JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 5045−5056

pubs.acs.org/jcim

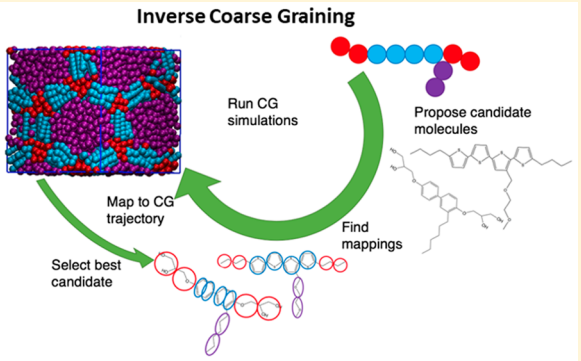
Framework for Inverse Mapping Chemistry-Agnostic Coarse-Grained Simulation Models into Chemistry-Specific Models

Christian Nowak, Mayank Misra, and Fernando A. Escobedo*

School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, United States

Supporting Information

ABSTRACT: Coarse-grained (CG) models have allowed molecular simulations to access large enough time and length scales to elucidate relationships between macroscale properties and microscale molecular interactions. However, an unaddressed inverse-design problem concerns the identification of an optimal chemistry-specific (CS) molecule that the generic CG model represents. This has been addressed here by introducing new tools for automatically generating and refining the mapping of CS-molecule candidates to the constraints of a CG model, based on representative optimization criteria. With these tools, for each CS-molecule from a candidate group, the best mapping of that molecule onto the CG model is found and their fit is assessed by an objective function designed to emphasize matching key properties of the CG model. We employ this methodology to a range of CG models from small solvent molecules up to block copolymer systems to show its ability to find optimal candidates and to uncover the underlying length scale of some of the CG models. For instances where the identity of the CG model is known a priori, the methodology identifies the correct AA chemistry. For instances where the identity is unknown and a pool of candidates is provided, the method selects a chemistry that aligns well with physical intuition. The best candidate chemistry is also found to be sensitive to changes to the CG model.



INTRODUCTION

The ever-expanding availability of computational resources has fueled a fast growth in the size and scope of the molecular simulations currently used for property prediction. Indeed, with resources such as XSEDE, an NSF-funded collection of computational resources,[1] simulations involving hundreds of processors and millions of atoms are potentially viable.[2−5] Despite these advances, there still exist many physical and chemical processes whose length scales are beyond the reach of computationally accessible time scales, such as those involving large biomolecules and other macromolecules. Indeed, such simulations often encounter rugged free-energy landscapes and kinetic trapping in deep metastable basins. To address these kinetic barriers, techniques such as parallel tempering/replica exchange,[6−9] metadynamics,[10−12] transition-path sampling,[13,14] and kinetic Monte Carlo[15,16] have been developed and used.

One of the most successful approaches to speed up molecular simulations has been the use of coarse-grained (CG) models, i.e., models that bundle groups of beads from a more detailed model into single beads to thus eliminate microscopic degrees of freedom that are not essential to resolve structural details above a certain length scale. CG models have smoother potential energy surfaces that are easier to sample ergodically compared to their all-atom (AA) counterparts, whose rougher potential energy landscape can create kinetic traps.[17] Indeed, studies using CG models have been able to access such collective properties as the self-assembly behavior for systems where traditional AA models would be intractable with typically available computational resources.[18−21]

The CG models for macromolecules can be broadly classified into two categories: (i) "chemical" models if mapped directly from a chemistry-specific (CS) polymer and (ii) "physical" models if intended to describe a broad class of polymers. In the former directly mapped CG or "DCG" models, their parameters are calibrated to match selected results of properties obtained from experiments or AA simulations of the material of interest. In the latter case, one begins with a relatively small simulation of the AA molecule and a CG molecule with a specific CG mapping (i.e., a recipe for the way how atoms in the AA molecule are mapped into the different CG beads); such a mapping is often guided by physical intuition. Once suitable functional forms have been selected for the bonded and nonbonded interaction potentials, the model can be parametrized by such methods as iterative Boltzmann inversion,[22,23] force matching,[24,25] or relative entropy.[26,27] The goal of the parametrization is to construct CG molecules such that their behavior mimics that of the known molecule at a prescribed level of detail. The degree of

Received: March 17, 2019
Published: November 19, 2019

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

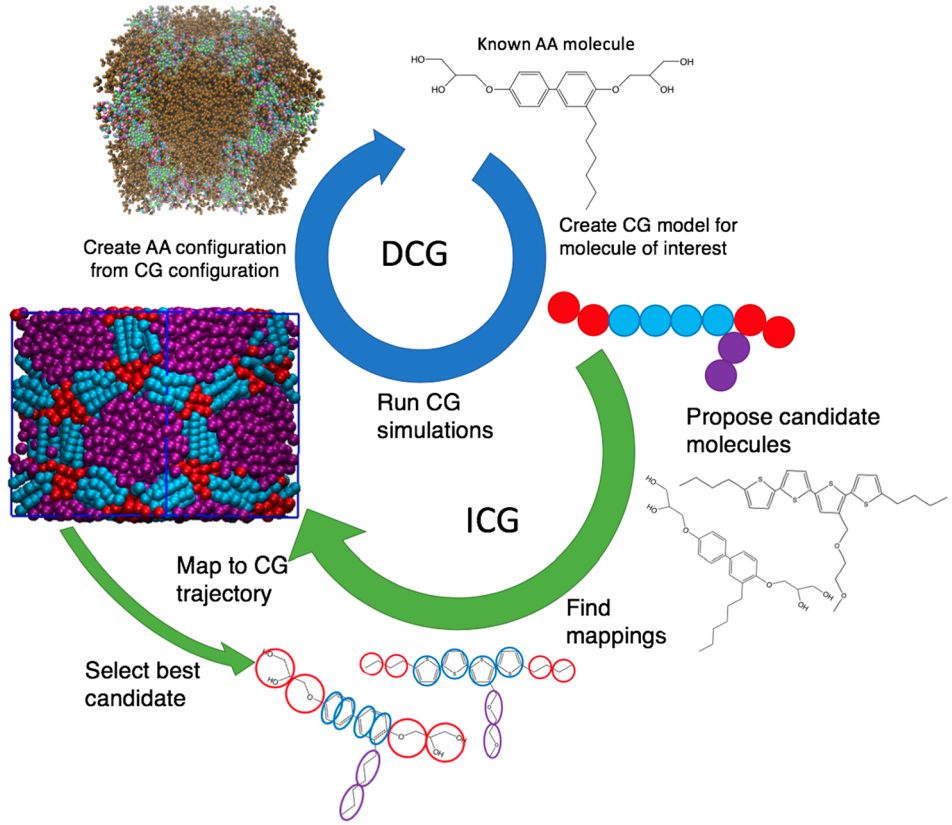


Figure 1. Schematic of the DCG (blue arrow) and ICG (green arrows) processes. DCG begins with parametrizing a CG model based on simulations of the original AA CS-molecule. The CG trajectory can then undergo reverse coarse-graining (RCG) to return an AA structure. In ICG, only the CG model is known and candidate molecules are proposed. Each CS-molecule has an optimal mapping that most closely replicates the original CG model. The trajectory of each CS-molecule is coarse-grained using this mapping into a trajectory of the CG model. These trajectories are compared to yield a best candidate.

coarsening (DOC); i.e., how much detail is averaged out, can range from simply integrating out the hydrogens like in typical united atom (UA) models,[28] to lumping entire monomers/amino-acids or even long sections of polymer coils into single beads as in models originally used with dissipative particle dynamics (DPD).[29,30]

Due to the simplicity and computational efficiency of CG models with greater DOC, many studies have adopted generic, physical CG models intended to capture the typical behavior of a class of polymers rather than that of any specific polymer chemistry, such as the widely used bead−spring chain models introduced by Kremer and Grest.[31] The use of chemistry-agnostic models is common not only to molecular simulations but also to theoretical polymer physics[32] where the goal is to broadly describe polymer behavior rather than the behavior of specific chemistries. While many of these CG simulation studies are able to reproduce experimentally known physical trends of some macromolecules, others can also reveal new or unusual results. In the latter case, it would be of interest to identify specific molecular chemistries (i.e., AA models) that could be good candidates to capture the predicted behavior of the generic CG molecules investigated. The process of determining the identity of these AA molecules is henceforth referred to as inverse coarse-graining (ICG).

