Automated Parameterization of Coarse-Grained Polyethylenimine Under Martini Framework

Subhamoy Mahajan and Tian Tang*

Department of Mechanical Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada

*Corresponding author, Email: tian.tang@ualberta.ca

Abstract

As a versatile polymer in many applications, synthesized polyethylenimine (PEI) is polydisperse with diverse branched structures that attain pH-dependent protonation states. Understanding the structure-function relationship of PEI is necessary for enhancing its efficacy in various applications. Coarse-grained (CG) simulations can be performed at length- and timescales directly comparable with experimental data while maintaining the molecular perspective. However, manually developing CG forcefields for complex PEI structures is time-consuming and prone to human errors. This article presents a fully automated algorithm that can coarse-grain any branched architecture of PEI from its all-atom (AA) simulation trajectories and topology. The algorithm is demonstrated by coarse-graining a branched 2 kDa PEI, which can replicate the AA diffusion coefficient, radius of gyration, and end-to-end distance of the longest linear chain. Commercially available 25 kDa and 2 kDa Millipore-Sigma PEIs are used for experimental validation. Specifically, branched PEI architectures are proposed, coarse-grained using the automated algorithm, and then simulated at different mass concentrations. The CG PEIs can reproduce existing experimental data on PEI's diffusion coefficient and Stokes-Einstein radius at infinite dilution, as well as its intrinsic viscosity. This suggests a strategy where probable chemical structures of synthetic PEIs can be inferred computationally using the developed algorithm. The coarse-graining methodology presented here can also be extended to other polymers.

1. Introduction

Polyethylenimine (PEI) is a versatile polymer useful in many applications, including nanofiltration,[1,2] carbon capture,[3,4] fuel cells,[5,6] solar cells,[7,8] biosensing,[9,10] and gene therapy[11,12]. Each application requires different properties like molecular weight (MW), dispersity (the ratio between weight average MW (M_w) and number average MW (M_n)), degree of branching (DB), and protonation ratio (PR). For example, it is desirable to have PEIs with high M_w and DB for nanofiltration,[13] linear PEIs for carbon capture,[3] and PEIs with moderate M_w[14,15] and DB[16] for gene therapy. Therefore, identifying the structure-function relationship of PEI is crucial for optimizing its performance.

Experimentally, identifying structure-function relationships is difficult due to the diverse MW, chemical structures, and PR of synthesized PEIs. A PEI solution can have a wide range of MW, such as 0.5-360 kDa for the commercially available M_w = 25 kDa PEI from Millipore-Sigma.[15] Fractionation can reduce the range, but not below 6 kDa.[14,15] Synthesis techniques usually dictate the chemical structures, for example linear[17] vs. branched[18], but branched PEIs in a solution can be very diverse. Nuclear Magnetic Resonance (NMR) can provide some insights, such as the percentage of primary (1°), secondary (2°), and tertiary (3°) amines, but not the precise PEI structures.[19–22] The average PR of a PEI solution is typically determined by curve-fitting titration data.[23] While NMR can provide information on protonated substructures,[22] it is difficult to discern the PR of individual PEI molecules or the relationship between PR and MW or DB.

On the other hand, computational studies can model monodisperse PEI with an exact chemical structure and PR, which is ideal for studying structure-function relationships. Several all-atom (AA) forcefields have been developed for linear PEI,[24–26] but only one can handle branched structures[27]. AA molecular dynamics (MD) simulations can typically access length scales of 10-

20 nm and timescales of 100-500 ns, limiting their applications to low MW PEIs. A few coarse-grained (CG) forcefields for PEI can simulate timescales of 1-10 μs and length scales of 20-100 nm;[28–30] however, they only include parameters for low MW or linear PEIs. High MW PEIs, especially those with branched structures, can potentially be simulated using the CG approach and compared with experiments, but they remain unreported. To our knowledge, the Martini-based CG PEI forcefield developed in our previous work[28,31] is the only one in the literature that models both linear and branched PEIs. However, this CG forcefield at the current stage only accounts for 12% (98 out of 834) of all possible bonded interactions (nonbonded interactions are fully parameterized).[28] Manually determining new bonded interaction parameters is a lengthy, inefficient, and inaccurate process, calling for an automated parameterization methodology.

Automated coarse-graining includes three steps: (i) determining the AA to CG mapping scheme, (ii) modeling CG interactions, and (iii) validating the model against AA or experimental data. Automated AA to CG mapping can be based on AA trajectories, examples include Delaunay triangulation to capture a molecule's moment of inertia,[32] principal component analysis to identify essential degrees of freedom,[33] and cost minimization for mapping under the Martini framework[34]. These techniques are sensitive to AA phase space sampling and can be computationally expensive. The graph-based approach of Webb et al.[35] uses AA topology to determine a mapping that preserves the chemical topology. This technique is efficient and was recently modified by Potter et al.[36] for Martini.

The CG interactions can be described using tabulated data, fixed functional forms, or neural networks, and their parameters can be determined using bottom-up or top-down approaches. Readers are directed to the excellent reviews by Noid[37] and Jin et al.[38] Briefly, top-down approaches aim to model specific structural or thermodynamic properties with interaction

potentials that do not contain extensive details and chemical specificity.[39] In contrast, bottom-up techniques can determine CG interaction potentials of arbitrary complexity from reference simulations (AA or CG with finer resolution). The CG interaction potentials can be learned from reference static pair correlations (e.g., bond length, bond angle, and dihedral angle distributions) using automated techniques such as iterative Boltzmann inversion[40] (and its variations[41–43]), inverse Monte Carlo,[44,45] and particle swarm optimization[46–48]. Also, comprehensive variational techniques such as force matching,[49,50] ultra-coarse-graining,[51] and relative entropy minimization[52] can non-iteratively determine CG interaction potentials as a multi-body potential of mean force from reference simulations.

