

# A Graph-based Approach to Systematic Molecular Coarse-graining

Michael A. Webb,<sup>†</sup> Jean-Yves Delannoy,<sup>‡</sup> and Juan J. de Pablo<sup>\*,†,¶</sup>

<sup>†</sup>*Institute for Molecular Engineering, University of Chicago, Chicago, IL 60637, USA*

<sup>‡</sup>*Simulation, Modeling and Artificial Intelligence team, Solvay, Bristol, PA 19007, USA*

<sup>¶</sup>*Institute for Molecular Engineering and Materials Science Division, Argonne National  
Laboratory, Lemont, Illinois 60439, USA*

E-mail: [depablo@uchicago.edu](mailto:depablo@uchicago.edu)

## Abstract

A novel methodology is introduced here to generate coarse-grained (CG) representations of molecular models for simulations. The proposed strategy relies on basic graph-theoretic principles, and is referred to as graph-based coarse-graining (GBCG). It treats a given system as a molecular graph, and derives a corresponding CG representation by using edge contractions to combine nodes in the graph, which correspond to atoms in the molecule, into CG sites. A key element of this methodology is that the nodes are combined according to well defined protocols that rank-order nodes based on the underlying chemical connectivity. By iteratively performing these operations, successively coarser representations of the original atomic system can be produced to yield a systematic set of CG mappings with hierarchical resolution in an automated fashion. These capabilities are demonstrated in the context of several test systems, including toluene, pentadecane, a polysaccharide dimer, and a rhodopsin protein. In these examples, GBCG yields multiple, intuitive structures that naturally preserve the chemical topology of the system. Importantly, these representations are rendered from algorithmic implementation rather than an arbitrary ansatz, which, until now, has been the conventional approach for defining CG mapping schemes. Overall, the results presented here indicate that GBCG is efficient, robust, and unambiguous in its application, making it a valuable tool for future CG modeling.

## 1 Introduction

Molecular simulations are widely used in the study of physical, chemical, and biological systems. However, many interesting phenomena—protein folding, macromolecular self-assembly, polymer rheology, etc.—involve spatiotemporal scales that are generally inaccessible by naive simulations with atomistic models. To reach such scales, a number of coarse-graining strategies have been proposed in which groups of atoms are combined or lumped into individual interaction sites.<sup>1-5</sup> Coarse-grained (CG) models are widely used to explore behaviors at

mesoscopic length and time scales, to engage in high-throughput studies in both chemical and state space, and to make insightful interpretations of experimental observables.<sup>6</sup> Coarse-grained models are also critical components of multi-scale simulation techniques that actively employ atomistic and CG representations in the same simulation, such as in adaptive resolution and hybrid resolution modeling.<sup>7-9</sup>

Development of CG models generally involves two interrelated challenges.<sup>10,11</sup> The first is to define a set of *representations* that describe the system components, and the second is to define a set of *interactions* that describe how those components are influenced by each other and/or by external conditions. While some CG models aim to capture essential physics and provide general insights, permitting abstract representations that have simple, phenomenological interactions like the Gō model for proteins,<sup>12</sup> the dynamic bond percolation model for ion transport,<sup>13</sup> or the bond-fluctuation model of polymers, many CG descriptions aim to retain chemical specificity and provide quantitative predictions. The representations and interactions in these models are frequently derived from higher-resolution models, generally atomistic, as part of bottom-up or multiscale modeling strategies.<sup>1</sup>

Over the last two decades, significant efforts have been devoted to the issue of determining interactions in CG descriptions. This is highlighted by the advent of a wide array of methods<sup>14-19</sup> that compute coarse-grained potentials from more detailed system descriptions and the development of various software packages that facilitate application of these methods.<sup>20-24</sup> In contrast, the issue of generating suitable CG representations, or mapping schemes, has received relatively little attention, despite it being a keystone to any coarse-graining problem.<sup>25</sup>

In the majority of applications, CG representations are introduced as an arbitrary ansatz that maps groups of atoms to CG sites for which appropriate coarse-grained potentials can then be developed. The mapping schemes are frequently based on physical intuition, chemical intuition, or even convenience. CG sites may represent functional groups, residues, or monomers; or, they might target a specific resolution, such that each CG site represents

$n$  atoms, or  $m$  CG sites represent a particular molecular moiety. Because the procedure for defining representations is not rigorously defined, different representations can be separately conceived for the same system, leading to possible ambiguities and issues with reproducibility and the general usefulness of the model. Importantly, the choice of a CG mapping can also affect the transferability and predictive capabilities of a model.<sup>25</sup> Even in cases where the CG mapping appears straightforward, the actual tasks of creating and applying a CG mapping can be tedious and error-prone.<sup>22</sup> Existing tools<sup>22</sup> that facilitate producing ansatz CG representations become cumbersome for macromolecules and polymers. As discussed below, few tools or methodologies are available for automated, systematic generation of CG representations.

Efforts to develop systematic methodologies for defining CG representations for simulations have largely been restricted to applications concerning large biomolecules.<sup>26–31</sup> In general, the essence of these methods is to determine the positions and connectivity among a chosen number of sites as part of an optimization problem. Approaches like shape-based coarse-graining, which aims to reproduce the shape and moment of inertia of a target molecule via adaptive Delaunay triangulation,<sup>27</sup> are based mostly on structure. Meanwhile, approaches such as the essential-dynamics coarse-graining method,<sup>28</sup> which variationally determines CG sites to approximate modes from a principal component analysis of an atomistic MD trajectory, are based on dynamic analysis of higher-resolution data. Because the optimization can be mathematically complex and/or require higher-resolution data that may be infeasible to generate,<sup>30</sup> several approaches<sup>26,29,30</sup> impose the framework of elastic network models, which are already coarsened representations, to derive the distribution of CG sites. Through iterative normal-mode analysis of elastic network representations, the recently developed Decimate method<sup>31</sup> generates a consistent hierarchy of models, which is an attractive feature for potential CG modeling. Although these strategies are viable for biomolecular complexes, strategies for general molecular coarse-graining are still needed.

In this work, a graph-based coarse-graining (GBCG) approach is proposed to system-

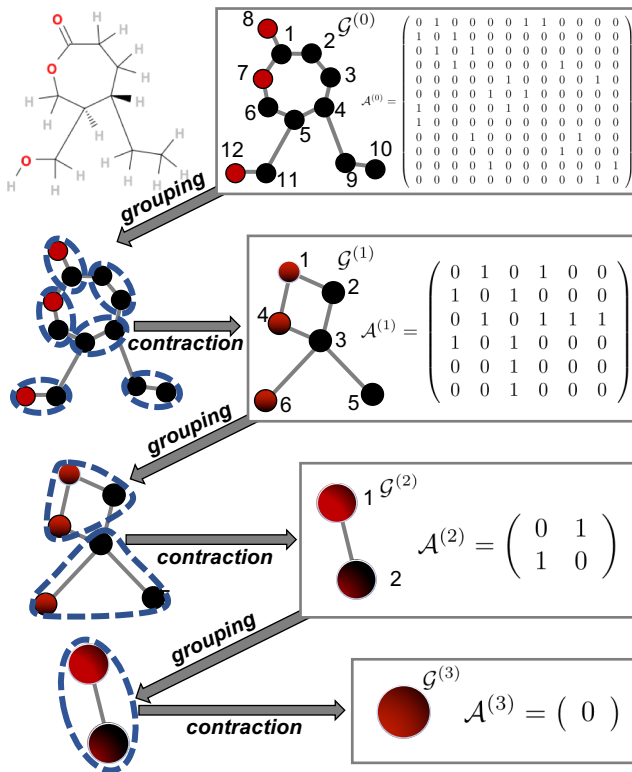
atically generate CG representations. The essence of the methodology is to represent the chemical connectivity of a molecule as a molecular graph, mapping atoms (or CG sites) to nodes and bonds to edges, and to derive successively coarser representations through the basic graph operation of edge contraction. The result of this procedure is a consistent, hierarchical set of possible CG representations that may be further parameterized and employed in stand-alone CG simulations or in multi-resolution modeling applications that can exploit the use of CG representations with varying spatial resolution. One attractive feature of GBCG is that the CG representations naturally preserve the chemical topology of higher-resolution representations because the graph-reduction process is intrinsically tied to the chemical connectivity. Consequently, GBCG typically produces intuitive representations, but as a byproduct of a simple, robust, systematic, and unambiguous protocol rather than by ansatz. Several illustrative examples are presented here, ranging from simple molecules to complex polymers, to demonstrate the generic capabilities and implementations of the method.

