

Swarm-CG: Automatic Parametrization of Bonded Terms in Coarse-Grained Models of Simple to Complex Molecules via Fuzzy Self-Tuning Particle Swarm Optimization

Charly Empereur-Mot^{*,1}, Luca Pesce¹, Davide Bochicchio¹, Claudio Perego¹ & Giovanni M. Pavan^{*,1,2}

1. *Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Galleria 2, Via Cantonale 2c, CH-6928 Manno, Switzerland.*

email: charly.empereur-mot@supsi.ch

2. *Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.*

email: giovanni.pavan@polito.it

Abstract

We present *Swarm-CG*, a versatile software for the automatic parametrization of bonded parameters in coarse-grained (CG) models. By coupling state-of-the-art metaheuristics to Boltzmann inversion, *Swarm-CG* performs accurate parametrization of bonded terms in CG models composed of up to 200 pseudo-atoms within 4h-24h on standard desktop machines, using an AA trajectory as reference and default settings of the software. The software benefits from a user-friendly interface and two different usage modes (default and advanced). We particularly expect *Swarm-CG* to support and facilitate the development of new CG models for the study of molecular systems interesting for bio- and nano-technology. Excellent performances are demonstrated using a benchmark of 9 molecules of diverse nature, structural complexity and size. *Swarm-CG* usage is ideal in combination with popular CG force fields, such as *e.g.* MARTINI. However, we anticipate that in principle its versatility makes it well suited for the optimization of models built based also on other CG schemes. *Swarm-CG* is available with all its dependencies via the Python Package Index (PIP package: *swarm-cg*). Tutorials and demonstration data are available at: www.github.com/GMPavanLab/SwarmCG.

1. Introduction

In many research fields, innovation passes through the design and development of new types of functional materials and molecular systems with controllable properties. The shape and functions of such complex nanostructures typically originates from the collective behavior of a large number of interacting molecules, as it is the case *e.g.* in lipid membranes,^{1,2} supramolecular polymers,^{3–5} crystals,^{6–8} cages,^{9–11} etc. The investigation of these molecular systems at a sufficiently high (submolecular) resolution is most often a prohibitive task both experimentally and computationally, especially when these are composed of large, soft and flexible macromolecules in solution.

Alongside with experimental studies, molecular modeling techniques such as Molecular Dynamics (MD) or Monte-Carlo simulations have turned out to be cornerstone tools to this purpose.^{12–20} Recent advances in computational hardware and simulation software have made possible to study and model increasingly larger molecular systems, allowing the investigation of their structural properties with great (atomistic-level) detail. However, the large number of degrees of freedom of these calculations still limits classical all-atom MD simulations (AA-MD) to the study of systems with a maximum of $\sim 10^6$ atoms (including the solvent, in *e.g.* explicit solvent simulations) and within the timescales of nano- to microseconds.^{12,13} Furthermore, AA-MD may typically suffer of limited sampling, especially in the simulation of complex molecules, with the risk of entrapment and oversampling of local minima and metastable states.^{21,22} As a consequence, AA-MD simulations cannot be practically employed for the observation of many crucial phenomena and molecular events occurring on long characteristic timescales.

A typical approach to overcome these limitations is coarse-graining, which consists in simplifying the description of the molecular model, reducing the resolution by grouping several atoms in coarse-grained (CG) beads (aka pseudo-atoms). The objective of CG modeling is to reduce the number of degrees of freedom to be treated in the simulations, while still providing a physically-relevant representation of the molecular systems. Different CG frameworks have gained popularity by allowing to simulate complex molecular systems, and their dynamical properties, such as lipid bilayers,^{23,24} vesicles,^{24–26} proteins,^{27–29} and various types of nanomaterials.^{14,30} Perhaps one of the most widely used CG schemes is the MARTINI force field,³¹ which maps molecular fragments composed of ~ 3 –4 heavy atoms into each CG bead. In principle, in this scheme each CG bead interacts with the others in the system (solute-solute and solute-solvent interactions) via a non-directional 12-6 Lennard-Jones (LJ) potential, parametrized according to the estimated partitioning of its associated molecular fragment between aqueous and hydrophobic environments. MARTINI provides a predefined set of CG beads spanning a range of polarities, together with a matrix of LJ interactions between bead types (*i.e.* the MARTINI force field). This offers a relative transferability of the FF, which is useful to create CG molecular models for various types of molecules, from biomolecules such as lipids,^{32–34} peptides and proteins,^{35,36} to synthetic molecules such as polymers,^{37,38} fullerenes,^{39,40} etc.

In the MARTINI scheme, building an adequate CG model for a molecule of interest requires to map the constitutive molecular fragments in the AA model to CG beads, which types are opportunely chosen based on chemical analogy and polarities of the groups, setting up the non-bonded interactions in the CG model. The non-bonded parameters can then be refined to obtain accurate pairwise interactions between the molecular species in the system. For example, enhanced sampling methods such as umbrella sampling^{41,42} or metadynamics^{43,44} can be applied to calculate dimerization free-energies between pairs of molecular species in the CG vs. AA models, which allows to optimize the CG models by matching the two observables.^{14,30,45} When available, experimental data can also be exploited.^{12,46} Then, the user has to determine appropriate intramolecular bonded interactions between CG beads, namely the bonds, angles and dihedrals parameters, in terms of equilibrium values and force constants.^{31,46} Noteworthy, while the optimization of bonded parameters may seem to some extent less important than that of non-bonded ones for the correct modeling of inter-molecular interactions, it is in principle just as relevant, since

properties such as molecular flexibility, shape and size (directly controlled by the bonded terms) are crucial for how the molecules interact between them and with the surroundings (also solvent).

The MARTINI force field requires users to manually tune the bonded parameters specifically for their molecular models.^{12,46} To optimize the bonded interactions of a CG model, usual practice is to first perform a (well-sampled) AA-MD simulation of the molecule at the relevant conditions of temperature, pressure, solvent, etc., using a reliable AA force field.^{12,31,46} The resulting AA-MD trajectory can then be mapped to CG beads (hereinafter referred to as the “AA-mapped” trajectory) and used as a reference to tune bonds, angles, and dihedrals parameters in the CG model, to ensure that the geometric features (*i.e.* bonds, angles and dihedrals distributions) are consistent with the AA model. To this purpose, multiple MD simulations of the CG model are performed iteratively, while manually tuning the bonded parameters in a trial-and-error approach. This task makes the development of reliably CG models a time-consuming process, especially for complex molecular architectures.^{12,46} For molecules composed of 20 to 200 CG beads and containing symmetrical, partially symmetrical, flexible or planar parts, efficient and automatic tools for the parametrization of bonded terms would be of great help, making the process more robust and reliable.^{12,47} Recent efforts towards automatized CG models parametrization have mostly focused on the refinement of non-bonded parameters,^{48–50} leveraging on the current diffusion of machine learning techniques. To the best of our knowledge, at present only two software implement automatic approaches to assist the bonded parametrization of CG models: *AutoMARTINI*⁵¹ and *PyCGTOOL*.⁵² Although these tools are well suited to treat small, drug-like molecules, they suffer of limited applicability to larger molecular systems. *AutoMARTINI*⁵¹ can be applied to molecules up to 20 heavy atoms maximum. *PyCGTOOL*⁵² does not suffer from such a strict limitation, but relies exclusively on a single-pass Boltzmann inversion,^{53,54} which considerably limits its accuracy when applied to larger complex molecules.

Here we introduce *Swarm-CG*, a general and easy-to-use tool that combines Boltzmann inversion^{53,54} (BI) and Particle Swarm Optimization^{55,56} (PSO) to automatically parametrize bonded interactions in CG models, within CG frameworks such as the MARTINI force field. The methods need only a reference AA-MD trajectory and a preliminary mapped CG model of the molecule of interest. *Swarm-CG* makes a first guess of the equilibrium bonded parameters via a single-pass BI, then automatically refines them via iterative CG-MD runs and PSO, until the distributions of the bonds, angles and dihedrals in the CG model are in good agreement with those of the AA model. To prove the robustness of this approach, we challenged *Swarm-CG* on a diverse molecular dataset including small to large molecules of different nature and shape: (i) flexible and symmetric self-assembling monomers generating supramolecular polymers in solution: water-soluble 1,3,5-benzenetricarboxamide (BTA) with amphiphilic side chains,⁵⁷ C3-symmetric benzotrithiophene (BTT) decorated by L-phenylalanine (BTT-F) and octaethylene glycol side-chains,⁵⁸ naphthalene diimide (NDI)⁵⁹ and Zn-porphyrin based self-assembling monomers,⁴⁵ (ii) cyclic structures: β -Cyclodextrin⁶⁰ and pillar[5]arene,⁶¹ and (iii) complex hyper-branched polymers: a spermine dendron⁶² and poly(amidoamine) dendrimers of generation 1 and 2 (PAMAM G1 and G2)^{63–66} (Figure 1). Benchmarking results demonstrate that *Swarm-CG* readily performs comparably to expert molecular modelers and systematically yields CG models that exhibit reliable behavior in solvent environment, within 4h-24h on standard desktop machines (wall time). Notably, such execution times allow to quickly explore different CG representations of the molecule of interest using different AA-to-CG (atoms-to-beads) mappings and topologies. The approach is perfectly suited for building and optimizing CG models based on widely used CG frameworks such as MARTINI. At the same time, *Swarm-CG* workflow is general and can be applied in principle to any CG framework and any CG passage through scales.

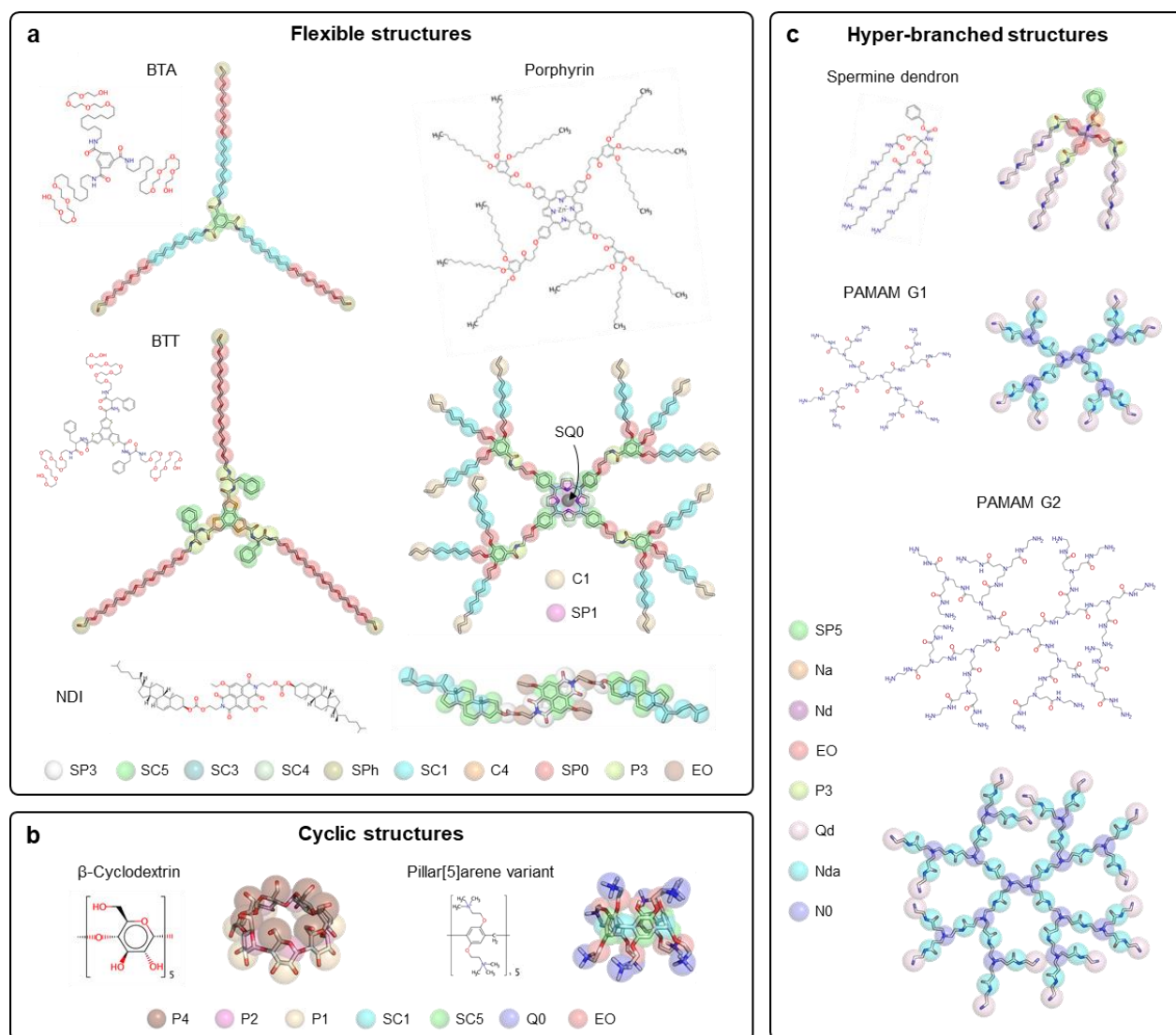


Figure 1. Molecules used to benchmark *Swarm-CG*. Each molecule is represented by its molecular structure and AA model with superimposed CG MARTINI beads mapping. (a) Flexible and symmetric molecular structures generating supramolecular polymers: water-soluble 1,3,5-benzenetricarboxamide (BTA) with amphiphilic side chains,⁵⁷ C3-symmetric benzotrithiophene (BTT) decorated by L-phenylalanine (BTT-F) and octaethylene glycol side-chains,⁵⁸ naphthalene diimide (NDI)-based⁵⁹ and Zn-porphyrin-based molecules.⁴⁵ (b) Examples of cyclic structures: β -Cyclodextrin⁶⁰ and a pillar[5]arene.⁶¹ (c) Complex hyper-branched polymer structures: spermine dendron⁶² and poly(amidoamine) dendrimers of generation 1 and 2 (PAMAM G1 and G2).^{63–66} Each panel indicates the color coding of the CG MARTINI beads types (see SI for exact mapping data).

2. Algorithm

The algorithm implemented in *Swarm-CG* is designed to automatically optimize the parameters of bonded interactions in a CG molecular model, namely the parameters of the potential functions used by the force field for bonds, angles and dihedrals in a user-provided CG molecular topology file. The software needs the user to preliminary define the AA-to-CG mapping and choose the CG bead types (defining the non-bonded interactions). The functional form of bonded interaction potentials can generally be described as:

$$V_{bonded} = \sum_i^{bonds} \frac{1}{2} k_{bi} (l_i - l_{0i})^2 + \sum_j^{angles} \frac{1}{2} k_{aj} (\theta_j - \theta_{0j})^2 + \sum_k^{dihedrals} \frac{1}{2} k_{dk} (1 \pm \cos(n_k \phi_k - \phi_{0k})) \quad (1)$$

where the first term is the potential associated to the length l_i of each bond, the second term is the potential of the angles θ_j and the third term is the potential of the dihedral angles ϕ_k . k_{bi} , k_{aj} and k_{dk} are respectively the bond, angle and dihedral force constants, l_{0i} indicates the equilibrium bond lengths, θ_{0j} the equilibrium angle values, n_k the periodicity and ϕ_{0k} the phase shift of each dihedral.

In the MARTINI framework, an *AA-mapped* trajectory of the target molecule in the solvent, at a chosen thermodynamic state (given conditions of temperature, pressure, etc.), is used as reference to tune bonded parameters (BP), namely the parameters in Eq. 1, and to obtain matching distributions of the bond lengths, angle values and dihedral torsions (hereafter referred to as “*geoms*”) in the CG representation. Usually, the tuning of BP in the CG models is performed manually, by repeating simulations of the CG system until the agreement between CG and *AA-mapped* models is deemed satisfactory.

Provided that a well-sampled AA-MD reference trajectory is available, the bonded parametrization of CG models is essentially an optimization problem, which requires to iterate CG-MD while the BP are optimized. To free molecular modelers from manually performing this time-consuming operation, *Swarm-CG* is built around a state-of-the-art metaheuristic, Fuzzy-Self Tuning Particle Swarm Optimization (FST-PSO),⁶⁷ which is employed to iteratively tune the BP to improve geometrical features of the CG model throughout successive CG-MD simulations. At each iteration step (*i.e.*, in each successive CG-MD run), a scoring function evaluates the current set of BP by comparing the resulting *geoms* distributions with those of the *AA-mapped* reference trajectory. After a defined number of iteration steps, the best matching set of BP is selected. The workflow implemented in *Swarm-CG* is summarized in Figure 2. The following sections describe the protocol and heuristics implemented in *Swarm-CG* that allowed to make the software parameters-free and versatile to deal with different modeling requirements and designs.

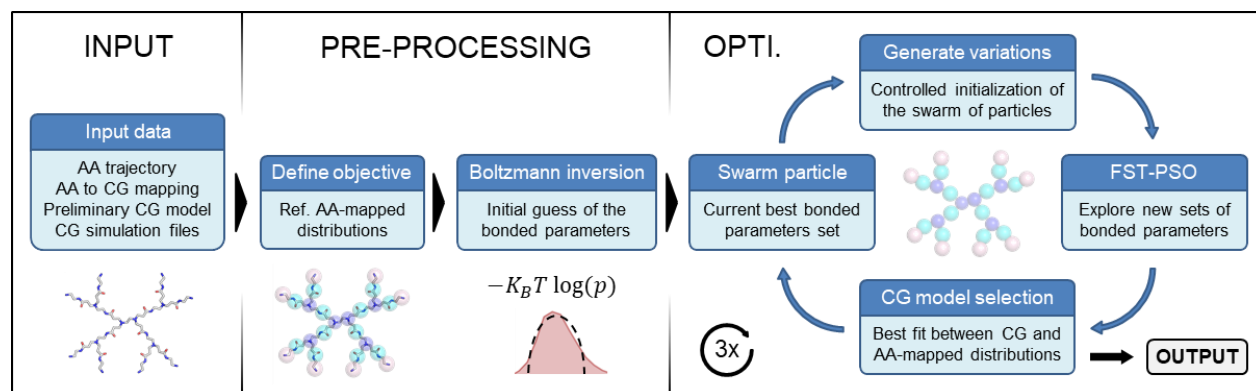


Figure 2. General workflow of *Swarm-CG*. This can be schematized into three phases. (i) Preparation of the input: the software requires a reference AA-MD trajectory, a pre-defined AA-to-CG mapping and a preliminary CG model, where the non-bonded interactions are pre-defined (CG bead types and interactions), (ii) Pre-processing: an *AA-mapped* reference model is built, computing the bond, angle and dihedral distributions of the reference *AA-mapped* MD trajectory, and an initial guess of bonded CG parameters is made (to be then optimized). (iii) Optimization process: Iterative CG-MD simulations are performed, while at each iteration *Swarm-CG*, starting from a “swarm particle” (a set of BP), changes the BP to optimize the consistency with the reference AA-MD trajectory. The resulting set of CG bond parameters is then obtained as the output.

2.1. Input

Swarm-CG is designed for usage with the GROMACS^{68,69} MD engine. In this paper we demonstrate *Swarm-CG* performances to optimize CG models built based on the well-known MARTINI force field. However, the workflow of *Swarm-CG* is general and it can be used for refining basically any CG model, provided that non-bonded parameters and a mapping scheme are defined, and a reliable reference trajectory is available. The necessary input can be divided in two groups: (i) AA data used to define the target of the optimization and (ii) preliminary CG data used to perform the model optimization.

The AA input data (i) include a well-sampled MD trajectory of the AA molecular model to be used as reference and its pre-defined mapping to CG beads. We note that while a few automatic AA-to-CG mapping schemes have been already proposed (*e.g.* in the MARTINI formalism),^{51,52} these typically work only for small molecules. The search of methods suggesting the best CG representation for accurately treating the dynamics and structural features of molecules is a subject of great scientific debate.^{70–73} Here, for sake of a broader practical utility of *Swarm-CG*, we preferred to leave the AA-to-CG mapping to the user, who is free to choose the preferred CG scheme (the MARTINI force field or other pre-set schemes),^{18,74,75} while the software will optimize the bonded terms accordingly.

Input CG data (ii) include a preliminary CG model, together with its non-bonded force field parameters, and simulation setup for the iterative MD simulations that will be used for the model refinement: *i.e.* the starting molecular configuration and the MD parameter files (*cf.* section 6.1). The starting molecular configuration will be minimized and pre-processed at each iterative MD simulation step using new sets of BP. The preliminary CG model needs to contain relevant information on the CG beads (*e.g.* type, charge, mass), the bonded potential topology and functional forms (which define the form of Eq. 1), while equilibrium values and force constants are arbitrarily initialized (*e.g.* to 0). Symmetries of the CG topology can be specified in the preliminary model file to improve the quality of the reference AA sampling and to reduce the number of free parameters to optimize. To this end, the bonds (or angles, dihedrals) that are structurally equivalent due to molecular symmetries can be gathered in “groups” so that: (i) their distributions are averaged in the analysis and (ii) they share the same BP (*cf.* section 6.1). Groups of *geoms* are directly indicated by the user in the preliminary model file. *Swarm-CG* provides detailed documentation and uses a set of default filenames for easier arguments handling.

2.2. Scoring function

To attribute a score to the BP set of a CG model (namely, how good/bad this performs compared to the reference AA model), the reference AA trajectory is first mapped to its CG representation to generate a “target/reference” *AA-mapped* trajectory, that the optimized CG model aims at reproducing. In this perspective, the geometrical features of the CG model can be evaluated by comparing the CG-MD trajectory to the *AA-mapped*, on two scales: (i) “global” structural molecular properties, *e.g.* the radius of gyration (Rg) and solvent accessible surface area (SASA) and (ii) “local” conformation and flexibility, which can be assessed *via* the distributions of *geoms*. Since multiple sets of BP can produce similar Rg or SASA values, it is not possible to directly use such global structural properties as feedback for the optimization process, as the results would be locally inaccurate. Therefore, *Swarm-CG* uses a scoring function based on the differences between the corresponding *geoms* distributions obtained from the CG and *AA-mapped* trajectories (reported in Figure 3a). Differences are evaluated using the Earth Mover’s Distance⁷⁶ (EMD, aka Wasserstein metric), which solves the optimal transport problem⁷⁷ to quantify the amount of “work” necessary to transform one distribution into another. The set of BP selected by *Swarm-CG* as the outcome of the optimization process is the one that minimizes the scoring function, while Rg and SASA are monitored during the execution and ultimately used for *a posteriori* model validation.

In the present context, using the EMD offers several advantages over other f-divergences. Notably, the EMD: (i) quantifies the difference between *geoms* distributions in interpretable units (Å, degrees), (ii) is well suited for comparison of multimodal distributions in this application case and (iii) it allows to correctly

handle dihedrals distributions by using a periodic distance matrix. A scaling factor C is applied to EMD obtained for bonds distributions to allow comparison with EMD obtained for angles and dihedrals, since units are different.