In Martini,[53,54] nonbonded interaction parameters (CG bead type) are determined non-iteratively from experimental free energies (top-down), and bonded interactions are determined iteratively from static pair correlations (bottom-up) using fixed functional forms. When modeling PEI, the Martini framework[53] presents several advantages compared to other coarse-graining techniques. Purely top-down approaches are insufficient due to the lack of detailed experimental data on PEI's chemical structure. Among the bottom-up techniques, the use of tabulated potentials can accurately describe the effective interaction between beads from a specific PEI. However, tabulated potentials cannot be repurposed to model another PEI with a different MW, DB, or PR. In contrast, CG potentials with fixed functional forms can achieve high adaptability. While variational bottom-up techniques are theoretically rigorous, Martini[53] is preferred when considering the balance between accuracy and parameterization time. The combined top-down and bottom-up approaches in Martini allow the reproduction of quantitative trends observed in experiments without forcefield reparameterization for every new simulation condition.[54] Furthermore, Martini contains several models for solvents, biomolecules,[55–57] and polymers[58].

Parameterizing PEI under this framework will ensure its compatibility with these existing CG models and facilitate the simulation of PEI with other molecules (e.g., DNA and RNA in gene delivery applications).

Several recent automated parametrizations within the Martini framework are worth highlighting. Potter et al.[36] used a top-down approach ("experimental-torsion knowledge distance geometry" method[59]) to determine equilibrium bond length and bond angle parameters, whereas generic Martini parameters are used for force constants and dihedral angle parameters. Empereur-Mot et al.[48] parameterized bonded interactions using a fuzzy self-tuning particle swarm optimization technique that minimizes a cost function based on the earth mover's distance[60,61] between the AA and CG bonded distributions. Bereau and Kremer[34] determined bead types using the ALOGPS[62] prediction of water/octanol partitioning free energies and hydrogen bonding capability predicted from RDKit[63]. The same methodology was followed by Potter et al.[36] Empereur-Mot et al.[47], on the other hand, determined bead types by minimizing the earth mover's distance[60,61] between the AA and CG radial distribution functions (RDF). There are challenges associated with automated Martini parametrization of branched PEIs using these existing algorithms. Since PEI contains repeating units of -(C-C-N)- as its building block, it is desirable to map this unit to a single bead. Such a mapping cannot be achieved using Potter et al.[36] because the underlying algorithm progressively groups two heavy atoms[35]. More importantly, the -(C-C-N)- building block has an intrinsically asymmetric architecture, which, without additional care, is lost after CG mapping. For example, the probability distributions of some bond angles would implicitly be assumed to be the same at the CG level, while they can be significantly different at the AA scale.[28] CG bead asymmetry is not unique to PEI; for example, backbone beads in Martini proteins[64] also have asymmetric atomistic architecture. However, such asymmetry is not

recognized in the existing automatic topology generators.[35,36] Finally, existing automated parametrization techniques are resource intensive. To overcome these limitations, we report a fully automated software package, *coarsen*, to automatically generate CG topology and iteratively parameterize CG PEIs with significantly lower resource requirements. By carefully considering bead asymmetry and its implication for bonded interactions, our algorithm generates CG parameters that accurately reproduce local and global PEI properties in an aqueous solution.

2. Methods

2.1 Automated Coarse-graining Algorithm

While the package *coarsen* is capable of simultaneously parametrizing multiple PEIs in the same system, the algorithm is explained below using one PEI as an example. The AA topology of a PEI with repeating units of -(C-C-N)- is used to create an *AA network*, where nodes represent atoms with attributes such as atom name, charge, mass, and residue name, while undirected edges represent chemical bonds (**Figure 1**, **Figure 2a**). For each PEI, the unique carbon bound to three hydrogens is designated as the *head node* (Carbon marked with * in **Figure 1**). Starting with the head node, every -(C-C-N)- and its associated hydrogens are grouped into a CG bead, forming the AA to CG mapping scheme. The trajectory from an AA simulation is mapped to the *reference CG trajectory* using the centers of mass of atoms in the CG beads. The last configuration of the AA simulation is used to initialize the CG simulation.

Figure 1. Schematic representation of the automated coarse-graining algorithm. Components in black are active when new CG parameters are being derived, whereas they are in blue when no new CG parametrization is required. The components in red are always active. In the image shown above "CG trajectory", orange, red, and cyan beads represent beads of PEI, chloride ions, and water, respectively.

The *CG network* is generated from the AA network (**Figure 1**), where nodes and edges, respectively, represent CG beads and bonds (**Figure 2b**). The total mass and charge of atoms included in a CG bead are assigned as the bead's mass and charge (converted to an integer), respectively. The names pq, p, sq, s, tq, and t denote beads containing charged primary, primary, charged secondary, secondary, charged tertiary, and tertiary amines, respectively. The name of a given bead is determined by its charge and the total number of hydrogen atoms bound to the bead's nitrogen (three for pq, two for p and s, one for s and tq, and zero for t). Based on our previous study in polarizable water,[28] the Martini bead type P1 is assigned to t, s, and p beads, and Qd is assigned to tq, sq, and pq beads.[28,31,53] The same bead types are applicable for refined polarizable water because the solvation and partitioning free energies of all bead types remain unaltered.[65]

Figure 2. (a) AA structure of a 2 kDa PEI and (b) its corresponding CG structure.