Figure 1 schematically compares DCG and ICG. If it is known that a given AA molecular model has a specific property of interest, then DCG would be a suitable approach to explore perturbations in behavior in a close proximity of compositional space. If the goal is to widely explore a potentially novel type of behavior, ICG would be a suitable methodology because generic CG molecular modes are typically "coarser" and more computationally efficient. Once a CG model has generated results of interest,[33−35] candidate AA CS-molecules need to be determined to guide experimental efforts toward realizing such predictions. ICG can thus become a powerful strategy in materials design, complementary to the existing DCG method and well aligned with the objectives of the Materials Genome Initiative.[36] We note that DCG and ICG are somewhat related but not the same as bottom-up and top-down CG approaches, respectively. DCG can be implemented via either a top-down or bottom-up approaches. In a top-down CG approach, experimental data are typically used as the target data to reproduce, and in ICG one can view the observables from the CG model simulations as the input data playing a role akin to "experimental data". But importantly, ICG is not about coarse-graining but rather "fine-graining". On the one hand, a main advantage of ICG over DCG is that the former avoids the difficulty of finding the optimal number of particles and the topology of the low-resolution model, which can present significant challenges. On the other hand, ICG is restricted by the availability of suitable high-resolution models.

ICG has not been as well studied as DCG, partly due to the ill-posed nature of the ICG problem compared to DCG. Indeed, for a given AA CS-molecule and CG model, there exists in principle one optimal set of model parameters, but for a given CG model, many different AA CS-molecules can be mapped onto the same CG molecule. In this context, ICG is much more dependent on the specific criterion adopted to

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

determine the goodness of fit. This important difference between reproducing the behavior of a well-defined chemistry (DCG) and reproducing the behavior of a virtual, "fuzzy" CG model by assigning chemical identity (ICG) is the impetus for this work. While some work exists that tried to generalize a set of chemistries to a single CG model,[37−40] a directed evolutionary approach would be highly desirable, so that the candidates for the optimal AA CS-molecule can be evaluated and evolved toward the best fit of the CG model, ideally, in an automated way (e.g., aided by machine-learning techniques).[41−46] Previous work that employed methodologies similar to the ICG process have candidate pools upward of 5×10^5 molecules,[47,48] further stressing the need for an automated process.

As a materials discovery strategy, ICG could be used to optimize a CG model by tweaking its parameters and extensively mapping associated phase diagrams, until an interesting or unique behavior is seen or enhanced. This leverages the high computational efficiency of CG models. After a CG model with the sought-after behavior is established, ICG would carry out the task of finding a CS molecule that can reproduce that behavior, even if only qualitatively. Two recent examples serve to illustrate how chemistry-agnostic CG models have been used to predict new mesophase behavior that is yet to be mapped to any CS models. Both a binary blend of CG particles exhibiting "positive mixing"[49] and bolaamphiphiles having three chemical blocks[20] have been shown to form complex mesophases, often with 3D networks of different domain types. If such periodic structures were mapped into suitable chemistries, they could have highly appealing optical, electronic, or catalytic properties. For example, a CG model of bolaamphihiles has predicted the formation of a single diamond phase and a single plumber's nightmare phase,[20] both of which had not been realized by CS models or experiments.[50,51] While tackling such phase mappings is beyond the scope of this work, it provides motivation for taking an initial step toward ICG strategies.

ICG STRATEGY. In this work we propose a flexible ICG framework for determining which AA CS-molecule from a given pool of candidate molecules is the best fit for a target CG model based on minimizing an objective function. Our approach is similar to that used in the relative entropy (RE) methodology.[27] Shell et al.[26] used the concept of RE to iteratively parametrize a CG model in DCG, where at each iteration a system of the CG molecules is equilibrated under a given set of potential parameters (defining the CG Hamiltonian $\mathcal{H}_{CG}$) to get the "true" CG trajectory, T_t. Then a separate equilibrated trajectory of the AA CS-molecular system is mapped to return a new CG trajectory, T_{AA}. $\mathcal{H}_{CG}$ and a function f which depends on $\mathcal{H}_{CG}$ are evaluated for each frame of T_t and T_{AA}, effectively "simulating" both trajectories under the CG force field. The objective function, ϕ, measures the degree to which T_t and T_{AA} differ by evaluating $\phi = f_t - f_{AA}$, where f_i is the average value of the function f obtained when "simulating" T_i. This difference determines the changes to be made to $\mathcal{H}_{CG}$ (i.e., the CG model parameters). This process is repeated until $\mathcal{H}_{CG}$ converges as ϕ is minimized.

In ICG, $\mathcal{H}_{CG}$ is fixed (i.e., the CG model does not change), and what changes is the candidate CS-molecule being considered, which can be thought of as changing $\mathcal{H}_{AA}$. For this purpose, for each candidate CS-molecule a system is simulated and coarse-grained according to a mapping which best satisfies the constraints of a desired CG molecule to give T_i^{AA}. The procedure to generate these mappings is described below. At this point we can calculate f_i^{AA} and f_{CG} for all T_i^{AA} and T_t, respectively, where the candidate CS-molecules which minimize

$$\phi = \mathrm{abs}(f_{CG} - f_{AA}^i) \tag{1}$$

will be kept for further application of a machine learning algorithm to propose new candidate CS-molecules. A schematic comparison of our method to the original RE method is given in Figure 2.

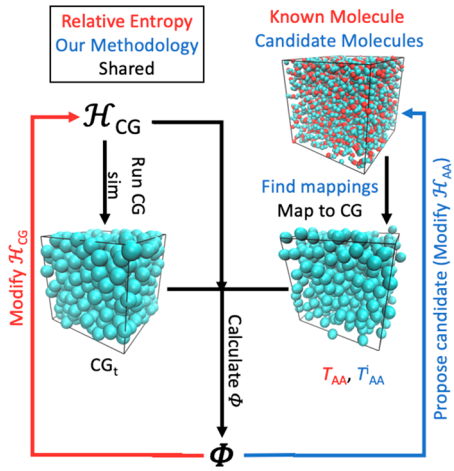


Figure 2. Schematic comparison between the relative entropy (RE) framework and the proposed methodology, where items belonging to just RE, just ICG, or both are colored in red, blue, and black, respectively. With a given $\mathcal{H}_{CG}$, a CG simulation is run to give a "true" trajectory of the CG model, T_t. In RE the AA CS-molecule is known, while in ICG multiple CS candidates are proposed and the trajectory of each is mapped onto the CG model to give T_{AA}. In RE this mapping is known a priori, while in ICG the optimal mappings must be found. Once mapped, T_t and T_{AA} are used to calculate the objective function ϕ. In RE the ϕ values are used to modify $\mathcal{H}_{CG}$, while in ICG they are used to identify the best candidate CS-molecules, which can in turn be used to propose new candidates.

Before different AA CS-molecules can be compared, the way in which the AA atoms are partitioned (mapped) to the CG molecule must be determined. There exist multiple mappings that can satisfy the constraints and several studies have explored how different mappings affect the ability of the CG molecule to reproduce properties of the AA CS-molecule.[52,53] As stated previously, mappings must satisfy key constraints of the CG model such as number of beads and bond topology as otherwise a comparison to the CG model across different chemistries would not be possible. In this study, the mapping needs to best reproduce selected properties of the CG molecule, not vice versa. Every mapping, "s", of a given CS-molecule onto the target CG model will return a different value of the objective function, ϕ_s, so only the "optimal" mapping which minimizes this value should be used when comparing across candidate AA CS-molecules. To facilitate finding this optimal mapping, a methodology to automate the generation of initial mappings is proposed. For each candidate CS-molecule, a process similar to ICG is followed except that, we now know the CS-molecule and CG model so what is modified

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

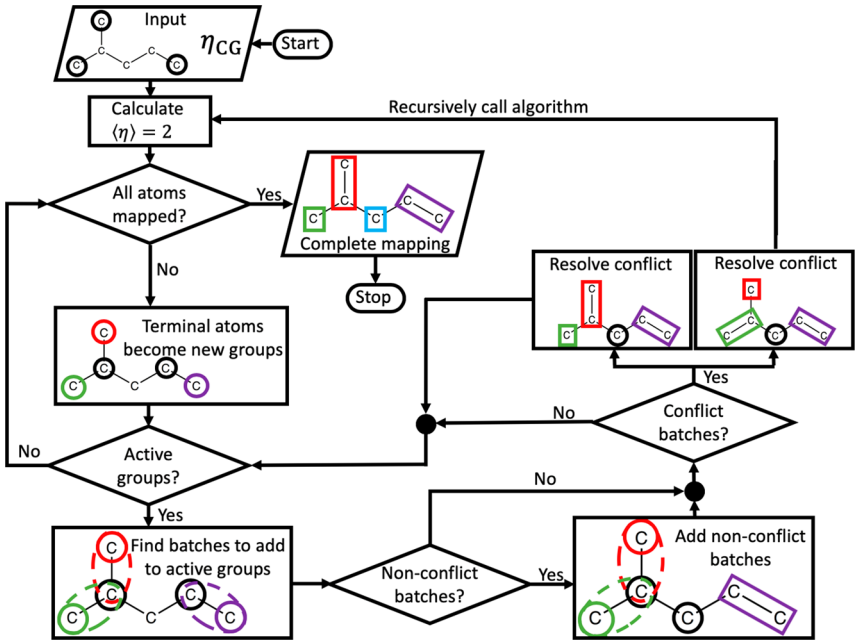


Figure 3. Depiction of algorithm to automatically generate CG mappings for a candidate molecule. It begins with (a) stripping the molecule of hydrogens to give the UA representation and placing the first initial groups starting at the terminal atoms. (b) Groups are grown until $\eta_i \leq \langle\eta\rangle$; however, if there is an atom that has been assigned to two or more groups, then new algorithm calls are initiated for each permutation where the conflicting atom is assigned to one of the groups. (c) Once all "active" groups are grown, atoms which are bonded to an atom already assigned to a group are labeled as terminal atoms and the algorithm is iterated.

through the algorithm is the mapping itself, until ϕ is no longer minimized.