## 2 Theory and Methods

### 2.1 Basic GBCG algorithm

The proposed graph-based coarse-graining (GBCG) begins with the representation of a molecule as a molecular graph.<sup>32,33</sup> In particular, the initial graph is defined as  $\mathcal{G}^{(0)} = (\mathcal{V}^{(0)}, \mathcal{E}^{(0)})$  where  $\mathcal{V}^{(0)}$  denotes a set of  $N$  vertices, or nodes, that correspond to atoms in the molecule and  $\mathcal{E}^{(0)}$  denotes a set of  $M$  edges that correspond to chemical bonds between atoms. The  $i$ th vertex is associated with the  $i$ th atom, which has a mass  $m_i$  and a position  $r_i$ . The number of edges incident to the  $i$ th vertex is quantified by its degree  $d_i$ . Because bonding occurs between pairs of atoms,  $\mathcal{G}^{(0)}$  is an undirected graph that is conveniently represented with its vertex adjacency matrix  $\mathcal{A}^{(0)}$ —a square, symmetric  $N \times N$  matrix with elements  $\mathcal{A}_{ij}^{(0)} = \mathcal{A}_{ji}^{(0)} = w_{ij}$  if atoms  $i$  and  $j$  are bonded and zero otherwise; equivalently,

one may utilize adjacency lists to encode the same topological information. The undirected graph representation for an example molecule is shown at the top of Figure 1. For convenience, all molecules are initially represented as hydrogen-depleted graphs, i.e., the initial graph is already a united-atom model.



**Figure 1:** The basic GBCG approach applied to an example chemical structure with SMILES string [C(=O)1OCC(CO)C(CC)C1]. The figure shows the process of (1) assigning vertex groups (indicated by the dashed blue lines) and (2) performing edge contractions, which result in the structures shown in the gray boxes. Note that the hydrogen atoms have already been subsumed into united atoms prior to the first grouping operation, and the edge weights  $w_{ij} = 1$  for all bonds. The numbers next to the nodes on the graph indicate their index in the adjacency matrix.

With this basic foundation, we obtain coarse-grained representations by alternately performing two operations:

- (i) *Assigning vertex groups.* The objective of this operation is to identify which vertices (atoms) are to be combined into a new vertex (CG site). We choose to construct vertex groups solely using a combination of graph-theoretic and atomic properties, as opposed to utilizing dynamical information that might be generated from higher-

resolution models; this makes the assignment based entirely on the chemical topology of a molecule. Moreover, we require that a vertex group be formed among nodes that are topologically contiguous, and that any given atom be assigned to only one vertex group. Despite these simplifying restrictions, there are still numerous, reasonable procedures that might be followed, depending on the application. A few, specific implementations are discussed in Section 2.2.

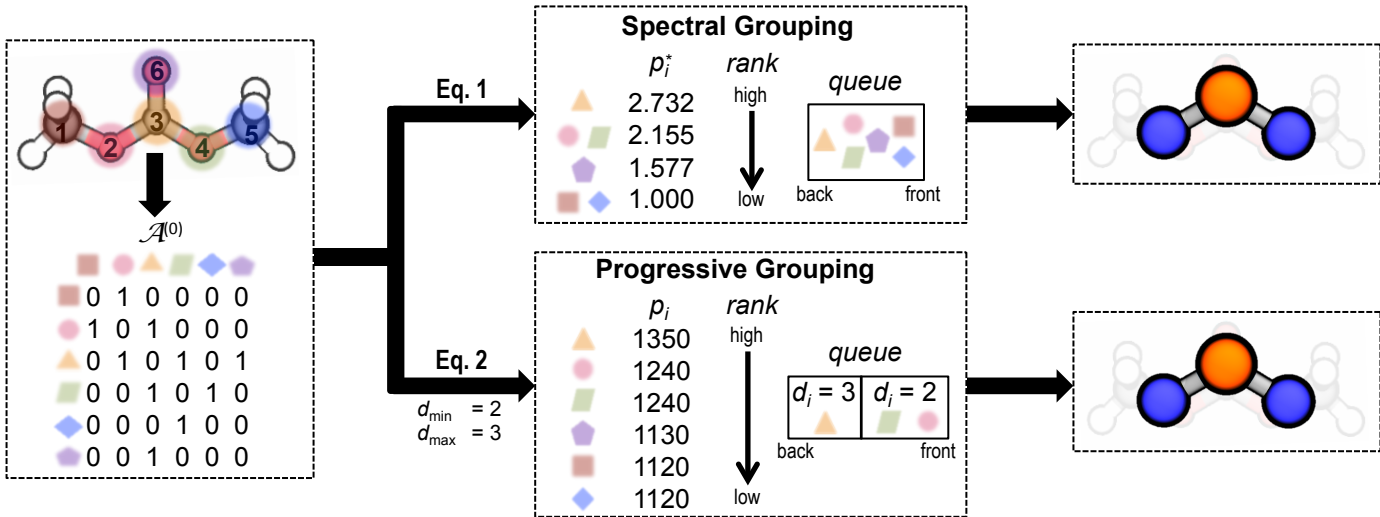
- (ii) *Performing edge contractions.* The objective of this operation is to generate a new CG representation by reducing the dimensionality of the graph and reconstituting its edges. Specifically, all the edges between nodes within a given vertex group are removed, the vertices are combined to define a CG site with the composite properties of its constituent atoms, and the adjacency matrix (or adjacency list) is resolved to reflect the new connectivity.

The  $k$ th iteration of these operations produces a graph  $\mathcal{G}^{(k)}$  represented by a reduced  $N_k \times N_k$  adjacency matrix  $\mathcal{A}^{(k)}$ , where  $N_k$  indicates the number of CG sites in the present representation, which presumably has a reduced number of edges  $M_k$ . The graph  $\mathcal{G}^{(k)}$  can then be used to seed the  $(k + 1)$ th iteration, with each iteration producing increasingly coarser representations of the original molecular graph  $\mathcal{G}^{(0)}$ . This iterative process is schematically depicted in the remainder of Figure 1.

## 2.2 Vertex Group Assignment

Between the two operations highlighted above (Section 2.1), identifying vertex groups is the more complicated procedure. There are multiple plausible ways to identify vertex groups, such as the use of clustering techniques or any number of graph-coloring approaches, or the design of algorithms that preserve certain dynamical or structural quantities. Here, we highlight two simple protocols, referred to as spectral and progressive grouping, which assign vertex groups using only properties of the current molecular graph and its constituents. Both

protocols first establish a queue based on priorities assigned to the nodes in the graph, and then construct vertex groups through analysis of a node and its connected neighbors, i.e., node pairings with nonzero elements in the adjacency matrix. The two protocols differ by when nodes are allowed to enter the queue and how nodes are determined to comprise a vertex group. The essential features of these protocols are described in the next two sections, and Fig. 2 schematically depicts the process of ranking and queuing nodes, for both protocols, to yield a CG representation for dimethyl carbonate.