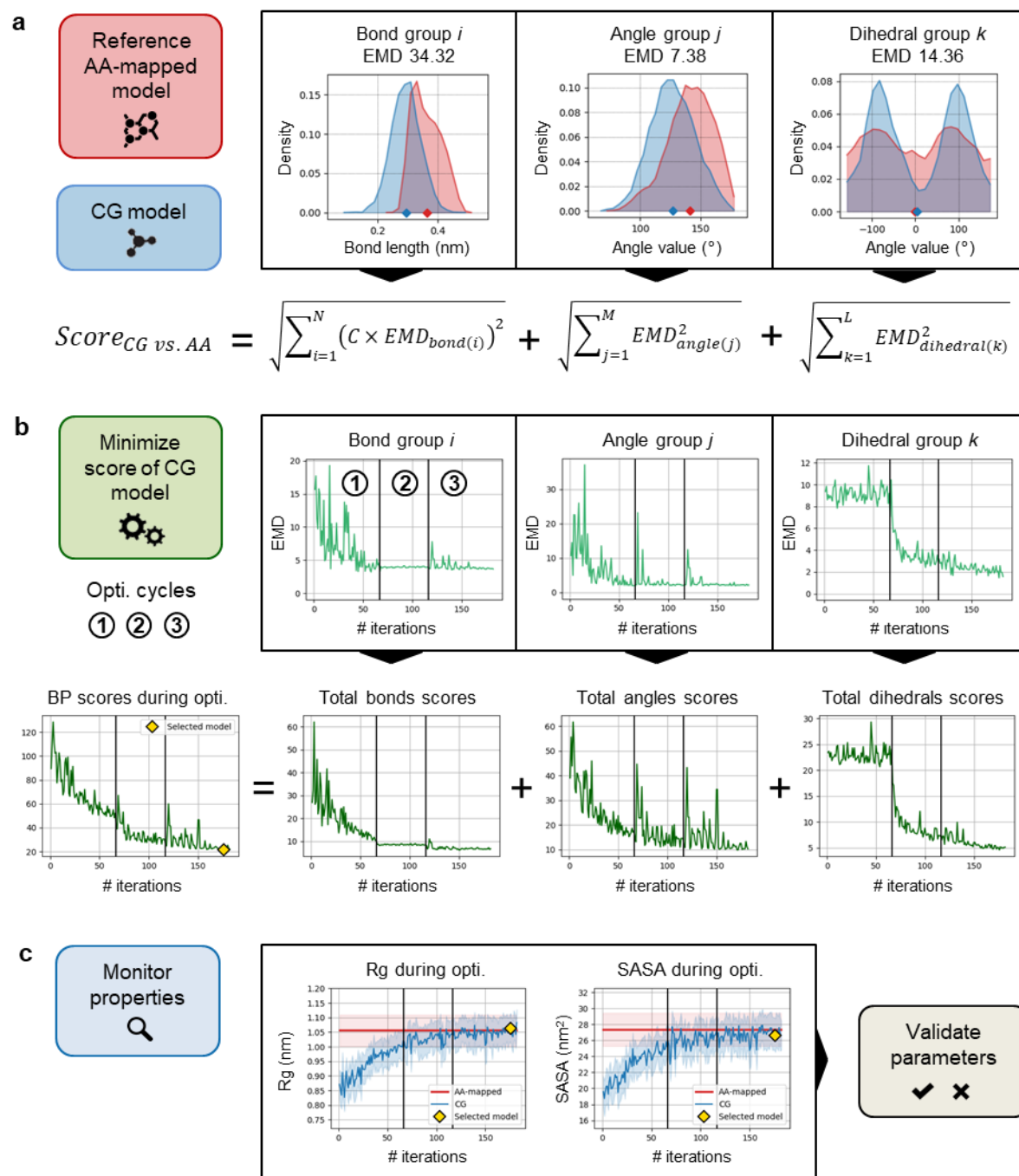


Figure 3. Overview of the scoring function and iterative optimization procedure used in *Swarm-CG* to automatically tune the BP of a preliminary CG model (using illustrative data). (a) Single model scoring: the scoring function evaluates the matching between pairwise distributions of N groups of bonds, M groups of angles and L groups of dihedrals from CG vs. AA models trajectories using the Earth Mover's Distance. C is a scaling factor applied to EMD of bonds. (b) Iterative model optimization: The procedure generates new sets of BP to minimize the differences between CG and reference AA-mapped distributions. (c) Quality control: Radius of gyration (Rg) and solvent accessible area (SASA) monitored during optimization.

By default, we set $C = 50$, meaning that an EMD of 0.4 Å between bond distributions is equivalent to an EMD of 20 degrees between angle or dihedral distributions. To better penalize large mismatches between distributions and respect the weight of each *geom*, we do not normalize score components by the number of *geoms* defined in the topology. Therefore, it is important to note that the scores can be compared exclusively between trajectories of CG models generated in similar conditions (*i.e.*, identical topology and non-bonded interactions parameters, but also simulation parameters) and with respect to a (well-sampled) reference AA-mapped trajectory. The components of the scoring function can be considered separately to exclusively evaluate the matching of bonds, angles or dihedrals distributions during the optimization procedure. *Swarm-CG* performs EMD calculations via PyEMD.^{76,78}

2.3. Iterative optimization procedure

PSO^{55,56} is a population-based global optimization algorithm (aka metaheuristic) inspired by the collective movement of birds flocks and fish schools. In PSO, a swarm of individuals (referred to as “particles”, each representing a set of values to be optimized) moves iteratively inside a bounded search space and cooperates to identify the best solution for a problem, according to an objective function. Usually, there are two groups of settings in PSO that control the cooperation within the swarm: (i) social attraction, which favors the collaboration among particles and (ii) cognitive attraction, which prompts a particle to rely on its individual experience. The swarm of particles can be initialized either randomly or from known approximate solutions. Metaheuristics such as PSO are particularly suited for solving black-box optimization problems and effectively handle noisy data.

To refine BP of CG models, *Swarm-CG* relies on FST-PSO,⁶⁷ a recently introduced PSO variant. FST-PSO exploits fuzzy logic to dynamically adjust PSO settings independently for each particle during optimization, making it a more efficient, parameters-free and versatile PSO variant.⁶⁷ Nonetheless, the performance of all PSO algorithms is greatly affected by the initial positioning of the swarm in the search space.⁵⁶ If the initial candidate solutions are positioned close to the basin of attraction of a local minimum of the objective function, the swarm might converge prematurely and be unable to move out of that region. To systematically achieve global optimization while minimizing execution times, *Swarm-CG* uses an iterative procedure that include 3 successive optimization cycles calibrated to complement each other (Figure 3b). Notably, BP of the CG model are optimized from higher to lower *geoms* vibrational frequencies. Exclusively bonds and angles are tuned in cycle 1. Angles and dihedrals are then optimized in cycle 2. Finally, all parameters are refined altogether in optimization cycle 3 (see Table 1). Accordingly, the scoring function is adapted for each cycle to include relevant components exclusively. *Swarm-CG* also calibrates each initialization of the swarm of particles to maximize FST-PSO performances. At start of cycle 1, initialization is performed using a single-pass BI (*cf.* SI section 1.1) to guess BP of the CG model, for one swarm particle, which is used as a reference to generate variations and initialize the rest of the swarm. At start of cycles 2 and 3, the best set of BP obtained in previous cycles is chosen as a reference particle to generate the rest of the swarm. For each bonded parameter, variations around the reference particle are generated randomly within adaptive ranges, which are decreased as the procedure progresses through optimization cycles (Table 1). Adaptive ranges also take into account the EMD previously obtained for each pairwise CG and AA-mapped *geoms* distributions, which directs the optimization procedure towards reducing first the largest discrepancies between models (*cf.* SI section 1.2). The two first optimization cycles allow a quick exploration of relevant sets of BP using short simulation times (10 ns by default), while the third optimization cycle uses longer simulation times to perform a final merging and refinement step (25 ns by default).

Table 1. Default settings used to perform 3 cycles of bonded parameters optimization of a CG model in *Swarm-CG*.

Opti. cycle	<i>Geoms</i> optimized			Reference swarm particle initialization	Variations around reference swarm particle	Simulation time of production runs ^(c)
	Bonds	Angles	Dihedrals			
1	Yes	Yes	No ^(b)	BI	Large	10 ns
2	No ^(a)	Yes	Yes ^(b)	Best from cycle 1	Medium	10 ns
3	Yes	Yes	Yes ^(b)	Best from cycle 2	Small	25 ns

^(a) In cycle 2, bonds parameters are fixed to those of the best scored model obtained during cycle 1.

^(b) Dihedral parameters are applied for simulation and optimized only in cycles 2 and 3, if dihedrals topologies are provided in the input preliminary CG model.

^(c) Default settings, simulation times can be increased by the user for very large molecules.

In all PSO algorithms, the procedure terminates after a pre-defined number of steps or when improvements over the objective function become minimal. To allow a parameters-free usage of *Swarm-CG*, simple heuristics enable automatic selection of a relevant swarm size and number of swarm iterations to perform in each cycle of optimization. Swarm size (S_{size}) is defined according to the dimension (D) of the search space as $S_{size} = 2 + \sqrt{D}$ and number of swarm iterations (S_{iter}) as $S_{iter} = 8 + \sqrt{D/2}$. An optimization cycle is terminated prematurely if no improvement occurred within 6 swarm iterations. Default settings readily allow to perform accurate bonded parametrization of virtually any CG model, as long as the provided topology and potential functions are relevant. The accuracy and execution times of *Swarm-CG* are expected to satisfy molecular modelers requirements for optimization of up to approximately 100 free parameters,⁶⁷ which represent approximately 50 groups of bonds, angles and dihedrals (*i.e.*, many more *geoms* in symmetrical molecules). Beyond that, users can easily access *Swarm-CG* parameters, for example, to increase the number of optimization steps or add more optimization cycles.

2.4. Execution modes

The software provides two execution modes, which conform to two different CG modeling philosophies. Using execution *mode 1*, all equilibrium values (l_{oi} , θ_{oj} and ϕ_{ok} in Eq. 1) are optimized together with the force constants (k_{bi} , k_{aj} and k_{dk} in Eq. 1), for each group of bonds, angles and dihedrals. This procedure, based on the BI philosophy, allows for a fully automatic and relatively easy usage of the software, which precisely reproduces *geoms* distributions from an AA-mapped reference trajectory. However, using *Swarm-CG* in execution *mode 1* as a black box may also have undesired effects. For instance, an insufficient conformational sampling in the reference AA-MD trajectory may automatically introduce artifacts in the optimized CG model. Indeed, poor sampling can attribute excessive statistical weight to some molecular conformation, which will affect the resulting CG model and limit its accuracy. For example, the folding of flexible hydrophobic molecules in polar solvents into metastable compact conformations may be typically oversampled in AA-MD simulations. While the folding is a consequence of solvophobic interactions, it may result in a CG model in which the output BP encode the bending of straight linear chains (*e.g.* long alkyl groups, formed by a straight chain of CG beads) in the form of spurious angles equilibrium values (different from *e.g.* 180°), producing shorter bonds, etc. Similarly, limited sampling can affect the modeling of symmetric molecules (*e.g.* the branched molecules of Figure 1), by enforcing non-symmetric parameters that emerge by oversampled local minima, in contradiction with the chemical structure of the molecule. In such cases, this may eventually result into having an “effect” emerging from the AA models and encoded into the bonded parameters of the CG models, which may then affect the way molecules interact between them, their flexibility, transferability across different molecular environments, etc. However, *Swarm-CG* is well equipped to mitigate such spurious effects of limited MD sampling, by averaging in the AA-mapped reference the distributions obtained for structurally symmetric/identical parts of the molecule. It is also

worth underlining that such possible issues emerging from using a not properly sampled AA-MD trajectory as the reference is not specific of *Swarm-CG*, but rather a general drawback of the BI approach. Therefore, while BI is standard and widely used, one should always be careful and check that the AA-MD trajectory is sufficiently well-sampled to ensure that the observed properties of the optimized CG models are reliable. While enhanced sampling approaches such as Replica Exchange Molecular Dynamics (REMD)^{79,80} and metadynamics^{43,44} may be useful in this sense, said issues can be mitigated by using the second execution mode (*mode 2*) of *Swarm-CG*.

Execution *mode 2* is identical to *mode 1*, with the exception that equilibrium values for angles and dihedrals can be pre-defined in the preliminary CG model and conserved during the optimization, while only their force constants are optimized (along with all bonds parameters). For example, in the case of flexible molecules containing long solvophobic chains (*e.g.* alkyl chains in water), execution *mode 2* allows to manually pre-define chemically-relevant equilibrium angles values between the alkyl CG beads (*e.g.* 180°) and to obtain an accurate folding propensity of the molecule (*i.e.* the correct folding effect) exclusively by softening angles force constants, without biasing the equilibrium conformation (and without encoding such bias in the optimized CG models). Compared to *mode 1*, *mode 2* requires more experience of the user and some knowledge of the molecular system, as well as an initial hands-on setup.

2.5. Usage

Swarm-CG allows users to quickly verify the setup and progress of an optimization procedure. At the outset of an optimization, *Swarm-CG* produces a graphical summary of the *geoms* distributions used as target for optimization, notably allowing to verify the consistency of the mapping and choice of potential functions. At any point during the iterative execution process, the best identified set of BP is readily provided as an output in a CG model ITP file, while the progress of the procedure can be monitored by producing a graphical summary similar to the one presented in Figure 3. Separate modules allow to perform these actions independently from the optimization. For manual editing of models, such as further modifying CG bead types and evaluating impact on bonded parametrization, the routine for model evaluation via scoring function is also available as a separate module.

3. Results

The following sections describe the results obtained for automatic bonded parametrization of CG models included in the *Swarm-CG* benchmark (Figure 1), using default settings of the program. Here we focus on synthetic structures as: (i) being the core activity of our group, we have a good amount of available data to test the performance and accuracy of *Swarm-CG* and (ii) because it is for the simulation of synthetic molecular systems that the development of *de novo* AA and CG models from scratch is most often required, while the accuracy of such models is clearly critical for the reliability of the results that these can provide. This can be a time-consuming activity, in which the advantages of an automatic tool such as *Swarm-CG* combined with a general CG force field such as MARTINI are more evident. The benchmark synthetic molecules that we use herein were selected for their structural diversity in terms of molecular flexibility, symmetry and complexity (*cf.* section 6.2). We first created AA models and generated AA-MD trajectories (up to 1 μ s of simulation) for each single molecule in explicit solvent (while most of the cases studied herein are in water, the approach is versatile to treat molecules in various solvents, as it is shown in the case of the NDI and porphyrin based structures of Figure 1, studied in methyl cyclohexane),⁴⁵ guaranteeing well-sampled references for the automatic bonded parametrization of the CG models. CG models were already available from the literature for 7 of the 9 benchmarking molecules (BTA,⁵⁷ BTT,⁵⁸ NDI,⁵⁹ Zn-porphyrin based molecule,⁴⁵ β -Cyclodextrin,⁶⁰ PAMAM dendrimers of generation G1 and G2⁶³), which allowed us to challenge *Swarm-CG* performances with respect to manually parametrized CG-models previously developed by expert molecular modelers. To this end, we used the available CG models as

provided (notably for what pertains to mapping and non-bonded interactions) and allowed *Swarm-CG* to modify exclusively the equilibrium values and force constants of each bonded potential function defined in the available CG topology (*cf.* section 6.2). For the 2 other molecules (pillar[5]arene⁶¹ and spermine dendron⁶²), CG models were built in the framework of MARTINI and optimized using AA models previously reported by our group (*cf.* section 6.2). For all models, bonded parametrizations were evaluated using: (i) *Swarm-CG* scoring function, which assesses the local geometrical features of a CG model, together with (ii) Rg and SASA, which provide a global evaluation of its molecular dynamics.

We first tackle relatively small and flexible molecules forming supramolecular polymers in solution, that we use as examples to discuss in detail the differences between execution *mode 1* vs. *mode 2*. Then we report the results of *Swarm-CG* for the parametrization of relatively rigid cyclic molecular structures. Finally, we increase molecular complexity by parametrizing complex hyper-branched directional and non-directional macromolecules, such as dendrons and dendrimers. Since the simulation times used in the optimization runs (10-25 ns by default in the examples reported herein) might be insufficient in some cases to get well-converged Rg and SASA data at each step of the optimization process, Rg and SASA values presented for the selected (*i.e.* best scored) set of BP are all issued from 200 ns validation simulations that are conducted at the end of the optimization procedure. This also allowed to verify that all optimized models are stable in CG simulation using a standard integration time step of 20 fs. All average Rg values obtained for optimized CG models using execution *mode 1* are summarized in Table 2 and compared to the available manually parametrized CG models. The number of iterative optimization steps used for each model is determined according to the formula previously described in section 2.3, while execution times are reported in Table 2 (hardware is detailed in Table S2).

3.1. Small flexible molecules generating supramolecular polymers

Since we have a good benchmark of AA and CG models for (relatively) small and flexible molecules that generate supramolecular polymers in different environments,^{30,45,57–63,81} we started from here in showing the potential of *Swarm-CG*. We chose the examples reported in Figure 1a. These are relatively flexible molecules that generate supramolecular structures in water (*e.g.* BTA and BTT)^{57,58} or in organic solvents (*e.g.* BTA, Porphyrin, NDI-based units).^{30,45,58,59} These molecules show an intrinsic symmetric character, having three (BTA and BTT), two (NDI) or four (Porphyrin) structurally identical arms originating from the molecule center. Thus, these are the typical motifs which, in a blind BI approach, may suffer of a spurious different parametrization of identical groups given by insufficient sampling. In this sense, they represent the ideal ground to test *Swarm-CG*. For these systems, we compared the results obtained using *Swarm-CG* execution *modes 1* vs. *mode 2*. As previously mentioned (*cf.* section 2.4), execution *mode 2* allows the user to pre-set conserved equilibrium values for angles and dihedrals, while *Swarm-CG* then optimizes the corresponding force constants to have the CG model behaving consistently with the reference AA model. For the CG modeling of BTA, BTT, NDI and porphyrin-based motifs, we relied on previously developed AA and CG models, where intramolecular non-bonded interactions were accurately tuned using state-of-the-art enhanced sampling techniques.^{14,30,45,57–59} Considering the non-bonded parameters of these CG models as reliable, we used *Swarm-CG* exclusively to optimize their bonded terms.

3.1.1. BTA

We first comment the optimization case of the CG model of the water-soluble BTA.⁵⁷ The main results of both execution modes are reported in Figure 4a. Geometrical features of the optimized models are compared to both those calculated from AA-mapped data and from the literature CG model³⁰ (*cf.* section 6.2).

Using execution *mode 1*, the single-pass BI coupled to distributions averaging within groups of similar *geoms* yielded an already appropriate set of BP at the very first step of the optimization process (Figure S2), here also validated by additional 200 ns of MD simulation. For BTA, which can be considered a

structurally (relatively) “simple” case with respect to the rest of the benchmark (*i.e.* composed of 3 core CG beads, linear side arms and a 3-fold symmetry), the BI implemented in *Swarm-CG* with *geoms* averaging proved very efficient. Optimization still reduced small mismatches in *geoms* distributions (Figure S4) and BP scores decreased from 23.3 to 16.1 without modifying average Rg of the CG model, which was found in good agreement with AA-mapped data for both sets of BP (*i.e.* both errors < 5%). The set of BP which obtained the lowest score during the optimization procedure is considered as the most relevant (Figure 4a, yellow diamonds), with respect to AA-mapped reference data, and was further validated in a 200 ns simulation. This longer simulation validates that the optimized CG model correctly reproduces both local and global geometrical features calculated from the AA-mapped trajectory. BP optimization converged within 257 steps (Table 2).

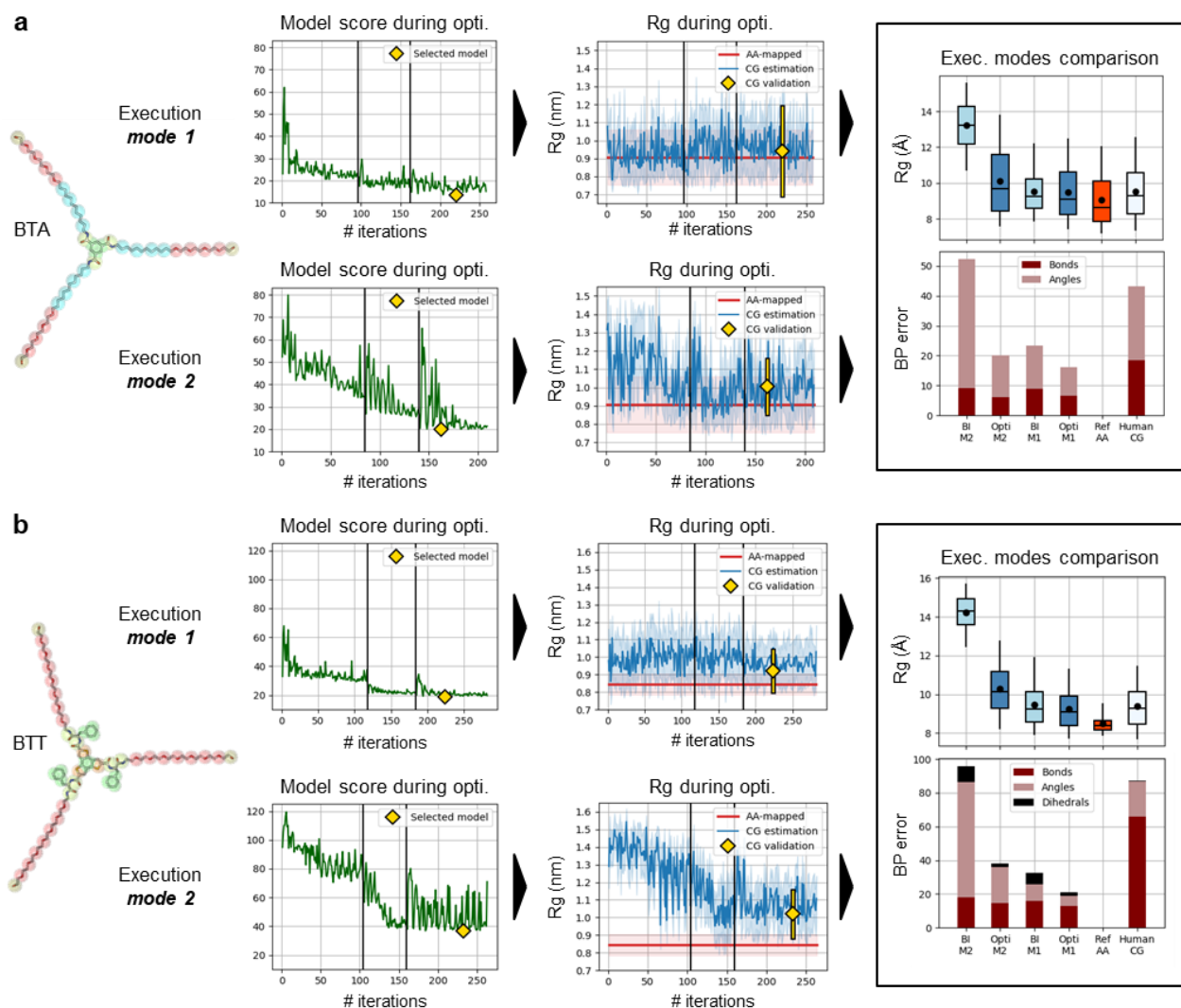


Figure 4. Results of *Swarm-CG* for the optimization of bonded parameters of two C_3 -symmetric flexible structures using execution modes 1 (M1) and 2 (M2) with default settings: (a) BTA model.⁵⁷ (b) BTT model.⁵⁸ From left to right we report: (i) the molecular structure, (ii) the evolution of the scoring function, where green lines show the score attributed to candidate BP during optimization. Yellow diamonds indicate the score of the selected model. (iii) the evolution of Rg, in which blue lines show average Rg estimates at each iteration of the CG model optimization (light blue intervals represent \pm standard deviation), and red horizontal lines show the average Rg of AA-mapped reference trajectories (light red intervals represent \pm standard deviation). Yellow diamonds and lines show averages and standard deviations obtained from 200 ns simulations. (iv) the comparison of Rg and BP errors in different models

in 200 ns simulations (BI: step 1, Opti: selected model). Boxplots and whiskers display percentiles 5, 25, 50, 75 and 95 of Rg values. Black dots show average Rg values. Stacked barplots show each component of the scoring function, the sum of which amounts to the BP score.

We also tested the manually parametrized CG model of BTA⁵⁷ in a 200 ns simulation. BP score and average Rg error for this model were 43.2 and 4.5% with respect to AA-mapped data (Figure 4a, right plots). Both BP sets obtained via BI and optimization using execution *mode 1* fixed small mismatches observed in local geometrical features of the manually parametrized model (Figure S6), without substantially improving the average Rg error.

Using execution *mode 2*, all equilibrium values for the angles between the CG beads representing the three side chains of the BTA were fixed at 180°. As expected, the BI initially produced an imperfect set of BP (Figure S3 and cf. SI section 1.1), which were quickly tuned as the optimization approached convergence. BP scores decreased from 52.2 to 20.0. The optimized set of BP produced more “loosely” adjusted overlap of the distributions for some angle groups that used equilibrium values at 180° (Figure S5, angle groups 7, 8 and 9), providing increased flexibility of the CG molecular model with respect to the optimized BP obtained via execution *mode 1* and AA-mapped data (i.e. average Rg increased just by 0.65 Å, or 7 points, reaching 10.1 Å). BP scores first decreased slowly during optimization cycle 1, then only angles distributions were further optimized during cycle 2 and all BP were refined during cycle 3 using longer simulation times. At start of cycles 2 and 3 the swarm is reinitialized around the best scored set of BP obtained in previous cycles, using calibrated variations (cf. section 2.3) which allowed to escape a local minima of the objective function and produced the fluctuations of BP scores observed after steps 95 and 160. BP optimization converged within 209 steps. The selected maximum number of optimization steps is reduced compared to execution *mode 1* because equilibrium angle values are provided by the user, reducing the dimensionality of the problem.

