Investigating Reversibility in Unwrapping Schemes for Molecular Dynamics Simulations

Martin Kulke and Josh V Vermaas

MSU-DOE Plant Research Laboratory and Department of Biochemistry and Molecular Biology, Michigan State University, 612 Wilson Rd, East Lansing, MI 48824, United States of America.

E-mail: vermaasj@msu.edu

Abstract

Molecular simulation technologies have afforded researchers a unique look into the nanoscale interactions driving physical processes. However, a limitation for Molecular Dynamics (MD) simulations is that they must be performed on finite-sized systems in order to map onto computational resources. To minimize artefacts arising from finite-sized simulation systems, it is common practice for MD simulations to be performed with periodic boundary conditions (PBC). In periodic systems, atoms near simulation cell edges interact with others across the periodic boundary, effectively mimicking a much larger simulated volume and eliminating edge effects from the simulation system. As a side-effect, particles exiting the simulation cell reenter on the opposing side, leading to a natural coordinate representation where all values are restrained to within the PBC boundaries. However, in order to calculate specific physical properties, such as mean square displacements to calculate diffusion coefficients, continuous particle trajectories where the atomic movements are continuous and do not jump between cell faces are required. In these cases, modifying atomic coordinates through unwrapping schemes are an essential post-processing tool to remove these jumps. Here, two established trajectory unwrapping schemes are applied to 1 µs wrapped trajectories for a small water box. The existing schemes can result in spurious diffusion coefficients, long bonds within unwrapped molecules, and inconsistent atomic coordinates when coordinates are rewrapped after unwrapping. We determine that prior unwrapping schemes do not account for changing periodic box dimensions, and introduce an additional correction term to the existing displacement unwrapping scheme by von Bülow et al. to correct for these artifacts. After treatment with this new unwrapping scheme, molecular geometries are correct even after long simulations that have become commonplace. We develop implementations for this new scheme in multiple PBC handling tools, and look forward to its broader adoption.

Introduction

Molecular dynamics simulations are a fantastic tool to study interactions and mechanisms in the nanoscale regime. Applications range broadly from material studies,\(^1\) to disease mechanism\(^5\) and trying to understanding how life itself works.\(^10\) Recent high profile applications include a recent study the complete HIV-1 empty capsid was simulated revealing biological implications based on its physical properties.\(^7\) Similarly, the simulation of SARS-CoV-2’s spike protein provided a unique inside into the role of its glycosylations in evading detection.\(^8\) The trend in the field is towards larger system sizes that cover longer simulation times, with the first reported billion atom biologi-
c al systems having already been simulated. From the simulation of a crowded bacterial cytoplasm, model for whole bacterial cells are under development.

Inherent to the success of molecular dynamics, there are certain common methodological assumptions made in system construction. Long range electrostatics is essential to accurately modeling these systems. The truncation of long range electrostatic interactions with cutoffs leads to wrong structures for charged proteins, DNA deforms and lipid bilayers have an incorrect area per lipid. To calculate long-range electrostatics through particle mesh Ewald, and to mitigate edge effects for finite sized-systems, periodic boundary conditions (PBC) are imposed onto the molecular simulation systems. With PBC the volume can be maintained during constant volume simulations and it allows barostats to adjust the pressure during simulations by adjusting the box volume.

A consequence of simulating with PBC is that atoms move across edges for a periodic cell, exiting across one cell face and reentering the cell from the opposite face. Most molecular dynamics simulation packages keep track of the atomic positions only within the origin cell and output frames to trajectories accordingly. Thus, most analysis is performed on a trajectory where the atomic coordinates are "wrapped" within a minimal unit cell. However, some analysis requires that particle positions account for cell face crossing events, such that atomic coordinates are "unwrapped", as though PBC wrapping had not been applied. A common use case for unwrapped trajectories is to calculate diffusion coefficients from atomic mean square displacements (MSD), or to keep drifting multidomain proteins in contact for visualization or analysis purposes.

Since the advent of periodic boundary conditions in molecular simulation, simulation and analysis packages have determined the unwrapped coordinates at timestep \( i + 1 \) \( \left( x_{i+1}^u \right) \) by tracking the number of periodic unit cell vectors that need to be added to account for crossing events across the unit cell boundary. Concretely, the unwrapped coordinate is calculated every step with the unwrapped coordinates of the previous step \( x_i^u \), the current wrapped position \( x_{i+1}^w \) and periodic unit cell dimensions \( L_{i+1} \):

\[
x_{i+1}^w = x_{i+1}^w - \left( x_{i+1}^w - x_i^w \right) - \left[ \frac{x_i^w - x_{i+1}^u}{L_{i+1}} + \frac{1}{2} \right] L_{i+1}
\]