**Figure 2:** A comparison between the rank-ordering and queuing process for spectral grouping (top arrows) and progressive grouping (bottom arrows), as applied to dimethyl carbonate. The left box shows dimethyl carbonate and its adjacency matrix, given the atomic indices shown; the colored shapes provides a non-numerical means of atom identification. For spectral grouping, diagonalization of the adjacency matrix (Eq. 1) yields eigenvector centrality scores  $p_i^*$ , which are used to rank-order the nodes as shown. Rankings are used to generate a queue with lowest-ranking nodes enqueued first. Applying the vertex grouping rules defined in Section 2.2.1 then yields the coarse-grained representation at the right. For progressive grouping, Eq. 2 is used to compute heuristic scores  $p_i$ , which leads to the rank-ordering shown. Nodes are sorted by their degree  $d_i$ , and rankings are used to generate a queue with highest-ranking nodes first; given  $d_{\min} = 2$  and  $d_{\max} = 3$ , only nodes with  $d_i = 2$  or  $3$  are queued. Applying the vertex grouping rules defined in Section 2.2.2 then yields the coarse-grained representation shown at the right. In this case, both grouping schemes yield the same coarse-grained representation.

### 2.2.1 Spectral Grouping

The essence of the spectral grouping scheme is to rank-order all nodes using a graph-theoretic measure, and then to systematically combine nodes that neighbor each other based on their ranking. The graph-theoretic measure is typically computed using the eigenvalues or eigenvectors of the graph adjacency matrix or the related Laplacian matrix.<sup>34</sup> Alternatively, measures that report on a node’s contribution to a suitably defined parametric graph entropy or chromatic information content could provide other reasonable means to rank-order nodes.<sup>35</sup> Here, we employ a measure based on *eigenvector centrality*, such that the ranking of a node is based on its contribution to the largest eigenvalue obtained after diagonalizing the adjacency matrix.<sup>33</sup> The motivation to use this measure is that the coarse-graining process can identify and focus on topologically important atoms. In particular, for the matrices defined by

$$\mathcal{A}^{(k)} = \mathcal{P}^{(k)} \mathcal{D}^{(k)} (\mathcal{P}^{(k)})^{-1}, \quad (1)$$

where  $\mathcal{P}^{(k)}$  is a matrix composed of the eigenvectors of  $\mathcal{A}^{(k)}$ ,  $(\mathcal{P}^{(k)})^{-1}$  its inverse, and  $\mathcal{D}^{(k)}$  a diagonal matrix with the corresponding eigenvalues, the components of the eigenvector corresponding to the largest eigenvalue,  $\mathbf{p}^* = (p_1^*, \dots, p_{N^{(k)}}^*)$ , are so-called centrality-scores that are used to rank nodes. The higher the centrality score  $p_i^*$  is for the  $i$ th node, the more important it is to the overall graph connectivity, which connects to its chemical arrangement.

Once the nodes are ranked, a queue to examine nodes and form vertex groups is established. To preserve the topology of the highest ranked nodes, we choose to order the queue from lowest-to-highest based on their centrality scores. All nodes with the same centrality score are considered simultaneously in the following manner. Considering the  $i$ th node, all neighboring nodes with equal or higher ranking are examined to potentially form a vertex group. If there are no such nodes, the  $i$ th node is assigned as its own vertex group. Otherwise, a vertex group is formed using the  $i$ th node and the neighboring node  $j$  with the *most similar* ranking score, i.e.  $\operatorname{argmin}_j [\operatorname{abs}(p_i^* - p_j^*)] \forall j$  that satisfy  $\mathcal{A}_{ij}^{(k)} \neq 0$ . In cases

when multiple neighboring nodes have equivalent similarity scores and simultaneously satisfy this condition, then all such nodes and the queued node become part of the same vertex group. Moreover, if the  $i$ th node shares a neighboring node with another node with the same centrality score, then that set of nodes is assigned to the same vertex group. Once a node is assigned to a vertex group, all nodes in the vertex group are removed from the queue, and those nodes cannot form vertex groups with any subsequently queued nodes. After assignment, the nodes with the next highest centrality score are considered, and the aforementioned assignment continues until the queue is exhausted, leaving all nodes assigned to vertex groups. Note that combining nodes with similarly small contributions to the largest eigenvalue and retaining nodes with large contributions tends to preserve the eigenvalue and the structure of the corresponding eigenvector, although no rigorous network renormalization is performed here.

### 2.2.2 Progressive Grouping

In the progressive grouping scheme, vertex groups are formed by separately generating queues for nodes with the same current degree and applying an analysis to construct vertex groups for that set of nodes. Specifically, each iteration is subdivided into rounds with a featured degree  $d$ . The featured degree for the first round is given by a minimum specified degree  $d_{\min}$ . At the completion of a round, the value of  $d$  is incremented by one, and the next round begins with that featured degree. This process continues through the completion of the last round, which features a maximum specified degree  $d_{\max}$  with  $d_{\min} \leq d_{\max}$ . Each round begins by first rank-ordering all nodes with the featured degree  $d$  using a specified metric; this metric could be based on eigenvector centrality, a heuristic formula, or any number of possibilities. This rank-ordering establishes a queue, and nodes are examined sequentially starting with the highest priority node. In this work, all applications featuring progressive

grouping use a heuristic score of

$$p_i = 1000n_i + 100d_i + 10 \sum_j d_j \mathcal{A}_{ij}, \quad (2)$$

where  $n_i$  is the number of composite “heavy” atoms in node  $i$ ,  $d_i$  is its current degree, and the third term depends on the degree of its neighboring nodes; the particular weighting coefficients are chosen to make Eq. (2) a reasonable hash function since the variables involved are all of order unity. Eq. (2) prioritizes more massive and highly connected nodes, but there is nothing particularly profound about its use. While we will show that it is effective in our applications, alternatives may sometimes be preferred or explored, and the GBCG framework is customizable in this fashion.

Irrespective of the way that nodes are ranked, the following steps will be the same. For the  $i$ th node, the set of neighboring nodes is examined to identify which nodes are not already part of a vertex group. If there are no such neighboring nodes, the  $i$ th node is identified as its own vertex group, just as in the spectral grouping scheme. Otherwise, the minimum degree among available neighboring nodes  $d_{\min}$  is recorded, and three options are pursued. First, if  $d_{\min} < d$ , then all neighbors with degree less than  $d$  are identified to form a vertex group with the  $i$ th node, and all nodes become exclusive to that group. Second, if  $d_{\min} > d$ , then the  $i$ th node is identified as its own vertex group. Third, if  $d_{\min} = d$ , then the node is tagged; tagged nodes are re-added to the end of the queue unless they have been previously tagged during the current round, in which case all neighbors with degree  $d$  form a vertex group with the  $i$ th node. Each round continues until the queue is exhausted, and all nodes with the featured degree are assigned to exclusive vertex groups. By construction, this protocol aims to preserve the overall chemical topology of the original system because the degree of a vertex group can only become less than the degree of the queued node if the vertex group contains a terminal node, i.e., a node with  $d = 1$ . It is important to note that nodes that are queued in an earlier round in the same iteration cannot form vertex groups in later rounds.

For example, if  $d_{\min} = 2$  and  $d_{\max} = 3$ , then no degree-2 nodes will form vertex groups with degree-3 nodes; however, an iteration with  $d_{\min} = d_{\max} = 3$  will permit this to happen.

## 2.3 Coarse-grained site types & Backmapping

The matter of assigning types to the new CG sites is ultimately an issue of determining interactions, since assigning a given type to multiple, similar CG sites may be desirable for parameterization. During GBCG, the identity of all atoms subsumed into a given CG site is tracked, and their connectivity can be preserved in the form of an internal adjacency matrix. Provided with a set of relative positions, this information can be used to initially back-map a CG representation into a higher-resolution model. For our purposes, this internal adjacency matrix or knowledge of composite atoms in the CG site is useful for assigning CG site types, akin to how elements are identified based on the composite set of subatomic particles. From a force-field perspective, CG sites may be further distinguished based on their chemical environment, just as not all carbon atoms are treated equally in various empirical force fields. An analogous procedure to determine types for all the CG sites could be applied here by using the internal adjacency matrix as the CG atom identifier.

