{ij}s_{ik}s_{il}s_{jl}s_{kl} - s_{ij}s_{ik}s_{il}s_{jk}s_{kl} - s_{ij}s_{ik}s_{il}s_{jk}s_{jl} \\ & + s_{ij}s_{ik}s_{il} + s_{ij}s_{jk}s_{jl} + s_{ik}s_{jk}s_{kl} + s_{il}s_{jl}s_{kl} \end{aligned} \tag{S7}$$

where $s_{\alpha\beta} = s^{2\mathrm{B}}(r_{\alpha\beta}|r_{\mathrm{in}}^{K}, r_{\mathrm{out}}^{K})$ with $\alpha, \beta = i, j, k, l$ which are indices of 1-mers in an 3-mer / 4-mer of type K, and r_{in}^{K} / r_{out}^{K} are the inner / outer cutoff for n-mer type K.

The rules for setting the cutoff variables used by MBX in the n-mer selection process, $r_{\mathrm{cutoff}}^{n\mathrm{B}}$, are summarized as follows:

- For all $n \geq 2$, $r_{\mathrm{cutoff}}^{n\mathrm{B}}$ should never exceed half of the shortest side of the periodic box in order to fulfill the minimum image convention (MIC). In addition, if $s^{3\mathrm{B}}$ adopts Eq. S5 or Eq. S6 and $s^{4\mathrm{B}}$ adopts Eq. S7, then $r_{\mathrm{cutoff}}^{3\mathrm{B}/4\mathrm{B}}$ should not exceed $\frac{1}{3}$ of the shortest side of the periodic box.

- For each $n \geq 2$, $r_{\mathrm{cutoff}}^{n\mathrm{B}}$ needs to be large enough so that if $s^{n\mathrm{B}} > 0$ for an n-mer, that n-mer should obey the CENTER-NEIGHBOR CRITERION (see definition from main text). Note that $s^{2\mathrm{B}}$, $s^{3\mathrm{B}}$ and $s^{4\mathrm{B}}$ of the forms discussed above fulfill this requirement only if $r_{\mathrm{cutoff}}^{n\mathrm{B}} > r_{\mathrm{out}}^{K}$, which implies that $r_{\mathrm{cutoff}}^{n\mathrm{B}}$ should be no smaller than the maximum of r_{out}^{K} among all n-mer types (K) involved in the system of interest.