The CG beads in PEI are asymmetric, and all bead asymmetries are aligned, i.e., the nitrogen of one bead bonds with the terminal carbon of another bead, forming -(C-C-N)-(C-C-N)-units. Units such as -(N-C-C)-(C-C-N)- or -(C-C-N)-(N-C-C)- do not appear in PEI. Thus, one can capture the bead asymmetry using directed edges that point from the nitrogen in one bead to the terminal carbon in another (**Figure 2b**), which provides a way to distinguish bond lengths, bond angles, and dihedral angles.[28] For example, t→sq ≠ sq→t because the tertiary nitrogen in the AA network is in the middle in t→sq but at the end in sq→t, while the situation is opposite for the location of the charge (see the AA and CG networks in **Figure 1**). These considerations can be extended to bond and dihedral angles. It should be noted that the directed edges in a bond or dihedral angle need not be in the same direction. For instance, a bond angle can have one directed edge in each direction (e.g. s←t→sq), which makes it symmetric (s←t→sq = sq←t→s). The bond angles sq←t→s and sq→t→s are different: in the former, both s and sq beads are bound to the same nitrogen in the t bead, while in the latter sq is bound to the terminal carbon and s to the nitrogen of the t bead (**Figure 1**). Similarly, the dihedral angle sq←t→t→pq is different from sq→t→t→pq. However, unlike bond angles, all dihedral angles are asymmetric. Finally, pq and p

8

beads only exist at the end of directed edges, that is, →pq and →p exist but not pq→ or p→. Likewise, only t beads can have directed edges in both directions, that is, ←t→ and ←tq→ exist but not ←sq→, ←s→, ←pq→ or ←p→.

A naming system is created for the bond lengths, bond angles, and dihedral angles based on the directed edges.[28] When all directed edges are aligned, they are referred to as *normal* bond lengths, bond angles, and dihedral angles because they each have one C-terminal and one N-terminal in the corresponding AA network (e.g., sq→t→s in **Figure 1**). In such a case, they are named along the direction of the edges, for example, bond length t→sq is named tsq, bond angle sq→t→s is named sqts, and dihedral angle sq→t→t→pq is named sqttpq. However, a bond or dihedral angle can have two N-terminals in the corresponding AA network (e.g., s←t*→sq and s←t*→sq→t in **Figure 1**). These are referred to as *N-type* bond or dihedral angles,[28] and the letter N is affixed to the beginning of their names. Since N-type bond angles are symmetric,[28] they can have two equivalent names. For example, the angle sq←t→s could be named Nsqts or Nstsq. To avoid ambiguity, a preference order of pq > p > sq > s > tq > t is used for the first bead, i.e., sq←t→s is uniquely named as Nsqts. The N-type dihedral angle is named such that the two edges that constitute a normal bond angle are placed at the end of the name. For example, the dihedral angle sq←t→t→pq is named Nsqttpq to ensure the last three beads, in sequence (t, t, and pq here), form a normal bond angle (t→t→pq). The probability distributions generated from the reference CG trajectories (**Figure 1**) are grouped by name and are referred to as the *reference bonded distributions*.

The CG network, known parameters,[28] and reference bonded distributions are used to generate the CG topology (**Figure 1**). Since the charge, mass, and type for each bead are available from the CG network, the nonbonded interactions are fully defined. Bonded interactions with the

same name are assigned the same parameters. The bond-stretching potential is harmonic, as shown in Eq 1, where r_{ij} is the distance between beads i and j, r_{eq} is the equilibrium bond length, and K_b is the force constant.

$$V_b(r_{ij}) = \frac{1}{2} K_b(r_{ij} - r_{eq})^2 \tag{1}$$

Similarly, the angle-bending potential is described using the cosine-harmonic potential in Eq 2, where θ_{ijk} is the angle between beads i, j, and k, θ_{eq} is the equilibrium bond angle, and K_a is the force constant.

$$V_a(\theta_{ijk}) = \frac{1}{2} K_a(\cos \theta_{ijk} - \cos \theta_{eq})^2 \tag{2}$$

For the dihedral-torsion potential, a sum of periodic functions is used as shown in Eq 3, where φ_{ijkl} is the dihedral angle between beads i, j, k, and l, $2W$ is the number of periodic functions, and $\lfloor \cdot \rfloor$ is the greatest integer function.

$$V_d(\varphi_{ijkl}) = \sum_{w=1}^{2W} K_{d,w} \left[1 + \cos\left(\left\lfloor \frac{w+1}{2} \right\rfloor \varphi_{ijkl} - \varphi_{eq,w}\right)\right] \tag{3}$$

The parameters $K_{d,w}$ and $\varphi_{eq,w}$ are the force constant and phase angle for the wth periodic function, respectively. Specifically, $\varphi_{eq,2n-1} \in \{0°, 180°\}$ and $\varphi_{eq,2n} \in \{-90°, 90°\}$ for an integer $n \geq 1$. Eq 3 is equivalent to the Fourier series in Eq 4, where $a_0 = \sum_{w=1}^{2W} K_{d,w}$, $a_w = K_{d,2w-1} \cos \varphi_{eq,2w-1}$, $b_w = K_{d,2w} \sin \varphi_{eq,2w}$, and integer $w \in [1, W]$.

$$V_d(\varphi_{ijkl}) = a_0 + \sum_{w=1}^{W} a_w \cos(w\varphi_{ijkl}) + b_w \sin(w\varphi_{ijkl}) \tag{4}$$

The specific forms of bond-stretching, angle-bending, and dihedral-torsion potentials are chosen to be compatible with the Martini force field.[53]

Bonded parameters that have been determined previously[28] are used directly without modification. For new bonded parameters, an initial guess is made by applying direct Boltzmann

inversion on the reference bonded distributions, i.e., extracting parameters in the corresponding potentials by least squares fitting. With the initial guess, a CG simulation can be performed, which generates bonded distributions referred to as the *CG bonded distributions*. The new bonded parameters are iteratively improved by minimizing a cost function (Eq 5), where χ_α is a vector of properties quantifying bonded distribution α and $|\cdot|$ represents the magnitude of a vector.