We will refer to equation 1 as the heuristic method. The heuristic method has some appealing features, such as only requiring the previous unwrapped coordinates to be kept in memory while replacing the wrapped coordinates with the unwrapped coordinates in place. The heuristic method is also invertible, as typical wrapping algorithms will arrive at the exact number of box vectors needed to regenerate the wrapped coordinates. However, von Bülow et al. demonstrated that the heuristic method yields incorrect diffusion constants during microsecond long simulations in NPT ensembles, especially for small boxes. The spurious diffusion coefficients were hypothesized to arise from periodic cell dimension fluctuations due to the applied barostat. When the unwrapped position becomes large relative to the unit cell dimensions, the number of unit cells to shift the wrapped position can vary depending on the instantaneous unit cell dimension at a given timestep. Based on these findings, von Bülow et al. proposed an alternative unwrapping scheme, which we label displacement method in this study:

\[
x_i^u = x_i^w + \left( x_{i+1}^w - x_i^w \right) - \left[ \frac{x_i^w - x_{i+1}^u}{L_{i+1}} + \frac{1}{2} \right] L_{i+1}
\]

Rather than counting the number of unit cell crossing events, the displacement scheme adds the wrapped displacement at every timestep, adjusting the displacement by the unit cell dimension, if the particle would otherwise travel more than half a box dimension between timesteps. As the unwrapped trajectory is determined by a sequence of displacement vectors, and the displacement vectors are typically
smaller than the unit cell length, the displacement scheme is more robust to unit cell fluctuations when in an NPT ensemble.

However, when applied to long simulation trajectories, the displacement scheme as written in Eq. 2 has two noticeable flaws. Under certain circumstances, the displacement scheme can generate long bonds indicative of a broken molecule. Furthermore, the displacement scheme is not reversible with conventional wrapping schemes. In this study, we examine the conditions where Eq. 1 and 2 fail, and propose an alternative unwrapping scheme that corrects Eq. 2 by explicitly accounting for fluctuating unit cell dimensions.

\[
x_{i+1}^u = x_i^u + \left( x_{i+1}^w - x_i^w \right)
- \left[ \frac{x_{i+1}^w - x_i^w}{L_{i+1}} + \frac{1}{2} \right] L_{i+1}
- \left[ \frac{x_i^w - x_i^u}{L_i} + \frac{1}{2} \right] (L_{i+1} - L_i)
\]

(3)

The corrected displacement scheme (Disp-correct) retains the best features from previous unwrapping methods, in that it is invertible by typical wrapping schemes and yields the correct diffusion coefficient without structural artifacts within a molecule.

**Methods**

To thoroughly explore the three different unwrapping schemes, we generate multiple wrapped trajectories through two molecular dynamics engines in both constant volume (NVT) and constant pressure (NPT) ensembles. After unwrapping, we evaluate the trajectories and measure diffusion coefficients and long bonds within all water molecules. To measure the reversibility for the unwrapping procedure, we subsequently also wrap the coordinates once more and compare with the initial coordinates. This provides three metrics with which we can evaluate the different unwrapping schemes for their suitability in general use.

**Molecular dynamics simulations**

510 TIP3P water molecules\(^{24,25}\) were placed in a \(2.5 \times 2.5 \times 2.5 \text{nm}^3\) simulation cell via the `gmx solvate` tool of the GROMACS 5.1 software package.\(^{26}\) The water box was then minimized and equilibrated for 70 ps in a NPT ensemble at 300 K and 1 atm prior to production simulations. The equilibrated water box starting point was simulated in both an NPT and NVT ensemble for 1 \(\mu\)s with wrapped trajectory snapshots saved every 20 fs. The frequent trajectory writes allowed for diffusion analysis to be performed using different lag times. Both ensemble simulations were performed with the software packages GROMACS 2020\(^{26,27}\) and NAMD 3.0a8.\(^{28}\) As an additional point of comparison, NAMD has the feature to write out unwrapped trajectory snapshots, which allowed us to simulate a reference trajectory for the NPT ensemble for our analysis.

In our GROMACS simulations, the equation of motion was integrated every 2 fs by a leap-frog integrator. Distances within water molecules are restrained with the SET-TLE algorithm.\(^{29}\) Intermolecular interactions are considered within 1 nm for Coulomb and van der Waals forces. The low cutoff values allowed the use of small box sizes even with large volume fluctuations in the NPT ensemble. Long-range electrostatic interactions are calculated with the fast smooth particle-mesh Ewald (PME) summation method, using a 0.12 nm grid spacing.\(^{20,21}\) The pair list is generated within 1.1 nm. Pressure and temperature are controlled by a Berendsen barostat\(^{30}\) to 1 atm and modified v-rescale thermostat\(^{31}\) to 300 K, respectively.

