ART-SM: Boosting Fragment-based Backmapping by Machine Learning

Christian Pfaendner,*,†,‡ Viktoria Korn,† Pritom Gogoi,† Benjamin Unger,†,‡ and Kristyna Pluhackova*,†,‡

†*Stuttgart Center for Simulation Science, Cluster of Excellence EXC 2075, University of Stuttgart, Universitätsstr. 32, 70569, Stuttgart, Germany*
‡*Artificial Intelligence Software Academy, University of Stuttgart, Stuttgart, Germany*

E-mail: christian.pfaendner@simtech.uni-stuttgart.de;
kristyna.pluhackova@simtech.uni-stuttgart.de

Abstract

In sequential multiscale molecular dynamics simulations, which advantageously combine the increased sampling and dynamics at coarse-grained resolution with the higher accuracy of atomistic simulations, the resolution is altered over time. While coarse-graining is straightforward, the reintroduction of the atomistic detail is a non-trivial process called backmapping. Here, we present ART-SM, a fragment-based machine learning backmapping framework that learns the Boltzmann distribution from atomistic data to switch from coarse-grained to atomistic resolution seamlessly. ART-SM requires minimal user input and goes beyond state-of-the-art fragment-based approaches by selecting from multiple conformations per fragment to simultaneously reflect the coarse-grained structure and the Boltzmann distribution. Additionally, we introduce a novel refinement step to connect individual fragments via optimization of specific bonds, angles, and dihedral angles in the backmapping process. We demonstrate that our algorithm accurately restores the atomistic bond length, angle, and dihedral angle distributions for various small molecules of up to three Martini coarse-grained beads and that the resulting high-resolution structures are representative of the original coarse-grained conformations. Moreover, the reconstruction of the TIP3P water model is fast and robust, and we illustrate that ART-SM can be, in principle, applied to larger molecules as well, indicating its potential extension to more complex molecules like lipids, proteins, and macromolecules in the future.

1 Introduction

In recent years, molecular dynamics (MD) simulations have gained popularity, especially in the fields of molecular biology,[1–8] chemical physics,[9,10] and materials science.[11,12] Between 2007 and 2017, the annual number of structural biology publications involving MD more than doubled, which can be attributed, among other factors, to the improved accuracy and performance of MD software and a greater number of experimentally determined protein structures,[13] in particular, cryoEM structures.[14] The unmatched spatial (sub-angstroms) and temporal resolution (femtoseconds) of MD simulations over the entire simulation period, which cannot be achieved through experimental methods, renders MD simulations a perfect complement to experiments.[1,6,15,16] Nevertheless, due to the substantial computational resources required, atomistic MD simulations, even though already neglecting quantum effects, are still limited to the microsecond and nanometer scales. Consequently, scientists often have to drastically

simplify biological systems or forego the study of large molecular complexes entirely.

Coarse-grained (CG) simulations are a popular strategy to mitigate the computational limitations by using a low-dimensional representation of the atomistic systems while preserving as much structural information as possible.[17] Even though different CG force fields with various granularity have been developed,[18] the most commonly used for biomolecular systems is Martini.[19–22] Nowadays, martinize2 and Vermouth[23] can be used for instance via the recently released Martini Database server[24] to automatically create the Martini CG representation of an atomistic structure, which is essential for large scale studies and high-throughput investigations.[25–29] Although CG simulations increase sampling and dynamics, they often lack the required accuracy and level of detail to properly analyze specific atomistic interactions like hydrogen bonds, which are key features in protein-protein and protein-lipid complexes.[3,6] Also, CG models are unable to describe individual water molecules and thus fail in the description of water-mediated hydrogen bonds[30] and in simulations of water passage in a single-file manner, e.g. through aquaporins.[15,31,32]

Sequential multiscale MD simulations combine the strengths of both approaches by initially simulating in CG resolution until an equilibrium or research-relevant conformation is reached.[33] This is followed by a process called backmapping, where the structure is reverse transformed to atomistic detail, enabling the continuation of the simulation at high resolution.[4,16,28,29] While coarse-graining is straightforward, accurate backmapping is a challenging task due to the loss of information about the atomistic structure underlying the CG representation, which must be reintroduced. Consequently, backmapping is a probabilistic process, and multiple valid solutions typically exist for the same CG structure.

State-of-the-art backmapping algorithms can roughly be categorized into geometric,[34–37] fragment-based,[38,39] and recently emerging machine learning (ML)[40–45] approaches. Geometry-based methods build an initial atomistic structure from a set of rules and aim to recover the Boltzmann distribution with a subsequent relaxation step, which typically consists of energy minimizations (EMs) and short MD simulations. Backward[34] is one of the most widely used backmapping algorithms and exactly follows this protocol. First, it places atoms at the weighted average of the corresponding CG beads, modifies the coordinates of specific atoms based on manually prepared mapping files to ensure correct stereochemistry and to avoid atom overlaps, and corrects the peptide bonds of the protein backbone. Afterward, the algorithm carries out two rounds of EM followed by four rounds of position restraint MD to further refine the structure and restore the Boltzmann distribution. While Backward is a highly adaptable and versatile method that generates accurate results for a broad spectrum of molecule types, developing new mapping files requires extensive expertise. The relaxation step can take a considerable amount of time for larger systems, and molecules can get trapped in local minima conformations. Moreover, the python2-based Backward struggles to accurately recover β-sheets and is restricted to GROMACS versions older than 2020, due to its dependence on the group cut-off scheme.

Another well-known geometric method was developed by Rzepiela *et al.*[35] that initially places the respective atoms randomly in a sphere with a radius of 3 Å around the CG beads and performs a simulated annealing (SA) protocol to obtain the final structure. First, the simulation begins at a high temperature to overcome energy barriers, and harmonic potentials hold the atoms in close proximity to the CG beads. Afterward, the temperature is gradually reduced to a biologically meaningful value, and the structure is relaxed by successively removing the potentials around the CG beads. SA is a flexible and straightforward approach directly integrated into the GROMACS software package, making it easy to use. Despite its advantages, a study by Wassenaar and colleagues[34] demonstrated that Backward outperforms SA in recovering molecular stereochemistry and protein secondary structures on the examples of YvoA, Aquaporin-1, and ASIC-1a.

Fragment-based approaches construct the initial atomistic structure by assembling individual fragments retrieved from a database rather than relying on geometric rules. This database can either be predefined, as in the case of CG2AT2,[39] or directly extracted from a reference structure.[38] Ideally, each fragment represents a minimum energy structure to obtain a reliable approximation of the overall energy minimum of every molecule and to reduce the runtime of the subsequent relaxation step compared to geometric algorithms. However, only one rigid conformation per fragment is usually available in the database. This one-size-fits-all approach may lead to structures that inadequately represent the CG conformation, get trapped in local minima during the subsequent relaxation step, and fail to reflect the underlying Boltzmann distribution accurately. Additionally, creating a comprehensive database encompassing the prevalent biomolecules and incorporating diverse mapping strategies is also challenging.

Apart from the conventional geometric and fragment-based approaches, ML methods have gained more interest in recent years and were successfully applied in backmapping.[40–45] When employing ML methods, it is crucial to avoid strictly modeling the probability density of the target as a unimodal function. Otherwise, the model may predict averaged structures as shown by An *et al.*[40] in their study on hexane, which has multiple valid atomistic conformations given the same CG structure. In general, particle coordinates can be either directly utilized as input features[40,42] or transformed to internal coordinates or voxels[41,45] first. The latter enables direct adaptation of generative adversarial networks (GANs) from the field of image generation and recognition to generate the atomistic structures conditioned on the CG conformations. For instance, Stieffenhofer *et al.* showed that their DeepBM framework successfully recovers the conformational space and the distributions of angles, dihedral angles, and Lennard-Jones energies for syndiotactic polystyrene.[45] Although ML frameworks have shown promising initial results, their application has been mainly limited to specific problems and types of molecules, making it challenging to compare their performance and choose the most suitable method for a particular use case. Furthermore, they have to be separately trained for each system, thus always requiring a matched set of atomistic and CG data. For large protein systems, this is computationally expensive and may even render the sequential multiscaling MD obsolete.

With the aforementioned advantages and drawbacks of state-of-the-art methods in mind, we developed **A**rtificial intelligence-based **R**everse **T**ransformation of **S**mall **M**olecules (ART-SM), a fragment-based ML framework that directly learns the Boltzmann distribution from atomistic simulations. Our method goes beyond traditional fragment-based approaches that employ only one rigid structure per fragment, as ART-SM identifies all main conformations from atomistic data and selects the most appropriate one based on the CG structure. Moreover, ART-SM optimizes bond lengths, angles, and the dihedral angle to connect individual fragments, rather than entirely relying on subsequent MD simulations to restore the Boltzmann distribution. Our algorithm relies on SMILES[46] to ensure correct stereochemistry. SMILES facilitates the ease and efficiency of comparing structures and their respective fragments by concisely representing a molecule's 3D structure in a single string The convenient installation via Python's package manager `pip`, applicability to any atomistic force field, the option to add new fragments to the database, and the possibility to automatically generate mapping files make ART-SM easy to use and avoid extensive user input. Here, we provide a proof-of-principle of our algorithm by backmapping small organic molecules of up to three CG beads from MARTINI to CHARMM36 and show that it has the potential to be extended to larger molecules in the future.

2 Methods

2.1 Prerequisites

2.1.1 Fragment Pair Definition

Given an atomistic molecule A, a fragment F is a substructure of A that is represented by a single bead in the corresponding coarse-graining scheme (see Figure 1). Let $\mathcal{F}_A$ denote the set of unique fragments derived from A, where fragments are considered identical if they can be represented by the same SMILES string. Then, a fragment pair P is a triple (F_1, F_2, C), where $F_1, F_2 \in \mathcal{F}_A$ and C is a connector linking F_1 and F_2. The connector C is a straight chain of four atoms (a_1, a_2, a_3, a_4), with $a_1, a_2 \in F_1$ and $a_3, a_4 \in F_2$. Analogously, fragments and fragment pairs can be derived from a set of molecules $\mathcal{A}$, instead of a single molecule A. The resulting sets are termed $\mathcal{F}$ and $\mathcal{P}$.