$$C = \sum_\alpha \left|\chi_\alpha^{CG} - \chi_\alpha^{ref}\right|^2 \tag{5}$$

Superscripts ref and CG affixed to χ_α indicate whether the distribution is from the reference or CG bonded distribution, respectively. For bond lengths and bond angles, we aim for the CG bonded distribution to accurately reproduce the mean and standard deviation of the corresponding reference bonded distribution. Therefore, $\chi_\alpha = \{\sqrt{\omega_\alpha}\mu_\alpha, \sqrt{1-\omega_\alpha}\sigma_\alpha\}$ for bond lengths and angles, where μ_α and σ_α are respectively the mean and standard deviation of the distribution, and $\omega_\alpha \in [0,1]$ is a weight that assigns the relative importance of μ_α and σ_α. For the dihedral angle, the prescribed dihedral-torsion potential (Eq 4) is a Fourier series containing multiple maxima and minima. Therefore, instead of mean and standard deviation, χ_α is chosen as a vector consisting of Fourier series constants $\{\tilde{a}_1, ..., \tilde{a}_W, \tilde{b}_1, ..., \tilde{b}_W\}$ determined by curve-fitting the reference or CG bonded distribution with $\exp(-V_d/k_B T)$ where V_d is the dihedral-torsion potential in Eq 4. Here $\tilde{\ }$ is used to distinguish the curve-fit constants from the forcefield parameters $\{a_1, ..., a_W, b_1, ..., b_W\}$. In other words, the forcefield parameters $\{a_1, ..., a_W, b_1, ..., b_W\}$ are iteratively adjusted to reduce the difference between $\{\tilde{a}_1, ..., \tilde{a}_W, \tilde{b}_1, ..., \tilde{b}_W\}$ extracted from the reference and CG bonded distributions. An adaptive gradient descent is used to minimize the cost function in Eq 5, and the details can be found in the supporting information (**SI) Section S1a**.

The entire parameterization workflow is shown using black and red arrows in **Figure 1**. For PEIs that do not contain unknown parameters, CG topology and initial structure are determined from CG PEI SMILE (Simplified Molecular-Input Line-Entry System[66]) strings (see **SI Section S1b** for details).

2.2   Simulations

A branched 2 kDa PEI with PR = 46% (**Figure 2a**) was simulated in TIP3P water with neutralizing salts (system AA2k in **Table 1**), based on the initial structure and forcefield (CHARMM General Force Field[67]) presented in Sun et al.[27,68] The CG2k PEI in **Table 1** and shown in **Figure 2b** was coarse-grained from **Figure 2a** and used to demonstrate the quality of the automated coarse-graining algorithm. **Table 1** lists several other CG simulations performed for PEIs with different MW, DB, and PR to validate the CG model against experiments. The initial structure of these CG PEIs was based on the forcefield parameters derived from the CG parametrizations (SI **Section S1b**). Technical details of AA and CG simulation can be found in **SI Section S1c** and **S1d**, respectively, whereas methodologies used to determine PEI properties are given in **SI Section S1e**.

Table 1: Summary of simulated systems. Each system is named after the simulation type, PEI structure, M_w, and the intended mass concentration of PEI (in g/L). The AA2k (same as CG2k) structure was proposed by Sun et al.[27] Other PEI structures are proposed in this work (SL: semi-linear; MB: moderately-branched; HB: hyper-branched; see **Section 3.3** for details).

System name	Simulation type	M_n (kDa)	Initial box size (nm³)
AA2k	All-atom	1.8	8.12×8.12×8.12
CG2k	Coarse-grained	1.8	8.12×8.12×8.12
CG SL2k c10, CG MB2k c10, CG HB2k c10	Coarse-grained	1.8	6.92×6.92×6.92
CG SL2k c20, CG MB2k c20, CG HB2k c20	Coarse-grained	1.8	5.50×5.50×5.50
CG SL2k c30, CG MB2k c30, CG HB2k c30	Coarse-grained	1.8	4.80×4.80×4.80
CG SL2k c37, CG MB2k c37, CG HB2k c37	Coarse-grained	1.8	4.50×4.50×4.50
CG SL25k c5, CG MB25k c5, CG HB25k c5	Coarse-grained	10	14.92×14.92×14.92
CG SL25k c10, CG MB25k c10, CG HB25k c10	Coarse-grained	10	11.84×11.84×11.84
CG SL25k c20, CG MB25k c20, CG HB25k c20	Coarse-grained	10	9.40×9.40×9.40
CG SL25k c30, CG MB25k c30, CG HB25k c30	Coarse-grained	10	8.21×8.21×8.21

3. Results

3.1 Automated coarse-graining of PEI

Analysis of the AA2k PEI revealed that parameters were known for 7 bond lengths, 12 bond angles, and 9 dihedral angles,[28] while unknown for 6 bond angles (Npqtt, Nsqtt, Nsqts, ttsq, sqtsq, and tsqt) and 21 dihedral angles (Nsqtspq, tsqtt, tsqts, Nsttt, tttsq, Npqttsq, tsspq, ttsqt, Nsqtsqt, sqtspq, tttt, ttss, Nsqtts, Nsqttpq, Nstsqt, tsqtpq, sqtsqt, Nttsqt, sqtss, Nttss, tsqtsq). The unknown parameters were determined over 30 iterations, with two additional sub-iterations for each bond angle to determine the Jacobian (see **SI Section S1a**) and eight periodic functions to model dihedral angles ($W = 4$ in Eq 4). Examination of five replicas of the automated parameterization confirmed the convergence of the algorithm and the uniqueness of the parameters it generated (shown below). The initial parameters were identical in all replicas because they were determined using direct Boltzmann inversion on the same reference bonded distributions. However, statistical variation among the replicas arose due to the different random seed values in CG simulations.