NAMD uses the velocity verlet integrator\(^{32}\) to advance atomic position and velocities every 2 fs, using the RATTLE algorithm\(^{29}\) to restrain bond lengths to hydrogen. Interaction pairs are generated every 10 steps between atom distances up to 12 Å. Short-range electrostatic and van der Waals interactions use an identical 10 Å cutoff as our GROMACS trajectories. Interaction forces are scaled down to zero with a switching function starting at 9 Å. Long-range electrostatics are considered with PME using
a grid spacing of 1 Å.\textsuperscript{20,21} A Langevin piston\textsuperscript{22} and Langevin thermostat\textsuperscript{33} were used to adjust the temperature to 300 K and the pressure to 1 atm.

Analysis

The heuristic method had been previously implemented in \texttt{pbctools} 3.0 and \texttt{fastpbc} 1.0, while \texttt{qwrap} 1.4\textsuperscript{34} provided an implementation for the displacement method accessible from within VMD 1.9.4a48.\textsuperscript{35} Our initial implementation for Eq. 3 was developed for \texttt{pbctools} 3.0. We later implemented Eq. 3 into \texttt{fastpbc} and \texttt{qwrap} to utilize faster execution times through gpu and cpu parallelisms. All implementations are provided in github repositories, or have already been integrated into the VMD source tree. Together with supporting libraries such as \texttt{numpy},\textsuperscript{36} \texttt{sklearn},\textsuperscript{37} and \texttt{matplotlib},\textsuperscript{38} we leveraged the VMD python API to gather, store, and visualize bond-length distributions, compute mean squared displacements, and compare coordinates between the original trajectories and the same trajectories after applying both wrapping and unwrapping operations. The mean square displacements were used to determine diffusion coefficients from the Einstein relation:\textsuperscript{39}

\[
D = \frac{\Delta(MSD)}{6\Delta t}
\]

and subsequently using the \texttt{sklearn}\textsuperscript{37} to determine a linear fit between MSD and time. The source code to build these analyses is available on ResearchGate.

Results and Discussion

Diffusion coefficients

The motivation for von Bülow et al.\textsuperscript{23} to propose alternative unwrapping schemes focused on how the heuristic scheme overestimates diffusion coefficients for simulations conducted in NPT ensembles. Analysis for our own independent simulations find similar results (Figure 1, S3-33), where the heuristic scheme can overstate the diffusion coefficient by up to a factor of 60. The unit cell size changes caused by the barostat within a NPT ensemble are clearly the underlying cause for this issue, as the diffusion estimates agree between unwrapping methods for NVT-generated trajectories.

One hypothesis proposed by von Bülow et al.\textsuperscript{23} is that the displacement and heuristic methods would agree once the trajectory frequency analysis rates trend towards zero. From Figure 1, we see that this is true only for GROMACS trajectories, and that instead NAMD trajectories have increased apparent diffusion after unwrapping with less time between frames. This counter-intuitive finding appears to be related to the variation in the unit cell dimensions within a NAMD trajectory, which vary more quickly than what was observed for a GROMACS trajectory (Figure S4). Faster variation in the box dimension would lead to more events where the particle crosses a PBC boundary that moves underneat it.

Bond length distributions

A correctly implemented unwrapping scheme should move the atomic positions only by multiples of the box dimension for a given frame. Consequently, bond lengths between atoms should not change during the unwrapping process, as a molecule should move together as one unit during simulation. To verify correctness for the existing heuristic (Eq. 1) and displacement (Eq. 4) schemes, we calculated the bond lengths over long trajectories (Figure 2), and compared them against our new corrected displacement scheme from Eq. 3. Since our system is entirely composed of water, bond lengths are expected to be uniformly less than 1 Å. However, for both the heuristic and displacement schemes, we observe longer bond lengths on the nanometer length scale, indicative of stretched and unphysical water molecule geometries (Figure 2).

The mechanism for generating the stretched bonds for the heuristic and displacement schemes are unique between the methods. As seen in Figure 2A, the spurious bond lengths for
the heuristic method are primarily quantized in increments of 2.5 nm, the PBC dimensions. Diagonal long bonds are also present, as evidenced by 3.5 and 4.3 nm bond lengths. It has been hypothesized that the discrete bonds are the result of atomic motion beyond half of a unit cell between trajectory frames. However, atoms have been measured to move around 2.5 Å on average between frames for the 2 ps trajectory frequency (Figure S5). While we cannot exclude that atoms moved more then half the box size in our simulations between trajectory writes, long bonds occur too frequently and only after a significant long simulated trajectory, suggesting that an alternative mechanism is at play.