It is worth re-underlining that in this case (*mode 2*), the BTA folding is not pre-encoded in the CG model as bonded terms, but it is exclusively a consequence of the spontaneous collapse of the molecule in the solvent (hydrophobic effect, bead-bead interactions), which, in a sense, is more physically correct. However, it is also worth noting that such a comparison between *mode 1* and *mode 2* shows that the two modes work substantially the same in these cases, demonstrating that the behavior of these CG models is mainly controlled by the non-bonded interactions between the CG beads in the models more than by the bonded terms, so that the difference between the 2 execution modes is globally negligible in this case (see Figure 4).

3.1.2. BTT

We then optimized the CG model of the three-branched BTT⁵⁸ motif, again using both execution modes. The main results are reported in Figure 4b. Geometrical features of the optimized models are compared to both those calculated from AA-mapped data and from the literature CG model⁵⁸ (cf. section 6.2).

Using execution *mode 1*, the BI again yielded an appropriate set of BP at the very first step of the optimization process, with essentially a single group of dihedrals for which distributions were not perfectly adjusted (Figure S8). These were fixed during optimization (BP scores down from 31.7 to 20.8) and allowed to retrieve a correct planar geometry of the core of BTT, along with slightly better adjusted *geoms* distributions (Figure S10). Notably, the average and spread of Rg values obtained for the optimized CG model are just slightly larger respect to those of the AA-mapped reference trajectory (Figure 4b: ΔRg of 0.7 Å). Although the error is substantially negligible, this is consistent with the higher dynamicity of CG models compared to the AA ones⁸² (this is more evident in BTT, as this motif allows stronger core-to-arms and arms-to-arms interactions compared to e.g. BTA). BP optimization converged within 282 steps (Table 2).

We tested the manually parametrized CG model of BTT in a 200 ns simulation. BP score and average Rg error for this model were 87.4 and 10.5% with respect to AA-mapped data. The set of optimized BP

obtained using execution *mode 1* fixed small mismatches observed in local geometrical features of the manually parametrized model (Figure S12), while producing a very similar average and spread of Rg values. Using execution *mode 2*, as expected the BI initially produced an inaccurate set of BP (Figure S9 and *cf.* SI section 1.1) which was then tuned during optimization. BP scores decreased from 95.6 to 38.0. Again, the optimized BP produced more “loosely” adjusted distributions overlap for angle groups that used equilibrium values at 180° (Figure S11, angle groups 3 to 8), providing increased flexibility of the CG molecular model with respect to the model optimized via execution mode 1 and AA-mapped data (i.e. average Rg increased by 1.10 Å or 12 points, reaching 10.3 Å). BP optimization converged within 261 steps.

3.1.3. NDI

We performed the same study for the NDI-based molecules, the results of which are reported in Figure 5a and compared to the CG model available from the literature.⁵⁹ Using execution *modes 1 and 2*, the initial BIs yielded inappropriate sets of BP, notably due to the several dihedral potentials used to maintain the planarity of the molecular core (Figures S14, S15). During optimization with *mode 1*, BP scores decreased from 182.2 to 41.3 and allowed to retrieve a correct planar geometry of the core of NDI, with correctly adjusted *geoms* distributions (Figure S14). Average Rg value of the optimized model is in perfect agreement with the AA-mapped reference trajectory (Figure 5a: $\Delta R_g = 0.5$ Å or 4.8%). BP optimization converged within 395 steps (Table 2). As a comparison, the manually parametrized CG model of NDI yielded a higher BP score of 74.42 in a 200 ns simulation, indicating some *geoms* distributions could be better adjusted (Figure S18), although average Rg was correct at 11.7 Å ($\Delta R_g = 0.7$ Å or 6.8%).

Using execution *mode 2*, BP scores went down from 350.2 to 52.4 and produced *geoms* distributions almost indistinguishable from those obtained using *mode 1*, since only the hinge junctions between core and arms (Figures S17, angle groups 10 and 11) used angles at 180° and those where already “loosely” adjusted with *mode 1* (although *Swarm-CG* selected an equilibrium value at 120°). In both resulting models, the low force constant associated to the hinge junctions should not restrict the full extension of the arms, and thus is not expected to significantly affect assembly simulations, but these specific flexibilities must be considered carefully. Average Rg was 11.8 Å ($\Delta R_g = 0.8$ Å or 7.8%). BP optimization converged within 301 steps.

3.1.4. Porphyrin

Lastly, we optimized the CG model of the porphyrin-based molecule of Figure 5b, which possesses four arms originating from a central core according to a square symmetry, as we did for the previous cases (Figure 5b). Again we compared the geometrical features of the optimized models with those calculated from *AA-mapped* data and the literature CG model.⁴⁵ Using execution *mode 1*, the BI initially produced inappropriate BP and mismatches between CG and AA-mapped *geoms* distributions were effectively reduced during optimization (Figure S20), notably allowing to obtain a relevant geometry of the porphyrin core in this more “complicated” case (i.e. nested network of bonds). BP scores went down from 45.2 to 32.6. Average Rg was also found in good agreement with the AA-mapped trajectory ($\Delta R_g = 0.4$ Å or 3%). The averaging of distributions within groups of similar *geoms*, coupled to CG modeling, produced a planar geometry of the porphyrin core which reduced the spread of Rg values with respect to AA-mapped data (Figure 5b, boxplots). BP optimization converged within 392 steps (Table 2). Using execution *mode 2*, as expected the BI initially produced an inaccurate set of BP (Figure S21 and *cf.* SI section 1.1) which was tuned during optimization. BP scores decreased from 80.6 to 40.2. Once again, optimized BP produced more “loosely” adjusted distributions overlap for angle groups that used equilibrium values at 180° (Figure S23, angle group 3 in particular), providing increased flexibility of the CG molecular model with respect to the optimized BP obtained via execution mode 1 and AA-mapped data. In this case, this allowed the model to adopt more folded conformations in solvent (i.e. average Rg decreased by 1.20 Å or 10 points, reaching 12 Å). BP optimization converged within 322 steps.

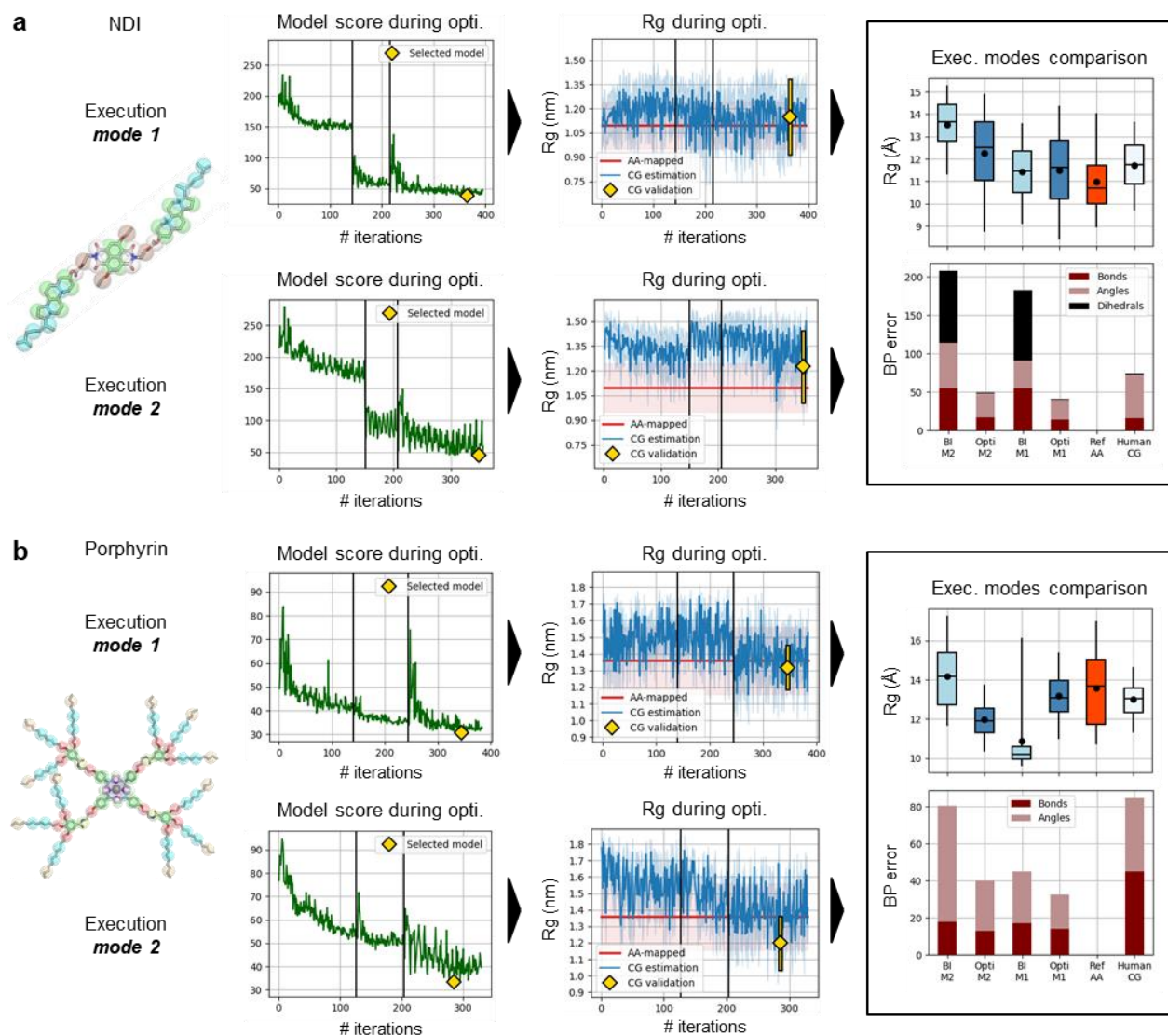


Figure 5. Results of *Swarm-CG* for the optimization of bonded parameters of other symmetric flexible structures in the benchmark, using execution *modes 1* (M1) and 2 (M2) with default settings: (a) NDI model.⁵⁹ (b) porphyrin-based monomer model.⁴⁵ From left to right we report: (i) the molecular structure, (ii) the evolution of the scoring function, where green lines show the score attributed to candidate BP during optimization. Yellow diamonds indicate the score of the selected model. (iii) the evolution of Rg, in which blue lines show average Rg estimates at each iteration of the CG model optimization (light blue intervals represent \pm standard deviation), and red horizontal lines show the average Rg of AA-mapped reference trajectories (light red intervals represent \pm standard deviation). Yellow diamonds and lines show averages and standard deviations obtained from 200 ns simulations. (iv) the comparison of Rg and BP errors in different models in 200 ns simulations (BI: step 1, Opti: selected model). Boxplots and whiskers display percentiles 5, 25, 50, 75 and 95 of Rg values. Black dots show average Rg values. Stacked barplots show each component of the scoring function, the sum of which amounts to the BP score.

3.2. Cyclic structures

We also challenged *Swarm-CG* in treating different types of molecular architectures – *i.e.* cyclic and symmetric molecules with a more rigid architecture, namely cyclodextrins and pillar[5]arene. In such cases, we report the results of execution *mode 1* (as for such relatively simple and rigid motifs, execution *mode 2* reported identical results).

3.2.1. β -Cyclodextrin

We first comment the optimization case of the CG model of β -Cyclodextrin,⁶⁰ which main results are reported in Figure 6a. Geometrical features of the resulting CG model are compared to both those calculated from AA-mapped data and from the CG model available from the literature⁶⁰ (*cf.* section 6.2). The mismatches between CG and AA-mapped *geoms* distributions were effectively reduced during optimization (Figure 6a, green line). However, in the 200 ns validation simulation, *geoms* distributions of the optimized CG model did not perfectly reproduce those calculated from the AA-mapped trajectory (Figure S27). A small issue in the topology of the model, related to the size of CG beads used in the MARTINI framework, prevents proper closing of angles between P4-P2-P2 beads (*i.e.* maroon-pink-pink and Figure S27, angle 4) and slightly affects other *geoms* distributions. These small mismatches in local geometry are acceptable at CG resolution, in the context of this molecular structure. According to the MARTINI framework, fixing them would require modifications such as scaling the bonds lengths between P4-P2 beads (*i.e.* maroon-pink), which would introduce other forms of error in the CG model, notably related to non-bonded parametrization. Nonetheless, the optimization process yielded appropriate BP in the context of the CG topology provided for β -Cyclodextrin and average Rg and SASA are in very good agreement with the AA-mapped trajectory (*i.e.* both errors < 1%). BP optimization converged within 206 steps (Table 2). We tested the manually parametrized CG model of β -Cyclodextrin in a 200 ns simulation. *Swarm-CG* scoring function yielded a BP score of 71.9 for this model, indicating larger discrepancies in local geometrical features (Figure S29) compared to the optimized CG model which obtained a BP score of 28.2, with respect to the AA-mapped reference trajectory. BP optimization produced marginal improvements on average Rg error (0.2 Å or 3.5 points) for this small and cyclic molecule, with respect to the available manual bonded parametrization (Table 2).

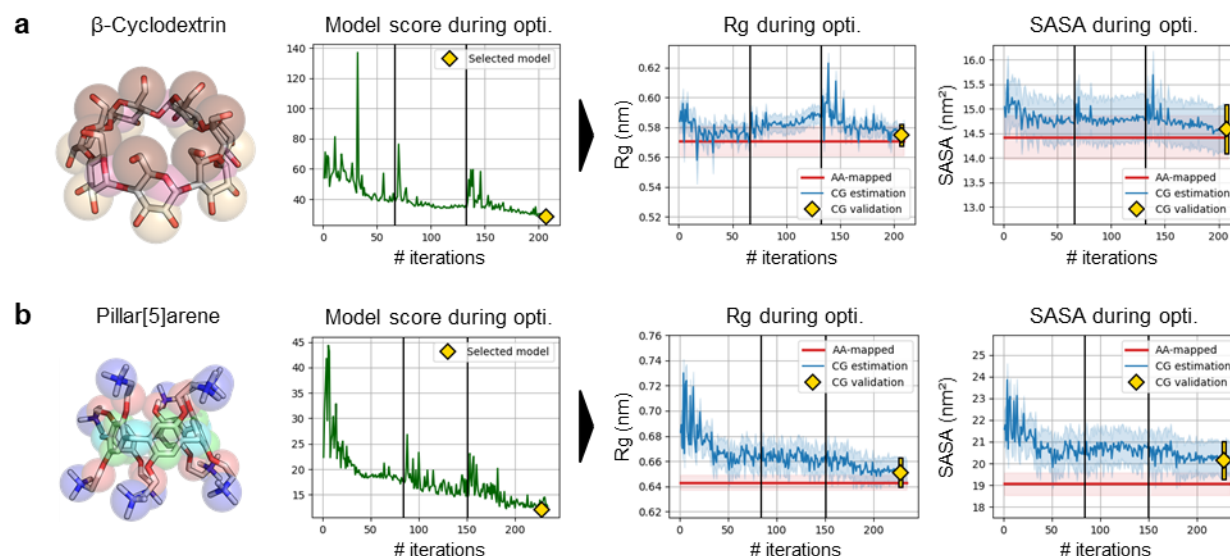


Figure 6. Results of *Swarm-CG* for the optimization of bonded parameters of two cyclic structures using execution *mode 1* with default settings. (a) β -Cyclodextrin model.⁶⁰ (b) Pillar[5]arene model.⁶¹ From left to right we report: (i) the molecular structure, (ii) the evolution of the scoring function, where green lines show the score attributed to

candidate BP during optimization. Yellow diamonds indicate the score of the selected model. (iii) the evolution of Rg, in which blue lines show average Rg estimates at each iteration of the CG model optimization. (iv) the evolution of SASA, in which blue lines show average SASA estimates at each iteration of the CG model optimization. Light blue intervals represent \pm standard deviation), and red horizontal lines show the average Rg/SASA of AA-mapped reference trajectories (light red intervals represent \pm standard deviation). Yellow diamonds and lines show averages and standard deviations obtained from 200 ns simulations.

3.2.2. Pillar[5]arene

Next, we optimized the CG model of a pillar[5]arene, for which main results are reported in Figure 6b. Geometrical features of the resulting CG model are exclusively compared to those calculated from AA-mapped data,⁶¹ since no manually parametrized CG model was available from the literature for this molecule. The mismatches between CG and AA-mapped *geoms* distributions were effectively reduced during optimization (Figure 6b, green line). The optimized CG model was further validated in a 200 ns simulation, in which *geoms* distributions correctly overlapped with those calculated from the AA-mapped trajectory (Figure S31). Average Rg and SASA values were also found in good agreement (errors: 1% and 5%). Since the CG model of the pillar[5]arene was prepared according to the MARTINI framework, bonds were rescaled between SC5 and EO beads (*i.e.* cyclic core to arms junctions) and the average Rg of the AA-mapped reference was rescaled accordingly (*cf.* SI section 2.2), while no transformation was applied to its average SASA value. Therefore, the small discrepancy observed between average SASA of the optimized CG and AA-mapped models is expected and the rescaled, average Rg is a better reference to assess this model's bonded parametrization. BP optimization converged within 230 steps (Table 2).

3.3. Complex hyper-branched macromolecules

Finally, we challenged *Swarm-CG* for the optimization of CG models of complex macromolecules, for which we use as case study a spermine-based dendron (flexible, small and directional branched macromolecule), and PAMAM dendrimers generations 1 (G1, symmetric/non-directional, small and flexible branched macromolecule) and 2 (G2, large, symmetric/non-directional and complex branched macromolecule). For these cases, only the results of execution *mode 1* are reported, as execution *mode 2* provided an analogous picture. In fact, while these complex architectures could suffer of reduced sampling at AA-MD level, their highly symmetric hyperbranched structure allows for a good analysis of the *geoms*, as *Swarm-CG* averages the behavior the identical/symmetric side branches.

3.3.1. Directional dendrons

We first comment the optimization case of the CG model of a spermine-functionalized dendron reported in Figure 7a, as an example of flexible, branched and directional molecule.^{62,81} Geometrical features of the resulting CG model are exclusively compared to those calculated from AA-mapped data, since no manually parametrized CG model was available from the literature for this molecule.

As depicted by the decreasing scores attributed to candidate sets of BP during optimization (Figure 7a, green line), the process effectively minimizes mismatches between CG and AA-mapped *geoms* distributions, which were reduced by a factor of 2 when comparing the initial set of BP (step 1) to those of the optimized model (step 249). BP optimization converged within 255 steps (Table 2). The optimized CG model was further validated in a 200 ns simulation. Distributions of the bonds and angles correctly overlap between the optimized CG and AA-mapped models (Figure S34) and, on a larger scale, average Rg and SASA values are also in very good agreement (*i.e.* both errors < 3%). Average SASA of the optimized CG model is slightly increased with respect to that of the AA-mapped, which is due to the scaling of bonds lengths between CG beads of the aromatic ring and is inherent to the MARTINI framework (*cf.* SI section 2.2). This scaling has little incidence on Rg values in this model and the AA-mapped average Rg was calculated without considering any offset.

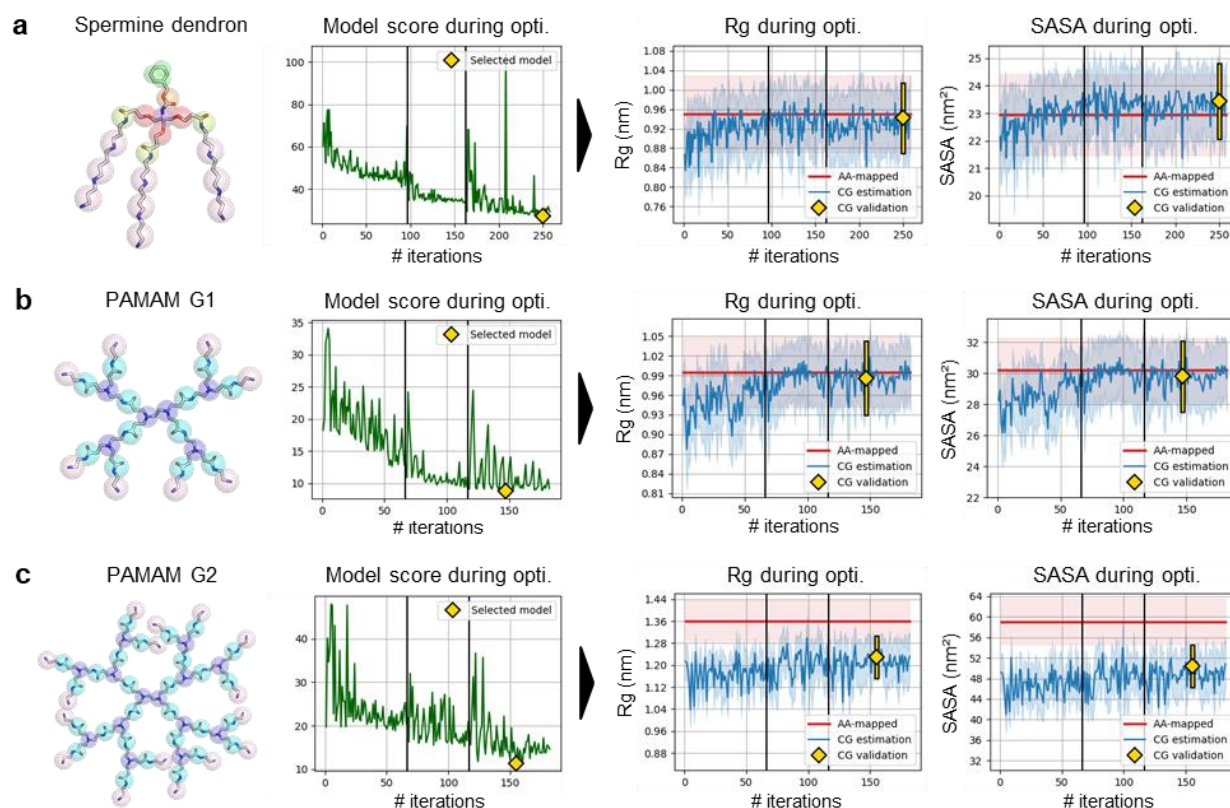


Figure 7. Results of *Swarm-CG* for the optimization of bonded parameters of three types of hyper-branched macromolecules (*i.e.* dendrons and dendrimers) using execution *mode 1* with default settings. (a) Spermine dendron model.⁶² (b) PAMAM G1 model.⁶³ (c) PAMAM G2 model.⁶³ From left to right we report: (i) the molecular structure, (ii) the evolution of the scoring function, where green lines show the score attributed to candidate BP during optimization. Yellow diamonds indicate the score of the selected model. (iii) the evolution of R_g , in which blue lines show average R_g estimates at each iteration of the CG model optimization. (iv) the evolution of SASA, in which blue lines show average SASA estimates at each iteration of the CG model optimization. Light blue intervals represent \pm standard deviation), and red horizontal lines show the average R_g /SASA of AA-mapped reference trajectories (light red intervals represent \pm standard deviation). Yellow diamonds and lines show averages and standard deviations obtained from 200 ns simulations.

3.3.2. PAMAM G1 dendrimer

Next, we optimized the CG model of a G1 PAMAM dendrimer, a small structural variant belonging to a well-known family of dendrimers,^{63,65,66} here used as an example of relatively flexible and small symmetric branched macromolecule, for which main results are reported in Figure 7b. Notably, a MARTINI CG model for PAMAM G5 is available from the literature⁶³ and could be adapted (truncated) to obtain a smaller PAMAM G1. Geometrical features of the resulting CG model are compared to both those calculated from AA-mapped data and from the CG model adapted from the literature⁶³ (*cf.* section 6.2). We show that *Swarm-CG* is capable of optimizing the CG parametrization for such type of molecules, which performs even better vs. AA models compared to the CG parameters available in the literature, and adapted for smaller dendrimer generations, in terms of behavior of the dendrimer in the solvent (water). In particular, the simple truncation to G1 of the available MARTINI model for G5 PAMAM dendrimers (maintaining the very same bonded and non-bonded literature parameters)⁶³ results into an overestimation of the R_g and SASA of the dendrimer in explicit water, which could be somewhat expected considered that Lee and Larson also highlighted similar slight size overestimations for G5 PAMAM when the model was developed.⁶³ By providing the same non-bonded terms as an input, however, here we show that *Swarm-*

CG is capable of easily optimizing the CG models to successfully improve the agreement with AA models for these branched macromolecules.