The total cost and the cost associated with the new parameters are shown in **Figure 3a** and **Figure 3b**, respectively. For each replica, an overall decreasing trend of cost vs. iteration was observed, with some fluctuations arising from the statistical nature of MD simulations. The decreasing trend vanished within 30 iterations, indicating convergence. A further decrease in cost was not expected. The cost associated with the new parameters (**Figure 3b**) decreased by ~5 times, whereas the total cost decreased by ~2 times (**Figure 3a**). The lowest total cost was achieved at iterations 27, 20, 25, 23, and 28 for replicas 1-5, respectively (dashed vertical lines in **Figure 3a**). The parameters associated with these iterations are referred to as the optimal parameters for the corresponding replica. More discussions on the replica simulations can be found in **SI Section S2**.

13
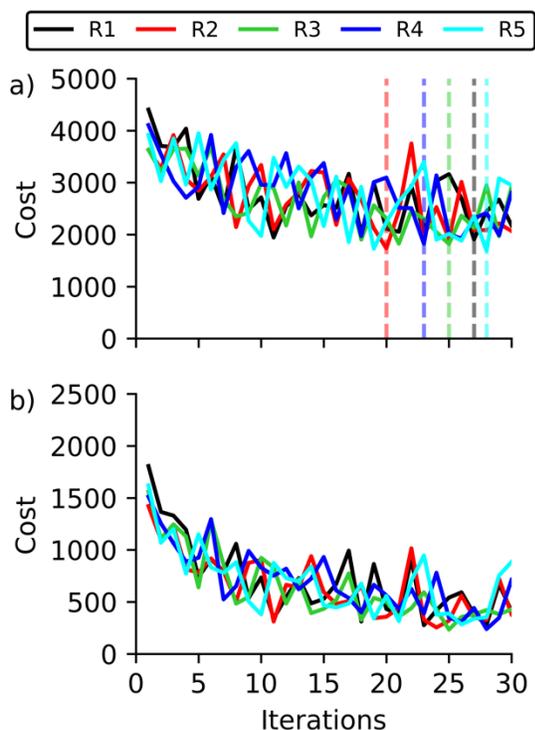
Figure 3. Convergence of automated coarse-graining of a branched 2 kDa PEI: (a) the total cost and (b) the cost associated with new parameters. R1-R5 are different replicas. The vertical lines in (a) denote the optimal iteration for each replica.

The new optimal parameters from different replicas are compared for bond angles in **Figure 4a** and dihedral angles in **Figure 4b-v**, where the optimized dihedral parameters $\{a_1, \dots, a_W, b_1, \dots, b_W\}$ in Eq. 4 were converted to $\{K_{d,1}, \dots, K_{d,2W}, \varphi_{eq,1}, \dots, \varphi_{eq,2W}\}$ in Eq. 3. **Figure 4a** demonstrates the uniqueness of the parameters for bond angles Npqtt, Nsqtt, Nsqts, ttsq, and tsqt. However, parameters were not unique for sqtsq because simultaneously increasing the equilibrium angle and decreasing the force constant resulted in a similar CG bonded distribution (see SI **Section S3**). For the optimal dihedral parameters (**Figure 4b-v**), uniqueness was observed for 14 dihedral angles ((b) Nsqtspq, (d) tsqtt, (e) tsqts, (h) Nsttt, (i) tttsq, (j) Npqttsq, (k) tsspq, (l) ttsqt, (m) Nsqtsqt, (o) sqtspq, (p) tttt, (q) ttss, (u) Nsqtts, and (v) Nsqttpq), with a small variation for 4 dihedral angles ((c) Nstsqt, (n) tsqtpq, (s) sqtsqt, (t) Nttsqt), and a moderate variation for 3 dihedral angles ((f) sqtss, (g) Nttss, (r) tsqtsq). Despite the variation in dihedral parameters, the

14

dihedral angle distributions were all similar to the corresponding reference bonded distributions (see SI **Section S3**). Therefore, the algorithm presented here ensures not only the accuracy of the CG forcefield but also the uniqueness of optimized parameters for most bonded interactions.
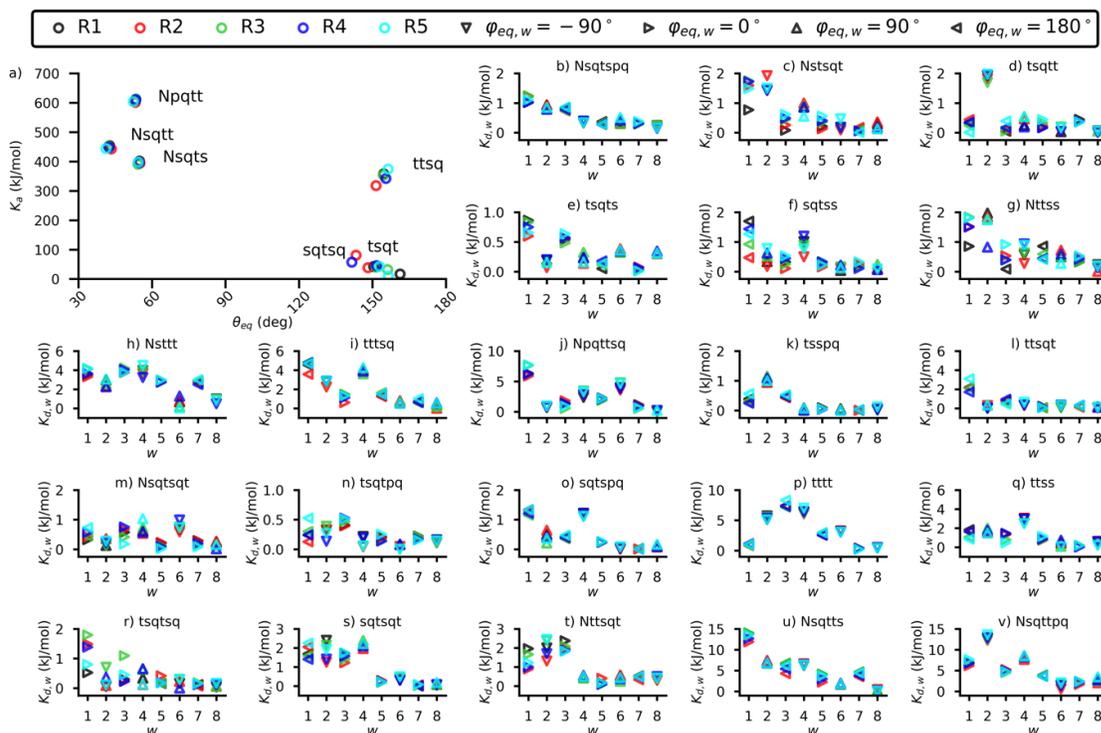
Figure 4. Optimal bonded parameters for CG2k, obtained from 5 replicas (R1-R5) of automated parameterization: (a) bond angles, and (b-v) dihedral angles. θ_{eq} and K_a are parameters associated with bond angles (Eq 2), and $\varphi_{eq,w}$, $K_{d,w}$ are parameters associated with dihedral angles (Eq 3). The values of $\varphi_{eq,w}$ are indicated by the symbols specified in the legend.