To demonstrate the origin for the quantized heuristic errors, we work through a concrete example observed from our own simulation trajectory. At a given time well into the simulation trajectory, the unit cell dimension is 25.24 Å. Two atoms $a_1$ and $a_2$ on the same molecule are at wrapped $x$-coordinates $x_1^w = 9.12$ Å and $x_2^w = 10.06$ Å, respectively, connected by a standard 1 Å bond. The unwrapped coordinates $x_1^u = 463.43$ Å and $x_2^u = 464.38$ Å are much larger, as the atoms have already traversed 18 simulation boxes. In the next time step, both atoms move backwards and exchange relative orientations to $x_1^w = 6.43$ Å and $x_2^w = 5.46$ Å. Simultaneously, the box length shrinks to 24.76 Å. When calculating the new unwrapped coordinates, the wrapped $x$-coordinate for atom $a_1$ is moved by $\left\lfloor \frac{463.43 - 6.43}{24.76} + \frac{1}{2} \right\rfloor = 18.96 = 18$ PBC dimensions. However, atom $a_2$ is moved by $\left\lfloor \frac{464.38 - 5.46}{24.76} + \frac{1}{2} \right\rfloor = 19.03 = 19$ PBC dimensions. In this demonstration, the molecule is not near a box edge in wrapped coordinates, and so crossing events themselves are not problematic. Instead, the error arises from mixing together coordinates expressed in different unit cell dimensions (Figure 3). More generally, in the limit when $x_i^u$ becomes large relative to $L_{i+1}$, small deviations in the box dimension can change the value for the floor function within Eq. 1 more or less at random. Common barostats scale the atomic positions when the unit cell changes size, and this is not taken into account in Eq. 1. Thus, it appears that in rare circumstances the heuristic method miscounts the number of simulation boxes to move an atom within our trajectories, in agreement with the findings by von Bülow et al. Since the unwrapped coordinates are cumulative with prior unwrapped steps, these errors propagate across time when unwrapping a simulation trajectory. When the box size changes quickly, as is the case for our NAMD trajectories with a fluctuating barostat (Figure S4), the floor func-
Figure 2: (A) Histograms of bond lengths after 1 µs for different trajectory frequencies in NAMD (red) and GROMACS (blue) using the three unwrapping implementations defined by Equations 1-3. Each row within the figure highlights a different combination of simulation engine and trajectory frequency, with the distribution colored according to the legend above. Each bin starts with a value of one to prevent the undefined logarithm of zero. (B) Fraction of bond lengths over 0.95 nm during the simulation.
tion will more frequently split apart molecules (Figure 2B).

Figure 3: Schematic of the heuristic method showing an unwrapping error occurring at step \( k+1 \) after a significant box size reduction. During the simulation atoms move between time steps \( s_{i-1} \) and \( s_i \). Afterwards the positions \( w_i \) are wrapped. In the post-processing analysis the coordinates \( u_i \) are unwrapped.

Unlike the quantized heuristic errors, the bond length distribution for GROMACS trajectories handled with the displacement scheme is effectively continuous (Figure 2A). The displacement method does not miscount unit cell traversals, as these are not actually tracked with the method. Each unwrapping timestep only checks if a box jump event occurred and shifts the coordinates accordingly by at most one box vector. However, the displacement scheme does not account for the coordinate scaling during the unwrapping process, and can under certain circumstances extend or shrink a bond by the difference in the unit cell dimensions. When aggregating the results over many thousands of frames, the small bond length deviations add up to larger errors. Interestingly, this is only a problem for the GROMACS trajectories. GROMACS, unlike NAMD, will split molecules when wrapping trajectories across a periodic boundary. Thus, the displacement vectors for GROMACS within a molecule can vary significantly between atoms, while the NAMD trajectories do not increase the bond length.