For the automatic generation of a CG molecule from a given AA CS-molecule, an existing approach is the CGTools plugin for VMD[54] wherein a neural network learns to map the AA molecule based on its "shape", i.e., the position and connectivity of the constituent atoms. Coarse-graining by the "shape" is also the basis of a dimensionality reduction method[55,56] which uses data graphs and diffusion maps, where in the context of coarse-graining our AA CS-molecule, the graph is the AA CS-molecule and the dimensionality reduction is the coarse-graining process. While these two approaches show promise, neither one operates under the key constraint that ICG imposes: that the final CG molecule is known while the CG mapping is not.

Even for the more well-posed problem of DCG, the process of parametrizing a CG model is laborious and software like VOTCA[57] has been developed to automate the process. We propose an equivalent toolkit that is focused on making the ICG process more automatic. The main capabilities of this toolkit are (i) finding an appropriate mapping for the AA CS-molecule (useful for both DCG and ICG), (ii) determining a one-to-one correspondence between the beads of a proposed mapping and the beads of the target CG molecule, and (iii) finding repeating motifs in a given molecule through a compression algorithm. These capabilities are used throughout the process of finding the best AA CS-molecule for a CG model but can also be used in other contexts, such as in the reverse coarse-graining (RCG)[58−61] process that maps atomistic detail onto the CG model. Codes to perform the RCG process often require the mapping of how the AA molecule fits onto the CG model, which can be tedious to produce. By automatically finding the mapping in the ICG process, this step is taken care of, circumventing the need to generate the mapping manually.

The rest of the manuscript is organized as follows. We describe our simulation model and the methods for proposing new mappings, for selecting optimal mappings for each candidate CS-molecule, and for comparing among different candidates. The following section describes the implementation of our methodology for finding the optimal CS-molecules to different CG models with varying DOC. We conclude by assessing the performance of our methodology and look at future avenues for improvement.

SIMULATION MODELS

A broad range of force fields were used which include the OPLS-AA/UA,[62] MARTINI,[63,64] DPD,[30,65] KG bead−spring model,[31] and new force fields derived for specific molecules using DCG.[66,67] For simulations using unscaled units, thermo/barostating was done using the Nosé−Hoover thermostat/barostat to maintain a temperature of 300 K and a pressure of 1 atm, with timesteps of 1 fs. The choices of 300 K and 1 atm are used as a baseline and can be readily changed to correspond the specific application of interest. For simulations using scaled (Lennard-Jones) units, simulations were run in the NVT ensemble using the Nosé−Hoover thermostat at $T^* = 1$, $\rho^* = 0.85$, and timesteps of 0.005τ. In all cases, a melt state is simulated by using a varying number of molecules depending on how large/small the molecules are. Detailed information about our simulation systems, parameters, and methodologies are given in the Supporting Information (SI; section 1.1).

FINDING MAPPINGS

While numerous mappings of the AA CS-molecule onto the CG model may exist, many of them can lead to stretched bonds, or to many more beads being mapped to one CG bead than another even when those CG beads are the same type in

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

the CG. As such, an automated process to generate mappings for evaluation is required. Additionally, since the process of finding a CG model for an AA molecule is often guided by intuition, it is important to implement a method that removes potential biases.

The proposed code requires two inputs, an AA information file containing one AA molecule and a similar file for the CG molecule. These files need to contain information regarding atom positions and types, as well as the bond structure of the molecules. An overview of the algorithm is presented in Figure 3. Importantly, although our main focus is on going from an atomistic level of detail (AA/UA models) to a CG model, this approach can be used to find mappings between any two models that differ in the level of atomic description.

The algorithm begins by finding η_{CG}, the ratio of total number of heavy atoms (non-hydrogen) to the number of desired CG beads. When rounded down, η_{CG} gives an average "size" of each CG bead, $\langle\eta\rangle$. By disregarding hydrogens, any atomic level description is reduced to the UA representation so the input model of the CS-molecule can be either AA or UA. In the next step, terminal atoms are assigned to their own groups (black circled atoms in Figure 3). A terminal atom is defined as one which is only bound to 1 other heavy atom which has not been assigned to a group yet (herein referred to as an "unmapped" atom). Because $\langle\eta\rangle$ is usually greater than 1, groups need to be "grown" to $\langle\eta\rangle$.

To grow the group with initial atom i, g_i, the "batch" of atoms which can be added to g_i is found. A batch is defined as all unmapped atoms, j, with a given bond separation number (the minimum number of bonds separating i and j, BS_{ij}). For example, the first (second) batch is all unmapped atoms with $BS_{ij} = 1(2)$. Batches are indicated by the atoms inside the dashed oval of the same color in Figure 3. A batch is only calculated for "active" groups having $\eta_i < \langle\eta\rangle$. Active groups are marked by nonblack solid circles/ovals in Figure 3. Groups with no atoms in the current batch or with $\eta_i = \langle\eta\rangle$ are labeled as "finished" and are no longer grown. Finished groups are marked by the nonblack rectangles in Figure 3. All atoms in a batch are added to g_i of any active group if $(\eta_i + \eta_{batch}) \leq \langle\eta\rangle$ and no atom in the batch is part of another batch.

When adding a batch to an active group, interbatch and intrabatch conflicts can arise. In the former case, an atom in the batch for g_i is also in the batch of at least one other active group, g_j. The strategy to circumvent this is to evaluate each way these conflicts can be resolved (i.e., the atom shared between two or more batches is assigned to only one of the groups). If N batches share the same atom, then there are N ways to resolve the conflict. For the first such resolution, the atom is assigned to the appropriate group. However, to evaluate all N ways to resolve the conflict the algorithm must be recursively called $N - 1$ times, each time carrying over information about the mapping, finished groups and how the conflict was resolved. This is illustrated in Figure 3 where the red and green groups have an interbatch conflict, which is resolved by assigning the shared atom to either group resulting in the original instance of the algorithm and one recursive call.

In the case of intrabatch conflicts, the size of the group, η_i, plus the size of the batch exceeds $\langle\eta\rangle$. Similar to interbatch conflicts, this is addressed by only adding a subset of the batch atoms such that, $\eta_i + \eta_{subset} = \langle\eta\rangle$, where permutations of the subset initiate a new loop of the algorithm (similar to interbatch conflicts). Once inter- and intrabatch conflicts have been resolved, the presence of any remaining active groups is checked. All active groups are grown simultaneously until no active groups remain, at which point any remaining unmapped heavy atoms are processed by creating and batch-wise growing new terminal groups. This cycle continues until all atoms are mapped, upon which the algorithm ends and the mapping is reported. A simple molecule is used in Figure 3 to illustrate the process, with an example of how the algorithm generates initial schemes for more complex chemistries in the SI (section 2), based on our previous work.[68]

By design this algorithm will always give a number of groups equal to or greater than n_{CG}. Thus, some groups may need to be merged so that the number of groups is equal to n_{CG}. Potential mergers are identified by pairing the smallest group(s) in the mapping with each of their smallest neighboring group(s). Each such a possible merger is tested by again recursively calling an algorithm similar to that used for finding the initial mappings. This process continues until the number of groups is equal to n_{CG}. This overall strategy of proposing many new mappings increases the likelihood of finding the optimal mapping because even a single misassigned atom may cause a mismatch with the bond topology of the desired CG molecule. While this algorithm for generating initial mappings is intended for automating ICG, it can also be helpful with DCG for objectively searching multiple feasible mappings, some of which may not have otherwise been considered by the researcher.