Local properties of AA2k PEI and five replicas of CG2k PEI are compared in **Figure 5** in terms of the RDFs between bead pairs P1-P1, P1-Qd, and Qd-Qd. The RDFs for AA2k were based on the reference CG trajectory ("ref-CG" in the legend), and those for CG2k were based on CG simulations using the optimal bonded parameters identified for the five replicas (R1-R5 in the legend). The RDF between 0-0.4 nm is shown in the top inset, 0.4-0.9 nm in the main plot, and 0.9-4 nm in the bottom inset. Bond lengths, N-type bond angles, normal bond angles, and dihedral angles made the highest contribution to the RDF at radial distances of 0-0.4, 0.4-0.5, 0.5-0.7, and 0.7-0.9 nm, respectively (comparison of bonded distributions can be seen in **SI Section S3**). The

contribution of 1-n ($n > 4$), i.e., nonbonded, interactions to the RDFs began at 0.8 nm and became prominent beyond 0.9 nm. The RDFs for CG2k can capture the general form of the RDFs for AA2k, with a few differences. First, in **Figure 5a** and **5b**, there were multiple peaks between 0 and 0.4 nm in the RDF of AA2k but only a single peak in the RDF of CG2k. This is typical of CG forcefields using a harmonic bond-stretching potential (e.g., Figure S3 in Song et al.[69]). The single peak, however, captured the spread of the multi-peak distribution and the location of the most prominent peak. Second, in **Figure 5b** and **5c**, the peak at 0.4 nm in the RDFs of AA2k was not captured by the RDFs of CG2k, which stemmed from the difficulty in capturing the N-type bond angle distributions due to large Lennard-Jones repulsion at small angles.[28] Atoms in AA simulations have more degrees of freedom that allow them to reduce the Lennard-Jones repulsions, which is difficult for the CG beads. Consequently, in the same figures, the peak at 0.45 nm was higher in the RDFs of CG2k. Third, in **Figure 5c**, the peak at 0.7 nm in the RDF of AA2k was not reproduced in the RDFs of CG2k, which is attributed to the previously parameterized bond angles, such as sqtpq[28] (see **SI Section S3**). Fully re-parametrizing a new PEI molecule is expected to increase the accuracy of local structure because angles with the same name but in different PEIs can be subjected to different steric hindrances. However, it is undesirable to perform a complete reparameterization for every new molecule because AA simulations of high molecular weight PEIs are computationally costly and impractical. Instead, our design strategy is to reuse parameters that were previously learned without additional modification. Admittedly, such an approach can reduce accuracy, such as the CG RDFs between 0.5 and 0.8 nm in **Figure 5c**. However, similar level of discrepancies has been reported in the literature, e.g., in Figure 3 of Izvekov and Voth[70]. To examine the impact of discrepancies caused by reusing existing bonded parameters, we evaluated

global properties of CG2k (presented next) as well as simulated a new 2 kDa PEI with different structure but without any reparameterization (**Section** Error! Reference source not found.).
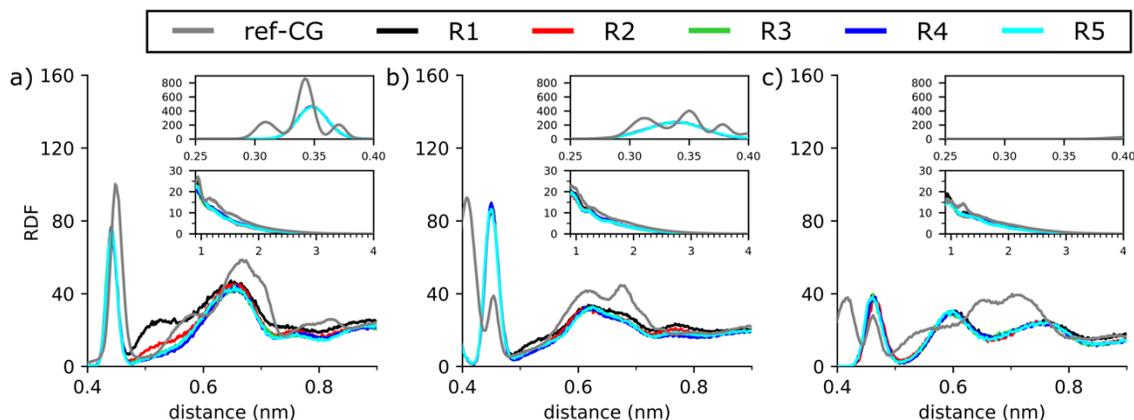
Figure 5. Validation of local properties of CG2k against AA2k (ref-CG). Radial distribution function between bead type pairs (a) P1-P1, (b) P1-Qd, and (c) Qd-Qd. CG simulations were performed with the optimal parameters obtained from five replicas (R1-R5) of automated coarse-graining.