To elucidate the displacement error, let us consider another concrete example from our simulation trajectory, again focusing on the behavior of two atoms. After around 40 ps in the simulation, the x coordinates of two bound atoms \( a_1 \) and \( a_2 \) are \( x_1 = 1.43 \text{ Å} \) and \( x_2 = 0.92 \text{ Å} \), respectively. The atoms did not move across the periodic box yet, and therefore the wrapped and unwrapped coordinates are identical. The bond length between the atoms is around 1 Å, with an x-contribution of \(|1.43 - 0.92| \text{ Å} = 0.51 \text{ Å}\). 2 ps later, the two atoms moved backwards, resulting in the jump of \( a_1 \) across the periodic box. The new wrapped coordinates are \( x_{1w} = 24.54 \text{ Å} \) and \( x_2 = 0.27 \text{ Å} \), with a box size of 25.13 Å, implying a bond length along the x-dimension of 0.86 Å. Using the displacement unwrapping scheme from Eq. 2 to unwrap the coordinates of \( a_1 \), the displacement between the current frame and the previous frame is added to the unwrapped coordinates from the previous frame, and the current box dimension is subtracted to account for the periodic jump that particle 1 undertook: \( x_{1w}^u = 1.43 \text{ Å} + (24.54 - 1.43) \text{ Å} - 25.13 \text{ Å} = -0.59 \text{ Å} \). At this point, the bond length along the x-dimension remains 0.86 Å for the unwrapped coordinates \((-0.59 - 0.27) \text{ Å} = 0.86 \text{ Å}\). However, after an additional 2 ps simulation time, the atoms moved again to the wrapped coordinates \( x_{1w} = 24.47 \text{ Å} \) and \( x_2 = 0.23 \text{ Å} \), and the box shrunk to 25.02 Å. Applying again the displacement unwrapping scheme to \( a_1 \) yields \( x_{1w}^u = -0.59 \text{ Å} + (24.47 - 24.54) \text{ Å} = -0.66 \text{ Å} \). Finally, we can compare the x contributions to the bond length again, yielding \(|-0.66 - 0.23| \text{ Å} = 0.89 \text{ Å} \) for the unwrapped and \(|(24.47 - 25.02) - 0.23| \text{ Å} = 0.78 \text{ Å} \), for the wrapped coordinates. The barostat scaling the atomic coordinates when adjusting the box volume causes this discrepancy. The error between the unwrapped and wrapped x coordinate difference, \(|0.89 - 0.78| \text{ Å} = 0.11 \text{ Å} \) is exactly the difference in the box dimension between the current and last frame \( |25.02 - 25.13| \text{ Å} = 0.11 \text{ Å} \). Recognizing this error yields Eq. 3, which was designed to eliminate these changes.
in distance between atoms near periodic boundaries caused by Eq. 2.

**Scheme Reversibility**

**Table 1:** Average differences in atomic positions between unwrapping methods and reference after certain simulation times. The reference positions are obtained from a NAMD trajectory without wrapping. To get coordinates for the different unwrapping methods, the reference trajectory is wrapped and subsequently unwrapped by the respective method.

| Time (fs) | RMSD ref. to unwrap. method |
|----------|----------------------------|
| 0 fs     | 0.e+00  0.e+00  0.e+00 |
| 0.2 ps   | 0.e+00  0.e+00  5.73e-08 |
| 1 ps     | 0.e+00  3.65e-02 7.70e-08 |
| 10 ps    | 0.e+00  4.27e-02 2.12e-07 |
| 0.1 ns   | 0.e+00  1.45e-01 9.73e-07 |
| 1 ns     | 0.e+00  5.70e-01 3.47e-06 |
| 10 ns    | 0.e+00  1.9e+00 7.53e-06 |
| 0.1 µs   | 4.95e+00 5.8e+00 1.34e-05 |
| 1 µs     | 1.84e+03 1.8e+01 2.16e-05 |

The application of wrapping and unwrapping should be a reversible operation that might be applied repeatably. To test the reversibility for the schemes introduced by Eqs. 1-3, the NAMD unwrapped NPT reference trajectory was wrapped followed by unwrapping with the respective method. Afterwards, the RMSD between the original and newly unwrapped coordinates were compared (Table 1). Both the heuristic and displacement scheme are observed to not preserve atomic positions, with significant RMSDs for the displacement method occurring after 0.1 ns, and the heuristic method failing after 100 ns when the atoms travel far from the origin. By contrast, the correction term preserves atomic positions to within the limit of single precision floating point arithmetic, with the increase in RMSD corresponding to the larger and larger atomic positions that reduce accuracy.

The key advance that enables the success for Equation 3 is the correction term that takes into account the scaling of atomic positions after a change in the box volume. The traversed boxes count is multiplied with the box size change between the current $L_{i+1}$ and last $L_i$ iteration, and subtracted from the atom positions as defined in Equation 3. The additional term corrects the diffusion overestimate for the heuristic scheme (Equation 1) observed in Figure 1. As seen in Figure 2, the bond lengths when applying Equation 3 are uniform, and their 0.1 nm length conforms to our expectation for water structure.

Table 1 also suggests that the correct choice for unwrapping schemes depends on the context. The heuristic scheme’s shortcomings only become readily apparent in a long simulation where the box size varies. Unwrapping NVT simulations would be expected to yield correct results under any scheme, so long as trajectory frames are taken frequently enough so that PBC crossing events are not missed. The lower memory footprint for the heuristic scheme may be beneficial in some instances. However, when unwrapping a trajectory with changing box dimensions, as is typically the case for simulations in a constant pressure ensemble, the heuristic scheme will eventually fail. In these instances, the corrected displacement scheme from Equation 3 is essential to capturing the correct molecular geometries and diffusive behavior. To facilitate wide adoption for the new scheme, we have implemented Equation 3 into the VMD packages pbctools, fastpbc and qwrap.