FINDING REPEATING MOTIFS

The method described above is suitable for molecules where it is computationally manageable to find all nonconflicting mappings. However, this is not the case for large macromolecules that would engender an intractably large number of initial mappings. Since these macromolecular systems often contain many repeat units, identifying them would greatly reduce the combinatorial redundancies as changes to the mapping within one repeat unit could then be propagated to all repeat units. Indeed, finding repeat units reduces the problem of finding the mapping of a macromolecule to the tractable problem of finding a mapping for a small molecular repeat unit. Identifying repeat units in a small molecule can also be helpful to speed up the process of enumerating possible mappings.

To identify these repeat units, we employ the simplified molecular-input line-entry system known as SMILES,[69,70] a methodology for representing the topology of a molecule as a linear string, called the "smile", where repeat units show up through a recurring pattern in the string. Our procedure for creating a smile for a given molecule follows the standard procedure, by tracing the molecular "backbone" determined as the path connecting the atoms with the maximum value of BS_{ij}. Due to the way the smile is constructed, each molecule has two smiles associated with it, created by starting at either end of the "backbone". One difference between the typical way a smile is created and our methodology is that we do not break up any ring but instead replace it by a "superatom". This modification avoids complications with how to open rings, and aligns with established CG techniques where rings are generally treated as single beads. With each smile, a string compression method is applied, similar to that used in the zip file format[71,72] as detailed in the SI (section 3).

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

AUTOMATED/MACHINE LEARNING MAPPING

Once a set of initial mappings with the correct number of groups has been created, an algorithm is used to ensure that the bond topology of a mapping matches that of the desired CG molecule. The algorithm begins with the calculation of BS_{ij} and the termination map TM_{ij} for both the mapping under evaluation (SBS_{ij} and STM_{ij}) and the desired CG molecule (DBS_{ij} and DTM_{ij}), respectively. The termination map is a metric of the network connectivity where for a given atom i in the molecule, TM_{ij} is the number of atoms, k, such that $BS_{ik} = j$ are also bonded to at least two heavy atoms.

We also create the permutation matrix, P, which is a binary matrix describing if atom i in the mapping and atom j in the desired CG molecule can ($P_{ij} = 1$) or cannot ($P_{ij} = 0$) be assigned to each other. A complementary correspondence matrix, C, is also created whose initial default entries are 0. A "correspondence" occurs when only an atom in the mapping, i, can be represented by a unique atom in the CG molecule, j:

$$\text{if } \sum_j P_{ij} = 1 \text{ and } P_{ij} = 1 \Rightarrow C_{ij} = 1 \tag{2}$$

A complete correspondence of beads has been found if there is one and only one entry of 1 in each row and column in P (all other entries being zero).

At this point several consistency checks are performed. The first check involves examining whether for each pair of i and j with $P_{ij} = 1$ that the termination map of both i and j are the same:

$$P_{ij} = \begin{cases} 1 & \text{if } STM_{ik} = DTM_{jk} \ \forall\ k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Then if $C_{ij} = 1$ a bond separation distance check is done:

$$\text{if } C_{ij} = 1 \text{ and } SBS_{ik} \neq DBS_{jl} \Rightarrow P_{lk} = 0 \tag{4}$$

This check is repeated until no more entries change in the permutation matrix. After these two checks, the only remaining atoms that can have more than one correspondence are symmetric atoms, i.e., atoms of the same type with the same neighboring bond topology. Examples of symmetric atoms in an AA model would be the hydrogens of a methyl group and ortho-position carbons, and in CG models would be the two end beads on a linear homopolymer chain. A decision must be made to break the symmetry and make the correspondence unique. After identifying a symmetric atom i, for each atom j such that $P_{ij} = 1$, the algorithm is called recursively, where the input is the permutation matrix where atom i only corresponds to atom j, and then the bond separation is rechecked. This process of selecting the correspondences of the symmetric atoms is repeated until no more symmetric atoms remain. Finally, if an atom i has no possible correspondences (i.e., $\sum_k C_{ik} = 0$), the mapping in question is discarded as it cannot map onto the desired CG molecule. Through this methodology, the algorithm learns which CG-AA mappings are possible from a prespecified (but potentially broad) chemical space.

PROPOSING NEW MAPPINGS FOR MINIMIZING ϕ

Once a set of initial mappings to the desired CG model have been found, we then proceed to optimize them by introducing sequential modifications. This is done by "shifting" the boundaries between two groups: If an atom in group i is bonded to an atom in group j, then that atom (and any terminal atoms bonded to it that are also in group i) becomes part of group j. Once the change has been made, it is checked that the topology of the CG molecule still matches the desired one and that all groups are contiguous.

OBJECTIVE FUNCTION

Regardless of the force field used for the AA and CG systems, the goal is to match the behavior of the AA CS-molecule to that of the CG molecule with respect to key geometrical and energetic details of the CG model. This may include matching such observables as the components of the nonbonded and bonded interaction energies. However, it is pointed out again that selection of a suitable objective function is an open-ended problem that requires further investigation; below we provide some simple illustrative choices.

When searching for the optimized mapping of a given AA CS-molecule onto the CG molecule we adopt as a starting point for the objective function f the average potential energy of the system, E_{pot}, mainly because most energy components of $\mathcal{H}_{CG}$ increase when the system deviates from equilibrium (e.g., arising from stretched bonds/angles), so that the optimal system tends to minimize E_{pot}. Thus, matching of E_{pot} is largely intended to penalize mappings lead to CG high-energy microstates. Additionally, since beads of the same type in a CG molecule are meant to have nearly identical constituent atoms, they should have similar masses. Hence, an additional factor in the functional form of f is included to penalize large variances in the masses of same-type CG beads (as determined from the coarse-graining of the AA CS molecule):

$$f = E_{pot} + K \sum_i \frac{\sigma_i}{\langle M_i \rangle} \tag{5}$$

where K is a penalty factor in energy units, M_i and σ_i are the average mass and standard deviation of the mass of CG beads of type i, respectively. This mass matching strategy steers the algorithm to search for mappings where the types of atoms mapped to a bead of type i are more similar to the types of atoms mapped to another bead of type i. In a broader sense, this constraint favors a uniform chemical identity for all the CG beads of the same type i. The value of K will vary based on the CG model, with larger values indicated for models with softer potentials. Again, this objective function is only applied when searching for the optimized mapping of a given candidate, to compare across candidates we implement a different objective function.

The use of scaled units, like Lennard-Jones units, in some CG models creates an ambiguity in their length scale. Because the bond types are known for the bonds in the CG molecule, the average bond length of each of these bonds can be found from the trajectories T_t and T_{AA}. Assuming all bond potentials are roughly harmonic and have approximately similar force constants, the bond scaling, χ, can be found as

$$\chi = \frac{\sum_i a_i c_i}{\sum_i a_i^2} \tag{6}$$

Where a_i and c_i are the average bond lengths for the bond of type i, calculated from T_{AA} and T_t, respectively. Once χ is calculated, it is used to scale the coordinates for all frames in T_{AA}. Once T_{AA} has been thus processed, $\mathcal{H}_{CG}$ is evaluated for each "simulated" frame to calculate f_{AA} and the same is done

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

Table 1. Table of ϕ Values for Candidate CS-Molecules (AA Models in First Column) Fitting onto a Set of UA Models[a]

	UA Models				
Candidates	Dimethyl sulfide	Acetonitrile	Acetone	Isobutylene	DMSO
Dimethyl sulfide	0.10	0.09			
Acetonitrile	0.28	0.05			
Acetone	0.17	0.06	0.22	0.21	0.12
Isobutane	0.98	0.36	1.00	1.00	1.00
Isobutylene	0.16	0.10	0.30	0.11	0.24
DMSO	0.22	0.06	0.27	0.27	0.09
Hexane	0.74	1.00			
Ethylamine	0.19	0.08			
Dichloromethylene	1.00	0.34			

[a]The best candidate for each UA model is highlighted in green. Candidate molecules which cannot be suitably mapped onto a given UA model are not given a value.