The mismatches between CG and AA-mapped *geoms* distributions were again reduced by a factor of 2 during optimization, when comparing the initial set of BP to those of the optimized CG model (Figure 7b, green line). The optimized CG model was further validated in a 200 ns simulation, in which *geoms* distributions correctly overlapped with those calculated from the AA-mapped trajectory (Figure S37). Average Rg and SASA values were also in perfect agreement (*i.e.* both errors < 2%). BP optimization converged within 182 steps (Table 2).

We then tested the available manually-parametrized CG model of PAMAM G1⁶³ via a 200 ns CG-MD simulation. The *Swarm-CG* scoring function yielded a BP score of 123.7 for this CG model (Figure S38). Noteworthy, the discrepancy in the local geometrical features with respect to the AA-mapped reference MD trajectory is considerably reduced in the optimized CG model produced by *Swarm-CG*, which yielded a BP score as low as 9.0 (Figure S37). With respect to manual parametrization, the average Rg error in the optimized CG model provided by *Swarm-CG* (calculated respect to the AA reference model) is also reduced by 2.5 Å (~25%) using the optimized set of BP (see Table 2).

3.3.3. PAMAM G2 dendrimer

We then optimized the CG model of PAMAM G2, for which main results are reported in Figure 7c. PAMAM G2 is here used as an example of a more complex, symmetric branched macromolecule. As for PAMAM G1, a MARTINI CG model for PAMAM G2 was adapted from the literature⁶³ and used for comparison. Also in this case, we show that *Swarm-CG* is capable of retrieving a reliable parametrization for such type of molecules, which performs even better vs. AA models compared to the existing CG parameters in terms of behavior of the G2 PAMAM dendrimer in the water, obtaining an improvement in the CG model performance in line with those discussed above for G1 PAMAM. Geometrical features of the resulting CG model are compared to both those calculated from AA-mapped data and from the CG model adapted from the literature⁶³ (*cf.* section 6.2).

The mismatches between CG and AA-mapped *geoms* distributions were effectively reduced during optimization (Figure 4c, green line). *Geoms* distributions of the optimized CG model correctly overlapped with those calculated from the AA-mapped trajectory in a 200 ns validation simulation (Figure S41). However, average Rg and SASA values were approximately 1.3 Å (10%) and 8 nm² (16%) lower with respect to the AA-mapped reference (all values are reported in Table 2). In this case, we can safely assume that the CG topology is valid. Thus, as we proved BP are correctly tuned according to a reference AA-mapped trajectory, the cause of the residual Rg and SASA offset cannot be directly attributed to errors in the bonded parametrization. In this perspective, this can be imputed: (i) to the non-bonded interaction parameters, which could possibly slightly underestimate interactions of the molecule with solvent or overestimate intramolecular interactions, or (ii) to sampling limitations in the AA-MD trajectory (oversampling of determined molecular configurations, possible for particularly complex molecules). We also underline that, since non-bonded terms used for the optimization of PAMAM G1 and G2 are identical, in the case of sub-optimal non-bonded parametrization *Swarm-CG* consequently adjusts the bonded terms to compensate and to reproduce at best in the CG model the behavior of the AA reference, exclusively based on *geoms* distributions. These observations highlight the potential of *Swarm-CG* as a diagnostic tool for CG modeling, which provides hints for model refinement beyond its primary purpose of automatic bonded parametrization.

We compared the automatically optimized CG model of G2 PAMAM with the manually parametrized one obtained using parameters available from the literature⁶³ via a 200 ns CG-MD simulation. *Swarm-CG* scoring function yielded a BP score of 123.8 for the manually optimized CG model. Again, the discrepancies in local geometrical features calculated with respect to the AA-reference trajectory were considerably reduced in the *Swarm-CG* optimized CG model (Figure S42), which provides a BP score of 23.0 (Figure S41).

With respect to manual CG parametrization, the average Rg error is also reduced by ~15 points using the optimized set of BP provided by *Swarm-CG* (see Table 2). BP optimization converged within 182 steps (Table 2). The selected maximum number of optimization steps is identical as for the optimization of PAMAM G1 owing to the hyper-branched structures of PAMAM molecules, for which topologies can be considered identical once similar bonds and angles have been grouped together. This is also why the manually parametrized models of PAMAM G1 and G2 obtained very similar BP scores (123.7 and 123.8), which are not exactly identical only due to the intrinsic statistical variability associated to the MD sampling. All these results provided for such a diverse set of different molecules, including these branched dendrimers, demonstrate that *Swarm-CG* has great potential to treat also molecular architectures of considerable complexity in an efficient and reliable way.

Table 2. Average Rg obtained for CG models of the benchmark optimized using execution *mode 1* and *Swarm-CG* default settings, the reference AA-mapped trajectories and manually parameterized CG models from the literature. All data points were obtained in 200 ns simulations. (+) Minimal bonds rescaling. (++) Important bonds rescaling. (+++) All bonds rescaled.

Molecule	Bonds scaling	Radius of gyration (Rg)			Optimization wall time ^(a)
		Ref. AA model [Å]	Optimized CG model error [$\Delta\text{\AA}$] ($\Delta\%$)	Manually parametrized CG model error [$\Delta\text{\AA}$] ($\Delta\%$)	
BTA	++	9.07	0.37 (4.1%)	0.41 (4.5%) ⁵⁷	6 h
BTT	++	8.45	0.73 (8.6%)	0.89 (10.5%) ⁵⁸	7 h
NDI	+++	10.98	0.53 (4.8%)	0.75 (6.8%) ⁵⁹	16.5 h
Porphyrin	++	13.58	0.40 (3.0%)	0.60 (4.4%) ⁴⁵	15.5 h
β -Cyclodextrin	n/a	5.71	0.15 (2.6%)	0.35 (6.1%) ⁶⁰	5 h
Pillar[5]arene	+	6.43	0.07 (1.1%)	n/a	6.5 h
Spermine dendron	+	9.50	0.27 (2.8%)	n/a	12 h
PAMAM G1	n/a	9.95	0.12 (1.2%)	2.62 (26.3%) ⁶³	5 h
PAMAM G2	n/a	13.61	1.26 (9.3%)	3.32 (24.4%) ⁶³	5.5 h

^(a) Using standard desktop machines, see SI and Table S2 for simulations parameters and hardware specifications.

4. Methodological considerations

To ease the development of CG models and increase their physical relevance and accuracy, we used a benchmark of wide structural diversity to demonstrate that *Swarm-CG* can be employed with default settings for tuning BP of a variety of CG models of diverse levels of complexity. *Swarm-CG* systematically yields appropriate sets of BP in the context of the provided model topology and reference AA trajectory, within wall times compatible with molecular modelers' requirements. Using execution *mode 1*, Rg values of the optimized models were systematically in agreement with the reference AA trajectory.

Importantly, *Swarm-CG* produces sets of BP via an optimization process which objective function is exclusively based on the user-provided AA reference trajectory. Therefore, the BP produced by *Swarm-CG* are optimized exclusively to reproduce the geometrical behavior observed in the AA trajectory for the molecule of interest. The process of averaging distributions within *geoms* groups does improve the quality of the reference sampling for large symmetrical molecules and yielded particularly accurate BP in the present benchmark. To this end, the step of defining relevant *geoms* groups in the preliminary CG model file is crucial. On the other hand, as mentioned before, care must be taken to provide AA reference trajectories that correctly sample the desired geometrical behavior of the molecule in solvent, or other environments.

For molecules that adopt folded conformations in AA-MD simulations, execution *mode 2* typically provides increased flexibility in the optimized CG models, which allows to reproduce the dynamical properties of the molecules as an “effect” rather than a pre-encoded condition in the models. This is particularly important when one may want to develop models for molecules composed of rather flexible groups, which then are supposed to interact between them or with other molecules in the simulations. In fact, this may limit spurious effects arising from too rigidly parametrized BP (eventually coming from the reduced timescales accessible by common AA-MD simulations) that may then affect how molecules interact between them. Although less automatic, *mode 2* has the advantage of not using the software as a black box, but at the same time it requires a prior knowledge of the molecular system that is not always accessible. In such a case, the user can always use *Swarm-CG* in *mode 1*, which in principle should provide the best accessible bonded parametrization. We included the option to select either usage mode to leave maximum freedom to the user, with standard or more advanced usage of the software.

All CG models resulting from the present benchmark could be run in 200 ns simulations using time steps of 20 fs. By default, *Swarm-CG* uses conservative maximum values for force constants, which maximizes the stability of the optimized models for usage in assembly simulations. To this end, the iterative optimization process aims at identifying an appropriate balance of force constants between all elements of the topology while reproducing AA-mapped *geoms* distributions in the CG model. For special needs, all parameters of the software can be modified, notably allowing users to increase the range of the force constants to be explored during optimization, although this might be detrimental to the model stability and should be used carefully. Simulations instabilities in optimized models would most likely be caused by issues in the topology definition (*e.g.* *geoms* or bond lengths scaling).

Interestingly, we demonstrated that *Swarm-CG* can also be used as a diagnostic tool, notably for large molecular structures for which both bonded and non-bonded parametrizations, as well as obtaining a sufficient AA-MD sampling to produce a reliable reference trajectory can be particularly complex (cf. section 3.3.3).

In principle, *Swarm-CG* can also be employed for tuning BP in higher-scale CG models (*e.g.* lower coarse-graining resolution, mapping more atoms into each CG bead).^{18,83,84} While demonstrating such a usage of the software is outside the scope of this paper, the workflow would remain substantially unchanged, except for: (i) the non-bonded parameters provided, (ii) the mapping file that would group multiple atoms into each larger CG bead and (iii) the number of CG beads and swarm iterations used for optimization, which could potentially be decreased to minimize execution times, without affecting the accuracy of the results. Moreover, in such a case the user would not be restricted to using an AA-MD trajectory as the reference, but also a finer CG-MD trajectory would work to this purpose (with the advantage of a speed up in the process and of an overall improved dynamical sampling). In this sense, the successive higher-scale parametrizations (finer-to-coarser CG optimization) would be less computationally expensive than the first one (AA-to-CG), while the user should at the same time consider that approximations intrinsically accompany every CG step, and that in multi-step approaches accuracy is key to avoid sum of errors. *Swarm-CG* could also be applied to the bonded parametrization of polarized CG models or elastic networks used in CG models of proteins (*e.g.* MARTINI),⁸⁵ via limited additional code developments. The code has been developed for immediate usage with the MARTINI CG force field (it works with both explicit and implicit solvent environments), although it is easily adaptable to be used also in combination with other CG frameworks, and as said potentially across various CG scales.

5. Conclusion

Leveraging FST-PSO,⁶⁷ an efficient PSO variant that frees users from parametrizing the optimization algorithm, here we designed *Swarm-CG*, a software that automatizes the parametrization and optimization of bonded parameters in CG molecular models within CG frameworks such as MARTINI. We

took particular care to provide a versatile software capable of systematically producing reliable results for virtually any CG model, from simple to complex molecular architectures, using default *Swarm-CG* settings. The software is versatile and can be used to cross different CG scales, provided that a well-sampled reference trajectory is available for the molecule to be coarse-grained. *Swarm-CG* can satisfy molecular modelers' requirements for routine building of CG models composed of up to 200 CG beads (in the MARTINI framework, this corresponds to molecular architectures containing at least ~600-800 heavy atoms), and possibly more, both in terms of accuracy and execution times. We particularly expect this tool to support the development of new CG molecular models for the study of synthetic molecular systems and their interaction with other (bio/non-bio) molecular targets, as it is becoming increasingly crucial in the various bio- and nano-technology fields. *Swarm-CG* is available via PIP (package: *swarm-cg*) with all its dependencies. Tutorials and demonstration data are also available at www.github.com/GMPavanLab/SwarmCG.

6. Methods

6.1. Input details

AA data used to set the target of the optimization procedure include a structure and trajectory files. The AA structure with atom types, connectivity, masses and charges can be provided via a GROMACS portable topology file (.tpr). The AA trajectory can be provided in any GROMACS format accepted by MDAnalysis^{86,87} (.xtc, .trr or else). Periodic boundary conditions (PBC) are handled internally if the trajectory file includes position and size of the simulation box at each time step. Otherwise, it is assumed PBC have already been handled and a warning is displayed at start of the program. The AA trajectory is mapped on-the-fly to allow faster experimentation with different mapping schemes. The mapping of atoms to CG beads must be provided as a GROMACS index file (.ndx). The weight of atoms that would be mapped to multiple CG beads will be split accordingly when performing the mapping and calculating all AA-mapped reference *geoms* distributions.

The preliminary CG model to be optimized must be provided as a GROMACS topology file (.itp), along with the non-bonded interactions force field to be applied in MD simulations. To better handle sampling in symmetrical molecules, users can easily form groups of bonds, angles and dihedrals in this topology file (using line returns or comments). AA-mapped distributions will be averaged within groups to create the references used as target of the optimization procedure, and shared parameters will be used and optimized for the *geoms* of each group. This also makes the optimization process more efficient by reducing the number of free parameters. For example, grouping 5 angles together reduces the number of associated free parameters from 10 to 2 when using execution mode 1 and GROMACS angles functions 2. *Swarm-CG* requires users to provide a GROMACS structure file (.gro) to be used as the starting conformation of each simulation step during the iterative optimization process. This structure will be (i) minimized and (ii) pre-processed before gathering data from the production run (iii), using three user-provided GROMACS MD parameters files (.mdp), one for each step. Only the simulation parameters of the production run will be modified to adapt its number of steps and the number of frames of the output trajectory to be analyzed, according to software parameters (1000 frames by default, within 10 or 25 ns). Although the starting conformation does not have to be perfectly accurate, as it will be minimized and pre-processed at start of each iteration, this conformation must allow running stable simulations while exploring different sets of bonded parameters. For example, molecular modelers can make use of an initial set of MARTINI bond parameters just stable enough to obtain a preliminary CG model file and a starting conformation to be used for the optimization phase.

The following GROMACS bonded potential functions^{68,69} are implemented, which should be necessary and sufficient for building CG models: constraints function 1, bonds function 1, angles functions 1 and 2, and dihedrals functions 1, 2, 4 and 9. *Swarm-CG* can effectively optimize parameters for dihedrals potential

functions with multiplicity greater than one, although these may be used carefully as they are known to easily trigger instabilities in simulations. During optimization, sets of BP (*i.e.* particles of the swarm) that cause simulations to terminate abruptly gets attributed the worst possible score the scoring function can yield, according to given topology and *geoms* domains.

6.2. Benchmarking data and models

To compare the performance of *Swarm-CG* for bonded parametrization with respect to manually parametrized CG models, we first collected several MARTINI models available from the literature and from previous results of our group. We selected the following molecular dataset based on CG data availability and structural diversity: 1,3,5-benzenetricarboxamides⁵⁷ (BTA) and benzotrithiophene (BTT) decorated by L-phenylalanine (BTT-F) and octaethylene glycol side-chains,⁵⁸ naphthalene diimide (NDI),⁵⁹ Zn-porphyrin based molecule,⁴⁵ β -Cyclodextrin⁶⁰ and poly(amidoamine) dendrimers of generation 1 and 2 (PAMAM G1 and G2).⁶³ Their CG topologies, selected bead types and non-bonded interactions force field were used as provided, except for β -Cyclodextrin and PAMAM G1 and G2 for which CG topologies were built by truncating the existing CG models of β -Cyclodextrin dimer⁶⁰ and PAMAM G5 models.⁶³ Additionally, two last molecules expand the benchmark and improve its structural diversity: a pillar[5]arene⁶¹ and spermine-functionalized branched dendron.⁶² For the spermine dendron and pillar[5]arene, we built CG models from scratch in the framework of MARTINI using the CG bead types presented in Figure 1. For the pillar[5]arene and the dendron there are no previously developed manually-optimized CG models to compare with, and the performance of *Swarm-CG* has been evaluated exclusively with respect to data from the AA-mapped trajectories in these cases.

For these 9 molecules, we created AA models and generated trajectories for each single molecule in solvent using time steps of 2 fs and extensive sampling (Table 3 and SI section 2.1), which are used as a trusted reference for bonded parametrization of CG models and benchmarking of *Swarm-CG*.

Table 3. Data used for the benchmarking of *Swarm-CG* using default settings.

Molecule	Reference AA trajectory			Solvent	Non-bonded parametrization ^(a)	Used for manual parametrization evaluation
	Simulation time	Number of frames	Force field			
BTA	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁵⁷	Yes
BTT	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁵⁸	Yes
NDI	1 μ s	5000	GAFF ⁸⁸	Cyclohexane	MARTINI 2.2 ⁵⁹	Yes
Porphyrin	1 μ s	5000	GAFF ⁸⁸	Cyclohexane	MARTINI 2.2 ⁴⁵	Yes
β -Cyclodextrin	1 μ s	5000	q4md-CD ⁹⁰ + TIP3P ⁸⁹	Water	MARTINI 2.1 ⁶⁰	Yes
Pillar[5]arene	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁹¹	No
Spermine dendron	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁹¹	No
PAMAM G1	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁶³	Yes
PAMAM G2	1 μ s	5000	GAFF ⁸⁸ + TIP3P ⁸⁹	Water	MARTINI 2.2 ⁶³	Yes

^(a) Where present, non-bonded interactions were further tuned as described in associated literature, starting from the cited force field.

All topologies defined for the CG models of the benchmark use exclusively bonds and angles defined between CG beads that are closely located on the AA molecular graph (*i.e.* no long-range bonds and angles between CG beads were used to artificially constraint flexibility of the CG models), except for molecules that contain flat cores and for the cyclic structures. For the BTT and porphyrin-based molecular models, longer range angles and dihedrals were defined to obtain flat structures of the cores. For β -Cyclodextrin and pillar[5]arene, longer-range angles were also used to obtain correct geometries of the central cyclic structures.

Acknowledgements

G.M.P. acknowledges the funding received by the Swiss National Science Foundation (SNSF grants IZLIZ2_183336 and 200021_175735) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 818776 - DYNAPOL). The authors also acknowledge the computational resources provided by the Swiss National Supercomputing Center (CSCS) and by CINECA.

References

- (1) Costerton, J. W.; Ingram, J. M.; Cheng, K. J. Structure and Function of the Cell Envelope of Gram-Negative Bacteria. *Bacteriol. Rev.* **1974**, *38* (1), 87–110.
- (2) Lugtenberg, B.; Van Alphen, L. Molecular Architecture and Functioning of the Outer Membrane of Escherichia Coli and Other Gram-Negative Bacteria. *Biochim. Biophys. Acta BBA - Rev. Biomembr.* **1983**, *737* (1), 51–115. [https://doi.org/10.1016/0304-4157\(83\)90014-X](https://doi.org/10.1016/0304-4157(83)90014-X).
- (3) Mitchison, T.; Kirschner, M. Dynamic Instability of Microtubule Growth. *Nature* **1984**, *312* (5991), 237–242. <https://doi.org/10.1038/312237a0>.
- (4) Whitesides, G. M.; Mathias, J. P.; Seto, C. T. Molecular Self-Assembly and Nanochemistry: A Chemical Strategy for the Synthesis of Nanostructures. *Science* **1991**, *254* (5036), 1312–1319. <https://doi.org/10.1126/science.1962191>.
- (5) Aida, T.; Meijer, E. W.; Stupp, S. I. Functional Supramolecular Polymers. *Science* **2012**, *335* (6070), 813–817. <https://doi.org/10.1126/science.1205962>.
- (6) Lin, N.; Dmitriev, A.; Weckesser, J.; Barth, J. V.; Kern, K. Real-Time Single-Molecule Imaging of the Formation and Dynamics of Coordination Compounds. *Angew. Chem. Int. Ed.* **2002**, *41* (24), 4779–4783. <https://doi.org/10.1002/anie.200290046>.
- (7) Theobald, J. A.; Oxtoby, N. S.; Phillips, M. A.; Champness, N. R.; Beton, P. H. Controlling Molecular Deposition and Layer Structure with Supramolecular Surface Assemblies. *Nature* **2003**, *424* (6952), 1029–1031. <https://doi.org/10.1038/nature01915>.
- (8) Gardner, G. B.; Venkataraman, D.; Moore, J. S.; Lee, S. Spontaneous Assembly of a Hinged Coordination Network. *Nature* **1995**, *374* (6525), 792–795. <https://doi.org/10.1038/374792a0>.
- (9) Olson, A. J.; Hu, Y. H. E.; Keinan, E. Chemical Mimicry of Viral Capsid Self-Assembly. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (52), 20731–20736. <https://doi.org/10.1073/pnas.0709489104>.
- (10) Suzuki, K.; Tominaga, M.; Kawano, M.; Fujita, M. Self-Assembly of an M6L12 Coordination Cube. *Chem. Commun.* **2009**, No. 13, 1638–1640. <https://doi.org/10.1039/B822311D>.
- (11) Bale, J. B.; Gonen, S.; Liu, Y.; Sheffler, W.; Ellis, D.; Thomas, C.; Cascio, D.; Yeates, T. O.; Gonen, T.; King, N. P.; Baker, D. Accurate Design of Megadalton-Scale Two-Component Icosahedral Protein Complexes. *Science* **2016**, *353* (6297), 389–394. <https://doi.org/10.1126/science.aaf8818>.
- (12) Frederix, P. W. J. M.; Patmanidis, I.; Marrink, S. J. Molecular Simulations of Self-Assembling Bio-Inspired Supramolecular Systems and Their Connection to Experiments. *Chem. Soc. Rev.* **2018**, *47* (10), 3470–3489. <https://doi.org/10.1039/C8CS00040A>.

- (13) Palma, C.-A.; Cecchini, M.; Samorì, P. Predicting Self-Assembly: From Empirism to Determinism. *Chem. Soc. Rev.* **2012**, 41 (10), 3713–3730. <https://doi.org/10.1039/C2CS15302E>.
- (14) Bochicchio, D.; Pavan, G. M. Molecular Modelling of Supramolecular Polymers. *Adv. Phys. X* **2018**, 3 (1), 1436408. <https://doi.org/10.1080/23746149.2018.1436408>.
- (15) Pesce, L.; Perego, C.; Grommet, A. B.; Klajn, R.; Pavan, G. M. Molecular Factors Controlling the Isomerization of Azobenzenes in the Cavity of a Flexible Coordination Cage. *J. Am. Chem. Soc.* **2020**. <https://doi.org/10.1021/jacs.0c03444>.
- (16) Wolde, P. R. ten; Frenkel, D. Enhancement of Protein Crystal Nucleation by Critical Density Fluctuations. *Science* **1997**, 277 (5334), 1975–1978. <https://doi.org/10.1126/science.277.5334.1975>.
- (17) Levitt, M.; Sharon, R. Accurate Simulation of Protein Dynamics in Solution. *Proc. Natl. Acad. Sci.* **1988**, 85 (20), 7557–7561. <https://doi.org/10.1073/pnas.85.20.7557>.
- (18) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, 109 (7), 2469–2473. <https://doi.org/10.1021/jp044629q>.
- (19) Šarić, A.; Chebaro, Y. C.; Knowles, T. P. J.; Frenkel, D. Crucial Role of Nonspecific Interactions in Amyloid Nucleation. *Proc. Natl. Acad. Sci.* **2014**, 111 (50), 17869–17874. <https://doi.org/10.1073/pnas.1410159111>.
- (20) Klein, M. L.; Shinoda, W. Large-Scale Molecular Dynamics Simulations of Self-Assembling Systems. *Science* **2008**, 321 (5890), 798–800. <https://doi.org/10.1126/science.1157834>.
- (21) Pavan, G. M.; Barducci, A.; Albertazzi, L.; Parrinello, M. Combining Metadynamics Simulation and Experiments to Characterize Dendrimers in Solution. *Soft Matter* **2013**, 9 (9), 2593–2597. <https://doi.org/10.1039/C3SM27706B>.
- (22) Three-Dimensional Directionality Is a Pivotal Structural Feature for the Bioactivity of Azabisphosphonate-Capped Poly(PhosphorHydrazone) Nanodrug Dendrimers | Biomacromolecules <https://pubs.acs.org/doi/full/10.1021/acs.biomac.7b01398> (accessed Jun 29, 2020).
- (23) Ingólfsson, H. I.; Melo, M. N.; van Eerden, F. J.; Arnarez, C.; Lopez, C. A.; Wassenaar, T. A.; Periole, X.; de Vries, A. H.; Tieleman, D. P.; Marrink, S. J. Lipid Organization of the Plasma Membrane. *J. Am. Chem. Soc.* **2014**, 136 (41), 14554–14559. <https://doi.org/10.1021/ja507832e>.
- (24) Risselada, H. J.; Mark, A. E.; Marrink, S. J. Application of Mean Field Boundary Potentials in Simulations of Lipid Vesicles. *J. Phys. Chem. B* **2008**, 112 (25), 7438–7447. <https://doi.org/10.1021/jp0758519>.
- (25) Smeijers, A. F.; Markvoort, A. J.; Pieterse, K.; Hilbers, P. A. J. A Detailed Look at Vesicle Fusion. *J. Phys. Chem. B* **2006**, 110 (26), 13212–13219. <https://doi.org/10.1021/jp060824o>.
- (26) Wu, S.; Guo, H. Dissipative Particle Dynamics Simulation Study of the Bilayer-Vesicle Transition. *Sci. China Ser. B Chem.* **2008**, 51 (8), 743. <https://doi.org/10.1007/s11426-008-0077-5>.
- (27) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab Initio Simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains. *Proc. Natl. Acad. Sci.* **2005**, 102 (7), 2362–2367. <https://doi.org/10.1073/pnas.0408885102>.
- (28) Zheng, W.; Schafer, N. P.; Davtyan, A.; Papoian, G. A.; Wolynes, P. G. Predictive Energy Landscapes for Protein–Protein Association. *Proc. Natl. Acad. Sci.* **2012**, 109 (47), 19244–19249. <https://doi.org/10.1073/pnas.1216215109>.
- (29) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* **1975**, 253 (5494), 694–698. <https://doi.org/10.1038/253694a0>.
- (30) Bochicchio, D.; Salvalaglio, M.; Pavan, G. M. Into the Dynamics of a Supramolecular Polymer at Submolecular Resolution. *Nat. Commun.* **2017**, 8 (1), 1–11. <https://doi.org/10.1038/s41467-017-00189-0>.
- (31) 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 (27), 7812–7824. <https://doi.org/10.1021/jp071097f>.