S4. TIMINGS

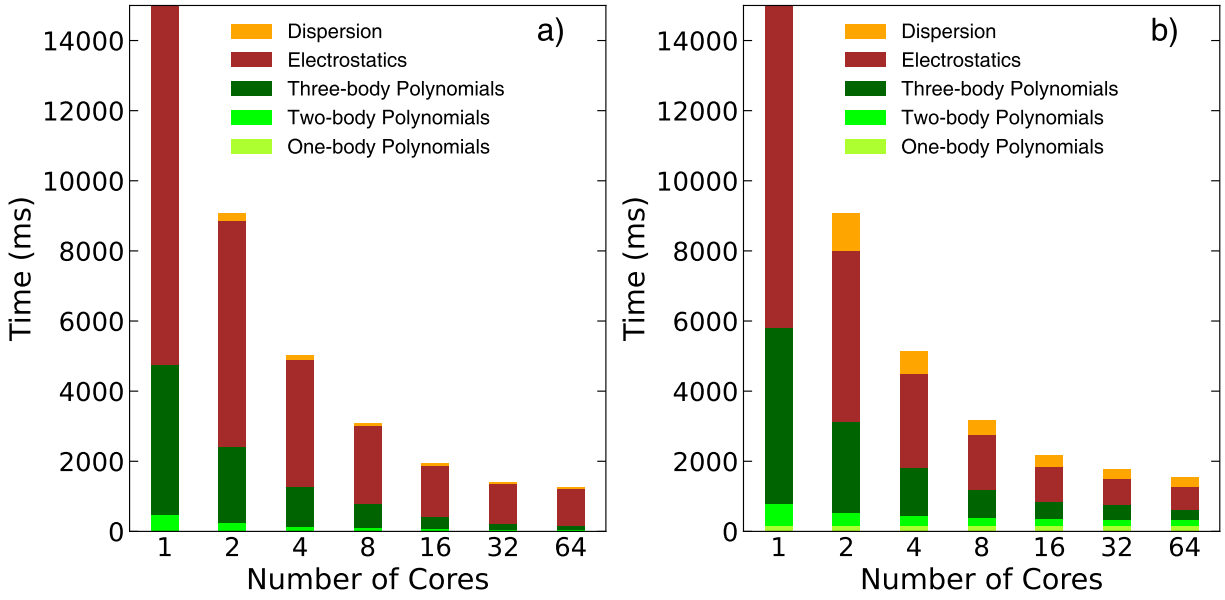

FIG. S1. Mean time to calculate all energies and gradients for a cubic box of 2048 water molecules in MBX in periodic boundary conditions. Calculations were each performed 100 times, and the average was taken. The times are presented as a function of the number of OpenMP threads used with MBX as a standalone code (a) and with LAMMPS using a single MPI rank (b). All the calculations were performed on a compute node with two sockets each with 64 2.6GHz AMD 7H12 Rome processors.

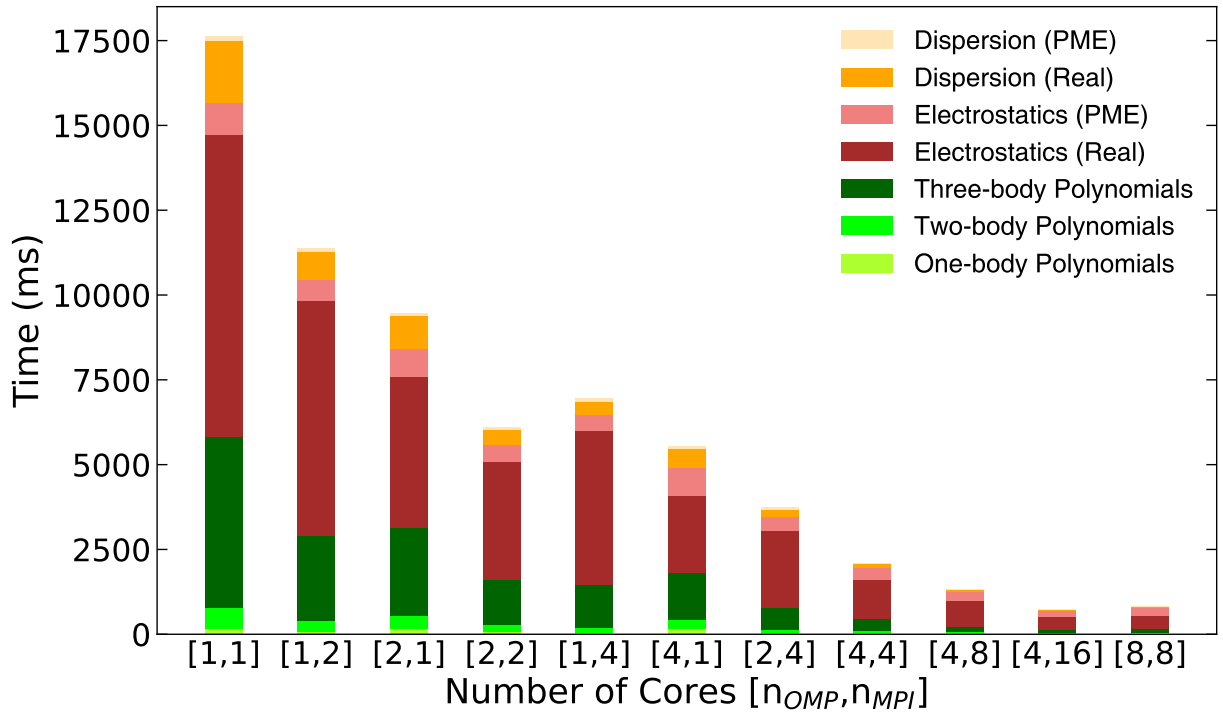

FIG. S2. Mean time to calculate all energies and gradients for a cubic box of 2048 water molecules in periodic boundary conditions using MBX interfaced with LAMMPS. Calculations were each performed 100 times, and the average was taken. The times are presented as a function of the number of OpenMP threads (n_{OMP}) per MPI rank and the number of MPI ranks (n_{MPI}). Calculations were performed on a compute node with two sockets each with 64 2.6GHz AMD 7H12 Rome processors.