**Conclusion**

In this study, we have demonstrated that the current implementations for the heuristic and displacement unwrapping methods lead to inaccurate unwrapped atomic positions through three different criteria; measured diffusion coefficients, bond lengths, and reversibility. The heuristic case (Equation 1) struggles if the simulated time scale leads to atomic positions that would be much larger than the periodic box size, leading to uncertainties for how many periodic boxes a molecule needs to be moved when the PBC dimensions change. The displacement
method (Equation 2) leads to changing internal geometries as the box dimensions change. We therefore proposed a correction to the displacement scheme Equation 3 that faithfully reproduces molecular geometries over time, and recommend its adoption when analyzing future simulations.

However, past conclusions from these simulations are unlikely to materially change, irrespective for the wrapping algorithm. Prior unwrapping schemes fail most obviously for pathological systems where a small fluctuating unit cell is simulated for an extended period of time. Only in a handful of studies likely meet these criteria, as diffusion coefficient estimates converge quickly enough that extended simulations are not required. Likewise, the trend towards larger simulation systems increases the timescales over which the heuristic approximation from Equation 1 is valid. However, with the increasing availability for graphical processing units or special purpose machines such as Anton 3 to deliver longer and longer trajectories, we foresee a future where box dimension changes must explicitly be accounted during trajectory analysis.

Acknowledgements

The authors thank Michigan State University (MSU) for providing startup funds. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 (project number TG-BIO210040). This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

Disclosure

The authors report no conflicts of interest in this work.

References

(1) Yang, K.; Xu, X.; Yang, B.; Cook, B.; Ramos, H.; Krishnan, N. M. A.; Smedskjaer, M. M.; Hoover, C.; Bauchy, M. Predicting the Young’s Modulus of Silicate Glasses using High-Throughput Molecular Dynamics Simulations and Machine Learning. Sci. Rep. 2019, 9, 8739, DOI: 10.1038/s41598-019-45344-3.

(2) Zepeda-Ruiz, L. A.; Stukowski, A.; Oppestrup, T.; Bulatov, V. V. Probing the limits of metal plasticity with molecular dynamics simulations. Nature 2017, 550, 492–495, DOI: 10.1038/nature23472.

(3) Lau, D.; Jian, W.; Yu, Z.; Hui, D. Nano-engineering of construction materials using molecular dynamics simulations: Prospects and challenges. Compos. Part B Eng. 2018, 143, 282–291, DOI: 10.1016/j.compositesb.2018.01.014.

(4) Massobrio, C.; Du, J.; Bernasconi, M.; Salmon, P. S. In Molecular Dynamics Simulations of Disordered Materials; Massobrio, C., Du, J., Bernasconi, M., Salmon, P. S., Eds.; Springer Series in Materials Science; Springer International Publishing: Cham, 2015; Vol. 215; DOI: 10.1007/978-3-319-15675-0.

(5) Arantes, P. R.; Saha, A.; Palermo, G. Fighting COVID-19 Using Molecular Dynamics Simulations. ACS Cent. Sci. 2020, 6, 1654–1656, DOI: 10.1021/acscentsci.0c01236.

(6) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. Neuron 2018, 99, 1129–1143, DOI: 10.1016/j.neuron.2018.08.011.

(7) Perilla, J. R.; Schulten, K. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. Nat. Commun. 2017, 8, 15959, DOI: 10.1038/ncomms15959.
(8) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S. et al. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* 2020, *6*, 1722–1734, DOI: 10.1021/acscentsci.0c01056.

(9) Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Ieong, P. U.; Dommer, A. C.; Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent. Sci.* 2020, *6*, 189–196, DOI: 10.1021/acscentsci.9b01071.

(10) Feig, M.; Yu, I.; Wang, P.-h.; Nawrocki, G.; Sugita, Y. Crowding in Cellular Environments at an Atomistic Level from Computer Simulations. *J. Phys. Chem. B* 2017, *121*, 8009–8025, DOI: 10.1021/acs.jpcb.7b03570.

(11) Bock, L. V.; Kolář, M. H.; Grubmüller, H. Molecular simulations of the ribosome and associated translation factors. *Curr. Opin. Struct. Biol.* 2018, *49*, 27–35, DOI: 10.1016/j.sbi.2017.11.003.

(12) Nawrocki, G.; Im, W.; Sugita, Y.; Feig, M. Clustering and dynamics of crowded proteins near membranes and their influence on membrane bending. *Proc. Natl. Acad. Sci.* 2019, *116*, 24562–24567, DOI: 10.1073/pnas.1910771116.