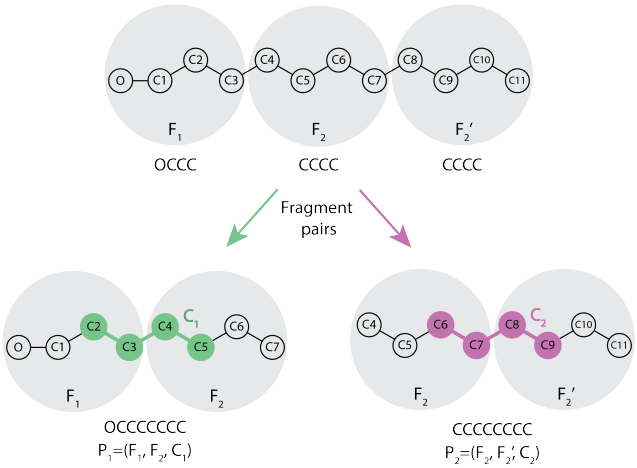

Figure 1: Fragmentation of molecules. The first row depicts the molecule undecan-1-ol along with a typical coarse-graining scheme for Martini (represented by grey spheres). The resulting fragments are denoted F_1, F_2, and F_2', whereby the latter are considered equal due to their identical SMILES representation CCCC. In the second row the corresponding fragment pairs P_1 and P_2 are displayed together with the respective connectors C_1 and C_2 (highlighted in green and magenta, respectively).

2.1.2 Main Conformations and Boltzmann Distribution

The Boltzmann distribution describes the probability p_i of a molecule, or a substructure of it, to be in conformation i. It is given by

$$p_i = \frac{\exp\left(-\frac{E_i}{\mathrm{k}T}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{E_j}{\mathrm{k}T}\right)}, \tag{1}$$

where E_i represents the energy of conformation i, n is the total number of conformations, T is the temperature, and k is the Boltzmann constant.[47] The energy E_i depends on various factors, including bond orders, dihedral angles, and interactions with the environment, resulting in a complex energy surface. Despite its complexity, the energy landscape often exhibits multiple local minima, which, according to the Boltzmann distribution, indicates that certain conformations are significantly more likely than others. We refer to these highly probable structures as main conformations (see Figure 2 and Supporting Information Figure S2). As the size of the molecules increases, it becomes progressively challenging to determine their main conformations based on atomistic data. We adopt a local approach to overcome this challenge by identifying the main conformations of F_1, F_2, and C for each fragment pair P separately, thereby approximating the overall main conformations.

2.2 Workflow of ART-SM

An overview of the ART-SM workflow is depicted in Figure 3. In the initial step, all fragment pairs are identified from a mapping file (box 0). Next, a database of fragment pairs is created from atomistic simulations. To this end, the atomistic data is preprocessed (arrow 1), hydrogens are removed from the atomistic structures (box 2), and for each fragment pair P, the main conformations of F_1, F_2, and C are determined (box 3). Subsequently, a random forest regressor (RFR) is trained for each P to learn the probabilities of each main conformation dependent on the center of mass (COM) distance between fragments (box 4). Bond lengths and angles are extracted from the atomistic data

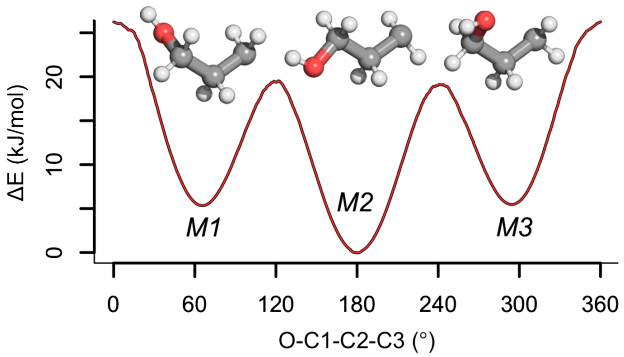

Figure 2: Relation between Boltzmann distribution and main conformations. The dependency of the change in the free energy E on the dihedral angle O-C1-C2-C3 is shown for the fragment F_1 of undecanol (see Figure 1). The three energy minima correspond to the most likely structures according to the Boltzmann distribution (see Supporting Information Figure S2) and are termed main conformations $M1$, $M2$, and $M3$. For each main conformation the respective structure is depicted at each energy minimum.

(arrow 5) to complete the database generation. This fragment database is used to recover atomistic details from a given CG structure. Thereby, ART-SM predicts the most suitable fragment conformations for the current CG structure using the previously trained RFRs (box 6) and optimizes bond lengths, angles, and dihedral angles to connect the individual fragments (box 7). In the last step, the hydrogen atoms are reintroduced (box 8), and optionally the all-atom system undergoes EM and short equilibration (arrow 9). Detailed descriptions of each step are provided in the following subsections.

2.2.1 Identification of Fragment Pairs

Box 0: Mapping files unambiguously define the 3D structure of atomistic molecules and their mappings to CG resolution. To this end, the atom connectivity is specified through bond lists, stereochemistry via SMILES, and fragmentation via matching sets of CG and atomistic particles (see Supporting Information Section S1 for an example). Mapping files are used to schematically split the specified molecules according to the provided coarse-graining model, resulting in a set of fragments $\mathcal{F}$. Subsequently, the set of fragment pairs $\mathcal{P}$ is derived from $\mathcal{F}$ by the provided bond lists of each molecule. Manually defining mapping files can be a tedious and error-prone task. Therefore, ART-SM can automatically generate mapping files given congruent atomistic and CG structures and a SMILES representation for each molecule.

2.2.2 Database Construction

Arrow 1: The database construction process is based on atomistic simulation data. Naturally, the conformations of molecules in subsequent time steps are dependent on each other. Therefore, snapshots are by default extracted every 500 ps to obtain independent training data. This time between the extracted snapshots is called sampling time in the subsequent sections, and a suitable value for our simulation systems is determined in Section 3.1.

Box 2: In this step, hydrogen atoms are removed from the atomistic structures since they provide negligible information on the overall conformations of the molecules.

Box 3: For each $P = (F_1, F_2, C) \in \mathcal{P}$, the main conformations of the individual fragments F_1 and F_2 and the connector C are determined via hierarchical clustering from the atomistic snapshots extracted in arrow 1. For brevity, this process is exclusively described for fragment F_1. Nevertheless, it applies analogously to F_2 and C. First, internal coordinates of heavy atoms are derived for F_1, which are preferable to cartesian coordinates due to their invariance to translation and rotation. Internal coordinates usually encompass bond lengths, angles, and dihedral angles. However, bond lengths and angle distributions are comparable across various conformations and thus ineffective in distinguishing different main conformations. Hence, only dihedral angles are extracted from the atomistic data. This approach results in an $m \times n$ data matrix D_{F_1}, assuming that the fragment pair P occurs m times in the atomistic data and the fragment F_1 can be described by n dihedral angles. Typically, $n \leq 2$ for the Martini force field. Note that $n = 1$ for the connector since it always consists of four atoms, allowing it to be

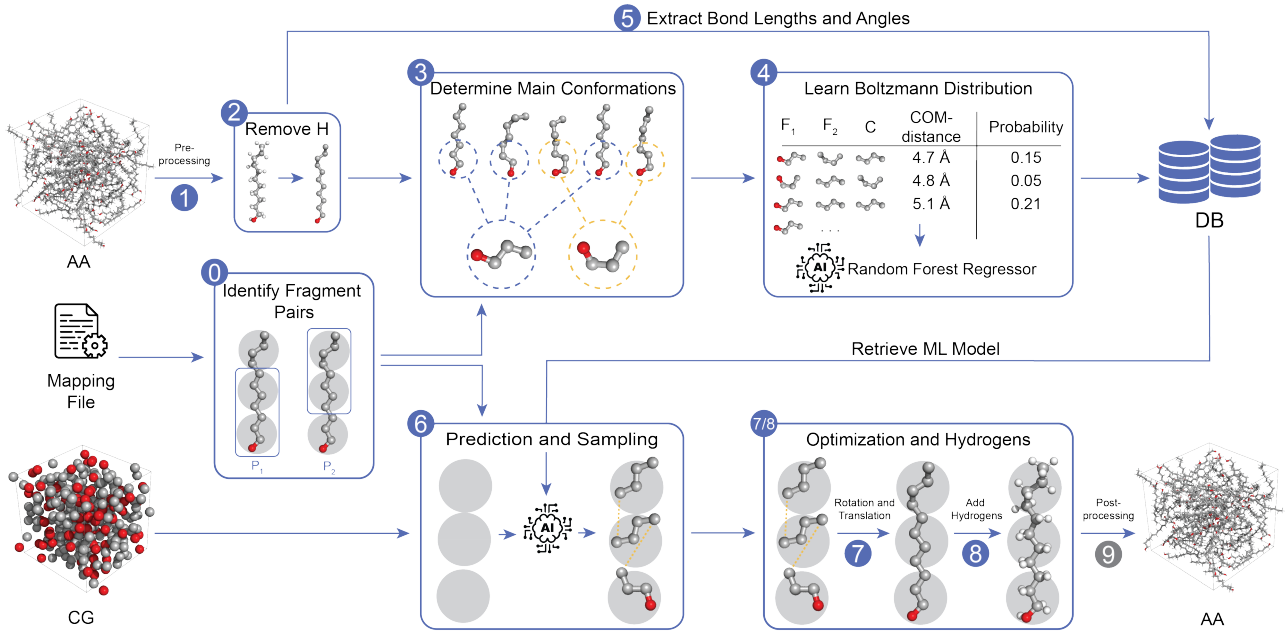

Figure 3: Workflow of ART-SM. The database construction process comprises steps 0 to 5 and the backmapping algorithm steps 1, and 6 to 8. Optionally, the backmapped structures are relaxed by energy minimization and short molecular dynamics simulation in step 9.

represented by a single dihedral angle.

Subsequently, agglomerative hierarchical clustering[48] is performed by default on 500 data points of D_{F_1} (see Section 3.1 for details on the optimal number of data points), using the average linkage criterion and an adjusted Euclidean distance, which additionally accounts for the periodicity of angles (see Supporting Information Section S2.1). Thereby, the clusters are identified automatically, and the medoid is chosen as the representative conformation for each cluster. These representative structures are called main conformations and are stored in the database. Note that different main conformations are calculated for each $P \in \mathcal{P}$, even if they contain identical fragments. For example, given $P_1 = (F_1, F_2, C_1)$ and $P_2 = (F_1, F_3, C_2)$ with $P_1, P_2 \in \mathcal{P}$, main conformations are determined for $F_1 \in P_1$ and $F_1 \in P_2$ separately. This distinction is necessary because the main conformations of fragments can vary depending on the particular fragment they are connected to.

Box 4: For each fragment pair $P \in \mathcal{P}$, an RFR is trained to predict the probabilities Y of its conformations. The predictors X consist of four variables: three categorical variables that indicate the main conformations of F_1, F_2, and C, and a numerical variable for the COM distances between F_1 and F_2. To obtain Y, the probability of each observation in X is determined. This involves identifying the number of observations with identical features and normalizing them by the total number of observations in X. Since the COM distance is a numerical and continuous variable, it is uniformly binned into 50 intervals first. Finally, a RFR from the `scikit-learn` package[49] is trained with default parameters on the dataset X with corresponding labels Y and stored in the database. To estimate the test error of the models, 10-fold cross-validation is performed and the errors are provided to the user via log files.