### 2.3.1 Additional Considerations

The previous sections characterize the grouping schemes employed in this manuscript, but there are additional considerations that deserve discussion for implementation. One question relates to the handling of nodes with the same priority. In spectral grouping, nodes with the same priority are considered simultaneously. In progressive grouping, however, these nodes are sorted in ascending order by their index in the molecule. Since the nodes are dequeued serially, it is possible for derived CG representations to exhibit an order dependence, i.e., indexing atoms differently in the molecule might yield slightly different CG representations (mirror images, for example). While this order-dependence is typically inconsequential, it is prudent to be mindful of the possibility when implementing the progressive grouping scheme.

Importantly, provided that the atoms in a molecule are consistently indexed, the mapping to CG representations will remain consistently generated using the prescribed approach.

As noted earlier, there are also several opportunities for customizing the GBCG implementation. The use of different rank-ordering schemes has already been mentioned. In general, modifications to the adjacency matrix provide a means to further customize or tailor GBCG for certain applications. For example, diagonal entries of the adjacency matrix might be populated with node weights to increase the importance of a given node, or bond edges could be changed away from unity, perhaps using renormalization or to encode chemical information such as bond strength, to highlight the importance of some bonds in the molecule. When modifying the adjacency matrix, at least for spectral grouping, the entries should remain strictly positive to ensure a unique largest eigenvalue and positive eigenvector by the Perron-Frobenius theorem. Another possibility is to limit the number of constituent atoms in a node or impose a restriction on an allowable mass, such that two nodes might not be an allowed vertex group if their combination exceeds a property threshold; this is an available option in our code implementation but is not utilized here. Finally, while applications considered in this study consider the molecule as a whole, the GBCG procedure could just as easily be applied to fragments of molecules or monomers; this approach is recommended for development of transferable representations.

## 2.4 Simulation Details

Section 3.1 utilizes molecular dynamics simulation to compare structural correlation functions obtained for various representations of toluene and pentadecane liquids. All parameters for the all-atom simulations are taken from the OPLS-AA force field;<sup>36</sup> a list of the required parameters is provided in the Supporting Information. Both all-atom (AA) and CG MD simulations are performed using the LAMMPS simulation package.<sup>37</sup> The equations of motion are evolved using the velocity-Verlet integrator with a 1 fs time step. A Nosé-Hoover thermostat (100 fs relaxation) and barostat (1000 fs relaxation) are used to

control the temperature and the pressure where appropriate. Particle-particle particle-mesh Ewald summation<sup>38</sup> is used to compute non-bonded interactions beyond a 12 Å cutoff in the AA simulations. For the CG simulations, non-bonded interactions are computed using tabulated potentials as described later. Electrostatic interactions are not computed because all CG sites are found to be charge-neutral, mostly as consequence of the local charge neutrality present in the atomistic force-field (see Supporting Information). Note, however, that charge-neutral CG sites are not guaranteed by GBCG in the absence of applied constraints, and electrostatic interactions must generally be considered at the force field level.

Parameterization of the CG models proceeds as follows. First, coarse-grained site types are determined based on the set of composite atoms forming the CG site. The set of required interactions is then determined, and parameterization of these interactions is carried out in two stages. In the first stage, bonded interaction potentials (stretching, bending, and torsional interactions) are obtained by direct Boltzmann inversion of probability distributions from the appropriate AA simulations. For convenience, the inverted potentials are then fit to standard analytical potentials, which are described in the Supporting Information. In the second stage, AA trajectories are mapped onto CG coordinates, and then two sets of pairwise, non-bonded interactions are computed in the form of tabulated potentials. The first set is obtained using iterative Boltzmann inversion (IBI),<sup>39</sup> and the second set is obtained using force-matching (FM).<sup>15,16</sup> In the case of IBI, the non-bonded interactions are obtained with the bonded interactions fixed as those determined in the first stage. For FM, the force contributions of bonded interactions are removed from the total force during the analysis. In the determination of both bonded and non-bonded interaction parameters, the mapping to CG coordinates is based on the center-of-mass of the composite atoms. No pressure corrections are applied to the non-bonded interactions. In general, we note that the capabilities of the developed force fields are limited, and the parameterization could be further improved by various means.<sup>40-44</sup> However, these more nuanced parameterization strategies are not pursued here since the focus of this work is rather on developing CG representation

strategies.

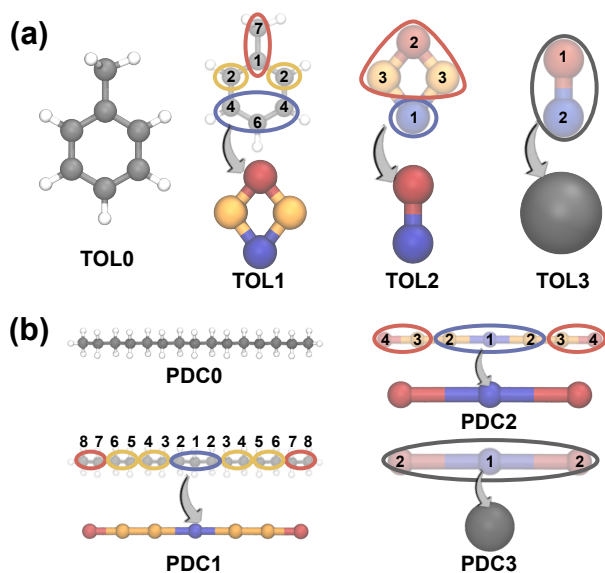
Following the generation of CG force-field parameters, coarse-grained simulations are run to obtain radial distribution functions for comparison to AA data. Initial configurations are obtained by mapping the last configuration of the AA trajectories to the CG coordinates. After this initial mapping, simulations are run for 20 ns at 298.15 K and constant volume; the last 10 ns are used to compute radial distribution functions.

## 3 Representative Applications & Discussion

### 3.1 Simple Molecules

We begin with application of GBCG to simulation of two molecules that form liquids at room temperature, namely toluene and pentadecane. These molecules are chosen based on their distinct chemical topologies, with toluene possessing an aromatic ring structure and pentadecane being a linear hydrocarbon, which combine to encompass a significant amount of typically simulated chemistries. For both molecules, three iterations of spectral grouping are used to generate three CG representations. During each iteration, the diagonal elements of the adjacency matrix are populated with the composite mass of the node normalized by the most massive node prior to diagonalization. The generated CG representations (and their parent structures) for both molecules are shown in Figure 3 along with labels (TOL0, TOL1, TOL2, and TOL3 for toluene and PDC0, PDC1, PDC2, and PDC3 for pentadecane) to distinguish the various representations.