- (32) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760. <https://doi.org/10.1021/jp036508g>.
- (33) Risselada, H. J.; Marrink, S. J. The Molecular Face of Lipid Rafts in Model Membranes. *Proc. Natl. Acad. Sci.* **2008**, *105* (45), 17367–17372. <https://doi.org/10.1073/pnas.0807527105>.
- (34) Tieleman, D. P.; Leontiadou, H.; Mark, A. E.; Marrink, S.-J. Simulation of Pore Formation in Lipid Bilayers by Mechanical Stress and Electric Fields. *J. Am. Chem. Soc.* **2003**, *125* (21), 6382–6383. <https://doi.org/10.1021/ja029504i>.
- (35) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834. <https://doi.org/10.1021/ct700324x>.
- (36) Leontiadou, H.; Mark, A. E.; Marrink, S. J. Antimicrobial Peptides in Action. *J. Am. Chem. Soc.* **2006**, *128* (37), 12156–12161. <https://doi.org/10.1021/ja062927q>.
- (37) Rossi, G.; Monticelli, L.; Puisto, S. R.; Vattulainen, I.; Ala-Nissila, T. Coarse-Graining Polymers with the MARTINI Force-Field: Polystyrene as a Benchmark Case. *Soft Matter* **2011**, *7* (2), 698–708. <https://doi.org/10.1039/C0SM00481B>.
- (38) Panizon, E.; Bochicchio, D.; Monticelli, L.; Rossi, G. MARTINI Coarse-Grained Models of Polyethylene and Polypropylene. *J. Phys. Chem. B* **2015**, *119* (25), 8209–8216. <https://doi.org/10.1021/acs.jpcc.5b03611>.
- (39) Qiu, L.; Liu, J.; Alessandri, R.; Qiu, X.; Koopmans, M.; Havenith, R. W. A.; Marrink, S. J.; Chiechi, R. C.; Koster, L. J. A.; Hummelen, J. C. Enhancing Doping Efficiency by Improving Host-Dopant Miscibility for Fullerene-Based n-Type Thermoelectrics. *J. Mater. Chem. A* **2017**, *5* (40), 21234–21241. <https://doi.org/10.1039/C7TA06609K>.
- (40) Monticelli, L. On Atomistic and Coarse-Grained Models for C60 Fullerene. *J. Chem. Theory Comput.* **2012**, *8* (4), 1370–1378. <https://doi.org/10.1021/ct3000102>.
- (41) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. [https://doi.org/10.1016/0021-9991\(77\)90121-8](https://doi.org/10.1016/0021-9991(77)90121-8).
- (42) Hansen, H. S.; Hünenberger, P. H. Using the Local Elevation Method to Construct Optimized Umbrella Sampling Potentials: Calculation of the Relative Free Energies and Interconversion Barriers of Glucopyranose Ring Conformers in Water. *J. Comput. Chem.* **2010**, *31* (1), 1–23. <https://doi.org/10.1002/jcc.21253>.
- (43) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562–12566. <https://doi.org/10.1073/pnas.202427399>.
- (44) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100* (2), 020603. <https://doi.org/10.1103/PhysRevLett.100.020603>.
- (45) Jung, S. H.; Bochicchio, D.; Pavan, G. M.; Takeuchi, M.; Sugiyasu, K. A Block Supramolecular Polymer and Its Kinetically Enhanced Stability. *J. Am. Chem. Soc.* **2018**, *140* (33), 10570–10577. <https://doi.org/10.1021/jacs.8b06016>.
- (46) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801–6822. <https://doi.org/10.1039/C3CS60093A>.
- (47) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of Information Limitations in Coarse-Grained Models. *J. Chem. Phys.* **2019**, *151* (24), 244105. <https://doi.org/10.1063/1.5129398>.
- (48) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, *9* (16), 4667–4672. <https://doi.org/10.1021/acs.jpcllett.8b01416>.
- (49) Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noe, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ArXiv181201736 Phys. Stat* **2019**.

- (50) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of Coarse-Grained Potentials via Multistate Iterative Boltzmann Inversion. *J. Chem. Phys.* **2014**, *140* (22). <https://doi.org/10.1063/1.4880555>.
- (51) Bereau, T.; Kremer, K. Automated Parametrization of the Coarse-Grained Martini Force Field for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11* (6), 2783–2791. <https://doi.org/10.1021/acs.jctc.5b00056>.
- (52) Graham, J. A.; Essex, J. W.; Khalid, S. PyCGTOOL: Automated Generation of Coarse-Grained Molecular Dynamics Models from Atomistic Trajectories. *J. Chem. Inf. Model.* **2017**, *57* (4), 650–656. <https://doi.org/10.1021/acs.jcim.7b00096>.
- (53) Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **1985**, *18* (3), 534–552. <https://doi.org/10.1021/ma00145a039>.
- (54) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates. *Acta Polym.* **1998**, *49* (2–3), 61–74. [https://doi.org/10.1002/\(SICI\)1521-4044\(199802\)49:2/3<61::AID-APOL61>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1521-4044(199802)49:2/3<61::AID-APOL61>3.0.CO;2-V).
- (55) Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*; 1995; Vol. 4, pp 1942–1948 vol.4. <https://doi.org/10.1109/ICNN.1995.488968>.
- (56) Sengupta, S.; Basak, S.; Peters, R. A. *Particle Swarm Optimization: A Survey of Historical and Recent Developments with Hybridization Perspectives*; preprint; MATHEMATICS & COMPUTER SCIENCE, 2018. <https://doi.org/10.20944/preprints201809.0007.v1>.
- (57) Bochicchio, D.; Pavan, G. M. From Cooperative Self-Assembly to Water-Soluble Supramolecular Polymers Using Coarse-Grained Simulations. *ACS Nano* **2017**, *11* (1), 1000–1011. <https://doi.org/10.1021/acsnano.6b07628>.
- (58) Casellas, N. M.; Pujals, S.; Bochicchio, D.; Pavan, G. M.; Torres, T.; Albertazzi, L.; García-Iglesias, M. From Isodesmic to Highly Cooperative: Reverting the Supramolecular Polymerization Mechanism in Water by Fine Monomer Design. *Chem. Commun.* **2018**, *54* (33), 4112–4115. <https://doi.org/10.1039/C8CC01259H>.
- (59) Sarkar, A.; Sasmal, R.; Empereur-mot, C.; Bochicchio, D.; Kompella, S. V. K.; Sharma, K.; Dhiman, S.; Sundaram, B.; Agasti, S. S.; Pavan, G. M.; George, S. J. Self-Sorted, Random, and Block Supramolecular Copolymers via Sequence Controlled, Multicomponent Self-Assembly. *J. Am. Chem. Soc.* **2020**, *142* (16), 7606–7617. <https://doi.org/10.1021/jacs.0c01822>.
- (60) López, C. A.; de Vries, A. H.; Marrink, S. J. Computational Microscopy of Cyclodextrin Mediated Cholesterol Extraction from Lipid Model Membranes. *Sci. Rep.* **2013**, *3* (1), 1–6. <https://doi.org/10.1038/srep02071>.
- (61) Beyeh, N. K.; Nonappa; Liljeström, V.; Mikkilä, J.; Korpi, A.; Bochicchio, D.; Pavan, G. M.; Ikkala, O.; Ras, R. H. A.; Kostianen, M. A. Crystalline Cyclophane–Protein Cage Frameworks. *ACS Nano* **2018**, *12* (8), 8029–8036. <https://doi.org/10.1021/acsnano.8b02856>.
- (62) Pavan, G. M.; Danani, A.; Pricl, S.; Smith, D. K. Modeling the Multivalent Recognition between Dendritic Molecules and DNA: Understanding How Ligand “Sacrifice” and Screening Can Enhance Binding. *J. Am. Chem. Soc.* **2009**, *131* (28), 9686–9694. <https://doi.org/10.1021/ja901174k>.
- (63) Lee, H.; Larson, R. G. Coarse-Grained Molecular Dynamics Studies of the Concentration and Size Dependence of Fifth- and Seventh-Generation PAMAM Dendrimers on Pore Formation in DMPC Bilayer. *J. Phys. Chem. B* **2008**, *112* (26), 7778–7784. <https://doi.org/10.1021/jp802606y>.
- (64) Pavan, G. M. Modeling the Interaction between Dendrimers and Nucleic Acids: A Molecular Perspective through Hierarchical Scales. *ChemMedChem* **2014**, *9* (12), 2623–2631. <https://doi.org/10.1002/cmdc.201402280>.

- (65) Tomalia, D. A.; Baker, H.; Dewald, J.; Hall, M.; Kallos, G.; Martin, S.; Roeck, J.; Ryder, J.; Smith, P. A New Class of Polymers: Starburst-Dendritic Macromolecules. *Polym. J.* **1985**, *17* (1), 117–132. <https://doi.org/10.1295/polymj.17.117>.
- (66) Tomalia, D. A.; Naylor, A. M.; Goddard, W. A. Starburst Dendrimers: Molecular-Level Control of Size, Shape, Surface Chemistry, Topology, and Flexibility from Atoms to Macroscopic Matter. *Angew. Chem. Int. Ed. Engl.* **1990**, *29* (2), 138–175. <https://doi.org/10.1002/anie.199001381>.
- (67) Nobile, M. S.; Cazzaniga, P.; Besozzi, D.; Colombo, R.; Mauri, G.; Pasi, G. Fuzzy Self-Tuning PSO: A Settings-Free Algorithm for Global Optimization. *Swarm Evol. Comput.* **2018**, *39*, 70–85. <https://doi.org/10.1016/j.swevo.2017.09.001>.
- (68) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91* (1), 43–56. [https://doi.org/10.1016/0010-4655\(95\)00042-E](https://doi.org/10.1016/0010-4655(95)00042-E).
- (69) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- (70) Kanekal, K. H.; Bereau, T. Resolution Limit of Data-Driven Coarse-Grained Models Spanning Chemical Space. *J. Chem. Phys.* **2019**, *151* (16), 164106. <https://doi.org/10.1063/1.5119101>.
- (71) Chen, J.; Chen, J.; Pinamonti, G.; Clementi, C. Learning Effective Molecular Models from Experimental Observables. *J. Chem. Theory Comput.* **2018**, *14* (7), 3849–3858. <https://doi.org/10.1021/acs.jctc.8b00187>.
- (72) Wang, J.; Chmiela, S.; Müller, K.-R.; Noé, F.; Clementi, C. Ensemble Learning of Coarse-Grained Molecular Dynamics Force Fields with a Kernel Approach. *J. Chem. Phys.* **2020**, *152* (19), 194106. <https://doi.org/10.1063/5.0007276>.
- (73) Giulini, M.; Menichetti, R.; Shell, M. S.; Potestio, R. An Information Theory-Based Approach for Optimal Model Reduction of Biomolecules. *ArXiv200403988 Cond-Mat Q-Bio* **2020**.
- (74) Devane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L. A Transferable Coarse Grain Non-Bonded Interaction Model For Amino Acids. *J. Chem. Theory Comput.* **2009**, *5* (8), 2115–2124. <https://doi.org/10.1021/ct800441u>.
- (75) Shinoda, W.; DeVane, R.; Klein, M. L. Coarse-Grained Molecular Modeling of Non-Ionic Surfactant Self-Assembly. *Soft Matter* **2008**, *4* (12), 2454–2462. <https://doi.org/10.1039/B808701F>.
- (76) Pele, O.; Werman, M. Fast and Robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*; 2009; pp 460–467. <https://doi.org/10.1109/ICCV.2009.5459199>.
- (77) Rao, S. S. *Engineering Optimization: Theory and Practice*; John Wiley & Sons, 2009.
- (78) Pele, O.; Werman, M. A Linear Time Histogram Metric for Improved SIFT Matching. In *Computer Vision – ECCV 2008*; Forsyth, D., Torr, P., Zisserman, A., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; Vol. 5304, pp 495–508. https://doi.org/10.1007/978-3-540-88690-7_37.
- (79) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57* (21), 2607–2609. <https://doi.org/10.1103/PhysRevLett.57.2607>.
- (80) Mishra, S. K.; Kara, M.; Zacharias, M.; Koča, J. Enhanced Conformational Sampling of Carbohydrates by Hamiltonian Replica-Exchange Simulation. *Glycobiology* **2014**, *24* (1), 70–84. <https://doi.org/10.1093/glycob/cwt093>.
- (81) Pavan, G. M.; Danani, A. The Influence of Dendron's Architecture on the “Rigid” and “Flexible” Behaviour in Binding DNA—a Modelling Study. *Phys. Chem. Chem. Phys.* **2010**, *12* (42), 13914–13917. <https://doi.org/10.1039/C0CP01124J>.
- (82) Meinel, M. K.; Müller-Plathe, F. Loss of Molecular Roughness upon Coarse-Graining Predicts the Artificially Accelerated Mobility of Coarse-Grained Molecular Simulation Models. *J. Chem. Theory Comput.* **2020**, *16* (3), 1411–1419. <https://doi.org/10.1021/acs.jctc.9b00943>.

- (83) Schindler, T.; Kröner, D.; Steinhauser, M. O. On the Dynamics of Molecular Self-Assembly and the Structural Analysis of Bilayer Membranes Using Coarse-Grained Molecular Dynamics Simulations. *Biochim. Biophys. Acta BBA - Biomembr.* **2016**, 1858 (9), 1955–1963. <https://doi.org/10.1016/j.bbamem.2016.05.014>.
- (84) Hybrid Approach for Highly Coarse-Grained Lipid Bilayer Models | Journal of Chemical Theory and Computation <https://pubs.acs.org/doi/abs/10.1021/ct300751h> (accessed Jun 2, 2020).
- (85) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **2009**, 5 (9), 2531–2543. <https://doi.org/10.1021/ct9002114>.
- (86) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, 32 (10), 2319–2327. <https://doi.org/10.1002/jcc.21787>.
- (87) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domański, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proc. 15th Python Sci. Conf.* **2016**, 98–105. <https://doi.org/10.25080/Majora-629e541a-00e>.
- (88) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, 25 (9), 1157–1174. <https://doi.org/10.1002/jcc.20035>.
- (89) 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 (2), 926–935. <https://doi.org/10.1063/1.445869>.
- (90) Cézard, C.; Trivelli, X.; Aubry, F.; Djedaïni-Pilard, F.; Dupradeau, F.-Y. Molecular Dynamics Studies of Native and Substituted Cyclodextrins in Different Media: 1. Charge Derivation and Force Field Performances. *Phys. Chem. Chem. Phys.* **2011**, 13 (33), 15103–15121. <https://doi.org/10.1039/C1CP20854C>.
- (91) Grunewald, F.; Rossi, G.; de Vries, A. H.; Marrink, S. J.; Monticelli, L. Transferable MARTINI Model of Poly(Ethylene Oxide). *J. Phys. Chem. B* **2018**, 122 (29), 7436–7449. <https://doi.org/10.1021/acs.jpcc.8b04760>.

Supplementary Information for:

Swarm-CG: Automatic Parametrization of Bonded Terms in Coarse-Grained Models of Simple to Complex Molecules via Fuzzy Self-Tuning Particle Swarm Optimization

Charly Empereur-Mot^{*,1}, Luca Pesce¹, Davide Bochicchio¹, Claudio Perego¹ & Giovanni M. Pavan^{*,1,2}

1. *Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Galleria 2, Via Cantonale 2c, CH-6928 Manno, Switzerland.*

email: charly.empereur-mot@supsi.ch

2. *Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.*

email: giovanni.pavan@polito.it

1. Algorithm details

1.1. Boltzmann inversion

To initialize a first swarm particle at the very first step of the PSO process, bonded parameters (BP) are estimated via BI to obtain initial values for bond, angle and dihedral equilibrium force constants. Bond lengths, angle values and dihedral torsions are set to their average values calculated from the *AA-mapped* MD trajectory. When using execution *mode 2*, force constants guessed via BI are applied together with angles and dihedrals equilibrium values defined in the preliminary CG model file. Accordingly, execution *mode 2* produces more flexible CG models with respect to the *AA-mapped* trajectories (*i.e.* CG distributions are typically broader than the AA ones with large Rg fluctuations) and models are corrected during optimization to reach appropriate average Rg value. Potential functions used for BI and optimization are all user-specified via the preliminary model file. Histograms are generated using 50 bins on the domains of bonds and angles, and 100 bins for dihedrals domains. For bonds and angles, three steps of exponential smoothing are applied using an alpha parameter of 0.55 and a window of 5 bins, while no transformation is performed for dihedrals. These histograms are then transformed using Boltzmann inversion^{1,2} as $BI(v) = -K_B T \log(v)$, with K_B the Boltzmann constant, T the temperature and v as bins values. GROMACS potential functions are then fit using SciPy³ to obtain guesses of the force constants. To ensure these initial guesses of bonded parameters will produce a stable enough model to initialize the optimization procedure, force constants are capped to 17000 kJ.mol⁻¹ for bonds using function 1, 1200 kJ.mol⁻¹ for angles function 2, 3.5 kJ.mol⁻¹ for dihedrals functions 1, 4 and 9 and 250 kJ.mol⁻¹.rad⁻² for dihedrals function 2 (see GROMACS manual⁴ sections 4.8 “Bonded interactions” and 5.2 “File formats”).

1.2. Iterative optimization process

Distributions of the bonds, angles and dihedrals calculated to apply the scoring function are obtained using bandwidths of 0.2 nm, 5 degrees and 5 degrees (*Swarm-CG* default settings). At the beginning of each optimization cycle, the swarm of particles is initialized in a supervised fashion to maximize FST-PSO⁵ performances. Variations around parameters included in the reference particle (*i.e.* the current best set of BP) are reduced as the optimization progresses through cycles. Additionally, the matching error (EMD distance) between CG and *AA-mapped geoms* distributions is taken into account in the calculation of the range allowed for random initialization of each parameter of each particle of the swarm at the beginning of cycles 2 and 3 (Table S1).

Table S1. Ranges of variations used for swarm initialization in the 3 cycles of optimization in *Swarm-CG*.

Opti. cycle	Geoms optimized			Variations around reference swarm particle	
	Bonds	Angles	Dihedrals ^(a)	Geoms equilibrium values range modifier	Force constants modifier
1	Yes	Yes	No	$\pm 100\%^{(b)} \times EMD_{best}/n^{(c)}$	$\pm 40\% \times EMD_{best}/n^{(c)}$
2	No ^(d)	Yes	Yes	$\pm 25\%^{(b)} \times EMD_{best}/n^{(c)}$	$\pm 30\% \times EMD_{best}/n^{(c)}$
3	Yes	Yes	Yes	$\pm 15\%^{(b)} \times EMD_{best}/n^{(c)}$	$\pm 20\% \times EMD_{best}/n^{(c)}$

^(a) Dihedrals parameters are optimized only in cycles 2 and 3, if dihedrals topologies are provided in the input preliminary CG model.

^(b) Applied to values of 0.025 nm for bonds or 10° for angles and dihedrals, to define the lower/higher range of variations around previous equilibrium values during swarm initialization (for optimization cycle 1, those are obtained *via* BI).

^(c) EMD_{best} is the EMD score obtained for a given *geom* in the previously obtained best scored model, while $n=2$ for bond and angle groups and $n=5$ for dihedral groups.

^(d) In cycle 2, bonds parameters are fixed to those of the previously obtained best scored model.

Additionally, one special particle is initialized at start of optimization cycles 2 and 3. Across optimization steps, each time the EMD score of an independent *geoms* group distribution improves, its associated subset of parameters is recorded (i.e. equilibrium value and force constant of the *geoms* group). For swarm initialization of the next optimization cycle, these are combined into a single special particle. This heuristic, which works like a “memory of independent best”, is expected to accelerate convergence of the iterative optimization procedure by providing either a new best solution, or useful information to the rest of the swarm.

2. MD Simulations

2.1. Parameters of MD simulations

All AA-MD simulations were performed by means of GROMACS 2018.^{6,7} The dynamics was integrated with a standard timestep of 2 fs. The temperature was maintained at T=300 K and the pressure at P=1 atm by coupling the dynamics with the V-rescale thermostat⁸ and the Berendsen barostat,⁹ with isotropic pressure scaling, respectively. Both algorithms were applied with a coupling time of 1 ps. Before the production runs all the systems were energy-minimized via steepest descent algorithm, and then a brief equilibration cycle of 200 ps, with an integration step of 1 fs, was performed to relax the initial configuration. Long range electrostatics is handled by means of Particle Mesh Ewald method.¹⁰

All CG-MD simulations were performed by means of GROMACS 2018.^{6,7} The dynamics was integrated with a standard timestep of 20 fs. The temperature was maintained at T=300 K and the pressure at P=1 atm by coupling the dynamics with the V-rescale thermostat⁸ and the Berendsen barostat,⁹ with isotropic pressure scaling, respectively. Both algorithm were applied with a coupling time of 1 ps. Again, before the production runs all the systems were energy-minimized via steepest descent algorithm, and then a brief equilibration cycle of 50 ps, with an integration step of 5 fs, was performed to relax the initial configuration. For the Coulomb and van der Waals interactions a cut-off of 1.1 nm was employed, using a relative dielectric constant of $\epsilon=15$, according to the MARTINI force-field scheme.¹¹

2.2. Bonds rescaling

Several bond lengths were rescaled in the benchmark CG models, notably between particles mapped to aromatic rings and molecular cores, according to the MARTINI framework.¹¹ Therefore, all Rg results of the main text are reported taking into account appropriate offsets for *AA-mapped* references. For SASA values, it is not possible to calculate a similar offset and SASA was not reported in cases where it was not relevant for evaluation. SASA were calculated using CG beads radii of 0.27 nm for standard MARTINI beads and 0.23 nm for small beads, using the double cubic lattice method¹² implemented in GROMACS.^{6,7} In the specific case of the NDI taken from the literature,¹³ all bond lengths were rescaled with respect to the *AA-mapped* reference trajectory for consistency between the CG representation of NDI aromatic rings and the CG representation of the cyclohexane solvent, which has rescaled bonds between CG particles in the MARTINI framework.¹¹ Non-bonded interactions were tuned accordingly in this study.¹³

2.3. Execution times details

Molecule	Number of CG particles	Opti. wall time	Number of opti. steps	Hardware (cpu + gpu)
BTA	1622	6 h	257	6x i7-5930K + 1x GTX 980 Ti
BTT	5682	7 h	282	9x i9-9980XE + 1x RTX 2080 Ti
NDI	7855	16.5 h	395	9x i9-9980XE + 1x RTX 2080 Ti
Porphyrin	8079	15.5 h	392	9x i9-9980XE + 1x RTX 2080 Ti
β -Cyclodextrin	439	5 h	206	9x i9-9980XE + 1x RTX 2080 Ti
Pillar[5]arene-	1231	6.5 h	230	6x X6 1090T + 1x GTX 580
Spermine-functionalized dendron	2110	12 h	255	8x FX-8120 + 1x GTX 580
G1 PAMAM dendrimer	1281	5 h	182	6x i7-5930K + 1x GTX 980 Ti
G2 PAMAM dendrimer	2387	5.5 h	182	4x i7-5930K + 1x GTX 980 Ti

Table S2. Execution times obtained for optimization of the molecules included in the *Swarm-CG* benchmark using execution mode 1, with short description of the hardware used.