for calculating f_{CG} from T_t. For force fields that would only lose a small amount of detail upon coarse-graining (like the UA model), or when searching for the optimal mapping of a given molecule, the previously defined form for ϕ in eq 1 is suitable. However, for CG models losing larger amounts of AA detail, it will be more difficult to discriminate between values of ϕ for each candidate. In such cases, a more discriminating form of ϕ is formulated which assesses the differences between the radial distribution functions, $g(r)$, from T_t and T_{AA}. A similar approach is used in iterative Boltzmann inversion where the difference between the two $g(r)$ functions is used to improve $\mathcal{H}_{CG}$. Here the difference between the two $g(r)$ functions is quantified using

$$\phi = \sum_{i}^{n} \sum_{j \geq i}^{n} \int_{r=0}^{r=r_{cut}} |g^{ij}_{CG}(r) - g^{ij}_{AA}(r)| \left(\frac{r^*}{r}\right) dr \quad (7)$$

where $g^{ij}_{CG}(r)$ and $g^{ij}_{AA}(r)$ are the $g(r)$ between beads of type i and j for the T_t and T_{AA} trajectories, respectively. The $g(r)$ function used here includes all nonbonded interactions which are calculated in the model. To better match $g(r)$ at low values of r, the difference between $g^{ij}_{CG}(r)$ and $g^{ij}_{AA}(r)$ is weighed by the ratio of r^* to r, where r^* is the first peak of $g^{ij}_{CG}(r)$. The integral is evaluated over all nonbonded interaction pairs i, j from $r = 0$ to $r = r_{cut}$ (r_{cut} is the potential cutoff). Again, as with the formulation of ϕ for unscaled systems, the chemistry which minimizes ϕ is the best fit. In the proposed framework, we begin with a single trajectory for the candidate AA molecule, but an alternative method would involve starting with the trajectory of the CG molecule and use the mappings generated by our algorithms and established backmapping techniques[59,73−75] to generate an AA configuration. A drawback of such an approach is that the initial configuration thus generated may not be representative of the AA model and lead to false positives.

While we are primarily focused on a structure-related metric in our studies, which emphasizes the static properties and local structure, one may desire to perform ICG with the objective of finding a good candidate for matching metrics which better capture structure on larger length scales or transport or dynamical properties of a CG model. Studies have used CG models to capture the trends in dynamic properties,[76] but due to the smoothing of the potential energy surface[77] the dynamics of CG models are much faster,[78] thus precluding direct quantitative comparisons. Of course, a ϕ function could be designed to capture the matching of scaled trends produced by the CG model based on the results of multiple simulations of the candidate molecules. For the examples presented in this work, however, we only focus on matching static properties. Our ICG framework, while basic, is also left open-ended to allow targeting different properties (for which the CG model has generated the predictions of interest). It is expected that this layout can be used as a base to later build on different strategies and improvements.

RESULTS

Small Molecule UA. To provide the most basic validation of the proposed method's consistency, we tested that the best candidate CS-molecule for a given UA model is the corresponding AA model. As seen in Table 1, two different topologies of UA molecules (linear and star) were simulated and the best fit for each UA molecule from a pool of candidate CS-molecules was found. We include the AA CS-molecule of each of the UA molecules so that the correct fit is known a priori. Note that some molecules cannot map onto the star UA molecules (e.g., dimethyl sulfide cannot map onto DMSO) so they are unsuitable candidates and are assigned a blank objective function value. Expectedly and consistently, our methodology returns the correct candidate molecule fit for each UA molecule tested. Due to the small degree of coarse-graining entailed by the UA model, the K factor in eq 5 was set to zero since any disparity in bead masses would lead to large increases in E_{bond}. All the values in the table are normalized to the maximum value calculated for a given UA model.

It is apparent from Table 1 that similar CS-molecules show similar values for ϕ. Taking the case of Acetone as an example, its UA model maps best to the (AA) candidate molecule Acetone. Note also that UA acetone maps similarly well onto the AA models for isobutylene and DMSO, as these molecules are similar to each other. Similarly, UA dimethyl sulfide maps significantly better to the AA model of dimethyl sulfide due to the long bond length between the sulfur and carbon in the molecule (1.81 Å), which precludes mapping onto the other linear molecules. The only other molecules that comes close to mapping well the UA model of dimethyl sulfide are acetone

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

and isobutylene which are larger molecules, so the bond lengths in the CG trajectory are similarly long.

Polymer CG Models. To further validate our methodology, we perform a similar comparison as in the previous section but using now CG models of polymers as test beds. While the difference between the AA and UA force-field representation of molecules was minimal, typical CG polymer models average out more atomistic details as one or multiple monomers are mapped onto single beads; hence making the process of finding the optimal AA molecule that fits onto the CG molecule more challenging. Due to the larger DOC, we use the second form of ϕ (eq 7) which assesses differences in $g(r)$. As before, we propose a group of candidate molecules (simulated using the UA model) to fit onto the CG models and calculate ϕ for each of them. We use three CG models: the popular MARTINI CG model for PEO and CG models for PS and PTFE.[79]

As shown in Table 2, the best candidates for the CG model of PS and PTFE are expectedly found to be the AA model of PS and PTFE, respectively. For the MARTINI CG model of PEO, however, both PE and PEO are found to be the best candidates, with a slightly worse fit for PEO (higher ϕ). This inconsistency can be attributed to the fact that the MARTINI PEO CG model was constructed and parametrized for the simulation of biological systems in an aqueous environment and was not intended to reproduce $g(r)$ of a dry environment, while our simulations describe the melt behavior of PEO. This issue notwithstanding, PEO and PE are comparatively the best candidates for this model, a reflection of the fact that the GC-to-AA mapping problem need not have a unique solution.

In case of a very large candidate pool, it is unlikely that a singular CS candidate could be identified as distinctively optimal; instead, a group of similar molecules would be expected to fit similarly well the CG model. In our tests with smaller pools, singling out one optimal candidate was achievable because of the chemical disparity among the chosen candidates was sufficiently large. In this context, ICG would generally be expected to only be able to narrow down the candidate chemical space.

Table 2. Table of ϕ Values for Candidate Polymer CS-Molecules (5-mers) Fitting onto Two Different Previously Developed CG Models[a]

	CG Polymer Models		
Candidates	MARTINI PEO	CG PS	CG PTFE
PEO	0.59	1.00	1.00
PS	0.84	0.54	0.46
PTFE	0.85	0.69	0.45
PMMA	0.96	0.59	0.59
PI	0.68	0.76	0.45
PP	0.67	0.66	0.45
PE	0.41	0.91	0.79
PVA	0.61	0.86	0.66
PDMS	1.00	0.67	0.52
PAN	0.54	0.73	0.54

[a]The best candidates for each CG model are highlighted in green. All values are scaled by the maximum value calculated for the respective CG model.

Generic CG Models. Having confirmed the validity of our methodology when looking at CG models having unscaled length units, we now examine the case of generic CG models originally developed using generically scaled length units. For this purpose, we test our methodology for the Kremer–Grest (KG) bead–spring model of a linear 10-mer AB diblock copolymer (DBC) (5 A beads, 5 B beads). This model has been shown to correctly represent the expected trends in equilibrium and transport properties of polymeric systems[80,81] and several KG-based variants have been used in many simulation studies.[33,34,76,82−84] Three cases of the model are examined: (i) the base case of a homonuclear, flexible DBC ("flexible" model), (ii) a similar DBC but with an added bending potential to increase the persistence length ("stiff" model), and (iii) the base DBC except the B beads are larger in diameter ($\sigma \rightarrow 1.5\sigma$) ("size-asymmetric" model). The force fields for these systems are detailed in the SI (section 1.2). All three cases reflect common modifications to the KG model and represent distinct challenges to our methodology. We used $K = 1$ in eq 5 when finding the optimal mapping for each candidate to have penalties on the same thermal energy scale as the LJ interactions.

The candidate CS-molecules (chemistries) are chosen from typical monomers used in DBCs and are constructed so that one chemical monomer maps onto a single bead in the KG model. Figure 4 is a graph bar showing the fitting scores for each candidate against the three different models in question.

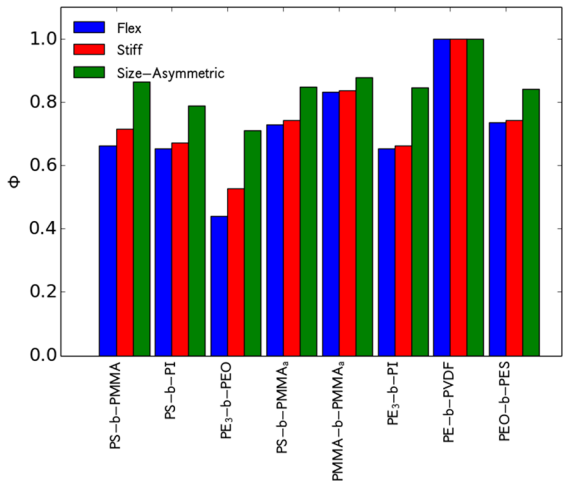


Figure 4. ϕ values for candidate DBC CS-molecules (10-mers, 5 monomers of each block) fitting onto 3 different forms of the KG model. $PMMA_a$ is the same as normal PMMA, except that the oxygens are replaced with carbons. PE_3 is the same as normal PE, except that a monomer is considered to have 3 CH_2 groups instead of 2. All values are scaled by the maximum value calculated for the respective CG model.