Arrow 5: Values for bond lengths and angles are required later in the optimization step of the backmapping algorithm (see the forthcoming Subsection 2.2.3). Instead of retrieving them from force field parameters or general databases, they are extracted directly from the atomistic simulations. For this purpose, up to 1000 data points are collected for each bond or angle type and the respective mean values are stored in the database.

2.2.3 Backmapping

The backmapping algorithm converts a CG structure to atomistic resolution given a corresponding mapping file (identical to the mapping file in the database construction step) and a fragment database constructed as shown in Section 2.2.2.

Box 6: For each molecule, the conformations of its fragments and their connectors are predicted in a stepwise manner rather than backmapping the entire molecule at once (see Figure 4 steps 1 and 2). Starting at a random boundary fragment pair $P = (F_1, F_2, C) \in \mathcal{P}$, i.e., F_1 is only connected to F_2 and to no other fragment, the corresponding regression model predicts the probabilities for all possible combinations of main conformations of F_1, F_2, and C given the distance between the CG bead (corresponds to the COM distance between F_1 and F_2 in the database construction process). For example, assume that the main conformations of F_1, F_2, and C are (f_{11}, f_{12}), (f_{21}), and (c_1, c_2, c_3) and the current bead distance is 4.8 Å. Then the probabilities are predicted for $(f_{11}, f_{21}, c_1, 4.8\,\text{Å})$, $(f_{12}, f_{21}, c_1, 4.8\,\text{Å})$, ..., $(f_{12}, f_{21}, c_3, 4.8\,\text{Å})$. The resulting probabilities $(p_1, p_2, \ldots, p_n)$, where n is the number of total combinations of main conformations, are transformed such that $p_i \in [0, 1]$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$. Afterward, a single combination of main conformations is selected by sampling according to the probabilities $(p_1, \ldots, p_n)$. Sampling is in accordance with the Boltzmann distribution since consistently selecting the most likely conformations would disregard the probabilistic nature of the system and result in a deterministic algorithm. This process is iteratively repeated for overlapping fragment pairs, with the difference that the conformation of the shared fragment is already fixed. Since the main conformations for the shared fragment were determined independently for different types of fragment pairs, they do not necessarily agree. Therefore, the fixed conformation of the shared fragment is compared with all possible main conformations for the current fragment pair via the metric given in Supporting Information Section S2.2, and the conformation closest to the fixed structure is chosen for the prediction step.

Box 7: The selected main conformations of all fragments are translated to the respective CG bead positions and connected by optimizing bond lengths, angles, and dihedral angles of the connectors (see Figure 4 steps 3, 4, and 5). The target values of the bond lengths and angles are directly obtained from the database. In contrast, the target values for the dihedral angle of the connectors are directly given by their predicted main conformations. The optimization consists of three steps: First, all fragments of the current molecule are rotated around their COM to achieve the target bond lengths as precisely as possible for all connectors. This provides a reasonable starting structure for the subsequent steps. Second, F_1 and F_2 of a randomly chosen boundary fragment pair $P = (F_1, F_2, C) \in \mathcal{P}$, where F_1 can only be connected to F_2 similar to the prediction step, are rotated and translated to obtain chemically reasonable angles and a dihedral angle of its connector. Third, the second step is repeated for overlapping fragment pairs in a stepwise manner. The position of the fragment shared with a previously optimized fragment pair was already modified and remains unchanged to prevent the generation of chemically incorrect structures. To obtain the optimal rotation angles and translation vectors for the individual fragments, the minimize function of `scipy`[50] with the method L-BFGS[51] is used. Moreover, the angle calculations are implemented in Cython[52] for improved runtime performance. A more detailed description of the optimization, which contains, among others, the respective objective functions, is given in Supporting Information Section S3.

Box 8: Finally, hydrogen atoms are added to the atomistic structure using the `Hydride`[53] package. This algorithm is based on a fragment database derived from all molecules of the Chemical Component Dictionary.[54] Selecting suitable fragments and superimposing them on the atomistic structure reliably predicts the hydrogen positions. Notably, we omit the relaxation step of `Hydride` as we observed that it resulted in marginal structural improvements while significantly increasing the runtime by approximately 50 %.

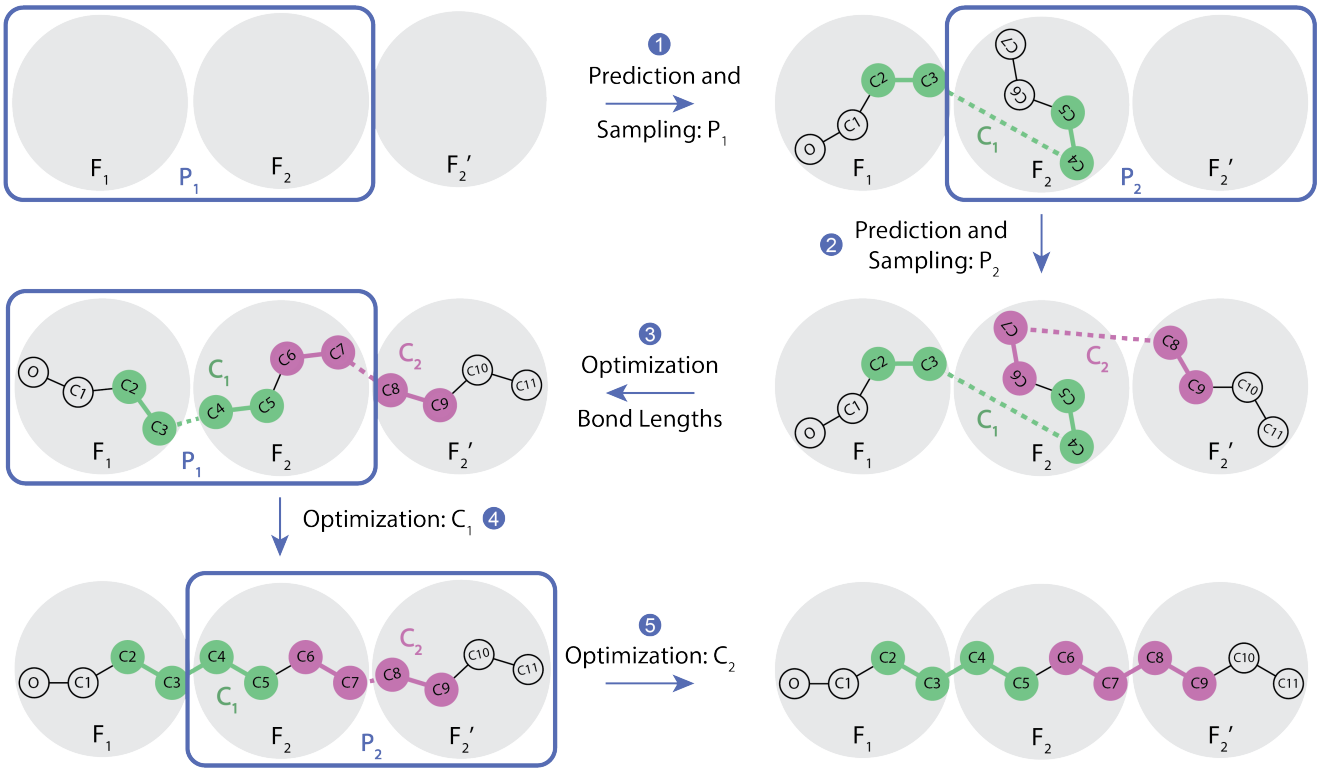

Figure 4: Illustration of the backmapping algorithm using undecan-1-ol. The main conformations of the individual fragments and the connector are predicted for $P_1 = (F_1, F_2, C_1)$ in step 1 and subsequently for $P_2 = (F_2, F_2', C_2)$ in step 2. Afterward, the bond lengths between the atoms C3-C4 and C7-C8 are simultaneously optimized by rotating all fragments F_1, F_2, and F_2' around their COMs (step 3). Next, the fragments of $P1$ are rotated and translated such that proper angles C2-C3-C4 and C3-C4-C5 and an accurate dihedral angle C2-C3-C4-C5 are received while maintaining proper bond lengths between C3-C4 and C7-C8 (step 4). This step is repeated for P_2 with the difference that only the position of F_2' is modified (step 5).

Arrow 9: We recommend relaxing the resulting structure by energy minimization and short MD simulation while restraining the atoms to the CG beads from which they were recovered. The optimal number of relaxation steps is determined in Supplementary Information Subsection S5.4. Note that this step has yet to be directly implemented in ART-SM and, at the moment, has to be manually performed by the user. In this way, the user can decide on his preferred force field and simulation program. Nevertheless, we plan to provide a default option with GROMACS in the future.

We emphasize that together with this article, we provide a custom database built from a selection of common organic molecules, which enables users to bypass the database construction process (boxes 2 to 4 and arrow 5 in Figure 3) if they are working with identical molecules or molecules that share the same fragment pairs. The respective molecules and the corresponding CG models are shown in Supporting Information Figure S6. Moreover, the user can extend the database with new fragment pairs, avoiding repeatedly building the entire database.

2.3 One-bead Molecules

Molecules represented by a single bead in the corresponding CG model do not have fragment pairs. The workflow for these short molecules can, therefore, be greatly simplified: First, the identification of fragment pairs (box 1), the removal and addition of hydrogen atoms in the database construction and the backmapping step (box 2 and box 8), the extraction of bond lengths and angles (arrow 5), and the optimiza-

tion of connectors (arrow 7) can be omitted. Hydrogen atoms can be kept during the entire workflow because their total number is fixed for one-bead molecules. This is not the case for larger molecules with multiple fragments. For instance, heptan-1-ol consists of one fragment pair with SMILES OCCCCCCC and 16 hydrogens, while undecan-1-ol has the same fragment pair with 15 hydrogens. Second, the internal coordinates and the main conformations, determined via hierarchical clustering, are calculated for whole molecules compared to the individual fragments and the connector of fragment pairs (box 3). Note that they are still derived using only heavy atoms, even though hydrogen atoms are not removed in the preprocessing step. Third, probabilities are computed for each main conformation from the clustering results instead of training an RFR for each fragment pair (box 4). Assuming that the main conformation i is representing the cluster C_i of size $|C_i|$, its probability p_i is given by

$$p_i = \frac{|C_i|}{\sum_{j=1}^{n} |C_j|}, \tag{2}$$

where n is the total number of clusters. In the backmapping step, the main conformations are randomly sampled according to their respective probabilities and translated to the positions of the corresponding CG beads (box 6). Finally, each conformation is randomly rotated around its COM to avoid the formation of crystal-like structures.