3. Optimization results details

Supplementary Files include all topology (ITP) files of the optimized models obtained using execution *modes 1* and *2* (in the cases where *mode 2* was used), formatted to allow a comprehensive reading of the geoms grouping choices according to symmetries, with the help of AA models with superimposed CG particles mappings and numbering (Figures S1, S7, S13, S19, S25, S29, S32, S35, S39).

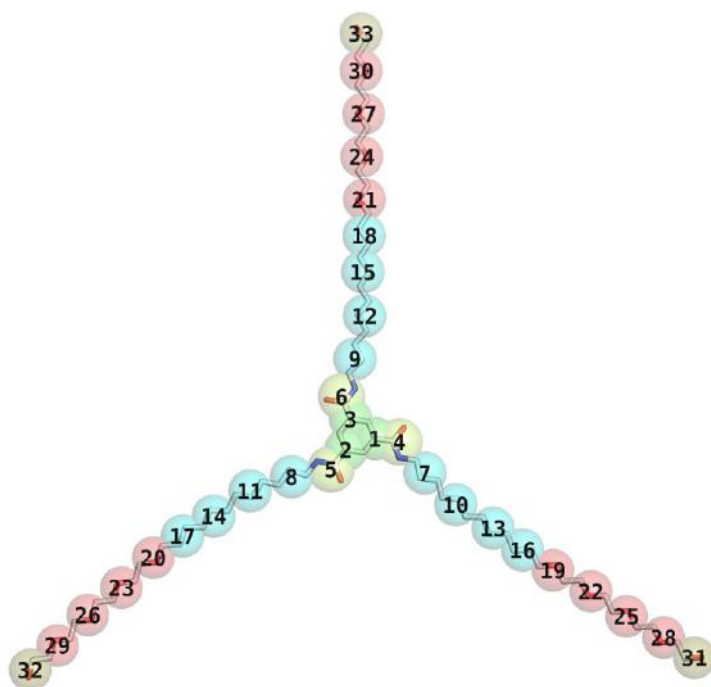


Figure S1. Molecular structure of the water-soluble BTA molecule studied herein¹⁴: AA model in sticks with superimposed transparent CG beads showing AA-CG mapping and numbering.

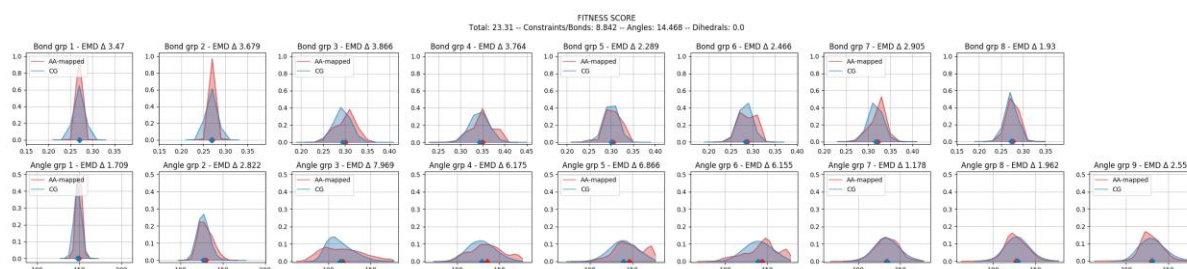


Figure S2. Geoms distributions for 200 ns of MD simulation of BTA¹⁴ using parameters obtained through Swarm-CG initial BI via execution mode 1 (initial guess only: CG in blue vs. reference AA model in red).

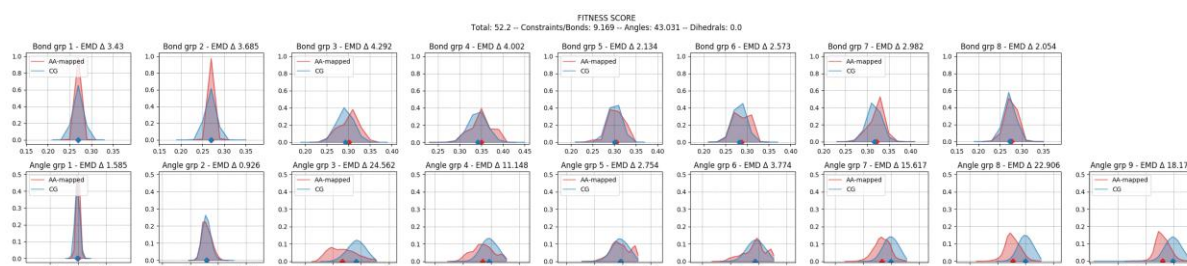


Figure S3. Geoms distributions for 200 ns of MD simulation of BTA¹⁴ using parameters obtained through Swarm-CG initial BI via execution mode 2 (initial guess only: CG in blue vs. reference AA model in red).

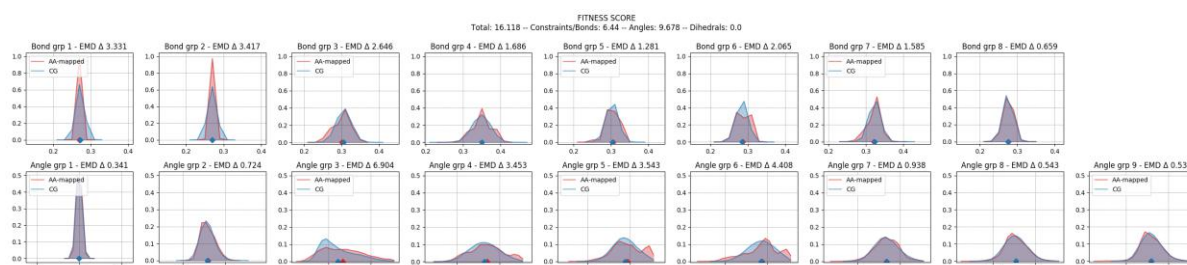


Figure S4. Geoms distributions for 200 ns of MD simulation of BTA¹⁴ using parameters obtained through Swarm-CG optimization process via execution mode 1 (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

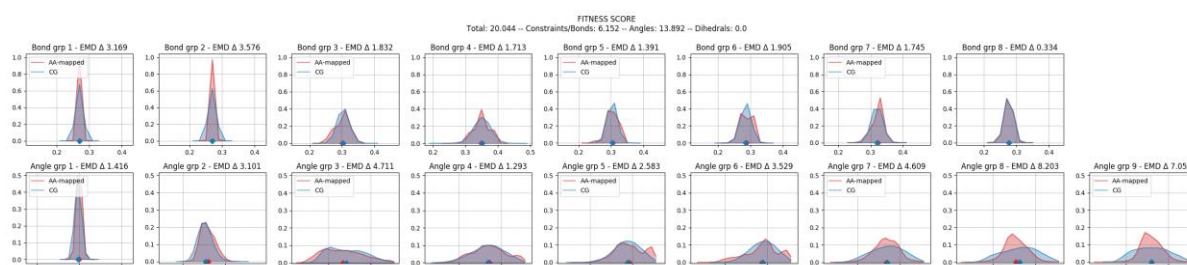


Figure S5. Geoms distributions for 200 ns of MD simulation of BTA¹⁴ using parameters obtained through Swarm-CG optimization process via execution mode 2 (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

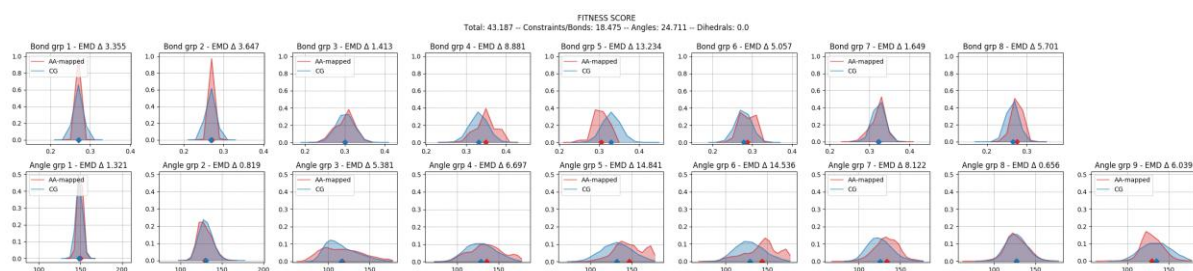


Figure S6. Geoms distributions for 200 ns of MD simulation of BTA¹⁴ using parameters obtained from the literature¹⁴ (manual CG parametrization in blue vs. reference AA model in red).

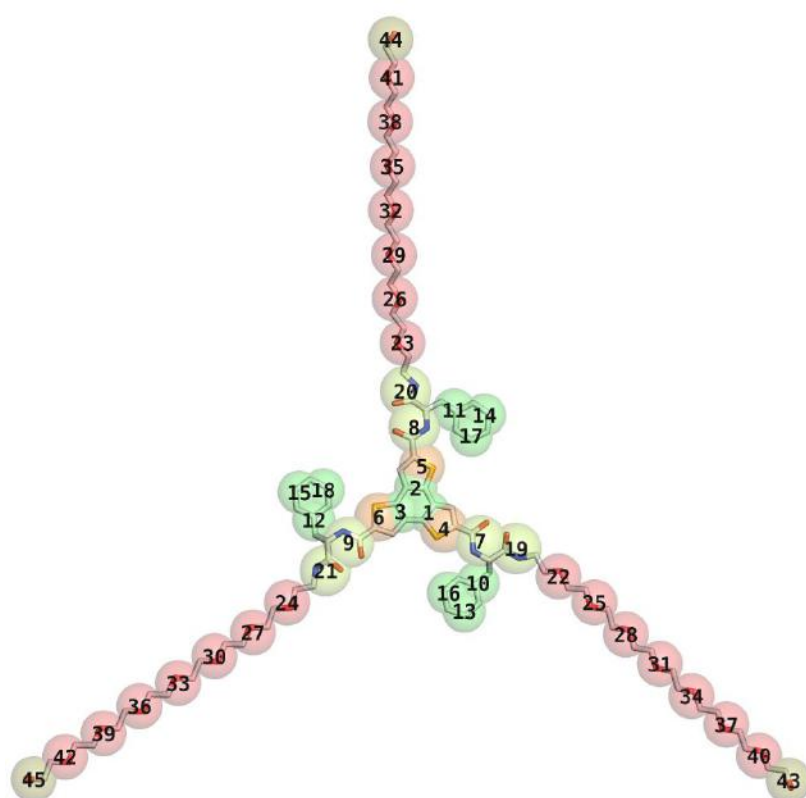


Figure S7. Molecular structure of the BTT-based molecule studied herein¹⁵: the AA model is represented in sticks with transparent superimposed CG beads with mapping and numbering.

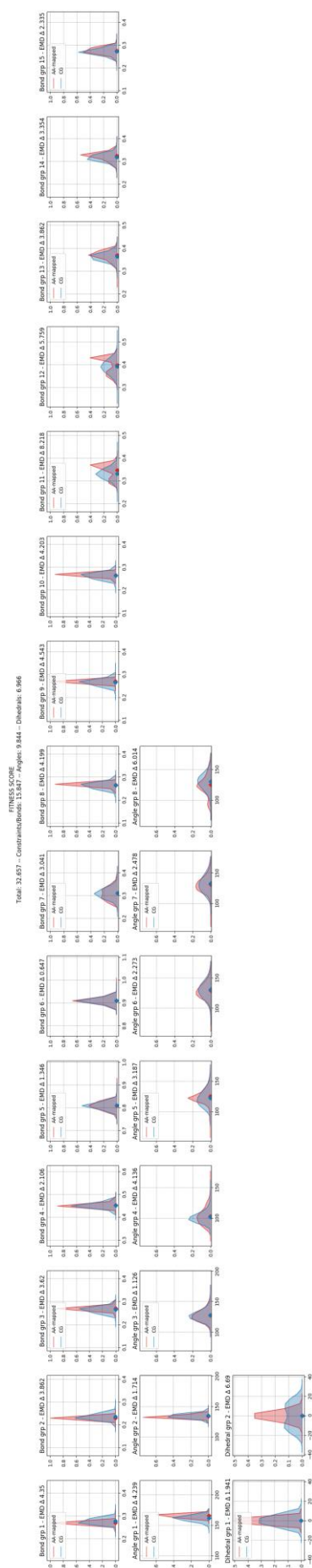


Figure S8. Geoms distributions for 200 ns of MD simulation of BTT¹⁴ using parameters obtained through Swarm-CG initial BI via execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

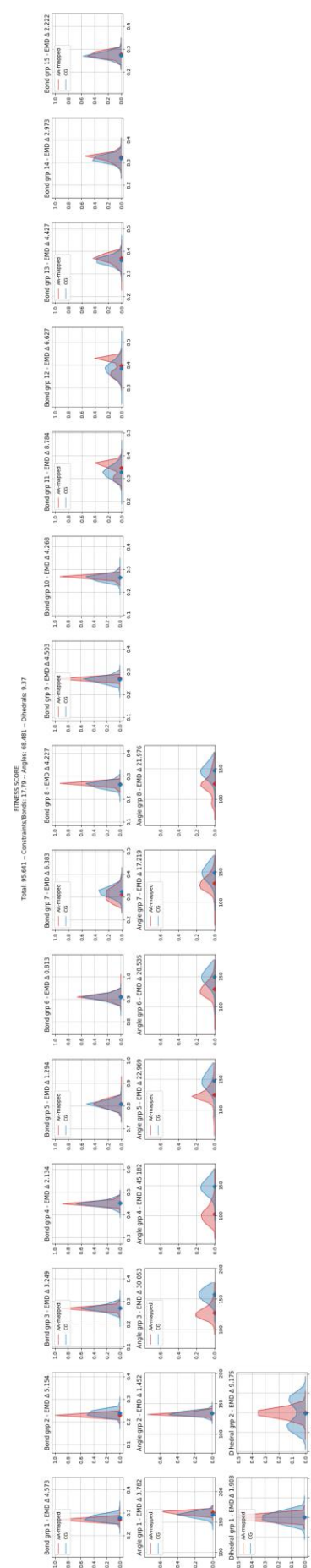


Figure S9. Geoms distributions for 200 ns of MD simulation of BTT¹⁴ using parameters obtained through Swarm-CG initial BI via execution *mode 2* (initial guess only: CG in blue vs. reference AA model in red).

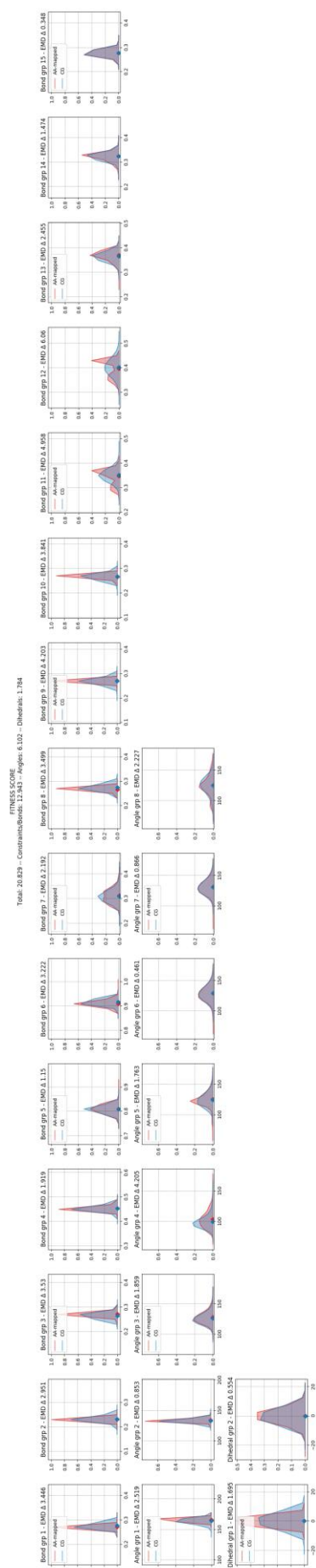


Figure S10. Geoms distributions for 200 ns of MD simulation of BTI¹⁴ using parameters obtained through Swarm-CG optimization process via execution mode 1 (final PSO result: optimized CG parametrization in blue, vs. reference AA model in red).

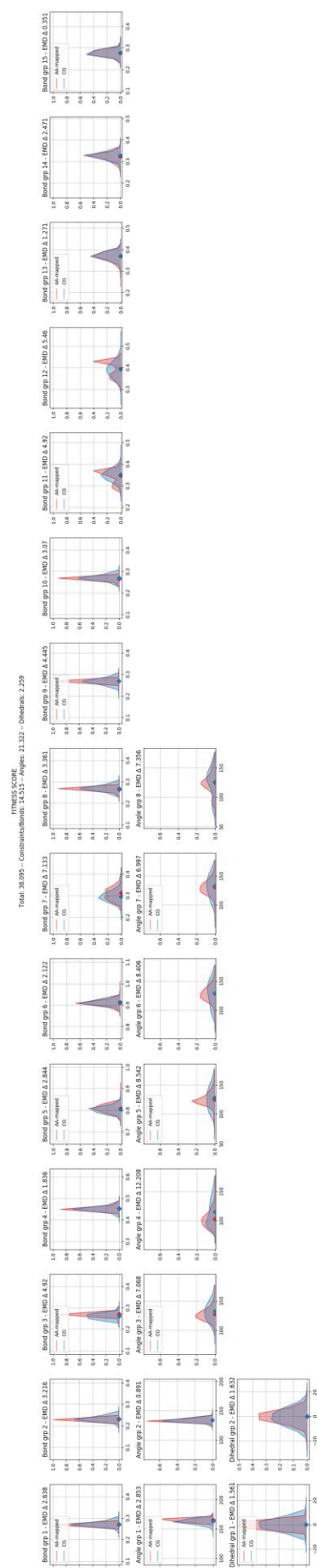


Figure S11. Geoms distributions for 200 ns of MD simulation of BTI¹⁴ using parameters obtained through Swarm-CG optimization process via execution mode 2 (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

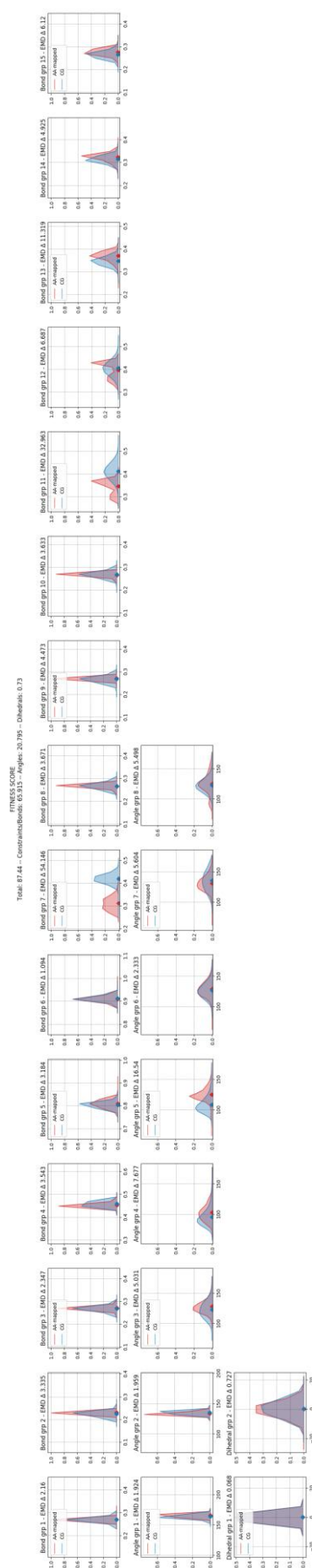


Figure S12. Geoms distributions for 200 ns of MD simulation of BTI¹⁴ using parameters obtained from the literature¹⁴ (manual CG parametrization in blue vs. reference AA model in red).

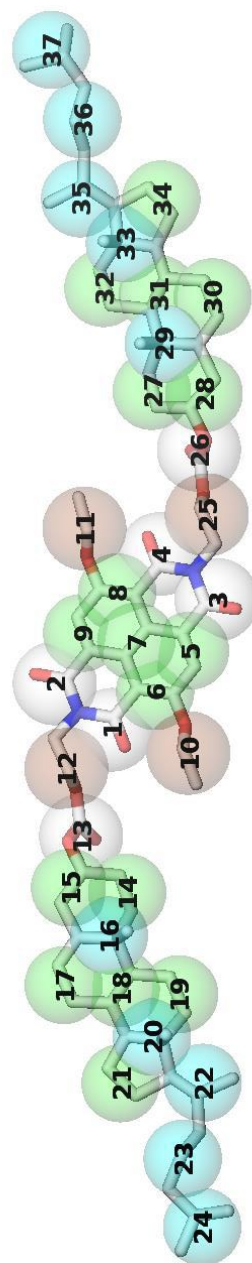


Figure S13. Molecular structure of the NDI-based molecule studied herein¹²; AA model in sticks with superimposed transparent CG beads with mapping and numbering.

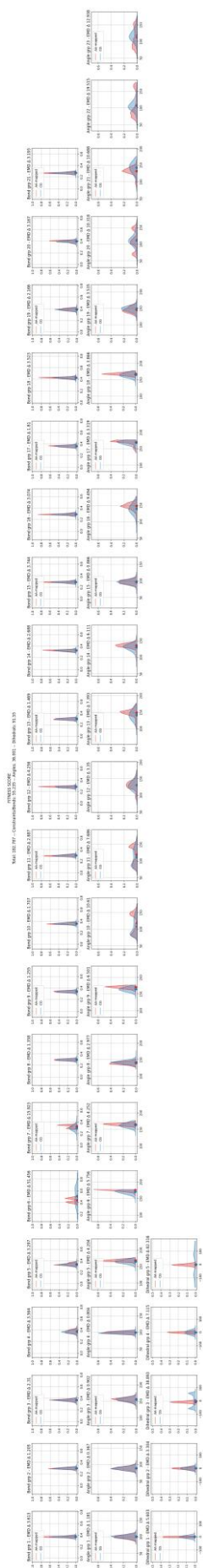


Figure S14. Geoms distributions for 200 ns of MD simulation of ND1¹² using parameters obtained through Swarm-CG initial BI via execution *mode 1* (initial guess only): CG in blue vs. reference AA model in red).

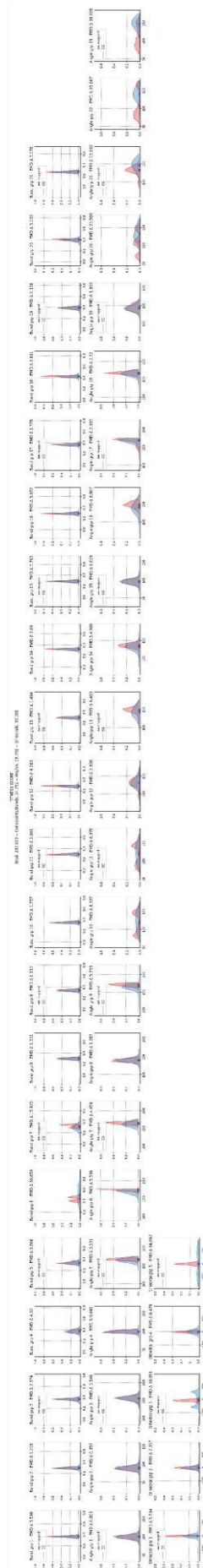


Figure S15. Geoms distributions for 200 ns of MD simulation of ND1¹² using parameters obtained through Swarm-CG initial BI via execution *mode 2* (initial guess only): CG in blue vs. reference AA model in red).