For the flexible model, PE_3-*b*-PEO is found to be the best candidate, which is appropriate given that both of the constituent polymers have the lowest persistence lengths among those tested. For the stiff model, PE_3-*b*-PEO ranks as the best candidate, likely because the increased backbone stiffness of the CG model is too small to penalize the flexibility of PE_3-*b*-PEO enough to make another candidate the best fit. Other good candidates for this model are PE_3-*b*-PI and PS-*b*-PI

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

which reflects that the backbone stiffness of the PI block helps better match the CG model as compared to PS-*b*-PMMA. For the size-asymmetric model, many of the proposed candidates show similar values of ϕ with PE_3-*b*-PEO again having the lowest (best) score. This aligns with idea that most DBCs have blocks where the monomers are of differing volumes, and so they should all fit a model with moderate size-asymmetry equally well. In this case, PE-*b*-PEO ranks as the best candidate likely due to the flexibility of the blocks, which also made it the best candidate for the flexible model.

As alluded to in previous sections where the PE monomer was treated as either two or three CH_2 groups, a particular issue with the mapping of polymeric systems onto unscaled models is not knowing how many monomers to map to a single bead in the CG model. Models with nonpenetrable beads like the KG model tend to be better represented with fewer monomers per bead, while models with soft nonbonded interactions such as the one typically used with the dissipative particle dynamics (DPD) model tend to be associated with more monomers per bead.[85,86] To verify this trend, we simulated CG homonuclear 5-mers using these two different models and tried to map polymers of varying DOC onto them, where DOC is now quantified as the number of monomers in the CS candidate molecule being mapped onto one CG bead (see SI section 1.3 for the DPD model). For the KG model, it is found that the minimum in ϕ for PEO and PI occurs for DOC = 1 and monotonically increases with DOC (see Figure 5, dashed lines), while for PE the ϕ minimum occurs for DOC

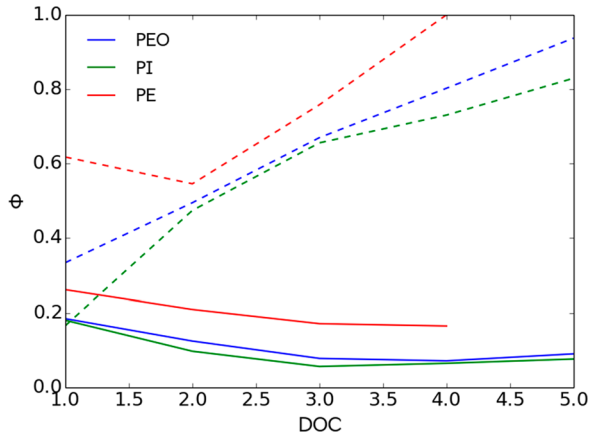


Figure 5. Plot of ϕ as a function of DOC for different polymers. Solid and dotted lines correspond to the DPD and KG models, respectively. All values are scaled by the maximum value calculated for any model.

= 2. This suggests that for PE the most suitable scale of KG coarse graining would correspond to ∼4 backbone atoms. Indeed, the minimum of ϕ for each molecule indicates that the best degree of coarse-graining corresponds to 3, 4, and 4 backbone atoms for PEO, PI, and PE, respectively. The persistence length of the flexible KG model is $\sim 1.5\sigma$,[87] which suggests that the contour length of 4 backbone atoms in PE should be similar to the persistence length of PE.[88] The contour length of 4 C−C bonds is approximately 6 Å, which is close to the experimental value of 6.5 Å. Similar agreement is found when considering the optimal DOC length scale we find and the persistence lengths for PEO[89] and PI.[90] This consistency will translate into a suitable mapping of several structural properties of the CS polymers known to correlate with the persistence length, such as the scaling with molecular weight of the average end-to-end distance and radius of gyration. Our results agree with previous studies suggesting the KG model is a suitable CG representation of many polymers.

In contrast, for the DPD model (Figure 5, solid lines) the ϕ curves show a minimum at around DOC = 3−4. These larger DOC values are consistent with the softer bead−bead potentials used in the DPD model. Our results are also consistent with findings from other studies[91,92] that place a DPD bead length-scale near the lower bound of the range of mesoscopic bead models.

The above examples explored the performance of ICG when either searching for AA candidates for CG models with relatively low DOC or determining the DOC with which a given chemistry is best represented by the target CG model. Here we explore how ICG performs when seeking to find the best candidate for a CG model when the anticipated DOC is large. To this end, we simulated a system of dimer A−B molecules, using the same model as the "flex" model in the previous examples. Plotted in Figure 6 is the values of ϕ for

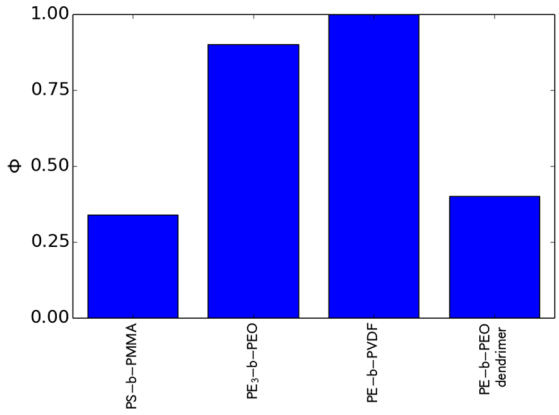


Figure 6. Plot of ϕ for different candidates for the dimer A−B model. The ϕ values are scaled to the maximum calculated value.

four different candidate molecules with different architectures ranging from linear polymers (PE_3-*b*-PEO, PE-*b*-PVDF), to linear polymers with side chain groups (PS-*b*-PMMA), up to dendrimer-like molecules (see the SI for details of the PE-*b*-PEO dendrimer chemistry). Except for the dendrimer, each block for all molecules contains five monomers of the given chemistry. It is clear that the linear polymers are poorer matches for this dimer model than molecules with sterically hindering groups. PS-*b*-PMMA has large bulky phenyl and methacrylate groups appended to the flexible chain backbones which means that when mapped to the CG model the beads show less overlap than they do in the linear polymers. When comparing just chain architecture, the PE-*b*-PEO dendrimer shows better performance than the linear counterpart in fitting the large DOC CG model considered here, which aligns with physical expectation.

CONCLUSIONS AND OUTLOOK

We have proposed algorithms to tackle an outstanding problem of molecular modeling concerning the inverse coarse-graining (ICG) process to find the best (chemically specific) molecules to map onto a known CG model of interest. To achieve this, several tools have been developed to facilitate the necessary steps in automating the ICG process, including the determination of the optimal CG mapping of a

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

candidate molecule onto the known CG model and a correspondence algorithm to uniquely determine if a given CG mapping results in a CG molecule whose topology is consistent with that of the desired CG molecule. While most available CG tools attempt to determine an "optimal" CG mapping in the absence of constraints on the topology of the CG molecule, both of the tools proposed here address the previously unresolved problem of identifying specific chemistries that best satisfy a pre-established CG molecule.

Our new mapping tools are used in conjunction with objective functions to screen for the best CG model fit from a pool of candidate molecules. Our objective functions are constructed based on metrics embodying energetic and structural metrics (including one previously used in the Iterative Boltzmann Inversion process). The methodology shows significant sensitivity of the optimal candidate identified to changes made to a CG model (Figure 4). Additionally, the method is sensitive to the degree of coarse-graining adopted, a property that was leveraged to quantify the optimal length scale of coarse-graining for two common generic CG model classes (KG bead−spring and DPD models).

Looking forward, we envision a few ways by which the performance of the proposed ICG methodology can be improved. First, the correspondence algorithm can be leveraged to fit predefined bead structures onto candidate molecules. These predefined structures would come from a database of literature-extracted data, or from a library of structural motifs (fragment library) collected from other users employing this method on their own molecules. This would allow the generation of initial structures that are closer to the optimal one by taking advantage of atom groupings and mappings already known to work well. Second, implementing a fully automated process will expedite the ICG process to allow it to continually advance without requiring user input. In this, an automated way of generating new AA candidates would be beneficial. Large molecular databases[93,94] could be used as an immediate pool of available candidate molecules, which can be supplemented with molecules outside these databases. The parametrization of force fields for these systems can be handled by automated software,[95,96] further reducing the need for user intervention. Third, more advanced objective functions can be developed to help better discriminate between candidates, target specific microstructures, and reduce the number of candidates that are kept for the next generation. Indeed, in the context of inverse design, effective but complex pair potentials describing coarse-grained mesoscale interaction sites have been identified that can assemble into desirable ordered crystals and mesophases.[97,98] In principle, objective functions incorporating information about the microstructure of such phases may help identify atomistically detailed or less coarse-grained building blocks[20,99] capable of realizing such assemblies. For long polymers, the packing length[100] and related metrics are key length scales that are known to correlate with many static and dynamic properties and should hence be taken into account when analyzing the results of ICG or in designing the objective function. Studies are already underway on some of these fronts. The proposed methodology is relatively simple and robust, and while further refinements are likely needed, it offers a suitable platform to tackle the challenging problem of inverse coarse-graining.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00232.