Figure 3 illustrates how GBCG yields progressively coarsened representations of the original atomic system. For toluene, the 15-atom molecule (TOL0) is first reduced to four CG sites (TOL1), then two (TOL2), then a single CG site (TOL3), and for pentadecane, the 47-atom molecule (PDC0) is first reduced to seven CG sites (PDC1), then three (PDC2), then a single CG site (PDC3). While the molecules here are simple enough that reasonable CG representations might be generated by hand, as exemplified by previous work on the

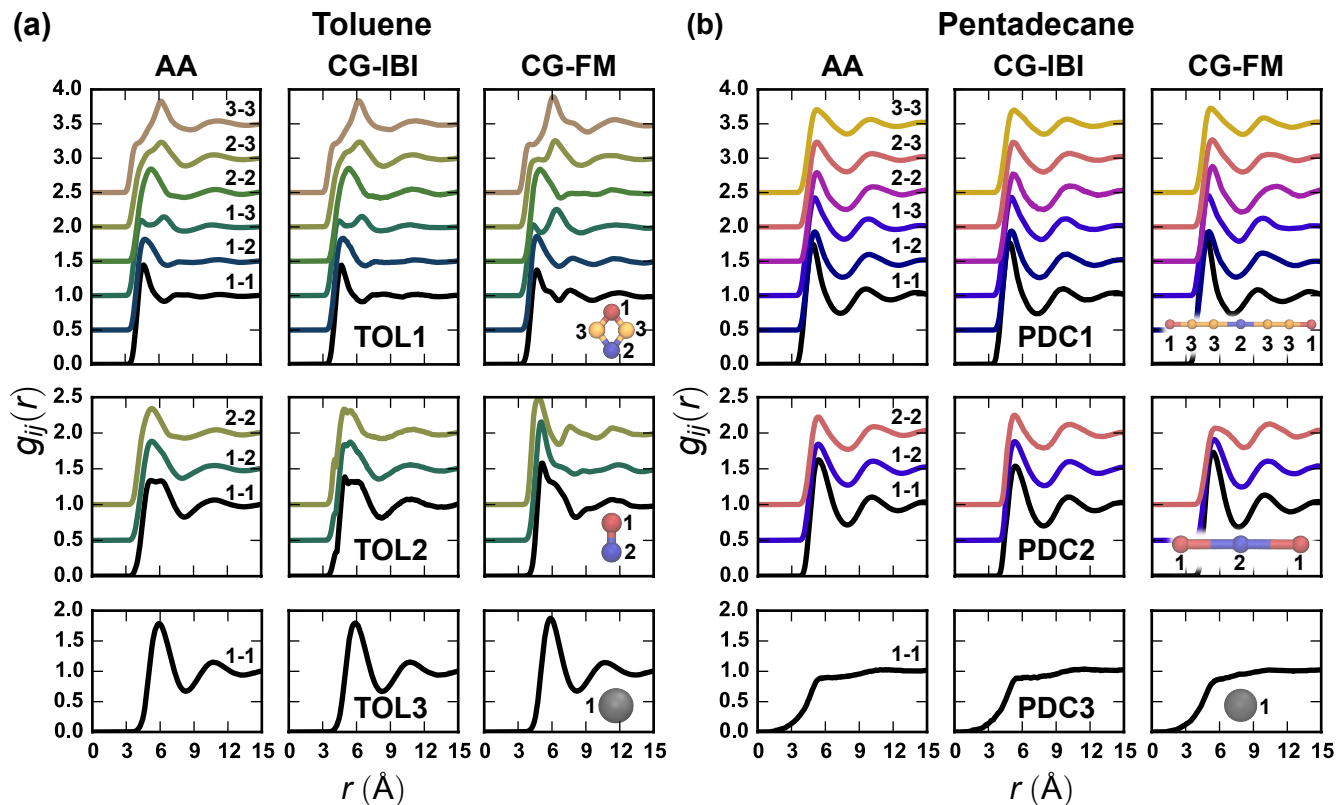


**Figure 3:** Coarse-grained representations for (a) toluene and (b) pentadecane generated using GBCG with the spectral grouping scheme. In both panels, the atomistic representations, labeled ‘TOL0’ and ‘PDC0’ for toluene and pentadecane, respectively, are shown at the left with progressively coarsened representations shown from left to right. Each CG representation is shown along with the parent structure (transparent coloring) from which it is derived. The numbers in the parent structure indicate the vertex priority obtained via diagonalization of the graph adjacency matrix, and the colored loops indicate the vertex groups identified by the algorithm described in Section 2.2.1, yielding the CG representation shown beneath the arrows. For a given CG representation, the sites are colored based on the set of composite atoms, and the placement of each CG site is based on the center-of-mass of its composite atoms.

same or similar molecules,<sup>41,45-47</sup> some advantages of using GBCG are that the generation of the CG representations is completely automated and the origin of those representations is clearly defined and traceable. In this case, the vertex groups (colored loops in the parent structure) are based on node rankings that are derived from the centrality-scores computed as described in Section 2.2.1; the scores are additionally provided in the Supporting Information. Importantly, GBCG still provides intuitive and reasonable CG representations, in this case, by forming CG sites from localized groupings of methine (-CH) groups, methylene bridges (-CH<sub>2</sub>), and methyl (-CH<sub>3</sub>) groups. Since the generated representations are not markedly different than those employed previously, the structures should be amenable to parameterization and subsequent simulation.

To demonstrate the ability to use GBCG-generated representations within a broader coarse-graining workflow, we developed basic force fields for all of the CG representations in Figure 3 using conventional approaches as described in Section 2.4. To facilitate parameterization, the various CG sites are assigned ‘types’ based on the number and kind of atoms that comprise the CG sites, such that sites with the same set of composite atoms are given the same type; the coloring of the various CG sites in Figure 3 indicate their assigned types. The site types, bonded force-field parameters, and non-bonded interaction potentials are provided in the Supporting Information. Note that while the CG representations are derived in sequence, each of the CG models is parameterized directly and only from the all-atom MD simulations. Consequently, any given CG model can be used without consideration of the other CG models, since its parameters are obtained separately and independently, in this case.

As a simple test of the efficacy of the parameterization, intermolecular pair radial distribution functions (RDFs) are computed for all pairwise combinations of CG site types for each of the CG representations in Fig. 3. Fig 4a presents the results for toluene with the RDFs for the highest-resolution CG representation at the top and the single-site representation at the bottom; Fig. 4b presents the same for pentadecane. For a given representation,



**Figure 4:** Pair radial distribution functions (RDFs) between coarse-grained site types for (a) toluene and (b) pentadecane. In both panels, results for a given CG representation are obtained from (left) all-atom simulations (AA), (middle) coarse-grained simulations with pair potentials computed from iterative Boltzmann inversion (CG-IBI), and (right) coarse-grained simulations with pair potentials computed using force-matching (CG-FM). In (a), the top row corresponds to the representation TOL1, the middle row corresponds to TOL2, and the bottom corresponds to TOL3. In (b), the same applies for PDC1, PDC2, and PDC3, respectively. The individual curves are labeled in the AA datasets, with the label  $i-j$  indicating the pair RDF between  $i$ th CG site type and  $j$ th CG site type; the site types are indicated by the inset of CG representations within the CG-FM datasets. Note that RDFs are vertically separated by 0.5 units within each set of axes to improve visibility. The axes labels are shared across different representations and simulation types. All results obtained from simulations at  $T = 298.15$  K.

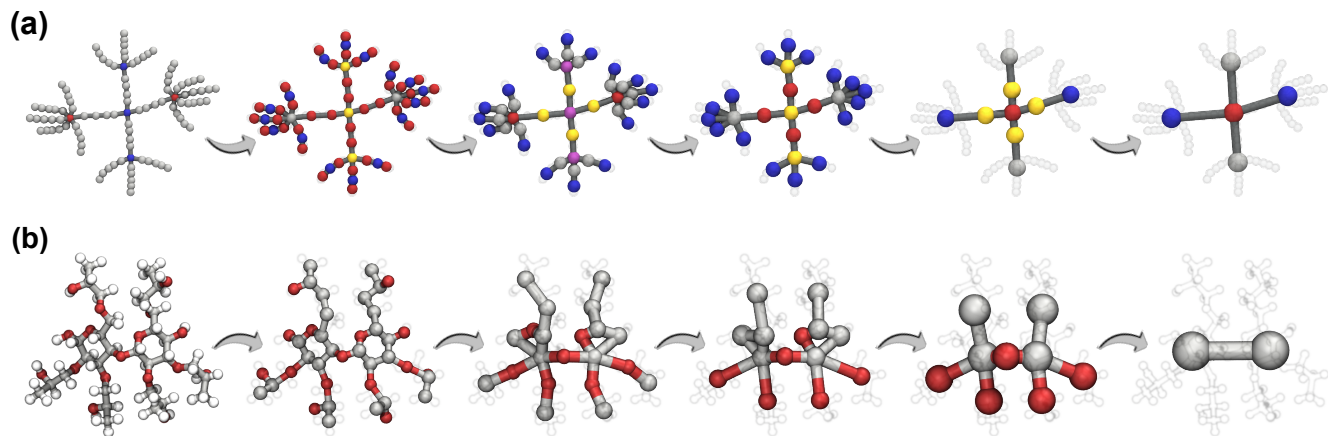
the figure compares results obtained from all-atom trajectories (labeled as AA) to those obtained from CG simulations using either the IBI or FM-generated pair potentials (labeled as CG-IBI and CG-FM, respectively); each curve reflects the RDF for two CG site types, which are indicated by the inset of the CG-FM datasets. In general, both sets of CG results are in good agreement with the AA results, irrespective of the particular CG representation or the specific pair RDF. For the CG-IBI results, this is unsurprising since the target function during parameterization is the AA RDF, but the CG-FM results also exhibit reasonable agreement with the AA RDFs despite having a target function based on the residual of the atomistic and projected CG forces rather than RDF itself. These results concretely show that typical parameterization strategies can be straightforwardly combined with CG representations generated via GBCG.