(13) Kubo, S.; Niina, T.; Takada, S. Molecular dynamics simulation of proton-transfer coupled rotations in ATP synthase FO motor. *Sci. Rep.* 2020, *10*, 8225, DOI: 10.1038/s41598-020-65004-1.

(14) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *Elife* 2016, *5*, DOI: 10.7554/eLife.19274.

(15) Maritan, M.; Autin, L.; Karr, J.; Covert, M. W.; Olson, A. J.; Goodsell, D. S. Building Structural Models of a Whole Mycoplasma Cell. *J. Mol. Biol.* 2021, 167351, DOI: 10.1016/j.jmb.2021.167351.

(16) Saito, M. Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. *J. Chem. Phys.* 1994, *101*, 4055–4061, DOI: 10.1063/1.468411.

(17) Loncharich, R. J.; Brooks, B. R. The effects of truncating long-range forces on protein dynamics. *Proteins Struct. Funct. Genet.* 1989, *6*, 32–45, DOI: 10.1002/prot.340060104.

(18) Norberg, J.; Nilsson, L. On the Truncation of Long-Range Electrostatic Interactions in DNA. *Biophys. J.* 2000, *79*, 1537–1553, DOI: 10.1016/S0006-3495(00)76405-8.

(19) Patra, M.; Karttunen, M.; Hyvönen, M. T.; Falck, E.; Lindqvist, P.; Vattulainen, I. Molecular Dynamics Simulations of Lipid Bilayers: Major Artifacts Due to Truncating Electrostatic Interactions. *Biophys. J.* 2003, *84*, 3636–3645, DOI: 10.1016/S0006-3495(03)75094-2.

(20) Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* 1921, *369*, 253–287, DOI: 10.1002/andp.19213690304.

(21) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* 1995, *103*, 8577–8593, DOI: 10.1063/1.470117.

(22) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* 1995, *103*, 4613–4621, DOI: 10.1063/1.470648.

(23) von Bülow, S.; Bullerjahn, J. T.; Hummer, G. Systematic errors in diffusion coefficients from long-time molecular dynam-
ics simulations at constant pressure. J. Chem. Phys. 2020, 153, 021101, DOI: 10.1063/5.0008316.

(24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 1983, 79, 926–935, DOI: 10.1063/1.445869.

(25) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. J. Chem. Phys. 1996, 105, 1902–1921.

(26) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015, 1-2, 19–25, DOI: 10.1016/j.softx.2015.06.001.

(27) Páll, S.; Zhmurov, A.; Bauer, P.; Abraham, M.; Lundborg, M.; Gray, A.; Hess, B.; Lindahl, E. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J. Chem. Phys. 2020, 153, 134110, DOI: 10.1063/5.0018516.

(28) Phillips, J. C.; Hardy, D. J.; Maia, J. D.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 2020, 153, 44130, DOI: 10.1063/5.0014475.

(29) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J. Comput. Chem. 1992, 13, 952–962, DOI: 10.1002/jcc.540130805.

(30) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 1984, 81, 3684–3690, DOI: 10.1063/1.448118.

(31) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. J. Chem. Phys. 2007, 126, 014101, DOI: 10.1063/1.2408420.

(32) Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids. Liq. Oxford Univ. Press. New York 1987, 18, 385, DOI: 10.2307/2938686.

(33) Grest, G. S.; Kremer, K. Molecular dynamics simulation for polymers in the presence of a heat bath. Phys. Rev. A 1986, 33, 3628–3631, DOI: 10.1103/PhysRevA.33.3628.

(34) Hénin, J.; Fiorin, G. jhenin/qwrap: Fast and accurate wrapping and unwrapping of MD trajectories. 2020; doi.org/10.5281/zenodo.3877030.

(35) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. J. Mol. Graph. 1996, 14, 33–38, DOI: 10.1016/0263-7855(96)00018-5.

(36) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J. et al. Array programming with NumPy. Nature 2020, 585, 357–362, DOI: 10.1038/s41586-020-2649-2.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830, DOI: 10.5555/1953048.2078195.

(38) Hunter, J. D. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 2007, 9, 90–95, DOI: 10.1109/MCSE.2007.55.

(39) Einstein, A. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann. Phys. 1905, 322, 549–560, DOI: 10.1002/andp.19053220806.
Supplementary Information

This PDF file includes

- Figures S1 to S33

Other supplementary information

- Dataset for the simulation on ResearchGate. Contains simulation input, topology and coordinate files to rerun the simulations, as well as python and tcl scripts to do the analysis and generate the images in this publication.

- pbctools implementation is on github.

- qwrap implementation is also on github.