S5. ENERGY CONSERVATION

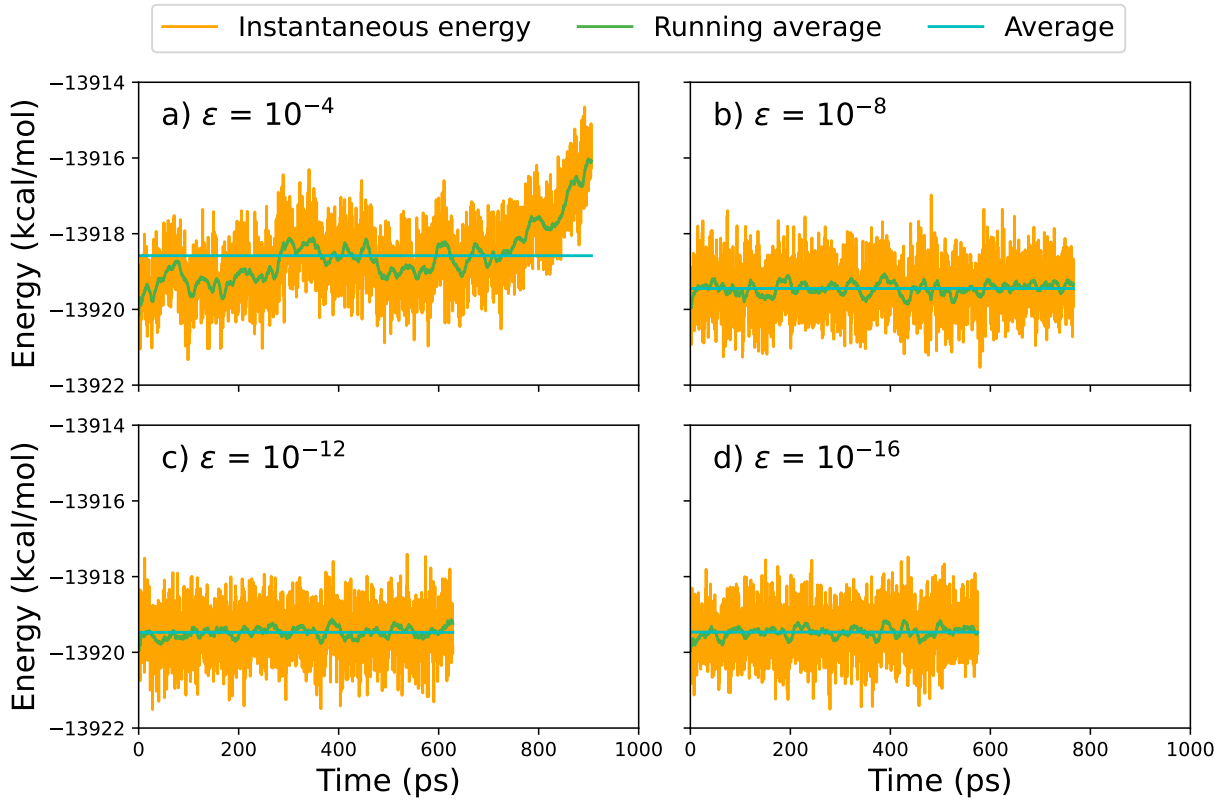

FIG. S3. Analysis of the energy conservation and associated energy fluctuations for simulations of 2048 water molecules in a periodic cubic box carried out in the microcanonical ensemble as a function of the convergence tolerance (ε)for the induced dipole moments. a) $\varepsilon = 10^{-4}$, b) $\varepsilon = 10^{-8}$, c) $\varepsilon = 10^{-12}$, d) $\varepsilon = 10^{-16}$.

REFERENCES

[1]A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comp. Phys. Comm. **271**, 108171 (2022).

[2]V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Corminboeuf, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, "i-PI 2.0: A universal force engine for advanced molecular simulations," Comput. Phys. Commun. **236**, 214–223 (2019).

[3]A. C. Simmonett and B. R. Brooks, "A compression strategy for particle mesh ewald theory," J. Chem. Phys. **154**, 054112 (2021).