## 3.2 Complex Topologies

The previous section featured molecules of relatively simple chemical topologies; however, GBCG is particularly well suited for generating CG representations for molecules with more complex structures, such as those with branch points or mixed linear/cyclic chemical moieties. To demonstrate these capabilities, we utilize the progressive grouping scheme (Section 2.2.2) to coarse-grain the structure of a model dendrimer and a dimer of hypromellose and the spectral grouping scheme (Section 2.2.1) to coarse-grain the structure of a rhodopsin protein.

### 3.2.1 Model Dendrimer

Fig. 5a depicts a “toy” model of a dendrimer and a series of CG representations generated via multiple progressive grouping iterations; each iteration is run with  $d_{\min} = 2$  and  $d_{\max} = 6$ , such that there are five rounds per iteration, although not all rounds produce new representations. The original dendritic structure is intentionally contrived to have subtle differences at each generation in terms of the number of branches and the number of particles



**Figure 5:** Coarse-grained representations for (a) a model dendrimer and (b) a hypromellose dimer generated using GBCG with the progressive grouping scheme described in Section 2.2.2. In both panels, the initial, full representation is shown at the left with progressively coarsened representations shown from left to right. Each CG representation shown is the structure derived at the end of each iteration; the full representation is transparently shown behind each CG representation for reference. In (a), the sites are colored based on the set of composite atoms for a given representation; however, in (b), the site colors are based on the atom with the highest degree in the group. In both (a) and (b), the placement of a CG sites is based on the center-of-mass of its composite atoms.

that form each branch. In particular, the dendrimer begins with a core particle with four branches of four particles; each branch is capped with a fifth particle, which is a second-generation branch point. Among the four second-generation branch points, two have three branches, and two have five branches; each of the second-generation branches consists of five particles.

It is apparent from Fig. 5a that GBCG preserves the overall topology of the dendrimer quite well. The first three CG representations possess the same number of branch points and branches as the original dendrimer; however, the total number of sites has decreased four-fold in the third CG representation compared to the starting structure. Eventually, the structural topology of the outer branches is lost in the fourth and fifth representations, but the overall shape of the original dendrimer remains well represented. In addition, the approach neatly preserves the symmetry of the original dendrimer by equivalently treating topologically identical portions of the dendrimer, as indicated by the coloring of the CG

sites within each representation. Although any single one of these representations might be generated by hand, GBCG not only removes the tedium associated with producing such representations but also provides a multitude of representations to choose from, some of which may be preferred depending on the application or desired resolution. For example, if the precise structure of the outer branches is unimportant, then the fourth or fifth representations may be more suitable for use in a CG simulation

### 3.2.2 Hypromellose Dimer

In another application, we consider the coarse-graining of a hypromellose (or hydroxypropyl methylcellulose) dimer. Hypromellose is a semisynthetic polysaccharide with numerous applications in foodstuffs, cosmetics, and pharmaceuticals.<sup>48</sup> Several simulation studies employing atomistic and CG simulations (at monomer resolution) have been performed to study the thermo-responsive aggregation behavior and rheology of structurally similar polymers.<sup>49–51</sup> Fig. 5b shows the atomistic representation of this dimer (left) along with a series of CG representations generated via multiple progressive grouping iterations. The progressive grouping iterations utilized  $d_{\min} = \{2, 2, 2, 3, 4\}$  and  $d_{\max} = \{2, 3, 3, 3, 4\}$ , with the basic strategy being to only increase  $d_{\max}$  or  $d_{\min}$  if leaving the variables unchanged would not further reduce the number of CG sites.

Fig. 5b shows how GBCG provides CG representations that preserve the overall structural features of hypromellose at varying levels of resolution. The dimer consists of two  $\beta(1\rightarrow4)$ -linked glucose units, which have the methoxy side chains substituted with  $\text{CH}_2\text{CH}(\text{OH})\text{CH}_3$  groups at the 2, 3, and 5 positions, for a total of 105 atoms. The first CG representation, which has 41 CG sites, is slightly more coarse than a typical united atom model, leaving intact the structure of nearly all functional groups. Meanwhile, the second and third CG representations, with 21 and 14 CG sites, reduce the amount of detail for the side chains, but both still represent the backbone with a reduced ring structure; similar CG representations of polysaccharides have been previously used.<sup>52,53</sup> In the fourth representation, the

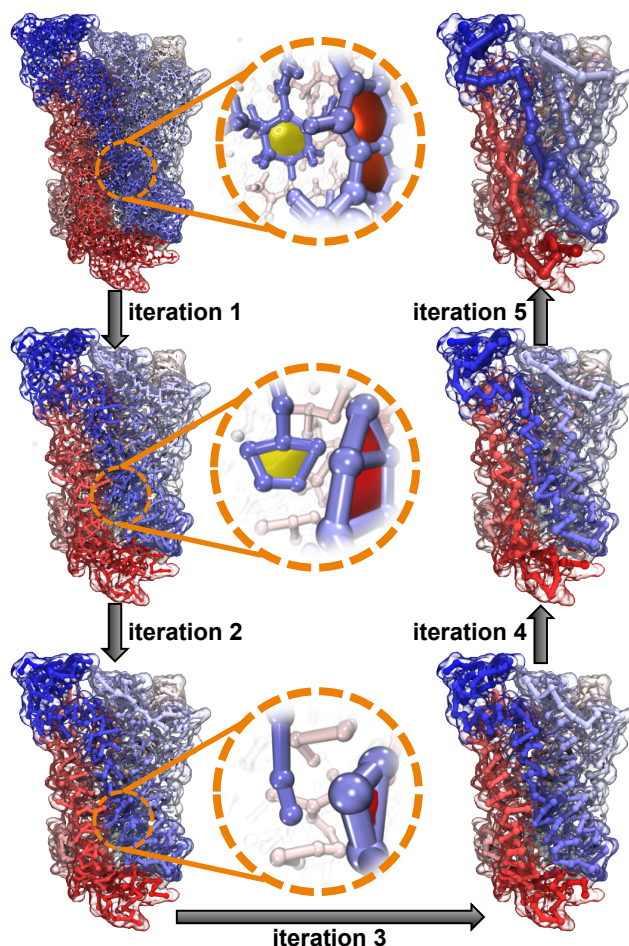
side chains remain explicitly represented with a single site each, but the glucose cycle has been eliminated in favor of a single site. Finally, in the fifth representation, the dimer is reduced to two CG sites, akin to common monomer-level CG representations that have been previously employed;<sup>49–51</sup> another iteration could further reduce the dimer to a single site for an even greater reduction in dimensionality. Overall, this illustrates how the GBCG offers a significant degree of flexibility in targeting CG simulations of a particular resolution.