2.4 Water Model

Water is a special case, as each CG bead represents four atomistic water molecules. Because of their uniqueness and wide usage in simulations, water models are not determined during the database's construction. Instead, predetermined main conformations of four water molecules, hereafter called main water groups because the term conformation usually describes the spatial arrangement of atoms within a single molecule, and their respective probabilities are provided for the TIP3P water model (see Supporting Information Figure S4). To this end, we analyzed the last 10 ns of a 100 ns atomistic simulation, consisting of over 2000 water molecules, as follows: For every 100 ps, 100 water molecules were randomly selected, and the three closest water molecules were identified for each of these molecules. The resulting 10 000 groups of four water molecules were encoded by the SOAP descriptor[55] with the parameters r_cut = 6, n_max = 6, and l_max = 6. Note that internal coordinates are not suitable in this case since the orientation of the individual water molecules to each other would be neglected. Subsequently, hierarchical clustering was performed with the Ward linkage criterion and the SOAP kernel as the distance metric. This analysis yielded six clusters, as shown in Supporting Information Figure S4, and the medoids were chosen as the main water groups. To validate the results, hierarchical clustering was compared to `DBSCAN`[56] via the adjusted Rand index (ARI),[57] whose values lie between -1 (complete mismatch) and 1 (perfect agreement). After removing noise points in DBSCAN, an ARI of 0.92 indicated a high degree of similarity (see Supporting Information Section S4.1 for details). The probabilities of all main water groups were determined in the same way as for one-bead molecules by normalizing the number of water groups in each cluster by the total number of water groups (see Equation (2)).

The backmapping proceeds iteratively. First, a main water group is sampled from the resulting probability distribution and transferred to the position of the corresponding CG bead. Second, the water group is rotated such that the distance between any two atoms in the system is ideally larger than 1 Å to avoid atom clashes. The optimization is performed using the minimize function of `scipy` with the method L-BFGS (see Supporting Information Section S4.2 for a detailed description).

2.5 Data Processing

We simulated 12 systems with different molecules in atomistic or CG resolution (Martini). See Supporting Information Section S5 for simulation parameters. Each simulation was split into training, validation, and test sets

Table 1: Overview of simulation systems. Mixed box consists of 100 molecules undecan-1-ol, propan-1-ol, S-2-bromo-2-chloropropan-1-ol, 1-hydroxypropan-2-one, each, and 200 molecules heptan-1-ol. Split indicates how the trajectory was divided into training, validation, and test sets. For instance, 60/20/20 means that the first 60 ns were used for training, the next 20 ns for validation, and the remaining 20 ns for testing. *Ala-Ala-Gly contains 27 water-soluted tripeptides consisting of the amino acids alanine, alanine, and glycine. Water was excluded from the conversion process since its mapping from atomistic to coarse-grained resolution is non-trivial.

System	**Resolution**	**No. Molecules**	**Sim. Time** (ns)	**Split** (ns)
Undecan-1-ol	Atomistic	104	100	60/ 20/ 20
Heptan-1-ol	Atomistic	190	100	60/ -/ 40
Ala-Ala-Gly*	Atomistic	27	100	60/ -/ 40
Propan-1-ol	Atomistic	461	100	60/ -/ 40
S-2-bromo-2-chloropropan-1-ol	Atomistic	326	100	60/ -/ 40
1-hydroxypropan-2-one	Atomistic	409	100	60/ -/ 40
Water - TIP3P	Atomistic	2022	100	100/ -/ -
1-10-decandiol	Atomistic	105	100	-/ -/100
Butyl-pentadecanoate	Atomistic	57	100	60/ -/ 40
Heptan-1-ol	Coarse-grained	190	105	-/ -/105
Mixed box	Coarse-grained	600	105	-/ -/105
Water	Coarse-grained	729	100	-/ -/100

as specified in Table 1. Atomistic simulations were coarse-grained with ART-SM using the command `artsm-coarse_grain` to generate matching atomistic and CG data. We call these simulations artificial coarse-grained simulations for later reference. To this end, mapping files were created by manually coarse-graining a single molecule and executing `artsm-mapping`.

Four fragment pair databases were built from the training sets with `artsm-build_db`. Three were only trained on individual molecules, namely undecan-1-ol to estimate the optimal number of data points for training RFRs (see Section 3.1), heptan-1-ol for backmapping other molecules with identical fragment pairs (see Section 3.2.3), and butyl-pentadecanoate to backmap molecules with more than three fragments (see Section 3.2.4). The fourth database was used for all other analyses and was generated from simulations of undecan-1-ol, heptan-1-ol, propan-1-ol, Ala-Ala-Gly tripeptide, S-2-bromo-2-chloropropan-1-ol, and 1-hydroxypropan-2-one. All databases were built with a sampling time of 500 ps to ensure that molecule conformations are independent of each other (see Section 3.1). The undecan-1-ol database was built with varying numbers of data points, whereas the other databases use at most 500 randomly selected data points per fragment pair (see Section 3.1).

Snapshots were extracted from the test sets of artificial or real CG simulations and subsequently backmapped with `artsm-backmap` or Backward. For our testcase, we use snapshots every 500 ps to obtain independent molecule conformations and enough data points (here more than 2000) to compare backmapped with atomistic distributions in a statistically meaningful manner. The properties of the resulting molecules were compared to the corresponding atomistic simulations. This workflow was repeated five times to obtain a comprehensive understanding of the robustness and variability of ART-SM and Backward's backmapping process and to ensure that our conclusions are not solely dependent on a single analysis run.

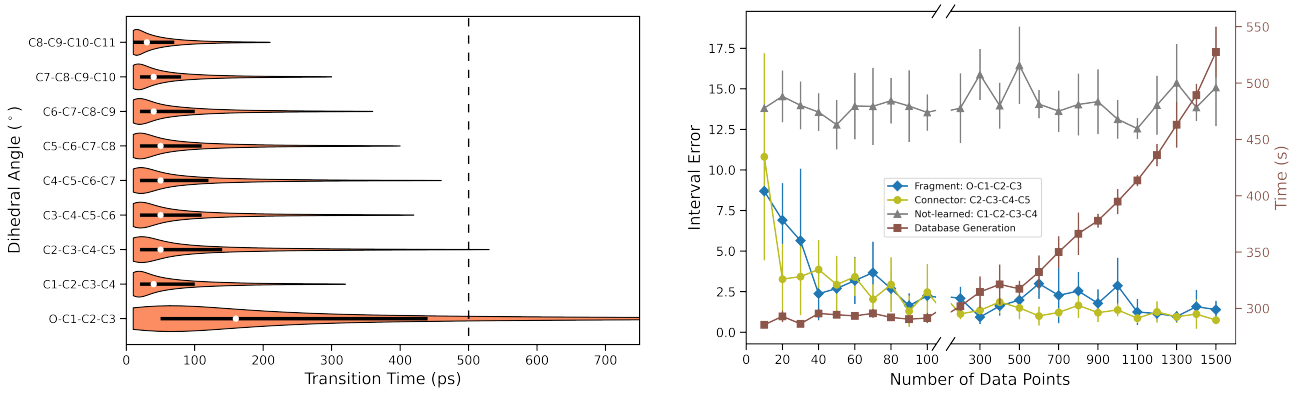

Figure 5: Violin plots of transition times of undecan-1-ol's dihedral angles (left) and estimation of the optimal number of data points for training (right). White circles in the violin plot represent the median values, while black bars indicate the interquartile ranges. The dashed vertical line denotes the chosen sampling time. The plot on the right depicts the mean interval error over five backmapping repetitions (40 snapshots each - see Section 2.5) for a dihedral angle of a fragment O-C1-C2-C3, a connector C2-C3-C4-C5, and neither of the two C1-C2-C3-C4. The mean time to build the corresponding fragment pair database is shown in brown. The error bars denote the corresponding standard deviations. Atom naming of the dihedral angles is consistent with Supporting Information Figure S1.

3 Results

3.1 Optimal Training Data Size

The performance and efficiency of ART-SM depend on the number of data points on which the RFRs of each fragment pair are trained, as well as on the sampling time. Selecting optimal values for both parameters is crucial to minimize both the simulation time required to generate the training data and the runtime to build the fragment pair database.

The sampling time is ideal if it is as small as possible while ensuring that the current molecule conformations do not depend on their conformations in previously selected time frames. To estimate it, we analyzed the transition time, i.e., the simulation time needed for dihedral angles to switch from one main conformation to another, for the undecan-1-ol training set. As illustrated in Figure 5, the transition time significantly varies depending on the types of atoms forming the dihedral angles. Specifically, the dihedral angle involving oxygen O-C1-C2-C3 (see Supporting Information Figure S1 for atom naming) requires more simulation time to undergo conformational changes than dihedral angles consisting solely of carbon atoms. In more detail, the median is 160 ps compared to 30–50 ps. For the subsequent analyses, we used a sampling time of 500 ps to account for possibly higher transition times for dihedral angles of other molecules. Note that even though this is a conservative choice, the transition time can be significantly larger than 500 ps as illustrated in Supporting Information Figure S7.

The number of training data points is ideal when training on more data points does not improve the performance of the backmapping algorithm. To evaluate this, we constructed the database and thus trained the RFRs using varying numbers of randomly selected data points from the undecan-1-ol training set. Subsequently, we backmapped 40 simulation snapshots from the validation set, i.e., subsequent frames are 500 ps apart, and the results were evaluated by computing the interval error, which compares the backmapped and atomistic distributions of each dihedral angle (see Supporting Information Section S6 and specifically S6.2 for a detailed explanation). As shown in Figure 5, the interval error rapidly decreases with the number of data points for the dihedral angles O-C1-C2-C3 and C2-C3-C4-C5, namely from approximately 10.0 to 2.5. The interval error is

not improving for more than 100 data points. Notably, the interval error for the dihedral angle C1-C2-C3-C4 fluctuates around approximately 14, independent of the number of data points. This behavior is expected since we are neither learning nor optimizing dihedral angles, whose atoms are neither fully part of a fragment nor form the connector of a fragment pair, in the backmapping algorithm (see Supporting Information Section S7 for a detailed explanation). The interval error progression over the number of data points for all dihedral angles of undecan-1-ol can be found in Supporting Information Figure S14. The runtime to build the database increases non-linearly with the number of data points (Figure 5), primarily due to the computationally expensive hierarchical clustering step. For the subsequent analyses, we used 500 data points to build our database of fragment pairs, which is a compromise between runtime and accounting for a potentially worse training behavior of different molecules. Similar to the sampling time selection, this is a conservative choice.