The global properties including the diffusion coefficient D (**Figure 6a**), end-to-end distance of the longest linear chain R_e (**Figure 6b**), and radius of gyration R_g (**Figure 6c**) are compared between CG2k replicas and AA2k for further validation. Note that no global properties were used in the cost functions for parameterizing CG2k. Based on the statistical z-test (SI **Section S4**), the CG and AA predictions were close for all replicas. Furthermore, CG simulations were able to reproduce the AA standard deviation of R_e and R_g. The AA standard deviation of D was reproduced by the first replica, while the other replicas produced a smaller standard deviation. Since the CG parameterization was based on matching AA structural properties (bonded distributions), it is not surprising that less accuracy was observed for dynamic properties such as D. It is worth pointing out that for comparing D, the CG simulation time was scaled by a factor that depended on the mass concentration and PR of PEI as well as the Martini water model. In SI **Section S5**, it was demonstrated that the time scaling factor of a PEI was not significantly affected by the molecular topology. That is, the CG time-scaling factor could be determined from AA and CG simulations of ethylamine in refined polarizable[65] or polarizable water[71], and the same factor

17

would apply to complex PEI structures at the same mass concentration and PR. This is fundamentally different from the existing time-scaling factor used in the literature,[53] which for every new molecule requires a new set of comparisons between AA and CG simulations.
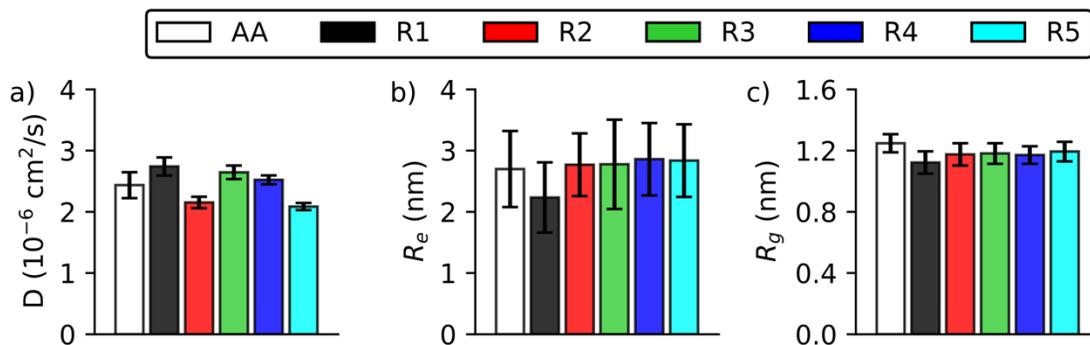
Figure 6. Validation of global properties of CG2k against AA2k. Comparison of (a) diffusion coefficient, (b) end-to-end distance of the longest linear chain, and (c) radius of gyration. CG simulations were performed with the optimal bonded parameters obtained from five replicas (R1-R5) of automated coarse-graining. Diffusion coefficients from CG simulations were scaled based on SI **Section S5**.

3.2 Transferability of CG parameters

To demonstrate the transferability of the CG parameters determined from CG2k, CG and AA simulations were performed for HB2k c37 (**Table 1**; **Section 3.3)**, which has a different DB, PR, and mass concentration. All bonded parameters were directly adopted from the CG2k model (replica R5) without modification. Additionally, to test the transferability of parameters across different Martini water models, CG simulations were performed in polarizable[71] (2.1P or 2.2P) and refined polarizable[65] (2.2refP) water. The local (RDF) and global (D, R_e, R_g) properties predicted from the CG simulations are compared against AA results in **Figure 7**. The accuracy of local properties (**Figure 7a-c**) was better than that of CG2k (**Figure 5**), whereas the accuracy of D (**Figure 7d**), R_e (**Figure 7e**), and R_g (**Figure 7f**) were comparable to **Figure 6**. The accuracy of predicting the mean and standard deviation of R_e and R_g was higher in polarizable water[71] than in refined polarizable water[65] (**Figure 7e-f**). However, all CG results were statistically similar to the