It is worth noting that the specific structures observed in Fig. 5b are partly tied to the particular protocol employed, and slightly different structures could be observed for a different protocol. For example, setting  $d_{\min} = 2$  for the first three iterations leaves the glycosidic linkage (the oxygen atom joining the two glucose units) explicitly represented; this atom, with  $d = 2$ , is assigned its own vertex group because both of the CG sites to which it is bonded have a higher degree. Explicit representation of the linkage enables description of the independent rotation of the two glucose units, which may be desirable and has been included in some previous CG models by intuition.<sup>52</sup> In this case, this topologically important bond is preserved as the result of an explicit protocol.

Compared to the spectral grouping scheme, the progressive grouping scheme involves more choices, since the protocol requires explicit definition of  $d_{\min}$  and  $d_{\max}$  for a series of rounds. To some extent, this is an opportunity to tailor or customize the resulting CG representations. Since the actual reduction process is automated and low-effort, many protocols can be tested and evaluated against each other. Importantly, no matter which protocol is selected, the procedure to produce the representations is well defined and communicable.

### 3.2.3 Rhodopsin

In a final example, we obtain a series of CG representations using spectral grouping for a rhodopsin protein, a light-sensitive G-protein-coupled-receptor that activates the signaling pathway central to vision.<sup>54</sup> Fig. 6 shows the mapping results for GBCG through five iterations. Beginning with the original atomistic system (5,501 total atoms, 5,587 bonds, 10,088



**Figure 6:** A series of coarse-grained (CG) representations generated using spectral grouping for the rhodopsin protein. The original atomic system is presented at the top left, and the CG representations generated after each iteration are shown following the labeled arrows. The CG representations have 1,348, 711, 385, 195, and 100 CG sites compared to 5,501 total atoms for the all-atom representation. The atoms/CG sites are colored from red to white to blue based on residue index. The placement of a CG site is based on the center-of-mass of its composite atoms. For reference, each structure is shown with the same transparent envelope surface that is generated based on the all-atom representation. A magnified portion of the protein, indicated by the dashed orange circles, accompanies the atomic system and the first two CG representations to illustrate the mapping of a ring and a fused-ring structure that are commonly occurring motifs on several side chains in the protein.

angles, and 15,497 dihedrals) reduced to a 2,760 site united-atom model, each GBCG iteration reduces the number of sites by approximately twofold, with the first iteration mapping to 1348 CG sites, the second to 711, and then to 385, 195, and 100 CG sites for the last three iterations. It is interesting to note that the second CG representation, with 711 CG sites, has roughly the same mapping resolution as that prescribed by the MARTINI force field,<sup>55</sup> which uses about a four-to-one heavy-atom to CG-site mapping and yields a 727-site model for rhodopsin.<sup>56</sup> Here, not only does the GBCG provide a similar mapping that preserves the topology of both the backbone and side chains, but through its progression, rhodopsin is eventually reduced to a nearly linear polymer as given by the final CG representation. By eleven iterations, rhodopsin would eventually be reduced to a single site in a systematic progression.

## 4 Conclusions

A graph-based coarse-graining strategy, GBCG, has been introduced here for systematic, unambiguous, automated coarse-graining of molecular models. The methodology is inspired by simple graph-theoretic principles. GBCG addresses a key challenge, the definition of a mapping scheme, that is critical to the success of any CG model or simulation. The automated nature of GBCG derives from using an algorithmically based and well defined protocol for combining atoms into CG sites, which is in contrast to the more typical mode of identifying CG sites by intuition. To this end, two relatively simple grouping schemes have been outlined, spectral grouping and progressive grouping, which were shown to be both effective and robust in a series of test cases ranging from the simple molecule, toluene, to a macromolecule, rhodopsin. For each system, GBCG successfully generated a series of CG representations that preserve the chemical topology of the original molecules at multiple levels of resolution. Importantly, by indicating the grouping scheme and the number of GBCG iterations, the coarse-graining process is fully specified and reproducible. These

aspects should facilitate proper communication of mapping schemes as is necessary to understand the model components, reproduce model data, or implement a model for further study.

This work illustrates that GBCG can be a valuable tool in a variety of applications and workflows that utilize CG models and simulations. For simple molecules, GBCG removes the tedium associated with defining and implementing a mapping. For complex molecules, GBCG makes the mapping procedure tractable. In either case, the same algorithmic machinery provides a spectrum of reasonable CG representations with limited user investment. The set of generated CG representations may be well suited for multi-scale simulation strategies, which utilized multiple resolution models, or a given mapping might be selected based on specific simulation targets or requirements. It is important to note that certain representations may be more transferable or possess more information content than others.<sup>25</sup> These and other issues will be explored in a future publication. Because the underlying operations of GBCG are computationally inexpensive, it is feasible to generate, inspect, and evaluate many different representations and protocols, which could prove useful in the context of materials screening or model optimization. Ongoing work involves using GBCG to find optimal resolution models for target properties and identifying metrics for driving the assignment of vertex groups that facilitate the generation of improved CG representations. Basic Python implementations of GBCG are available for download at <https://github.com/xmwebb/GBCG> and in the Supporting Information.

## Associated Content

Atomistic simulation details. Force-field parameters for all-atom simulations. Bonded parameters for coarse-grained simulations. Eigenvector-centrality scores for toluene and pentadecane representations. Tabulated potentials for coarse-grained simulations. Python scripts with basic GBCG implementations.

## Acknowledgement

This research was supported by Solvay and used computational resources at the LCRC of Argonne National Laboratory. Additional support for development of software was provided by the Midwest Center for Computational Materials (MICCOM), which is supported by the Department of Energy, Basic Energy Sciences, Materials Science and Engineering Division. The authors thank Nicholas E. Jackson for numerous helpful discussions. M.A.W. also acknowledges Thomas Dannenhoffer-Lafage and Alexander J. Pak for assistance with the force-matching methodology. Figures 2-5 utilized images rendered using VMD, which is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

## References

- (1) Voth, G. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press/Taylor and Francis Group: Boca Raton, FL, 2009.
- (2) Saunders, M. G.; Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annu. Rev. Biophys.* **2013**, *42*, 73–93, PMID: 23451897.
- (3) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12423–12430.
- (4) Kamerlin, S. C.; Vicatos, S.; Dryga, A.; Warshel, A. Coarse-Grained (Multiscale) Simulations in Studies of Biophysical and Chemical Systems. *Annu. Rev. Phys. Chem.* **2011**, *62*, 41–64, PMID: 21034218.
- (5) de Pablo, J. J. Coarse-Grained Simulations of Macromolecules: From DNA to Nanocomposites. *Annu. Rev. Phys. Chem.* **2011**, *62*, 555–574, PMID: 21219152.

- (6) Ingolfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 225–248.
- (7) Praprotnik, M.; Site, L. D.; Kremer, K. Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly. *J. Chem. Phys.* **2005**, *123*, 224106.
- (8) Fritsch, S.; Poblete, S.; Junghans, C.; Ciccotti, G.; Delle Site, L.; Kremer, K. Adaptive resolution molecular dynamics simulation through coupling to an internal particle reservoir. *Phys. Rev. Lett.* **2012**, *108*, 170602.
- (9) Sokkar, P.; Choi, S. M.; Rhee, Y. M. Simple Method for Simulating the Mixture of Atomistic and Coarse-Grained Molecular Systems. *J. Chem. Theory Comput.* **2013**, *9*, 3728–3739, PMID: 26584124.
- (10) Wagner, J. W.; Dama, J. F.; Durumeric, A. E. P.; Voth, G. A. On the representability problem and the physical meaning of coarse-grained models. *J. Chem. Phys.* **2016**, *145*, 044108.
- (11) Rudzinski, J. F.; Noid, W. G. Investigation of Coarse-Grained Mappings via an Iterative Generalized Yvon–Born–Green Method. *J. Phys. Chem. B* **2014**, *118*, 8295–8312, PMID: 24684663.
- (12) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289–316, PMID: 18573083.
- (13) Druger, S.; Nitzan, A.; Ratner, M. Dynamic bond percolation theory: A microscopic model for diffusion in dynamically disordered systems. I. Definition and one-dimensional case. *J. Chem. Phys.* **1983**, *79*, 3133–3142.
- (14) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comp. Chem.* *24*, 1624–1636.