3.2 Backmapping via ART-SM and its Comparison to Backward

After selecting a suitable sampling time of 500 ps, an optimal number of 500 data points per fragment pair to train the RDFs and estimating the number of steps for the EM and position restraint simulations to 200 and 5000, respectively, (see Supplementary Information Subsection S5.4) we examined the performance of ART-SM by backmapping a variety of CG structures: Artificial CG structures (Section 3.2.1), real CG simulations (Section 3.2.2), a molecule not used for training (Section 3.2.3), a larger five-bead molecule (Section 3.2.4), and water (Section 3.2.5). The chemical accuracy of the resulting reverse transformed structures was evaluated by comparing distributions of bond lengths and angles via the Bhattacharyya distance[58] (see Supporting Information Section S6.4 for details) and of dihedral angles via the Wasserstein distance and interval error to those of atomistic simulations. Furthermore, the backmapped snapshots were re-coarse-grained, and the distances of each bead to the corresponding beads in the original CG structure were computed. This comparison estimates how representative the reverse transformed structures are for the original CG ones. Moreover, we backmapped the same structures using the state-of-the-art method Backward[34] and compared the respective results to ART-SM. For the dihedral angle distributions of undecan-1-ol (see Figure 6) and the runtime analysis, we also compared different stages of the respective algorithms. More precisely, ART-SM has the consecutive stages (i) projection consisting of fragment prediction with ML and optimization, (ii) EM, and (iii) position restraint simulation, whereas Backward's wrapper initram performs after the projection itself the successive steps EM 1, EM 2, MD 1, MD 2, MD 3, and MD 4. A visual example of the molecule undecan-1-ol at the different stages is given in Figure 7). We compared the structures of ART-SM to Backward after projection, after EM and EM 2, and after position restraint simulation and MD 4. In the subsequent sections, we call these defined stages projection, EM, and final, respectively.

3.2.1 Artificial Coarse-grained Simulations

As shown in Table 2 section a) the mean Bhattacharyya distances for bonds and angles are below one for ART-SM at the final stage across all molecules, implying an optimal agreement with atomistic simulations. While Backward leads to similar values for angles, the Bhattacharyya distances for bonds are consistently above 60 as the distributions are too narrow compared to atomistic ones (see Supporting Information Figure S16). The distance to the original CG structure is between 1.0 and 1.3 Å for ART-SM and 1.3 to 2.1 Å for Backward. In particular, using Backward for the tripeptide Ala-Ala-Gly results in an exceptionally large mean distance of 2.03 Å.

Concerning dihedral angle distributions, ART-SM performs very well for undecan-1-ol, heptan-1-ol, and 1-hydroxypropan-2-one

Table 2: Evaluation of backmapped structures at the final stages of ART-SM and Backward. The initial coarse-grained (CG) systems were artificially generated from atomistic simulations a), stem from real CG simulations b), contain molecules that were not used for training c), consist of five-bead molecules d). Dihedral angle distributions of the backmapped structures (40 to 200 snapshots depending on the simulated system - see Table 1 and Section 2.5) were evaluated via the Wasserstein distance and interval error. The bond and angle distributions were evaluated via the Bhattacharyya distance. Atomistic data was always used as a reference. Additionally, the distance to the original CG structures was computed by re-coarse-graining the reverse transformed structures. The mean values and standard deviations were obtained from five repetitions.

	Molecules	Method	Wasserstein Distance (°)	Interval Error	Bhattacharyya Angle	Bhattacharyya Bonds	Distance to CG (Å)
a)	Undecan-1-ol	ART-SM	6.58 ± 2.90	3.13 ± 1.47	0.07 ± 0.03	0.02 ± 0.01	1.01 ± 0.03
		Backward	14.64 ± 9.62	7.50 ± 4.77	0.66 ± 0.42	70.51 ± 2.97	1.38 ± 0.06
	Heptan-1-ol	ART-SM	5.37 ± 2.24	2.45 ± 1.09	0.03 ± 0.01	0.03 ± 0.01	1.06 ± 0.03
		Backward	9.70 ± 6.18	4.74 ± 2.83	0.73 ± 0.56	70.05 ± 3.56	1.44 ± 0.06
	Propan-1-ol	ART-SM	10.46 ± 0.70	4.64 ± 0.34	0.05 ± 0.01	0.05 ± 0.02	1.27 ± 0.03
		Backward	60.20 ± 0.44	28.37 ± 0.22	0.20 ± 0.20	67.66 ± 4.26	1.63 ± 0.05
	Ala-Ala-Gly	ART-SM	21.09 ± 10.11	10.11 ± 7.14	0.66 ± 0.58	0.22 ± 0.24	0.97 ± 0.05
		Backward	21.05 ± 15.68	9.91 ± 8.35	1.87 ± 3.37	83.87 ± 31.02	2.03 ± 0.13
	1-hydroxypropan-2-one	ART-SM	2.50 ± 0.00	0.70 ± 0.11	0.01 ± 0.00	0.02 ± 0.00	1.07 ± 0.02
		Backward	4.34 ± 0.01	1.06 ± 0.17	0.15 ± 0.08	64.31 ± 8.26	1.35 ± 0.04
	S-2-bromo-2-chloropropan-1-ol	ART-SM	5.16 ± 0.02	2.22 ± 0.00	0.02 ± 0.02	0.02 ± 0.01	1.23 ± 0.03
		Backward	27.38 ± 0.08	11.70 ± 0.01	0.56 ± 0.36	77.05 ± 10.25	1.33 ± 0.04
b)	Heptan-1-ol	ART-SM	8.00 ± 2.20	3.68 ± 1.11	0.07 ± 0.01	0.03 ± 0.01	1.24 ± 0.03
		Backward	10.60 ± 6.06	5.00 ± 2.78	0.90 ± 0.59	69.98 ± 3.52	1.82 ± 0.06
	Mixed System	ART-SM	11.01 ± 3.99	4.95 ± 2.29	0.06 ± 0.04	0.15 ± 0.17	1.24 ± 0.02
		Backward	18.87 ± 14.02	8.75 ± 6.88	0.37 ± 0.33	70.49 ± 6.83	1.53 ± 0.04
c)	1-10-decandiol	ART-SM	7.00 ± 2.99	3.22 ± 1.54	0.05 ± 0.02	0.02 ± 0.01	1.00 ± 0.03
		Backward	12.59 ± 9.78	6.31 ± 4.80	0.60 ± 0.41	69.85 ± 3.82	1.22 ± 0.06
d)	Butyl-pentadecanoate	ART-SM	6.68 ± 2.87	3.26 ± 1.29	0.06 ± 0.06	0.05 ± 0.02	1.01 ± 0.03
		Backward	16.39 ± 14.34	9.32 ± 9.25	1.49 ± 2.26	70.84 ± 5.72	1.38 ± 0.07

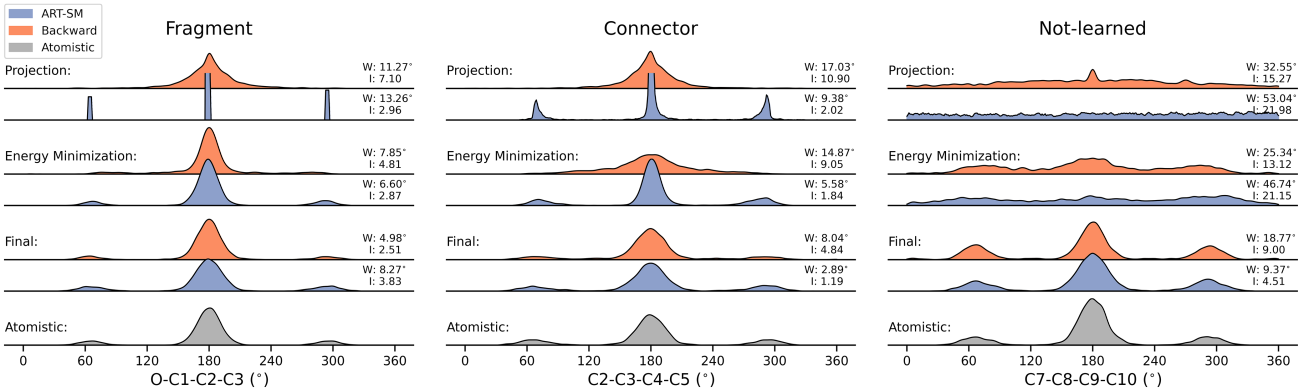

Figure 6: Ridge plots depicting the distributions of the fragment, connector, and not-learned dihedral angles O-C1-C2-C3, C2-C3-C4-C5, and C7-C8-C9-C10, respectively, for undecan-1-ol at the different stages of ART-SM (blue) and Backward (orange). The distributions are based on 40 snapshots containing 104 undecan-1-ol molecules. The atomistic reference is plotted in gray at the bottom row. The Wasserstein distance (W) and interval error (I) relative to the corresponding atomistic distribution are listed next to each distribution. The atom naming is consistent with Supporting Information Figure S1.

as the mean Wasserstein distances are below 7.0° and the standard deviations are lower than 3.0°. For comparison, Backward leads to higher Wasserstein distances ranging from 4.34° for 1-hydroxypropan-2-one to 14.64° for undecan-1-ol.

For undecan-1-ol, the distributions of one fragment, connector, and not-learned dihedral angle are exemplarily shown at different key stages in Figure 6 and for all dihedral angles in Supporting Information Figure S15. After the projection stage of ART-SM, the distributions of fragment and connector dihedral angles are reasonably close to those of the atomistic ones. Thus, the subsequent short EM is enough to restore the atomistic distributions accurately. For the dihedral angles that are neither part of a fragment nor a connector, the distributions are comparatively uniform and do not have distinct peaks. Therefore, to obtain accurate backmapping results, the entire 5000 steps of position restraint simulation are required, as expected from our analysis on the optimal number of relaxation steps. For Backward, the dihedral angle distributions of the projected structures are either uniform or peak at around 180°. The subsequent EMs and MD simulations of Backward restore the correct distributions only for the latter. Note that this behavior does not generalize to other molecules. The dihedral angle distributions after the projection with Backward and, thus, whether the atomistic distributions can be restored only depends on the provided mapping file (not identical to the ART-SM mapping file), which contains instructions on how to build the projected structure. Optimizing these mapping files requires a lot of trial and error and extensive expert knowledge and might still fail due to the limited functionality of Backward's mapping files.

For propan-1-ol, ART-SM at the final stage results in a slightly higher mean Wasserstein distance of 10.46° than for the previously analyzed molecules. Propan-1-ol has a single dihedral angle, and its atomistic distribution has three peaks: two less populated ones at around −60°, and 60°, and one significantly populated at 180° (see Supporting Information Figure S17). While the shape of the distribution is well reproduced, the small peaks are slightly overpopulated, which leads to the mentioned increase in the Wasserstein distance. This effect is far more pronounced for Backward. In fact, the peaks at −60° and 60° are higher than the peak at 180°, which leads to the large Wasserstein distance of 60.20°. This is because each dihedral angle is almost equally likely in the initially projected structure, and the subsequent EMs and MD simulations of Backward cannot fully recover the correct dihedral angle distribution.