Simulation methods, a more complex example of mapping, and code for compression (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Email: fe13@cornell.edu.

ORCID

Christian Nowak: 0000-0003-4524-4394

Mayank Misra: 0000-0002-2700-1228

Fernando A. Escobedo: 0000-0002-4722-9836

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Awards DMREF 1629369, DMREF-1922259, and CMMI 1435852. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

■ REFERENCES

(1) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifk, A. D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **2014**, *16*, 62−74.

(2) Xia, J.; Flynn, W. F.; Gallicchio, E.; Zhang, B. W.; He, P.; Tan, Z.; Levy, R. M. Large-scale asynchronous and distributed multi-dimensional replica exchange molecular simulations and efficiency analysis. *J. Comput. Chem.* **2015**, *36*, 1772−1785.

(3) Kitao, A.; Yonekura, K.; Maki-Yonekura, S.; Samatey, F. A.; Imada, K.; Namba, K.; Go, N. Switch interactions control energy frustration and multiple flagellar filament structures. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 4894−4899.

(4) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006**, *14*, 437−449.

(5) Vashishta, P.; Kalia, R. K.; Nakano, A. Multimillion atom simulations of dynamics of oxidation of an aluminum nanoparticle and nanoindentation on ceramics. *J. Phys. Chem. B* **2006**, *110*, 3727−3733.

(6) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910−3916.

(7) Swendsen, R. H.; Wang, J. S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607.

(8) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(9) Rhee, Y. M.; Pande, V. S. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* **2003**, *84*, 775−786.

(10) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(11) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Comput. Mol. Sci.* **2011**, *1*, 826−843.

(12) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(13) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **1998**, *108*, 1964−1977.

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

(14) Escobedo, F. A.; Borrero, E. E.; Araque, J. C. Transition path sampling and forward flux sampling. Applications to biological systems. *J. Phys.: Condens. Matter* **2009**, *21*, 333101.
(15) Beeler, J. R., Jr Displacement spikes in cubic metals. i. α-iron, copper, and tungsten. *Phys. Rev.* **1966**, *150*, 470.
(16) Chatterjee, A.; Vlachos, D. G. An overview of spatial microscopic and accelerated kinetic Monte Carlo methods. *J. Comput.-Aided Mater. Des.* **2007**, *14*, 253−308.
(17) Christen, M.; van Gunsteren, W. F. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J. Chem. Phys.* **2006**, *124*, 154106.
(18) Martinez-Veracoechea, F. J.; Escobedo, F. A. Monte Carlo Stabilization of complex bicontinuous phases in diblock copolymer systems. *Macromolecules* **2007**, *40*, 7354−7365.
(19) Padmanabhan, P.; Martinez-Veracoechea, F.; Escobedo, F. A. Simulation of free-energies of bicontinuous phases for blends of diblock copolymer and selective homopolymer. *Macromolecules* **2016**, *49*, 5232−5243.
(20) Sun, Y.; Padmanabhan, P.; Misra, M.; Escobedo, F. A. Molecular dynamics simulation of thermotropic bolaamphiphiles with a swallow-tail lateral chain: formation of cubic network phases. *Soft Matter* **2017**, *13*, 8542−8555.
(21) Dong, B. X.; Liu, Z.; Misra, M.; Strzalka, J.; Niklas, J.; Poluektov, O. G.; Escobdeo, F. A.; Ober, C. K.; Nealey, P. F.; Patel, S. N. Structure Control of a π-Conjugated Oligothiophene-Based Liquid Crystal for Enhanced Mixed Ion/Electron Transport Characteristics. *ACS Nano* **2019**, *13*, 7665−7675.
(22) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *J. Chem. Phys.* **2014**, *140*, 224104.
(23) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624−1636.
(24) Ercolessi, F.; Adams, J. B. Interatomic potentials from first-principles calculations: the force-matching method. *EPL (EuroPhys. Letters)* **1994**, *26*, 583.
(25) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.
(26) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
(27) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
(28) Dunfield, L. G.; Burgess, A. W.; Scheraga, H. A. Energy parameters in polypeptides. 8. Empirical potential energy algorithm for the conformational analysis of large molecules. *J. Phys. Chem.* **1978**, *82*, 2609−2616.
(29) de Pablo, J. J.; Jones, B.; Kovacs, C. L.; Ozolins, V.; Ramirez, A. P. The materials genome initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **2014**, *18*, 99−117.
(30) Groot, R. D.; Warren, P. B. Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **1997**, *107*, 4423−4435.
(31) Grest, G. S.; Lacasse, M. D.; Kremer, K.; Gupta, A. M. Efficient continuum model for simulating polymer blends and copolymers. *J. Chem. Phys.* **1996**, *105*, 10583−10594.
(32) Rubinstein, M.; Colby, R. H. *Polymer Physics*; Oxford, 2003.
(33) Nowak, C.; Escobedo, F. A. Stability of the Gyroid Phase in Rod−Coil Systems via Thermodynamic Integration with Molecular Dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 5984−5991.
(34) Nowak, C.; Escobedo, F. A. Optimizing the network topology of block copolymer liquid crystal elastomers for enhanced extensibility and toughness. *Phys. Rev. Materials* **2017**, *1*, 035601.
(35) Horsch, M. A.; Zhang, Z.; Glotzer, S. C. Simulation studies of self-assembly of end-tethered nanorods in solution and role of rod aspect ratio and tether length. *J. Chem. Phys.* **2006**, *125*, 184903.
(36) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(37) von Lilienfeld, O. A.; Tuckerman, M. E. AlChem. variations of intermolecular energies according to molecular grand-canonical ensemble density functional theory. *J. Chem. Theory Comput.* **2007**, *3*, 1083−1090.
(38) Bereau, T.; Andrienko, D.; Von Lilienfeld, O. A. Transferable atomic multipole machine learning models for small organic molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225−3233.
(39) Bereau, T.; Wang, Z. J.; Deserno, M. *J. Chem. Phys.* **2014**, *140*, 115101.
(40) von Lilienfeld, O. A.; Andrienko, D. Coarse-grained interaction potentials for polyaromatic hydrocarbons. *J. Chem. Phys.* **2006**, *124*, 054307.
(41) Arnold, F. H. Design by directed evolution. *Acc. Chem. Res.* **1998**, *31*, 125−131.
(42) Kuchner, O.; Arnold, F. H. Directed evolution of enzyme catalysts. *Trends Biotechnol.* **1997**, *15*, 523−530.
(43) Whitley, D. A genetic algorithm tutorial. *Statistics & Computing* **1994**, *4*, 65−85.
(44) Fox, R. Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J. Theor. Biol.* **2005**, *234*, 187−199.
(45) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(46) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(47) Menichetti, R.; Kanekal, K. H.; Bereau, T. Drug−Membrane Permeability across Chemical Space. *ACS Cent. Sci.* **2019**, *5*, 290−298.
(48) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 1242−1250.
(49) Kumar, A.; Molinero, V. Self-Assembly of Mesophases from Nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8*, 5053−5058.
(50) Liu, F.; Perhm, M.; Zeng, X.; Tschierske, C.; Ungar, G. Skeletal Cubic, Lamellar, and Ribbon Phases of Bundled Thermotropic Bolapolyphiles. *J. Am. Chem. Soc.* **2014**, *136*, 6846−6849.
(51) Zeng, X.; Prehm, M.; Ungar, G.; Tschierske, C.; Liu, F. Formation of a Double Diamond Cubic Phase by Thermotropic Liquid Crystalline Self-Assembly of Bundled Bolaamphiphiles. *Angew. Chem., Int. Ed.* **2016**, *55*, 8324−8327.
(52) Harmandaris, V. A.; Reith, D.; Van der Vegt, N. F.; Kremer, K. Comparison between coarse-graining models for polymer systems: Two mapping schemes for polystyrene. *Macromol. Chem. Phys.* **2007**, *208*, 2109−2120.
(53) Karimi-Varzaneh, H. A.; van der Vegt, N. F.; Müller-Plathe, F.; Carbone, P. How good are coarse-grained polymer models? A comparison for atactic polystyrene. *ChemPhysChem* **2012**, *13*, 3428−3439.
(54) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(55) Lafon, S.; Lee, A. B. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence* **2006**, *28*, 1393−1403.
(56) Lee, A. B.; Nadler, B. Treelets| A Tool for Dimensionality Reduction and Multi-Scale Analysis of Unstructured Data. *Artif. Intell. Stat.* **2007**, 259−266.
(57) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. Versatile object-oriented toolkit for coarse-graining applications. *J. Chem. Theory Comput.* **2009**, *5*, 3211−3223.
(58) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* **2014**, *10*, 676−690.
(59) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056