- (15) 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 method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.
- (16) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *J. Chem. Phys.* **2008**, *128*, 244115.
- (17) Mullinax, J. W.; Noid, W. G. A Generalized-YvonBornGreen Theory for Determining Coarse-Grained Interaction Potentials. *J. Phys. Chem. C* **2010**, *114*, 5661–5674.
- (18) Shell, M. S. In *Adv. Chem. Phys.*; Rice, SA and Dinner, AR., Ed.; Advances in Chemical Physics; Wiley-Blackwell, 2016; Vol. 161; pp 395–441.
- (19) Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodríguez-Ropero, F.; van der Vegt, N. F. A. Systematic coarse-graining methods for soft matter simulations – a review. *Soft Matter* **2013**, *9*, 2108–2119.
- (20) 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, PMID: 26602505.
- (21) Mashayak, S. Y.; Jochum, M. N.; Koschke, K.; Aluru, N. R.; Rühle, V.; Junghans, C. Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA. *PLOS ONE* **2015**, *10*, 1–20.
- (22) Mirzoev, A.; Lyubartsev, A. P. MagiC: Software Package for Multiscale Modeling. *J. Chem. Theory Comput.* **2013**, *9*, 1512–1520, PMID: 26587613.
- (23) Bevc, S.; Junghans, C.; Praprotnik, M. STOCK: Structure mapper and online coarsegraining kit for molecular simulations. *J. Comp. Chem.* **2015**, *36*, 467–477.

- (24) Dunn, N. J. H.; Lebold, K. M.; DeLyser, M. R.; Rudzinski, J. F.; Noid, W. BOCS: Bottom-up Open-source Coarse-graining Software. *J. Phys. Chem. B* **2018**, *122*, 3363–3377, PMID: 29227668.
- (25) Foley, T. T.; Shell, M. S.; Noid, W. G. The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **2015**, *143*, 243104.
- (26) Gohlke, H.; Thorpe, M. F. A Natural Coarse Graining for Simulating Large Biomolecular Motion. *Biophys. J.* **2006**, *91*, 2115–2120.
- (27) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure* **2006**, *14*, 1767 – 1777.
- (28) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfandtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073–5083.
- (29) Zhang, Z.; Pfandtner, J.; Grafmüller, A.; Voth, G. A. Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophys. J.* **2009**, *97*, 2327–2337.
- (30) Li, M.; Zhang, J. Z.; Xia, F. Constructing Optimal Coarse-Grained Sites of Huge Biomolecules by Fluctuation Maximization. *J. Chem. Theory Comput.* **2016**, *12*, 2091–2100, PMID: 26930392.
- (31) Koehl, P.; Poitevin, F.; Navaza, R.; Delarue, M. The Renormalization Group and Its Applications to Generating Coarse-Grained Models of Large Biological Molecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 1424–1438, PMID: 28170254.
- (32) Trudeau, R. *Introduction to Graph Theory*; Dover Books on Mathematics; Dover Publications, 2013.

- (33) Newman, M. The Structure and Function of Complex Networks. *SIAM Review* **2003**, *45*, 167–256.
- (34) Chung, F.; Graham, F.; Society, A. M.; on Recent Advances in Spectral Graph Theory, C. C.; of the Mathematical Sciences, C. B. *Spectral Graph Theory*; CBMS Regional Conference Series no. 92; American Mathematical Society, 1997.
- (35) Mowshowitz, A; Dehmer, M. Entropy and the Complexity of Graphs Revisited. *Entropy* **2012**, *14*, 559–570.
- (36) 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.
- (37) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.* **1995**, *117*, 1 – 19.
- (38) Brown, W. M.; Kohlmeyer, A.; Plimpton, S. J.; Tharrington, A. N. Implementing Molecular Dynamics on Hybrid High Performance Computers - Particle-Particle Particle-Mesh. *Comput. Phys. Commun.* **2012**, *183*, 449–459.
- (39) Reith, D.; Pütz, P.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comp. Chem.* *24*, 1624–1636.
- (40) Mullinax, J. W.; Noid, W. G. Extended ensemble approach for deriving transferable coarse-grained potentials. *J. Chem. Phys.* **2009**, *131*, 104110.
- (41) Dunn, N. J. H.; Noid, W. G. Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids. *J. Chem. Phys.* **2015**, *143*, 243148.
- (42) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *J. Chem. Phys.* **2014**, *140*, 224104.

- (43) Sanyal, T.; Shell, M. S. Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation. *J. Chem. Phys.* **2016**, *145*, 034109.
- (44) Rosenberger, D.; van der Vegt, N. F. A. Addressing the temperature transferability of structure based coarse graining models. *Phys. Chem. Chem. Phys.* **2018**, *20*, 6617–6628.
- (45) DeVane, R.; Klein, M. L.; Chiu, C.-c.; Nielsen, S. O.; Shinoda, W.; Moore, P. B. Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data. *J. Phys. Chem. B* **2010**, *114*, 6386–6393, PMID: 20426449.
- (46) Dunn, N. J. H.; Noid, W. G. Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures. *J. Chem. Phys.* **2016**, *144*, 204124.
- (47) Ruiz-Morales, Y.; Mullins, O. C. Coarse-Grained Molecular Simulations to Investigate Asphaltenes at the Oil–Water Interface. *Energy & Fuels* **2015**, *29*, 1597–1609.
- (48) Clasen, C.; Kulicke, W.-M. Determination of viscoelastic and rheo-optical material functions of water-soluble cellulose derivatives. *Prog. Polym. Sci.* **2001**, *26*, 1839 – 1919.
- (49) Huang, W.; Ramesh, R.; Jha, P. K.; Larson, R. G. A Systematic Coarse-Grained Model for Methylcellulose Polymers: Spontaneous Ring Formation at Elevated Temperature. *Macromolecules* **2016**, *49*, 1490–1503.
- (50) V., G. V.; L., S. R.; Wenjun, H.; G., L. R. Anisotropic self-assembly and gelation in aqueous methylcellulose—theory and modeling. *J. Polym. Sci. B* **2016**, *54*, 1624–1636.
- (51) Li, X.; Bates, F. S.; Dorfman, K. D. Rapid conformational fluctuations in a model of methylcellulose. *Phys. Rev. Materials* **2017**, *1*, 025604.

- (52) Markutsya, S.; Devarajan, A.; Baluyut, J. Y.; Windus, T. L.; Gordon, M. S.; Lamm, M. H. Evaluation of coarse-grained mapping schemes for polysaccharide chains in cellulose. *J. Chem. Phys.* **2013**, *138*, 214108.
- (53) Rusu, V. H.; Baron, R.; Lins, R. D. PITOMBA: Parameter Interface for Oligosaccharide Molecules Based on Atoms. *J. Chem. Theory Comput.* **2014**, *10*, 5068–5080, PMID: 26584387.
- (54) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Trong, I. L.; Teller, D. C.; Okada, T.; Stenkamp, R. E. et al. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **2000**, *289*, 739–745.
- (55) 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*, 819–834, PMID: 26621095.
- (56) Periole, X.; Huber, T.; Marrink, S.-J.; Sakmar, T. P. G Protein-Coupled Receptors Self-Assemble in Dynamics Simulations of Model Bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132, PMID: 17658882.

For Table of Contents Use Only