This is also the case for S-2-bromo-2-chloropropan-1-ol, where ART-SM leads to a Wasserstein distance of 5.16° compared to 27.38° for Backward. Moreover, it is important to note that in rare cases ($\leq 0.05\,\%$), the final molecules of Backward are in the R instead of the S configuration.

The mean and standard deviation of the Wasserstein distance of the tripeptide Ala-Ala-Gly are comparatively high at 21.09° and 10.11° for ART-SM at the final stage. One reason is that the cis-trans isomerism of peptide bonds is incorrect for a few molecules. Moreover, some dihedral angle distributions have additional peaks that are not present in the atomistic distributions. Backward also gives a high Wasserstein distance for Ala-Ala-Gly of 21.05° and additionally a higher standard deviation of 15.68°. Despite Backward recovering the cis-trans isomerism of the peptide bond, some dihedral angle distributions have shifted, or the number of peaks differs from the atomistic ones. These deviations are more pronounced than with ART-SM. Histograms for propan-1-ol, S-2-bromo-2-chloropropan-1-ol, and Ala-Ala-Gly are exemplarily shown for the final stages of ART-SM and Backward in Supporting Information Figures S16, S17, and S17, respectively. Note that we only discussed the Wasserstein distance in this section since the interval errors are highly correlated. Usually, they are approximately half of the Wasserstein distance, except for 1-hydroxypropan-2-one, where they amount to about one-fourth.

In summary, ART-SM outperforms Backward at reproducing atomistic dihedral angle and bond length distributions due to the more advantageous configurations after the projection step,

except for the tripeptide Ala-Ala-Gly. Here, ART-SM might profit from a specific adjustment to proteins in the future, for instance, by geometric corrections of the backbone. Furthermore, the reverse transformed structures better represent the original CG structures in the case of ART-SM.

3.2.2 Real Coarse-grained Simulations

Until this point, we backmapped structures that were prepared by re-coarse-graining atomistic snapshots. This is because we initially required matching sets of atomistic and CG structures for training RDFs and thus building the fragment pair database, and we tested the performance of ART-SM on data similar to that from training. However, real CG simulations are less accurate than atomistic simulations due to their lower particle resolution and, thus, overall structural information content. Consequently, structural characteristics such as bead distances and angles might differ between real and artificial CG snapshots, potentially leading to feature values not seen during training. To investigate the effect on backmapping, we simulated 190 molecules of heptan-1-ol and a mixed system consisting of 100 molecules of undecan-1-ol, propan-1-ol, S-2-bromo-2-chloropropan-1-ol, and 1-hydroxypropan-2-one, each, and 200 molecules of heptan-1-ol at CG resolution with the Martini force field for 105 ns (see Supporting Information Section S5.2 for simulation parameters). Subsequently, snapshots were extracted approximately every 500 ps, and the resulting 213 structures were backmapped with ART-SM and Backward.

The results at the final stage are presented in Table 2 section b). For the snapshots of pure heptan-1-ol, the Wasserstein distance increased from 5.37° to 8.00° for ART-SM and 9.70° to 10.60° for Backward compared to the artificial CG structures. An analogous increase can be observed for the interval error. Moreover, the distance to the original CG structure increased from 1.06 Å to 1.24 Å and 1.44 Å to 1.82 Å for ART-SM and Backward, respectively. The Bhattacharyya distances for bond lengths and angles remain unchanged. For the mixed system, the Wasserstein distance and interval error for ART-SM amount to 11.01° and 4.95, which is higher than for any individual molecule of the artificial CG structures. For Backward, the Wasserstein distance and interval error of 18.87° and 8.75 are comparable to the mean overall individual artificial CG molecules of 20.99° and 9.69°. Notably, the respective standard deviations of 14.02° and 6.88° are exceptionally large because Backward converts S-2-bromo-2-chloropropan-1-ol and propan-1-ol worse than other molecules. The final structures of ART-SM and Backward are slightly less representative of the original CG structures as the respective distances increased from 1.12 Å and 1.42 Å to 1.24 Å and 1.53 Å, respectively. Similar to pure heptan-1-ol, the Bhattacharyya values for the bonds and angles are comparable to the ones of the artificial CG structures for both methods. In summary, most evaluation criteria slightly worsened for the backmapped structures of real CG simulations compared to artificial CG simulations for ART-SM and Backward. However, the difference is relatively small.

3.2.3 Transferrability to Molecules with Identical Fragment Pairs

One strength of fragment-based approaches is that once a database has been built, any molecule with identical fragments, or in the case of ART-SM fragment pairs, can be backmapped. To verify this statement, we generated a small database only from the atomistic simulation of heptan-1-ol and backmapped the three-bead molecule 1-10-decandiol with ART-SM. As depicted in Table 2 section c) at the final stage, the Wasserstein distance amounts to 7.00°, the interval error to 3.22, the Bhattacharyya values for bonds and angles to 0.02 and 0.05, and the distance to the original CG beads to 1.00 Å. This implies an excellent agreement with the simulated system at an atomistic level and is comparable to previous results. Consistently with other systems, Backward has a higher Wasserstein distance and interval error for dihedral angles of 12.59° and 6.31, a higher Bhattacharya value of 69.85 for bonds, and deviates further from the original CG structure (1.22 Å).

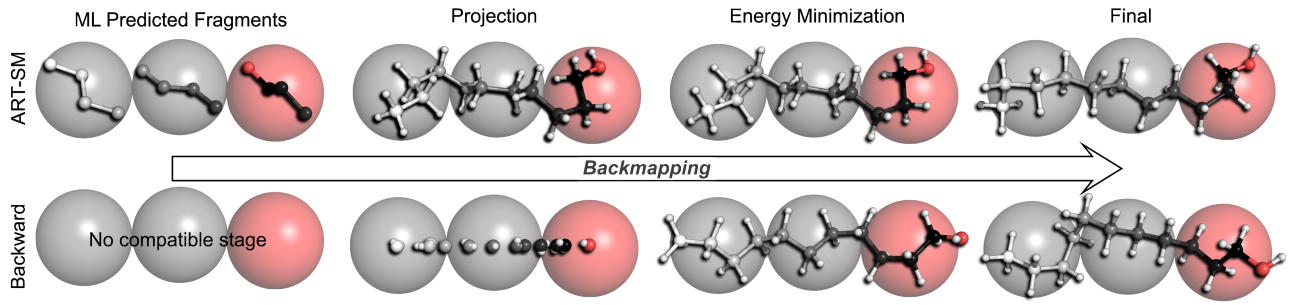

Figure 7: Exemplary visualisation of undecan-1-ol at different backmapping stages for ART-SM and Backward. Carbons are colored via a sequential grayscale, oxygens are depicted in red, and hydrogens in white. The corresponding Martini coarse-grained model is shown as transparent spheres.

3.2.4 Extension to Larger Molecules

Thus far, we have focussed on short molecules of at most three CG beads. This section examines whether ART-SM can also accurately backmap larger molecules despite learning only local features for each fragment pair. For this purpose, we have atomistically simulated, manually coarse-grained, and backmapped the five-bead molecule butyl-pentadecanoate and compared the results at the final stage with Backward. As shown in Table 2 section d), ART-SM provides a quality of backmapped structures that is similar to shorter molecules. Namely, bond lengths and angle distributions are correctly reproduced, the mean distance to the original CG structure is 1.01 Å, and the Wasserstein distance and interval error are 6.68° and 3.26 with standard deviations of 2.87° and 1.29, respectively, which indicates an excellent agreement with atomistic distributions across all dihedral angles. Similarly to other investigated molecules, Backward's structures of butyl-pentadecanoate are characterized by too-narrow bond length distributions, poorer dihedral angle distributions indicated by the higher Wasserstein distance and interval error of 16.39° and 9.32 together with high standard deviations of 14.34° and 9.25, and diminished representativeness of the original CG structure (mean distance of 1.38 Å).

3.2.5 Water

Water, a fundamental component of many simulations, is a special case because multiple water molecules are represented by a single CG bead. During the backmapping process, ART-SM places, similar to Backward, predefined groups of four water molecules at the position of the respective CG bead. However, ART-SM uses six different water groups extracted from an atomistic simulation compared to only one in Backward and rotates them in an optimization step to avoid atom clashes. We backmapped 201 snapshots consisting of 729 CG water beads with ART-SM and Backward. Subsequently, we determined the oxygen-oxygen radial distribution functions (RDF) using GROMACS's tool *gmx rdf* with a cutoff of 15 Å for each backmapped snapshot (see Figure 8). The RDF for the reference atomistic simulation is zero for distances smaller than 2.5 Å and rapidly rises afterward to the maximum of 2.8 at approximately 2.7 Å. Subsequently, it decreases to a local minimum of 0.9 at 3.8 Å and levels out at 1.0 for larger distances. ART-SM at the final stage reproduces the atomistic RDF accurately, with a slightly too pronounced first maximum of 3.5 and a local minimum of 0.8 at approximately 2.7 Å and 3.5 Å. On the other hand, Backward has a lower first maximum of 1.9 and minimum of 0.8, and the minimum is shifted to approximately 4.5 Å. Moreover, Backward's RDF has multiple maxima and minima for distances larger than 5.0 Å, which are not present in the atomistic RDF. This is due to the low default simulation temperature of 200 K for the MD simulations of Backward compared to 310 K for the atomistic and ART-SM simulations. We also compared

the RDFs after the EM and projection stages in the Supporting Information Section S4.3.

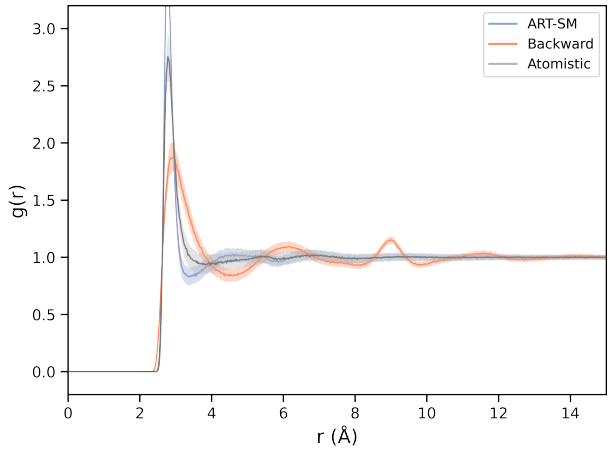

Figure 8: Oxygen-oxygen radial distribution functions after the final stages of ART-SM (blue) and Backward (red) for the TIP3P water model. The atomistic reference is shown in gray. The solid lines show the mean values over 200 snapshots and the shaded areas denote the corresponding standard deviations.