- fastpbc implementation, which has already been integrated into the VMD source tree.
Figure S4: (Top) Isotropic box length fluctuations during the NPT simulations for NAMD (left) and GROMACS (right) over the simulation time. (Bottom) Histograms for the box length change after 2 ps intervals. The box volume fluctuates significantly more during the NAMD simulations.
Figure S5: Histogram of the displacements after 2 ps intervals in the NAMD NPT reference simulation with unwrapped snapshots. The average displacement is around 0.25 nm, although rarely much higher displacements are possible. With increasing simulation time the molecules diffuse further away from the original simulation box. Due to the coordinate scaling of the barostat, the histograms get distorted with increasing simulation time, as molecules further away from the original simulation box are more affected by this scaling (compare left and right histogram). We cannot exclude that atoms moved more than half a box length during our simulation, resulting in an unwrapping error, but we assume that these events happen far less frequent than the other errors discussed in this manuscript.
Figure S6: Mean square displacements for the NPT simulation in GROMACS with the displacement method and a trajectory frequency of 2 ps.
Figure S7: Mean square displacements for the NPT simulation in GROMACS with the displacement method and a trajectory frequency of 1 ps.
Figure S8: Mean square displacements for the NPT simulation in GROMACS with the displacement method and a trajectory frequency of 0.5 ps.
Figure S9: Mean square displacements for the NPT simulation in GROMACS with the displacement method and a trajectory frequency of 0.2 ps.
Figure S10: Mean square displacements for the NVT simulation in GROMACS with the displacement method and a trajectory frequency of 0.2 ps.
Figure S11: Mean square displacements for the NPT simulation in GROMACS with the heuristic method and a trajectory frequency of 2 ps.
Figure S12: Mean square displacements for the NPT simulation in GROMACS with the heuristic method and a trajectory frequency of 1 ps.
Figure S13: Mean square displacements for the NPT simulation in GROMACS with the heuristic method and a trajectory frequency of 0.5 ps.
Figure S14: Mean square displacements for the NPT simulation in GROMACS with the heuristic method and a trajectory frequency of 0.2 ps.
Figure S15: Mean square displacements for the NVT simulation in GROMACS with the heuristic method and a trajectory frequency of 0.2 ps.
Figure S16: Mean square displacements for the NPT simulation in GROMACS with the corrected displacement method and a trajectory frequency of 2 ps.
Figure S17: Mean square displacements for the NPT simulation in GROMACS with the corrected displacement method and a trajectory frequency of 1 ps.
Figure S18: Mean square displacements for the NPT simulation in GROMACS with the corrected displacement method and a trajectory frequency of 0.5 ps.
Figure S19: Mean square displacements for the NPT simulation in GROMACS with the corrected displacement method and a trajectory frequency of 0.2 ps.
Figure S20: Mean square displacements for the NVT simulation in GROMACS with the corrected displacement method and a trajectory frequency of 0.2 ps.
Figure S21: Mean square displacements for the NPT simulation in NAMD with the displacement method and a trajectory frequency of 2 ps.
Figure S22: Mean square displacements for the NPT simulation in NAMD with the displacement method and a trajectory frequency of 1 ps.
Figure S23: Mean square displacements for the NPT simulation in NAMD with the displacement method and a trajectory frequency of 0.5 ps.
Figure S24: Mean square displacements for the NPT simulation in NAMD with the displacement method and a trajectory frequency of 0.2 ps.
Figure S25: Mean square displacements for the NVT simulation in NAMD with the displacement method and a trajectory frequency of 0.2 ps.
Figure S26: Mean square displacements for the NPT simulation in NAMD with the heuristic method and a trajectory frequency of 2 ps.
Figure S27: Mean square displacements for the NPT simulation in NAMD with the heuristic method and a trajectory frequency of 1 ps.
Figure S28: Mean square displacements for the NPT simulation in NAMD with the heuristic method and a trajectory frequency of 0.5 ps.
Figure S29: Mean square displacements for the NPT simulation in NAMD with the heuristic method and a trajectory frequency of 0.2 ps.
Figure S30: Mean square displacements for the NVT simulation in NAMD with the heuristic method and a trajectory frequency of 0.2 ps.
Figure S31: Mean square displacements for the NPT simulation in NAMD with the corrected displacement method and a trajectory frequency of 2 ps.
Figure S32: Mean square displacements for the NPT simulation in NAMD with the corrected displacement method and a trajectory frequency of 1 ps.
Figure S33: Mean square displacements for the NPT simulation in NAMD with the corrected displacement method and a trajectory frequency of 0.5 ps.
Figure S34: Mean square displacements for the NPT simulation in NAMD with the corrected displacement method and a trajectory frequency of 0.2 ps.
Figure S35: Mean square displacements for the NVT simulation in NAMD with the corrected displacement method and a trajectory frequency of 0.2 ps.
Figure S36: Mean square displacements for the reference simulation in NAMD with a trajectory frequency of 0.2 ps.
Graphical TOC Entry