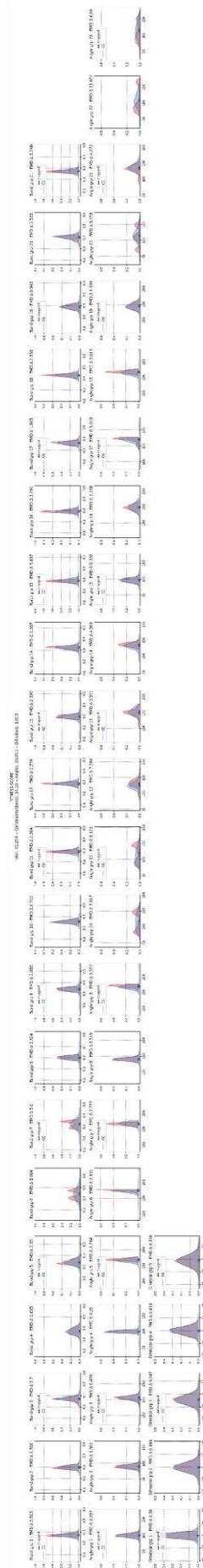


Figure S16. Geoms distributions for 200 ns of MD simulation of ND1¹² using parameters obtained through Swarm-CG optimization process via execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

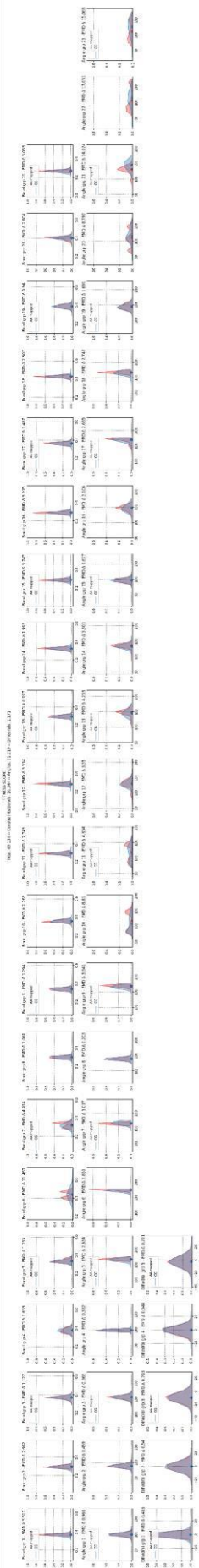


Figure S17. Geoms distributions for 200 ns of MD simulation of ND1¹² using parameters obtained through Swarm-CG optimization process via execution mode 2 (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

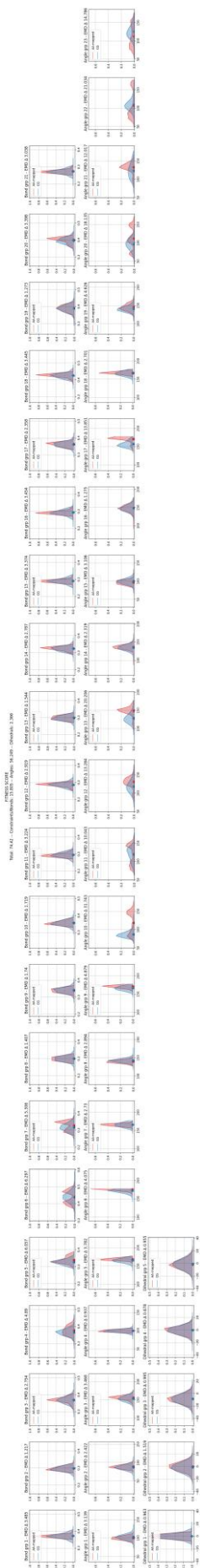


Figure S18. Geoms distributions for 200 ns of MD simulation of ND1¹² using parameters obtained from the literature¹² (manual CG parametrization in blue vs. reference AA model in red).

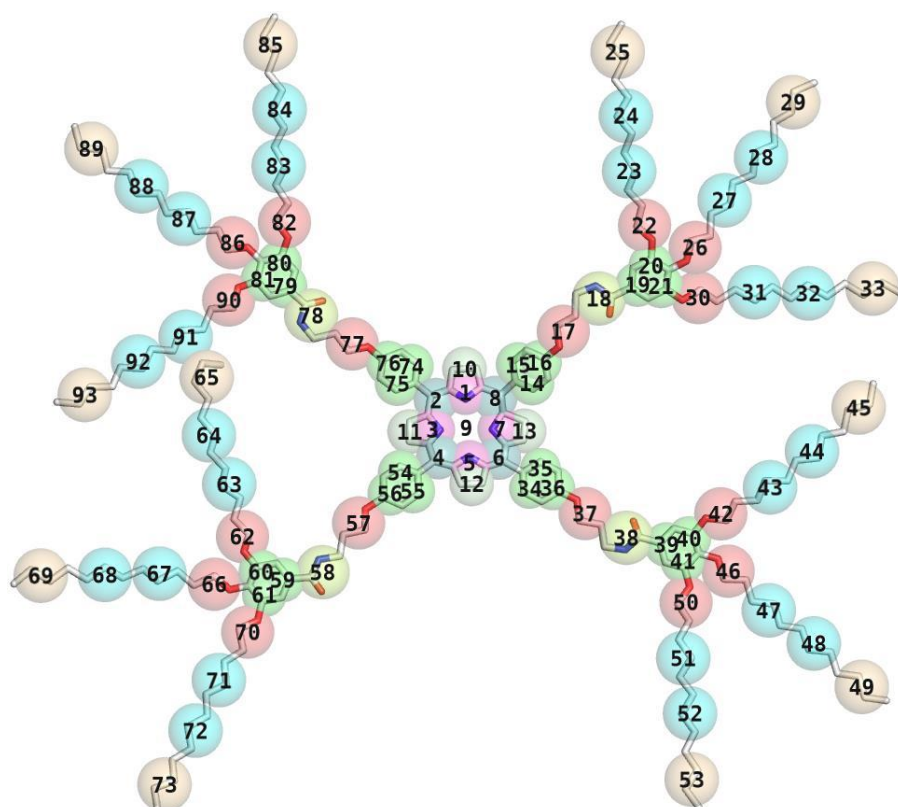


Figure S19. Molecular structure of the Zn-porphyrin molecule studied herein¹⁶: AA model in sticks with superimposed transparent CG beads with mapping and numbering.

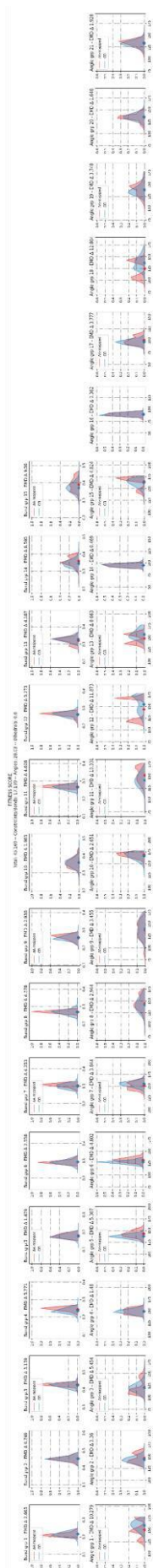


Figure S20. Geoms distributions for 200 ns of MD simulation of the Zn-porphyrin based monomer¹⁵ using parameters obtained through *Swarm-CG* initial BI via execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

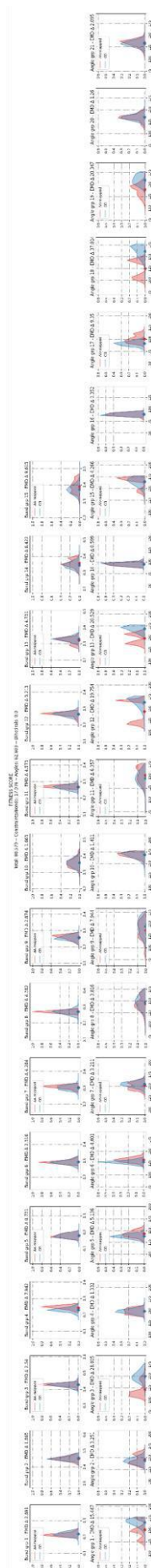


Figure S21. Geoms distributions for 200 ns of MD simulation of the Zn-porphyrin based monomer¹⁵ using parameters obtained through *Swarm-CG* initial BI via execution *mode 2* (initial guess only: CG in blue vs. reference AA model in red).

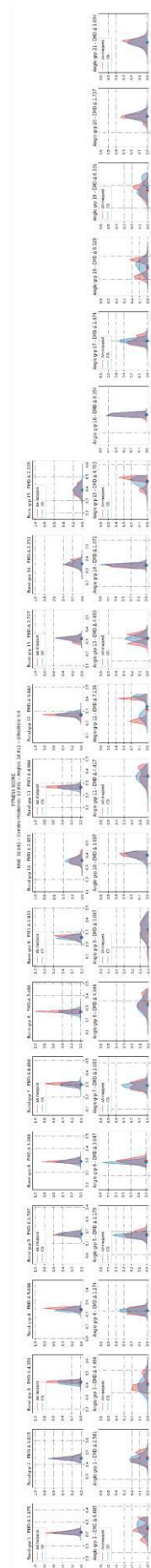


Figure S22. Geoms distributions for 200 ns of MD simulation of the Zn-porphyrin based monomer¹⁵ using parameters obtained through *Swarm-CG* optimization process via execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

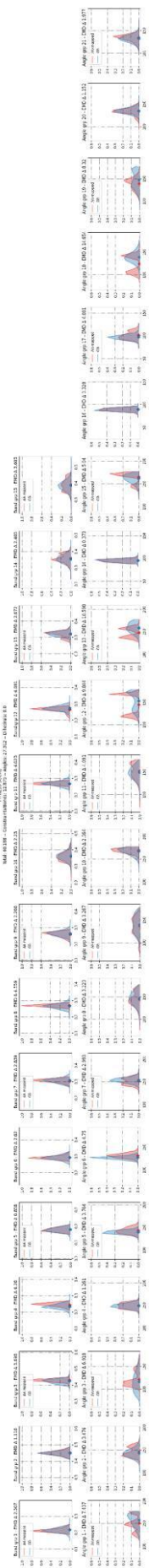


Figure S23. Geoms distributions for 200 ns of MD simulation of the Zn-porphyrin based monomer¹⁵ using parameters obtained through Swarm-CG optimization process via execution *mode 2* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

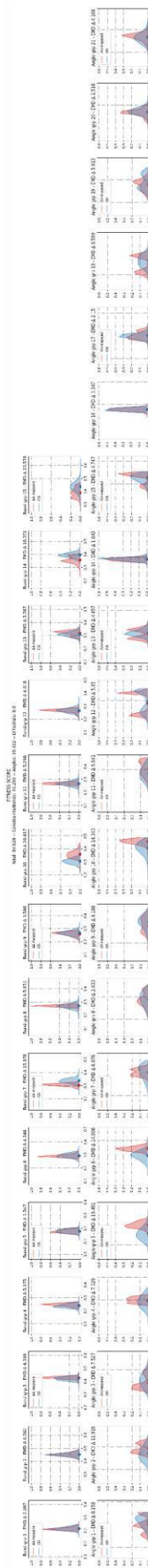


Figure S24. Geoms distributions for 200 ns of MD simulation of the Zn-porphyrin based monomer¹⁵ using parameters obtained from the literature¹⁵ (manual CG parametrization in blue vs. reference AA model in red).

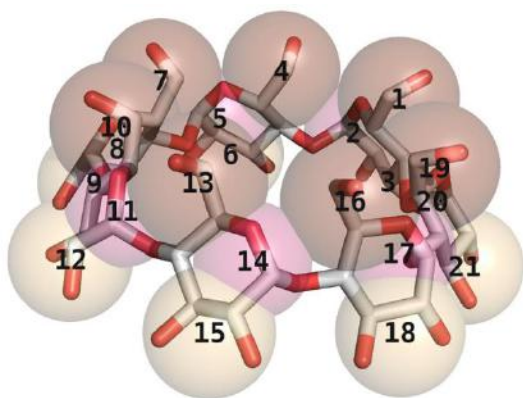


Figure S25. Molecular structure of the β -cyclodextrin studied herein¹⁷: AA model in sticks and superimposed transparent CG beads with mapping and numbering.

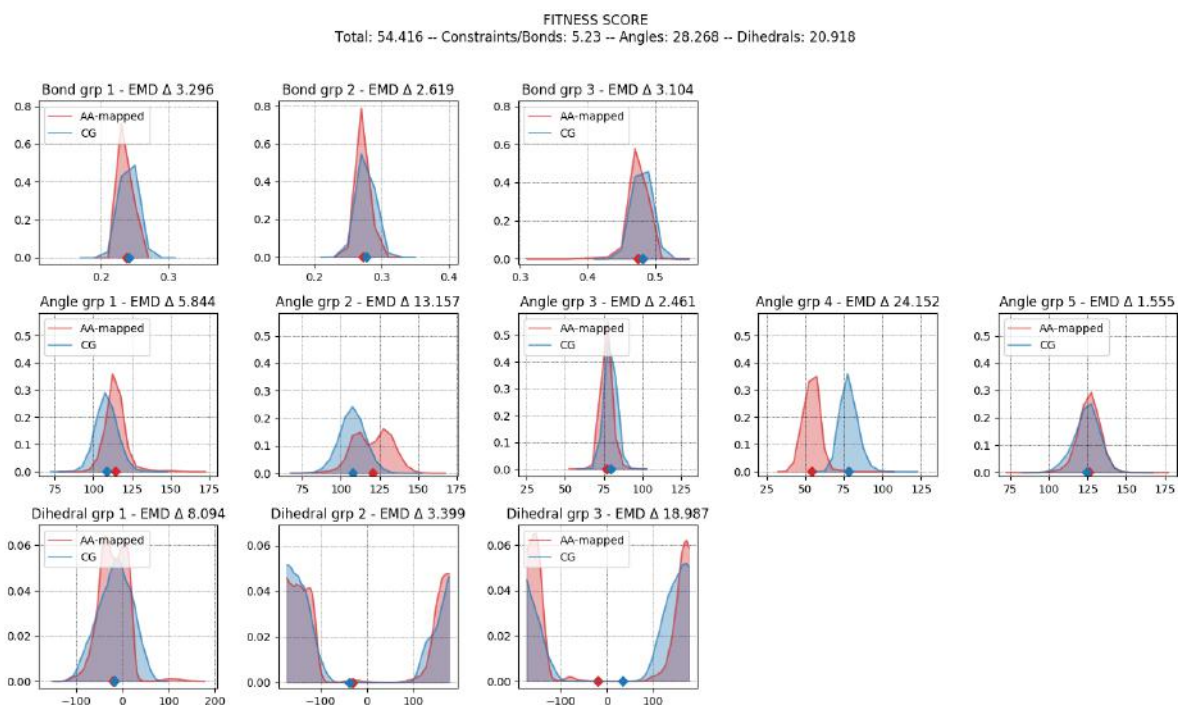


Figure S26. Geoms distributions for 200 ns of MD simulation of β -cyclodextrin¹⁷ using parameters obtained through *Swarm-CG* initial BI *via* execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

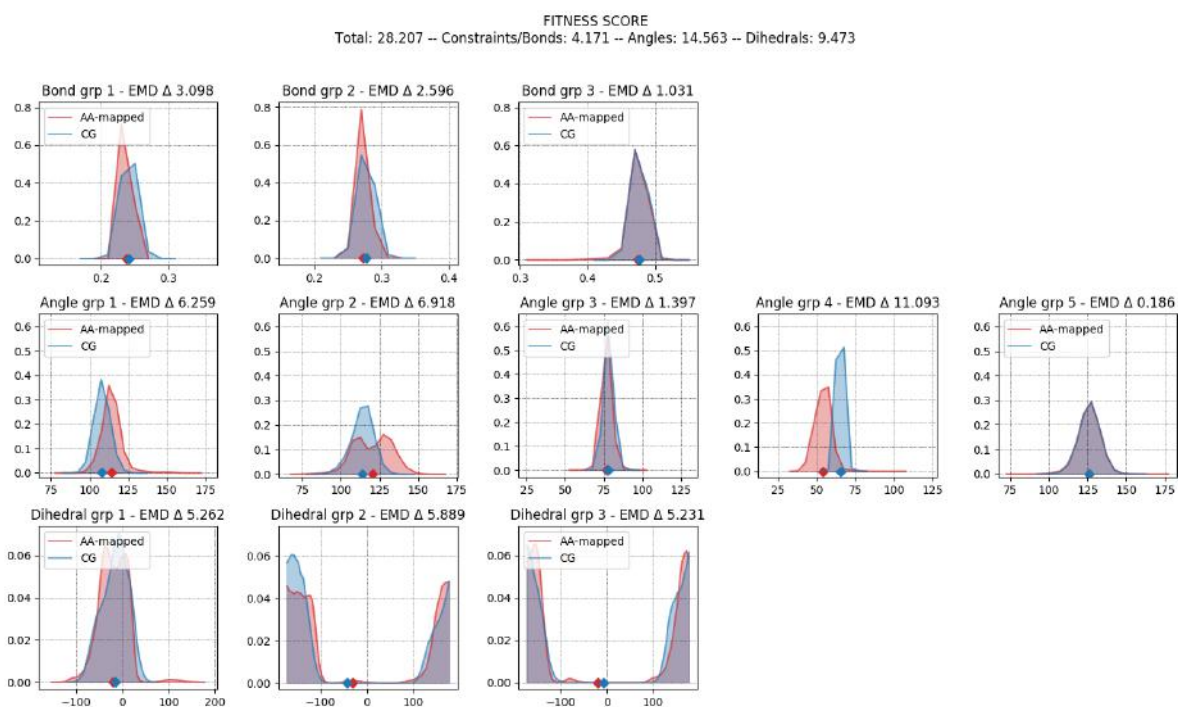


Figure S27. Geoms distributions for 200 ns of MD simulation of β -cyclodextrin¹⁷ using parameters obtained through *Swarm-CG* optimization process *via* execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

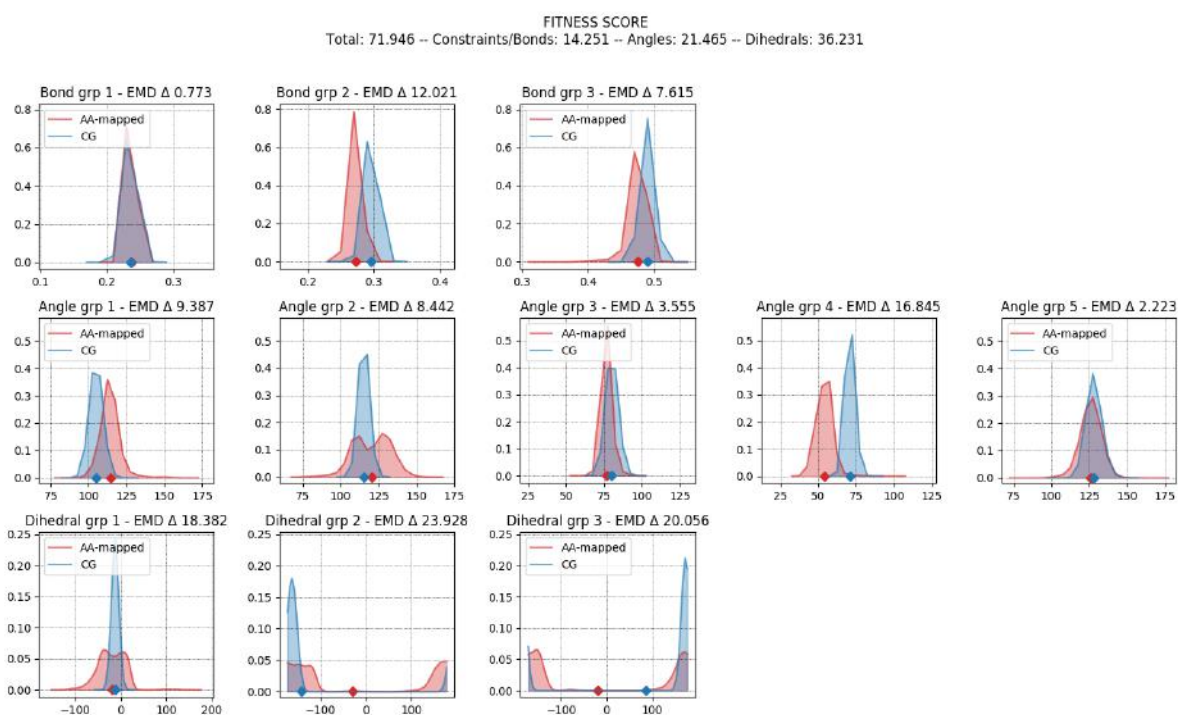


Figure S28. Geoms distributions for 200 ns of MD simulation of β -cyclodextrin¹⁷ using parameters obtained from the literature¹⁷ (manual CG parametrization in blue vs. reference AA model in red).

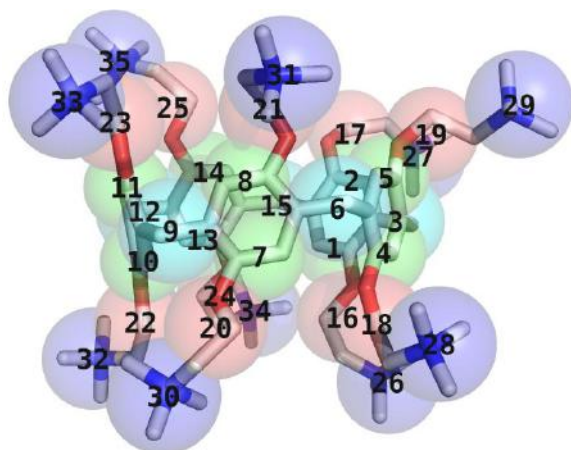


Figure S29. Molecular structure of the pillar[5]arene studied herein¹⁸: AA model in sticks and superimposed transparent CG beads with mapping and numbering.