3.2.6 Runtime

Finally, we examined the runtime performance of ART-SM compared to Backward by backmapping undecan-1-ol systems containing various numbers of CG beads, ranging from 312 to 67 392. This corresponds to 3744 up to 808 704 atoms in the backmapped atomistic structures. To generate the respective CG structures, we multiplied the smallest system in the x, y, and z directions and adjusted the simulation boxes accordingly. Subsequently, each structure was backmapped ten times on an Intel(R) Xeon(R) Gold 6226R CPU 2.90GHz with ART-SM and Backward using one core, respectively. Additionally, an NVIDIA GeForce RTX 3090 GPU was used for the EMs and position restraint simulations following the projections with ART-SM. As the EM 1 step of Backward utilizes energy exclusions to optimize individual molecules without considering the surrounding atoms, the group cutoff scheme has to be used, which is incompatible with GPUs. Also, due to the nearby clashes of the atoms resulting from the backward projection, EM 1 cannot be run in parallel.

Both algorithms scale linearly with the number of atoms (see Figure 9), whereby the slope is steeper for Backward than for ART-SM. Consequently, backmapping the smallest system with ART-SM is about two times faster than Backward with 15.3 compared to 29.3 seconds, while for the largest system, ART-SM is roughly three times faster with 39.6 compared to 114.9 minutes. We emphasize that the outlier for Backward at 808 704 atoms, which took exceptionally long, was thereby not considered. It is important to note here, that the projected structure of both algorithms may contain atom clashes or unfavorable conformations which lead to failure of the consecutive EM or position restraint simulation. In our test systems, ART-SM failed less often than Backward. In more detail, out of 60 conversions, 59 were successful for ART-SM, in contrast to only 37 for Backward. Thereby, the success rate of Backward declined as the number of atoms increased. For Backward, we observed that crashes usually happened during the second EM, i.e., after switching off the energy exclusions.

4 Conclusions

We developed ART-SM, an open-source fragment-based backmapping algorithm that exploits ML to learn the Boltzmann distribution from atomistic simulation data. Its user-friendliness is ensured by requiring minimal user input, having multiple supporting tools, for instance, to generate mapping files automatically, and the easy installation via `pip`. ART-SM improves traditional approaches, which only use one rigid conformation per fragment for backmapping, by selecting suitable fragment conformations based on the Boltzmann distribution and the underlying CG conformation. Furthermore, individual fragments are connected by rotating them in space, thus optimizing specific bond lengths, angles, and dihedral angles. This approach yields preferable conformations that can be easily relaxed within a few steps of an EM and a position restraint simulation.

In this study, we backmapped artificial

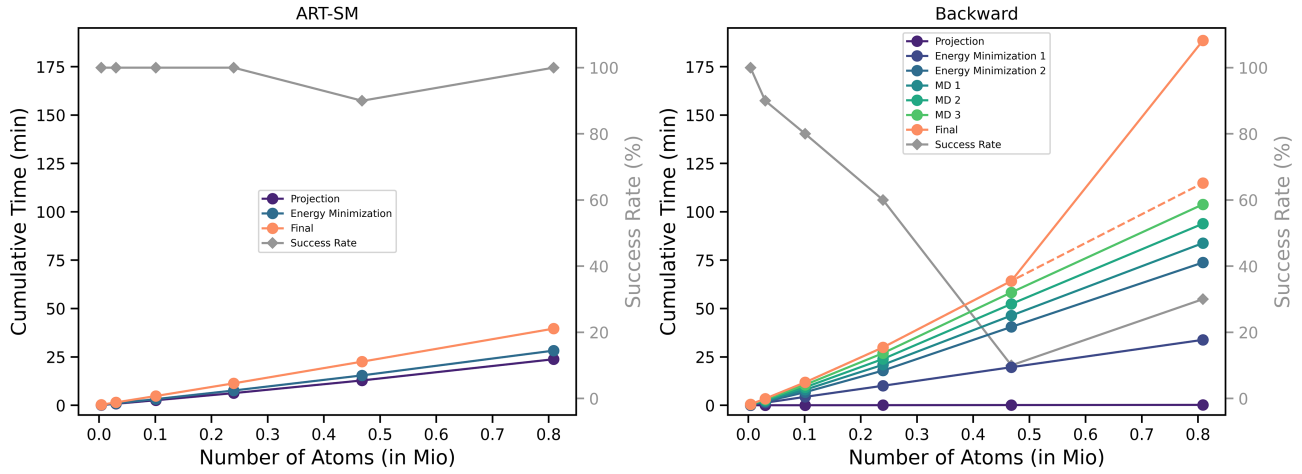

Figure 9: Average runtime comparison between ART-SM and Backward. The cumulative runtimes of the individual stages are plotted for ART-SM (left) and Backward (right). The success rates, i.e., the percentages of successful backmapping attempts, are plotted in gray. For about 808 704 atoms, one of the final molecular dynamics simulations of Backward required an exceptionally long computing time. The runtime without this outlier is plotted as a dashed orange line. The standard deviations are omitted to improve readability.

and real Martini CG structures containing short molecules of up to three beads, such as undecan-1-ol or S-2-bromo-2-chloropropan-1-ol, and showed that the resulting bond length, angle, and dihedral angle distributions closely match the atomistic reference (see Section 3.2.1 and 3.2.2). Re-coarse-graining of the backmapped structures revealed that they are representative of the original CG structures as the distance between original and re-coarse-grained beads was, on average, approximately 1.1 Å. ART-SM also generalized well to molecules not seen during training, given that they share identical fragment pairs with molecules included in the training data set (see Section 3.2.3). This was tested for 1-10-decandiol, which has the same fragment pairs as heptan-1-ol. Furthermore, ART-SM accurately backmapped the five-bead molecule butyl-pentadecanoate, indicating that the incremental pairwise backmapping approach can be applied to larger molecules as well (see Section 3.2.4). Nevertheless, ART-SM was not tested on large molecules with complex structural features such as branched and cyclic structures yet and should be used with caution in these cases. Additionally, a robust backmapping procedure was successfully developed and tested for the 3-point water model TIP3P. In all instances, ART-SM outperformed the widely used backmapping algorithm Backward[34] in terms of similarity to atomistic bond length and dihedral angle distributions, except for the tripeptide Ala-Ala-Gly where both algorithms provided comparable results (see Table 2). Moreover, the re-coarse-grained structures of ART-SM were closer to the original CG structures, and our algorithm was about two to three times faster with significantly fewer crashes than Backward (see Section 3.2.6).

In summary, we illustrated that our algorithm accurately backmaps Martini CG systems consisting of small molecules, as bond length, angle, and dihedral angle distributions are well recovered. Moreover, the resulting atomistic structures closely resemble the CG structures. Additionally, ART-SM outperformed Backward in terms of runtime, success rate, and our evaluation criteria in all studied cases except for the tripeptide Ala-Ala-Gly. In the future, we will extend our algorithm to lipids, proteins, and other macromolecules. Moreover, coarse-graining schemes of more than six heavy atoms per bead will be investigated.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC 2075 – 390740016, by the Stuttgart Center for Simulation Science (SC SimTech), and by the Ministry of Science, Research and the Arts Baden-Wuerttemberg in the Artificial Intelligence Software Academy (AISA). The computer time was provided by the BwForCluster BiNAC funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

Code Availability

The Python code for ART-SM and an introductory tutorial is available on GitHub[1]. A pre-built fragment pair database based on the molecules in this study (see Supplementary Information Figure S6) can be downloaded from the same repository.

Supporting Information Available

Section S1: Illustration of ART-SM's mapping files and an example for heptan-1-ol. Figure S1: Schematic drawings depicting the structure and atom names of molecules whose bonds, angles, or dihedral angles (e.g. O-C1-C2-C3) are directly described in the manuscript. Figure S2: Boltzmann distribution of the dihedral angle O-C1-C2-C3 of undecan-1-ol based on atomistic data. Section S2: Angle and main conformation metrics used in the backmapping process of ART-SM. Section S3 and Figure S3: Details on the optimization of connectors in the backmapping process of ART-SM. Includes connector definition, algorithmic steps, and optimization functions. Section S4.1, S4.2, and Figure S4: Description of the TIP3P hierarchical clustering together with the resulting dendrogram and cluster representatives. Section S4.3 and Figure S5: Oxygen-oxygen radial distribution functions for water after projection and energy minimization with ART-SM and Backward. Figure S6: Visualization of all molecules used in this study in atomistic and coarse-grained resolution. Figure S7: Illustration of the transition time differences between undecan-1-ol and S-2-bromo-2-chloropropan-1-ol. Section S5 and Figure S8: GROMACS simulation parameters for atomistic, coarse-grained, and flat-bottomed position restraint simulations and determination of optimal number of relaxation steps after projection by ART-SM. Figures S9, S10, and S11: Evaluation of energy minimized and final structures of ART-SM dependent on the number of steps, respectively. Analyzed were the distances to the original CG structure, the Wasserstein distance for bonds and angles, and the maximum force that occurred during the energy minimizations. Section S6 and Figure S12: Explanation of the comparison between atomistic and backmapped (ART-SM and Backward) bond length, angle, and dihedral angle distributions with the interval error, Wasserstein distance, and Bhattacharyya distance. Section S7 and Figure S13: Classification of dihedral angles into the categories fragments, connectors, and not-learned. Figure S14: Interval errors between ART-SM projected and atomistic structures dependent on the number of training data points used for undecan-1-ol. Figure S15: Dihedral angle distributions of undecan-1-ol for different stages of ART-SM and Backward depicted as ridge plots. Figures S16, S17, S18, and S19: Example histograms of the bond lengths (undecan-1-ol), angles (undecan-1-ol), and dihedral angles (propan-1-ol, S-2-bromo-2-chloropropan-1-ol, and Ala-Ala-Gly) at the final stage of ART-SM and Backward compared to the atomistic reference.

[1] https://github.com/chrispfae/ART-SM.git

References

(1) Degen, M.; Santos, J.; Pluhackova, K.; Cebrero, G.; Ramos, S.; Jankevicius, G.; Hartenian, E.; Guillerm, U.; Mari, S. A.; Kohl, B.; Müller, D. J.; Schanda, P.; Maier, T.; Perez, C.; Sieben, C.; Broz, P.; Hiller, S. Structural basis of NINJ1-

mediated plasma membrane rupture in cell death. *Nature* **2023**,

(2) Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **2017**, *117*, 139–155.

(3) Pluhackova, K.; Gahbauer, S.; Kranz, F.; Wassenaar, T. A.; Böckmann, R. A. Dynamic Cholesterol-Conditioned Dimerization of the G Protein Coupled Chemokine Receptor Type 4. *PLOS Computational Biology* **2016**, *12*, 1–25.