method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.
(60) Ghanbari, A.; Böhm, M. C.; Muller-Plathe, F. A simple reverse mapping procedure for coarse-grained polymer models with rigid side groups. *Macromolecules* **2011**, *44*, 5520−5526.
(61) Santangelo, G.; Di Matteo, A.; Muller-Plathe, F.; Milano, G. From mesoscale back to atomistic models: A fast reverse-mapping procedure for vinyl polymer chains. *J. Phys. Chem. B* **2007**, *111*, 2765−2773.
(62) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(63) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(64) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schafer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **2013**, *9*, 687−697.
(65) Espanol, P.; Warren, P. Statistical mechanics of dissipative particle dynamics. *EPL (EuroPhys. Letters)* **1995**, *30*, 191.
(66) 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*, 7436−7449.
(67) Spyriouni, T.; Tzoumanekas, C.; Theodorou, D.; Müller-Plathe, F.; Milano, G. Coarse-grained and reverse-mapped united-atom simulations of long-chain atactic polystyrene melts: Structure, thermodynamic properties, chain conformation, and entanglements. *Macromolecules* **2007**, *40*, 3876−3885.
(68) Dong, B. X.; Nowak, C.; Onorato, J. W.; Strzalka, J.; Escobedo, F. A.; Luscombe, C. K.; Nealey, P. F.; Patel, S. N. Influence of Side-Chain Chem. on Structure and Ionic Conduction Characteristics of Polythiophene Derivatives: A Computational and Experimental Study. *Chem. Mater.* **2019**, *31*, 1418−1429.
(69) Weininger, D. SMILES, a Chem. language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(70) Veselinovic, A. M.; Veselinovic, J. B.; Zivkovic, J. V.; Nikolic, G. M. Application of SMILES notation based optimal descriptors in drug discovery and design. *Curr. Top. Med. Chem.* **2015**, *15*, 1768−1779.
(71) Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337−343.
(72) Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530−536.
(73) Zhang, G.; Chazirakis, A.; Harmandaris, V. A.; Stuehn, T.; Daoulas, K. C.; Kremer, K. Hierarchical modelling of polystyrene melts: from soft blobs to atomistic resolution. *Soft Matter* **2019**, *15*, 289−302.
(74) Zhang, G.; Stuehn, T.; Daoulas, K. C.; Kremer, K. Communication: One size fits all: Equilibrating Chem.ly different polymer liquids through universal long-wavelength description. *J. Chem. Phys.* **2015**, *142*, 221102.
(75) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* **2014**, *10*, 676−690.
(76) Alshammasi, M. S.; Escobedo, F. A. Correlation between morphology and anisotropic transport properties of diblock copolymers melts. *Soft Matter* **2019**, *15*, 851−859.
(77) Whitehead, L.; Edge, C. M.; Essex, J. W. Molecular dynamics simulation of the hydrocarbon region of a biomembrane using a reduced representation model. *J. Comput. Chem.* **2001**, *22*, 1622−1633.
(78) Depa, P.; Chen, C.; Maranas, J. K. Why are coarse-grained force fields too fast? A look at dynamics of four coarse-grained polymers. *J. Chem. Phys.* **2011**, *134*, 014903.
(79) Pan, D.; Liu, C.; Qi, X.; Yang, Y.; Hao, X. A tribological application of the coarse-grained molecular dynamics simulation and its experimental verification. *Tribol. Int.* **2019**, *133*, 32−39.
(80) Morita, H. Applicable simulation methods for directed self-assembly-advantages and disadvantages of these methods. *J. Photopolym. Sci. Technol.* **2013**, *26*, 801−807.
(81) Sen, S.; Kumar, S. K.; Keblinski, P. Viscoelastic properties of polymer melts from equilibrium molecular dynamics simulations. *Macromolecules* **2005**, *38*, 650−653.
(82) Nowak, C.; Escobedo, F. A. Tuning the Sawtooth Tensile Response and Toughness of Multiblock Copolymer Diamond Networks. *Macromolecules* **2016**, *49*, 6711−6721.
(83) Kalra, V.; Mendez, S.; Escobedo, F.; Joo, Y. L. Coarse-grained molecular dynamics simulation on the placement within symmetric diblock copolymers under flow. *J. Chem. Phys.* **2008**, *128*, 164909.
(84) Martinez-Veracoechea, F.; Escobedo, F. A. Simulation of the Gyroid phase in off-lattice models of pure diblock copolymer melts. *J. Chem. Phys.* **2006**, *125*, 104907.
(85) Murat, M.; Kremer, K. From many monomers to many polymers: Soft ellipsoid model for polymer melts and mixtures. *J. Chem. Phys.* **1998**, *108*, 4340−4348.
(86) Ashbaugh, H. S.; Patel, H. A.; Kumar, S. K.; Garde, S. Mesoscale model of polymer melt structure: Self-consistent mapping of molecular correlations to coarse-grained potentials. *J. Chem. Phys.* **2005**, *122*, 104908.
(87) Svaneborg, C.; Karimi-Varzaneh, H. A.; Hojdis, N.; Fleck, F.; Everaers, R. Kremer-Grest models for commodity polymer melts: Linking theory, experiment and simulation at the Kuhn scale. *arXiv.org* **2018**, 1808.03509.
(88) Ramachandran, R.; Beaucage, G.; Kulkarni, A. S.; McFaddin, D.; Merrick-Mack, J.; Galiatsatos, V. Persistence length of short-chain branched polyethylene. *Macromolecules* **2008**, *41*, 9802−9806.
(89) Kienberger, F.; Pastushenko, V. P.; Kada, G.; Gruber, H. J.; Riener, C.; Schindler, H.; Hinterdorfer, P. Static and dynamical properties of single poly (ethylene glycol) molecules investigated by force spectroscopy. *Single Mol.* **2000**, *1*, 123−128.
(90) Faller, R.; Müller-Plathe, F.; Doxastakis, M.; Theodorou, D. Local structure and dynamics of trans-polyisoprene oligomers. *Macromolecules* **2001**, *34*, 1436−1448.
(91) Boek, E. S.; Van Der Schoot, P. Resolution effects in dissipative particle dynamics simulations. *International J. Modern Phys. C* **1998**, *9*, 1307−1318.
(92) Yamanoi, M.; Pozo, O.; Maia, J. M. Linear and non-linear dynamics of entangled linear polymer melts by modified tunable coarse-grained level dissipative particle dynamics. *J. Chem. Phys.* **2011**, *135*, 044904.
(93) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513−530 2018.
(94) Pence, H. E.; Williams, A. Chem Spider: an online chemical information resource. *J. Chem. Educ.* **2010**, *87* (11), 1123−1124 2010.
(95) Garberoglio, G. OBGMX: A web-based generator of GROMACS topologies for molecular and periodic systems using the universal force field. *J. Comput. Chem.* **2012**, *33* (27), 2204−2208.
(96) http://zarbi.chem.yale.edu/ligpargen/index.html.
(97) Torquato, S. Inverse optimization techniques for targeted self-assembly. *Soft Matter* **2009**, *5*, 1157−1173.
(98) Jadrich, R. B.; Lindquist, B. A.; Truskett, T. M. Probabilistic inverse design for self-assembling materials. *J. Chem. Phys.* **2017**, *146*, 184103.
(99) Mukhtyar, A.; Escobedo, F. A. Developing Local Order Parameters for Order-Disorder Transitions: From Particles to Block Copolymers: Methodological Framework. *Macromolecules* **2018**, *51*, 9769−9780.
(100) Fetters, L. J.; Lohse, D. J.; Milner, S. T.; Graessley, W. W. Packing length influence in linear polymer melts on the entanglement, critical, and reputation molecule weights. *Macromolecules* **1999**, *32*, 6847−6851.

DOI: 10.1021/acs.jcim.9b00232
J. Chem. Inf. Model. 2019, 59, 5045−5056