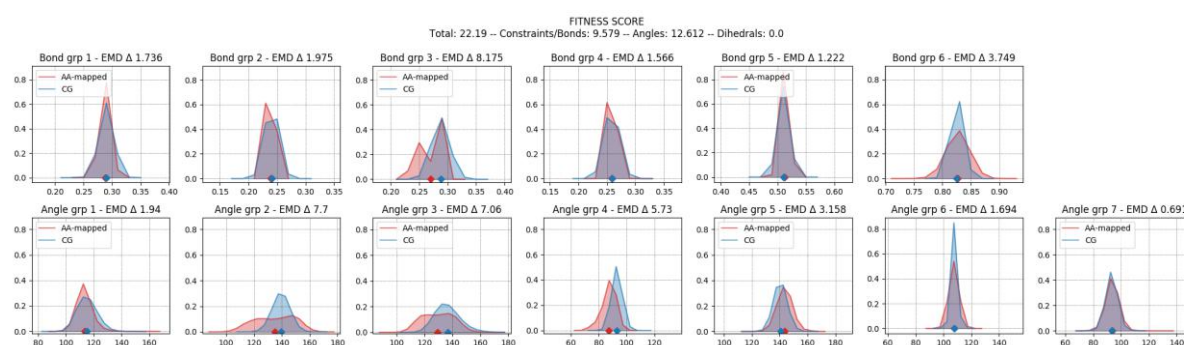


Figure S30. Geoms distributions for 200 ns of MD simulation of the pillar[5]arene¹⁸ using parameters obtained through *Swarm-CG* initial BI *via* execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

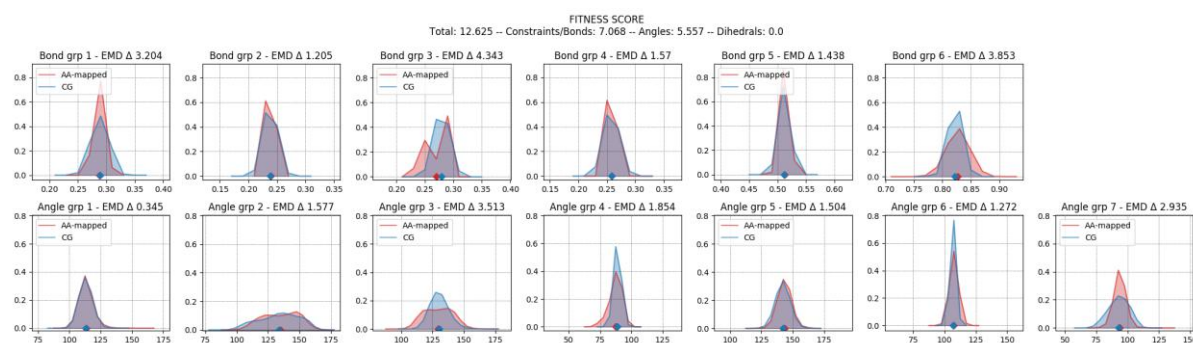


Figure S31. Geoms distributions for 200 ns of MD simulation of the pillar[5]arene¹⁸ using parameters obtained through *Swarm-CG* optimization process *via* execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

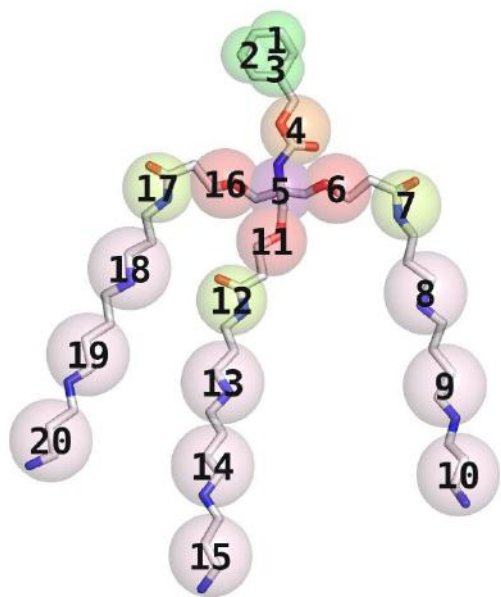


Figure S32. Molecular structure of the spermine-functionalized dendron studied herein^{19,20}: AA model in sticks and superimposed transparent CG beads with mapping and numbering.

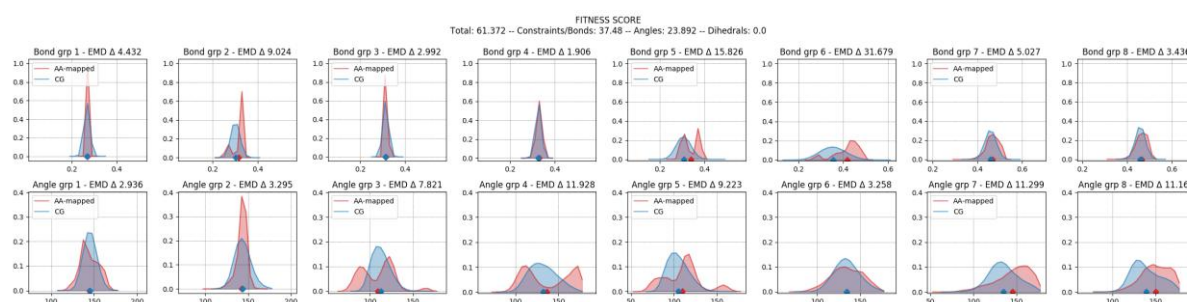


Figure S33. Geoms distributions for 200 ns of MD simulation of the spermine dendron¹⁹ using parameters obtained through *Swarm-CG* initial BI via execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

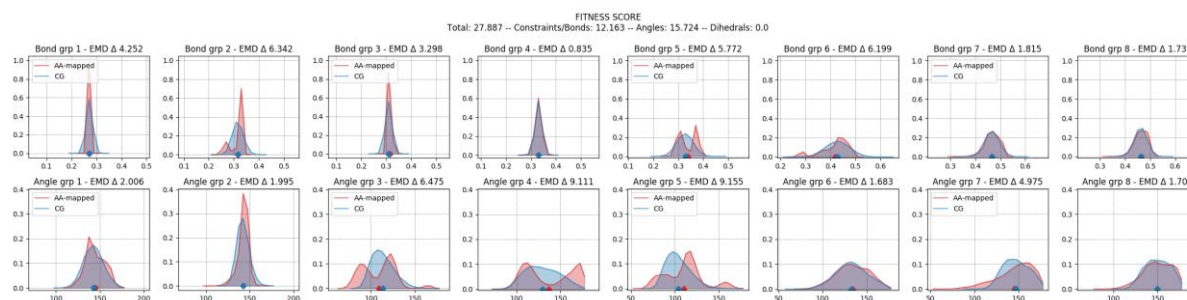


Figure S34. Geoms distributions for 200 ns of MD simulation of the spermine dendron¹⁹ using parameters obtained through *Swarm-CG* optimization process via execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

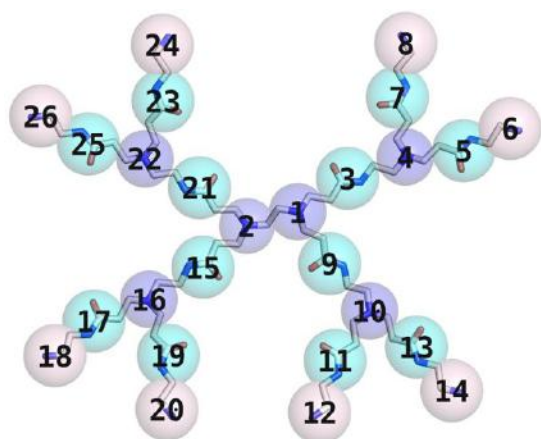


Figure S35. Molecular structure of PAMAM G1 dendrimer²¹: AA model in sticks and superimposed transparent CG beads with mapping and numbering.

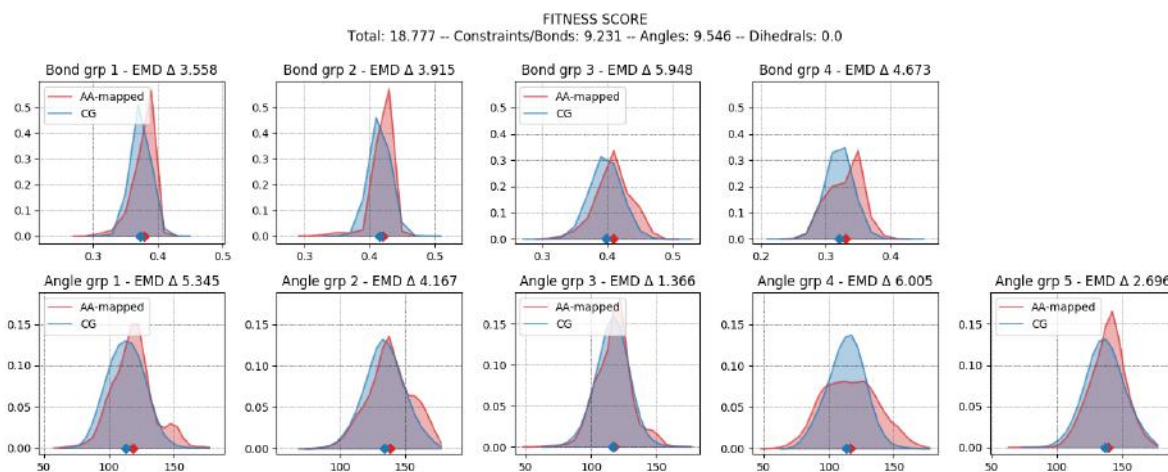


Figure S36. Geoms distributions for 200 ns of MD simulation of PAMAM G1²¹ using parameters obtained through *Swarm-CG* initial BI *via* execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

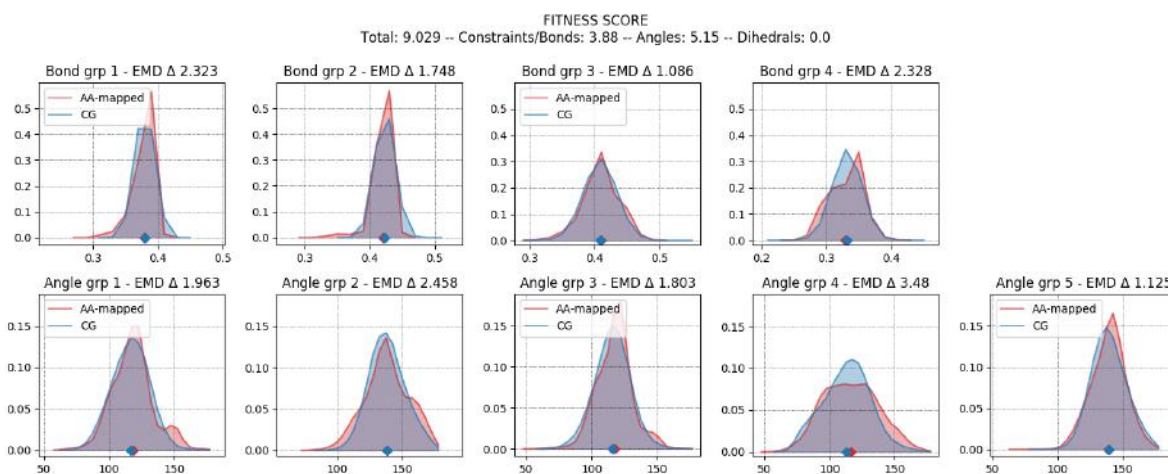


Figure S37. Geoms distributions for 200 ns of MD simulation of PAMAM G1²¹ using parameters obtained through *Swarm-CG* optimization process *via* execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

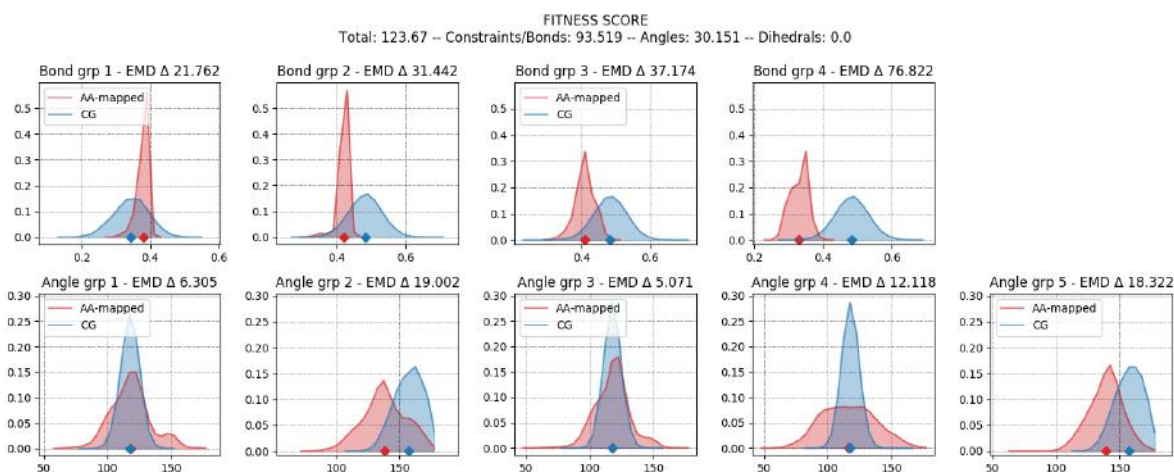


Figure S38. *Geoms* distributions for 200 ns of MD simulation of PAMAM G1²¹ using parameters obtained from the literature²¹ (manual CG parametrization in blue vs. reference AA model in red).

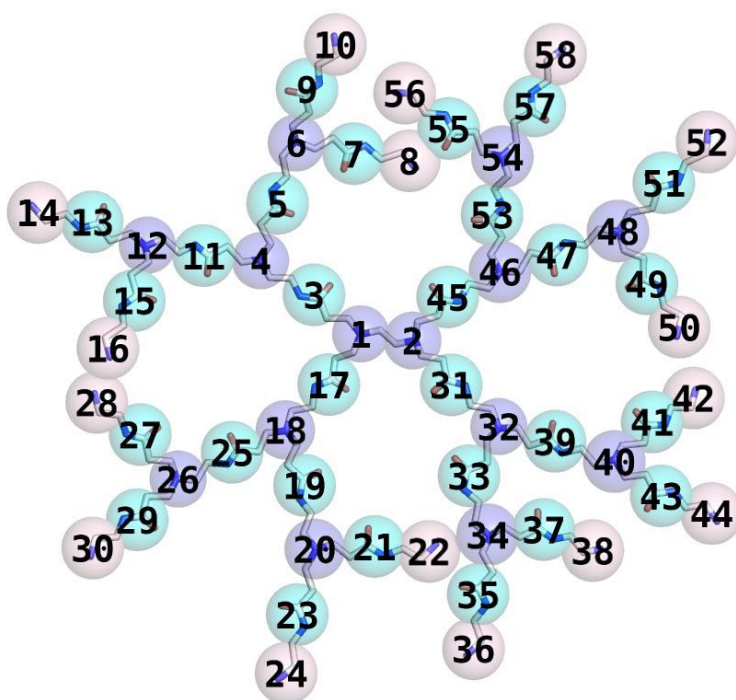


Figure S39. Molecular structure of PAMAM G2 dendrimer²¹: AA model in sticks and superimposed transparent CG beads with mapping and numbering.

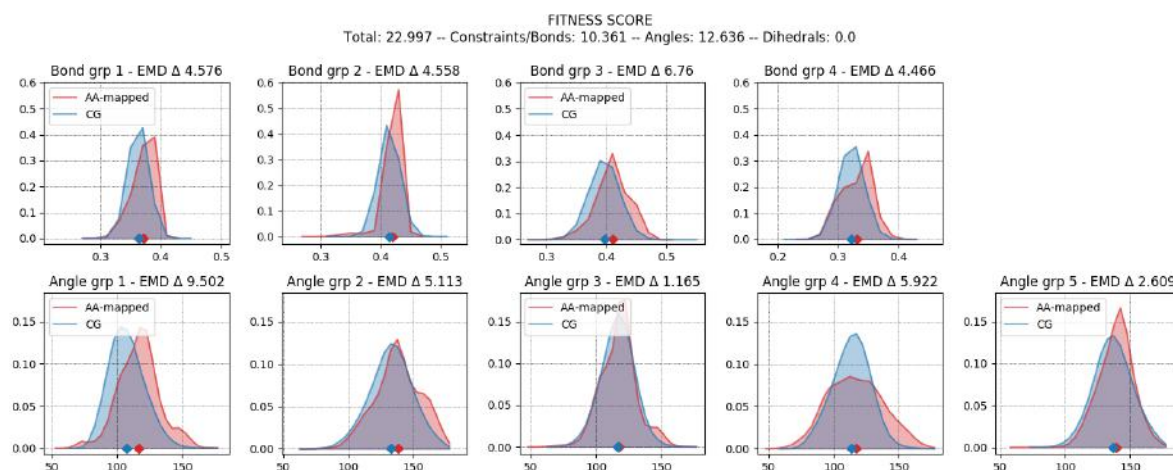


Figure S40. Geoms distributions for 200 ns of MD simulation of PAMAM G2²¹ using parameters obtained through *Swarm-CG* initial BI via execution *mode 1* (initial guess only: CG in blue vs. reference AA model in red).

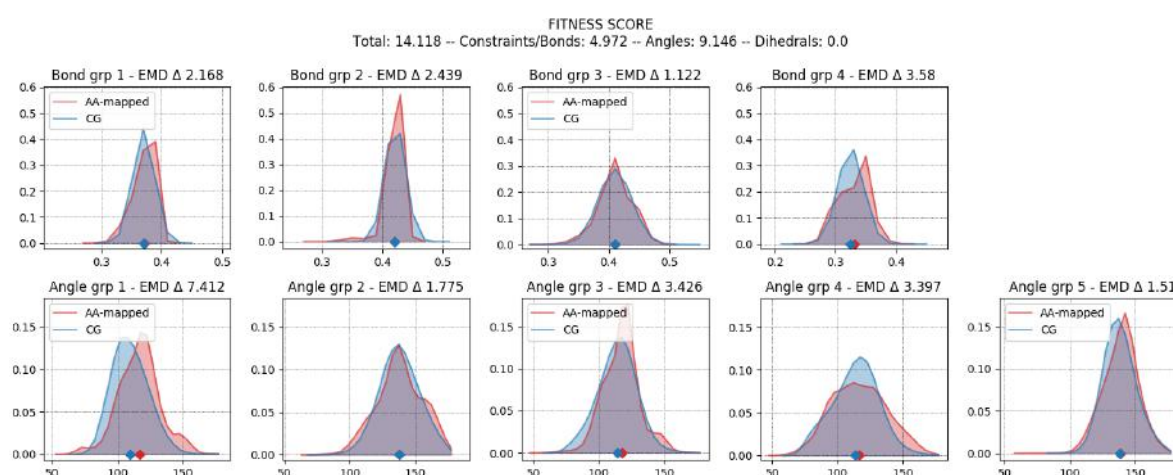


Figure S41. Geoms distributions for 200 ns of MD simulation of PAMAM G2²¹ using parameters obtained through *Swarm-CG* optimization process via execution *mode 1* (final PSO result: optimized CG parametrization in blue vs. reference AA model in red).

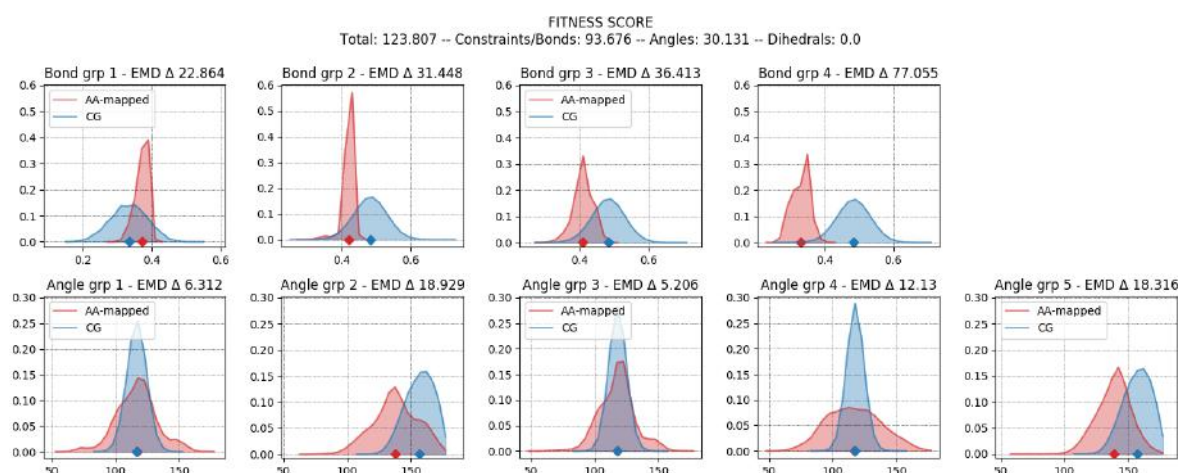


Figure S42. Geoms distributions for 200 ns of MD simulation of PAMAM G2²¹ using parameters obtained from the literature²¹ (manual CG parametrization in blue vs. reference AA model in red).

Supplementary References

- (S1) Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **1985**, *18* (3), 534–552. <https://doi.org/10.1021/ma00145a039>.
- (S2) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates. *Acta Polym.* **1998**, *49* (2–3), 61–74. [https://doi.org/10.1002/\(SICI\)1521-4044\(199802\)49:2/3<61::AID-APOL61>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1521-4044(199802)49:2/3<61::AID-APOL61>3.0.CO;2-V).
- (S3) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, 1–12. <https://doi.org/10.1038/s41592-019-0686-2>.
- (S4) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B.; the GROMACS development team. GROMACS User Manual Version 2019.
- (S5) Nobile, M. S.; Cazzaniga, P.; Besozzi, D.; Colombo, R.; Mauri, G.; Pasi, G. Fuzzy Self-Tuning PSO: A Settings-Free Algorithm for Global Optimization. *Swarm Evol. Comput.* **2018**, *39*, 70–85. <https://doi.org/10.1016/j.swevo.2017.09.001>.
- (S6) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91* (1), 43–56. [https://doi.org/10.1016/0010-4655\(95\)00042-E](https://doi.org/10.1016/0010-4655(95)00042-E).
- (S7) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, 1–2, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- (S8) Canonical sampling through velocity rescaling: The Journal of Chemical Physics: Vol 126, No 1 <https://aip.scitation.org/doi/10.1063/1.2408420> (accessed Jun 2, 2020).
- (S9) Molecular dynamics with coupling to an external bath: The Journal of Chemical Physics: Vol 81, No 8 <https://aip.scitation.org/doi/10.1063/1.448118> (accessed Jun 2, 2020).
- (S10) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. <https://doi.org/10.1063/1.464397>.
- (S11) 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* (27), 7812–7824. <https://doi.org/10.1021/jp071097f>.
- (S12) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* **1995**, *16* (3), 273–284. <https://doi.org/10.1002/jcc.540160303>.
- (S13) Sarkar, A.; Sasmal, R.; Empereur-mot, C.; Boichicchio, D.; Kompella, S. V. K.; Sharma, K.; Dhiman, S.; Sundaram, B.; Agasti, S. S.; Pavan, G. M.; George, S. J. Self-Sorted, Random, and Block Supramolecular Copolymers via Sequence Controlled, Multicomponent Self-Assembly. *J. Am. Chem. Soc.* **2020**, *142* (16), 7606–7617. <https://doi.org/10.1021/jacs.0c01822>.
- (S14) Boichicchio, D.; Pavan, G. M. From Cooperative Self-Assembly to Water-Soluble Supramolecular Polymers Using Coarse-Grained Simulations. *ACS Nano* **2017**, *11* (1), 1000–1011. <https://doi.org/10.1021/acsnano.6b07628>.
- (S15) Casellas, N. M.; Pujals, S.; Boichicchio, D.; Pavan, G. M.; Torres, T.; Albertazzi, L.; García-Iglesias, M. From Isodesmic to Highly Cooperative: Reverting the Supramolecular Polymerization Mechanism in Water by Fine Monomer Design. *Chem. Commun.* **2018**, *54* (33), 4112–4115. <https://doi.org/10.1039/C8CC01259H>.

- (S16) Jung, S. H.; Bochicchio, D.; Pavan, G. M.; Takeuchi, M.; Sugiyasu, K. A Block Supramolecular Polymer and Its Kinetically Enhanced Stability. *J. Am. Chem. Soc.* **2018**, *140* (33), 10570–10577. <https://doi.org/10.1021/jacs.8b06016>.
- (S17) López, C. A.; de Vries, A. H.; Marrink, S. J. Computational Microscopy of Cyclodextrin Mediated Cholesterol Extraction from Lipid Model Membranes. *Sci. Rep.* **2013**, *3* (1), 1–6. <https://doi.org/10.1038/srep02071>.
- (S18) Beyeh, N. K.; Nonappa; Liljeström, V.; Mikkilä, J.; Korpi, A.; Bochicchio, D.; Pavan, G. M.; Ikkala, O.; Ras, R. H. A.; Kostianen, M. A. Crystalline Cyclophane–Protein Cage Frameworks. *ACS Nano* **2018**, *12* (8), 8029–8036. <https://doi.org/10.1021/acsnano.8b02856>.
- (S19) Pavan, G. M.; Danani, A.; Pricl, S.; Smith, D. K. Modeling the Multivalent Recognition between Dendritic Molecules and DNA: Understanding How Ligand “Sacrifice” and Screening Can Enhance Binding. *J. Am. Chem. Soc.* **2009**, *131* (28), 9686–9694. <https://doi.org/10.1021/ja901174k>.
- (2S0) Pavan, G. M.; Danani, A. The Influence of Dendron’s Architecture on the “Rigid” and “Flexible” Behaviour in Binding DNA—a Modelling Study. *Phys. Chem. Chem. Phys.* **2010**, *12* (42), 13914–13917. <https://doi.org/10.1039/C0CP01124J>.
- (S21) Lee, H.; Larson, R. G. Coarse-Grained Molecular Dynamics Studies of the Concentration and Size Dependence of Fifth- and Seventh-Generation PAMAM Dendrimers on Pore Formation in DMPC Bilayer. *J. Phys. Chem. B* **2008**, *112* (26), 7778–7784. <https://doi.org/10.1021/jp802606y>.