(4) Pluhackova, K.; Wilhelm, F. M.; Müller, D. J. Lipids and Phosphorylation Conjointly Modulate Complex Formation of β2-Adrenergic Receptor and β-arrestin2. *Frontiers in Cell and Developmental Biology* **2021**, *9*.

(5) Zhang, H.; Pluhackova, K.; Jiang, Z.; Böckmann, R. A. Binding Characteristics of Sphingosine-1-Phosphate to ApoM hints to Assisted Release Mechanism via the ApoM Calyx-Opening. *Scientific Reports* **2016**, *6*, 30655.

(6) Mari, S. A.; Pluhackova, K.; Piperce-vic, J.; Leipner, M.; Hiller, S.; Engel, A.; Müller, D. J. Gasdermin-A3 pore formation propagates along variable pathways. *Nat. Commun.* **2022**, *13*, 2609.

(7) Han, J.; Pluhackova, K.; Böckmann, R. A. The Multifaceted Role of SNARE Proteins in Membrane Fusion. *Frontiers in Physiology* **2017**, *8*.

(8) Korn, V.; Pluhackova, K. Not sorcery after all: Roles of multiple charged residues in membrane insertion of gasdermin-A3. *Frontiers in Cell and Developmental Biology* **2022**, *10*.

(9) Wachlmayr, J.; Fläschner, G.; Pluhackova, K.; Sandtner, W.; Siligan, C.; Horner, A. Entropic barrier of water permeation through single-file channels. *Communications Chemistry* **2023**, *6*, 135.

(10) Zhang, J.; Clennell, M. B.; Sagotra, A.; Pascual, R. Molecular dynamics simulation and machine learning for predicting hydrogen solubility in water: Effects of temperature, pressure, finite system size and choice of molecular force fields. *Chemical Physics* **2023**, *564*, 111725.

(11) Alessandri, R.; Grünewald, F.; Marrink, S. J. The Martini Model in Materials Science. *Advanced Materials* **2021**, *33*, 2008635.

(12) Siwaipram, S.; Bopp, P. A.; Keupp, J.; Pukdeejorhor, L.; Soetens, J.-C.; Bureekaew, S.; Schmid, R. Molecular Insight into the Swelling of a MOF: A Force-Field Investigation of Methanol Uptake in MIL-88B(Fe)–Cl. *The Journal of Physical Chemistry C* **2021**, *125*, 12837–12847.

(13) Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129–1143.

(14) Callaway, E. The revolution will not be crystallized: a new method sweeps through structural biology. *Nature* **2015**, *525*, 172–174.

(15) Pluhackova, K.; Schittny, V.; Bürkner, P.-C.; Siligan, C.; Horner, A. Multiple pore lining residues modulate water permeability of GlpF. *Protein Science* **2022**, *31*, e4431.

(16) Laskowski, P. R.; Pluhackova, K.; Haase, M.; Lang, B. M.; Nagler, G.; Kuhn, A.; Müller, D. J. Monitoring the binding and insertion of a single transmembrane protein by an insertase. *Nature communications* **2021**, *12*, 1–11.

(17) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *WIREs Computational Molecular Science* **2014**, *4*, 225–248.

(18) Pluhackova, K.; Böckmann, R. A. Biomembranes in atomistic and coarse-grained simulations. *Journal of Physics: Condensed Matter* **2015**, *27*, 323103.

(19) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42*, 6801–6822.

(20) Marrink, S. J.; Monticelli, L.; Melo, M. N.; Alessandri, R.; Tieleman, D. P.; Souza, P. C. T. Two decades of Martini: Better beads, broader scope. *WIREs Computational Molecular Science* **2023**, *13*, e1620.

(21) Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A.; Kroon, P. C.; Melcr, J.; Nieto, V.; Corradi, V.; Khan, H. M.; Domański, J.; Javanainen, M.; Martinez-Seara, H.; Reuter, N.; Best, R. B.; Vattulainen, I.; Monticelli, L.; Periole, X.; Tieleman, D. P.; de Vries, A. H.; Marrink, S. J. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods* **2021**, *18*, 382–388.

(22) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(23) Kroon, P. C.; Grünewald, F.; Barnoud, J.; van Tilburg, M.; Souza, P. C. T.; Wassenaar, T. A.; Marrink, S.-J. Martinize2 and Vermouth: Unified Framework for Topology Generation. 2022.

(24) Hilpert, C.; Beranger, L.; Souza, P. C. T.; Vainikka, P. A.; Nieto, V.; Marrink, S. J.; Monticelli, L.; Launay, G. Facilitating CG Simulations with MAD: The MArtini Database Server. *Journal of Chemical Information and Modeling* **2023**, *63*, 702–710.

(25) Wassenaar, T. A.; Pluhackova, K.; Moussatova, A.; Sengupta, D.; Marrink, S. J.; Tieleman, D. P.; Böckmann, R. A. High-throughput simulations of dimer and trimer assembly of membrane proteins. The DAFT approach. *Journal of chemical theory and computation* **2015**, *11*, 2278–2291.

(26) Pluhackova, K.; Wassenaar, T. A.; Kirsch, S.; Böckmann, R. A. Spontaneous adsorption of coiled-coil model peptides K and E to a mixed lipid bilayer. *The Journal of Physical Chemistry B* **2015**, *119*, 4396–4408.

(27) Gahbauer, S.; Pluhackova, K.; Böckmann, R. A. Closely related, yet unique: Distinct homo- and heterodimerization patterns of G protein coupled chemokine receptors and their fine-tuning by cholesterol. *PLOS Computational Biology* **2018**, *14*, 1–30.

(28) Han, J.; Pluhackova, K.; Wassenaar, T. A.; Böckmann, R. A. Synaptobrevin Transmembrane Domain Dimerization Studied by Multiscale Molecular Dynamics Simulations. *Biophysical Journal* **2015**, *109*, 760–771.

(29) Friess, M. D.; Pluhackova, K.; Böckmann, R. A. Structural Model of the mIgM B-Cell Receptor Transmembrane Domain From Self-Association Molecular Dynamics Simulations. *Frontiers in Immunology* **2018**, *9*.

(30) Nygaard, R.; Valentin-Hansen, L.; Mokrosinski, J.; Frimurer, T. M.; Schwartz, T. W. Conserved Water-mediated Hydrogen Bond Network between TM-I, -II, -VI, and -VII in 7TM Receptor Activation*. *Journal of Biological Chemistry* **2010**, *285*, 19625–19636.

(31) Gössweiner-Mohr, N.; Siligan, C.; Pluhackova, K.; Umlandt, L.; Koefler, S.; Trajkovska, N.; Horner, A. The Hidden Intricacies of Aquaporins: Remarkable Details

in a Common Structural Scaffold. *Small* **2022**, *18*, 2202056.

(32) de Groot, B. L.; Grubmüller, H. Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science* **2001**, *294*, 2353–2357.

(33) Pluhackova, K.; Wassenaar, T. A.; Böckmann, R. A. In *Membrane Biogenesis*; Rapaport, D., Herrmann, J. M., Eds.; Methods in Molecular Biology; Humana Press, 2013; Vol. 1033; pp 85–101.

(34) Wassenaar, A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S.-J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **2014**, *10*, 676–690.

(35) Rzepiela, A. J.; Schäfer, L. V.; Goga, N.; Risselada, H. J.; De Vries, A. H.; Marrink, S. J. Reconstruction of atomistic details from coarse-grained structures. *Journal of computational chemistry* **2010**, *31*, 1333–1343.

(36) Brocos, P.; Mendoza-Espinosa, P.; Castillo, R.; Mas-Oliva, J.; Pineiro, Á. Multiscale molecular dynamics simulations of micelles: coarse-grain for self-assembly and atomic resolution for finer details. *Soft Matter* **2012**, *8*, 9005–9014.

(37) Heath, A. P.; Kavraki, L. E.; Clementi, C. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins: Structure, Function, and Bioinformatics* **2007**, *68*, 646–661.

(38) Li, M.; Teng, B.; Lu, W.; Zhang, J. Z. Atomic-level reconstruction of biomolecules by a rigid-fragment-and local-frame-based (RF-LF) strategy. *Journal of Molecular Modeling* **2020**, *26*, 1–14.

(39) Vickery, O. N.; Stansfeld, P. J. CG2AT2: an enhanced fragment-based approach for serial multi-scale molecular dynamics simulations. *BioRxiv* **2021**,

(40) An, Y.; Deshmukh, S. A. Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. *Chem. Commun.* **2020**, *56*, 9312–9315.

(41) Li, W.; Burkhart, C.; Polińska, P.; Harmandaris, V.; Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *The Journal of Chemical Physics* **2020**, *153*, 041101.

(42) Louison, K. A.; Dryden, I. L.; Laughton, C. A. GLIMPS: A Machine Learning Approach to Resolution Transformation for Multiscale Modeling. *Journal of Chemical Theory and Computation* **2021**,

(43) Peng, J.; Yuan, C.; Ma, R.; Zhang, Z. Backmapping from multiresolution coarse-grained models to atomic structures of large biomolecules by restrained molecular dynamics simulations using Bayesian inference. *Journal of chemical theory and computation* **2019**, *15*, 3344–3353.

(44) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **2019**, *5*, 1–9.

(45) Stieffenhofer, M.; Wand, M.; Bereau, T. Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *Machine Learning: Science and Technology* **2020**, *1*, 045014.

(46) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(47) Landau, L. D.; Lifshitz, E. M. *Statistical Physics: Volume 5*; Elsevier, 2013; Vol. 5.

(48) James, G.; Witten, D.; Hastie, T.; Tibshirani, R., et al. *An introduction to statistical learning*; Springer, 2013; Vol. 112.

(49) 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. *Journal of machine learning research* **2011**, *12*, 2825–2830.

(50) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **2020**, *17*, 261–272.

(51) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming* **1989**, *45*, 503–528.

(52) Behnel, S.; Bradshaw, R.; Citro, C.; Dalcin, L.; Seljebotn, D. S.; Smith, K. Cython: The best of both worlds. *Computing in Science & Engineering* **2011**, *13*, 31–39.

(53) Kunzmann, P.; Anter, J. M.; Hamacher, K. Adding hydrogen atoms to molecular models via fragment superimposition. *Algorithms for Molecular Biology* **2022**, *17*, 1–8.

(54) Westbrook, J. D.; Shao, C.; Feng, Z.; Zhuravleva, M.; Velankar, S.; Young, J. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* **2015**, *31*, 1274–1278.

(55) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

(56) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd. 1996; pp 226–231.

(57) Steinley, D. Properties of the hubert-arable adjusted rand index. *Psychological methods* **2004**, *9*, 386.

(58) Kashyap, R. The perfect marriage and much more: Combining dimension reduction, distance measures and covariance. *Physica A: Statistical Mechanics and its Applications* **2019**, *536*, 120938.