AA results. Overall, the CG parameters remain accurate despite a different branched structure, protonation state, and mass concentration.
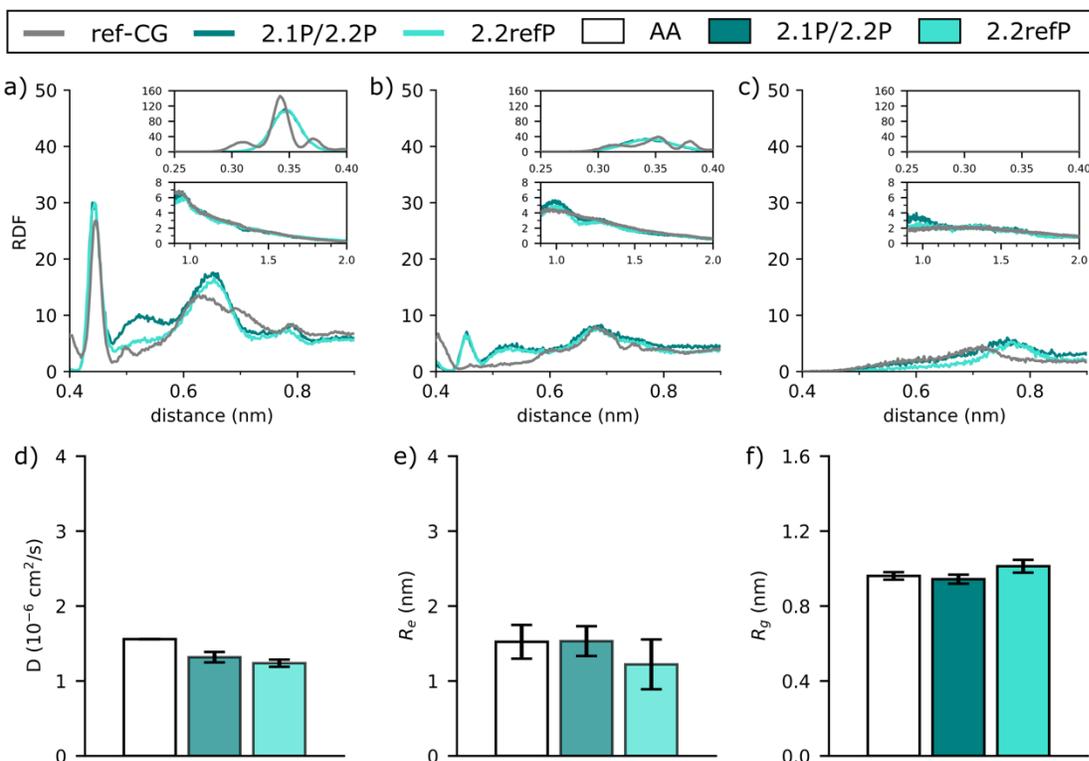
Figure 7. Validation of (a-c) local and (d-f) global properties of CG HB2k c37 against AA results, with parameters from replica R5. Radial distribution functions (RDFs) between bead type pairs (a) P1-P1, (b) P1-Qd, and (c) Qd-Qd and global properties (d) diffusion coefficient, (e) end-to-end distance of the longest linear chain, and (f) radius of gyration are shown. Diffusion coefficients from CG simulations are scaled as described in **Section S5**. Properties are determined from AA trajectories (AA), reference CG trajectories (ref-CG), CG simulations with polarizable water (2.1P/2.2P), and CG simulations with refined polarizable water (2.2refP).

3.3 Validation with experiments

Additional validation was conducted against experiments using the optimal parameters of replica 5 for the CG models because of its lowest total cost among all replicas (**Figure 3a**). For such comparisons, the chemical structure and solvent conditions in the CG simulations must closely match those in the experiments. Commercially available Millipore-Sigma PEIs with M_w = 25 kDa (M_n = 10 kDa) and 2 kDa (M_n = 1.8 kDa) were characterized by von Harpe et al.,[72] and we used the reported amine ratios (1º: 2º: 3º) directly (**Table 2**). The PR of M_w = 25 kDa PEI at pH 7 was determined from the buffering capacity data[72] using an extended theory of Suh and Hwang[23]

(see SI **Section S6**). Due to the lack of buffering capacity data for M_w = 2 kDa PEI, the same PR

as M_w = 25 kDa PEI was used. For ease of discussion, these PEIs are referred to as 25 kDa and 2

kDa PEIs based on their M_w.

Table 2. Characteristics of Millipore-Sigma PEIs, proposed CG PEIs (SL25, MB25, and HB25 for 25 kDa; SL2, MB2, and HB2 for 2 kDa), and PEIs previously reported experimentally in the literature (Roman numerals I to IX). When PR data was unavailable (indicated by ‡), it was predicted from the buffering capacity of 25 kDa Millipore-Sigma PEI[72] using an extended theory of Suh and Hwang[23] (**SI Section S6**).

PEI 25kDa	M_w (kDa)	M_n (kDa)	1°: 2°: 3°	PR (%)	pH	Solvent	Buffer/Salt
Millipore-Sigma[72]	25	10	1.03: 1.30: 1	31	7.0	Water	0 M NaCl
SL25	10	10	1.01: 1.30: 1	31	7.0	Water	0 M NaCl
MB25	10	10	1.01: 1.30: 1	31	7.0	Water	0 M NaCl
HB25	10	10	1.01: 1.30: 1	31	7.0	Water	0 M NaCl
I[73]	10.4	12.7	—	7‡	9.63	Water	0.109 M NaCl
II[74]	7.2	—	1.00: 2.00: 1	34	7.0	Water	0 M NaCl
III[75]	16.6	7.9	—	—	—	Water	1 M NaCl
IV[76]	25	10	1.03: 1.30: 1	26	7.4	Water	20 mM Hepes
V[77]	20	—	—	—	10	Water	0.086 M NaCl
VI[20]	25	10	1.03: 1.30: 1	—	—	Water	—
VII[78]	25	10	1.03: 1.30: 1	28‡	7.2	Water	0.05 M Tris-HCl
VIII[79]	25	10	1.25: 1.50: 1	—	—	Chloroform	—
IX[80]	25	—	—	31‡	7.0	Water	—
PEI 2kDa							
Millipore-Sigma[72]	2	1.8	1.20: 1.13: 1	31‡	7.0	Water	0 M NaCl
SL2	1.8	1.8	1.08: 1.15: 1	31	7.0	Water	0 M NaCl
MB2	1.8	1.8	1.08: 1.15: 1	31	7.0	Water	0 M NaCl
HB2	1.8	1.8	1.08: 1.15: 1	31	7.0	Water	0 M NaCl
I[73]	2.1	—	—	7‡	9.63	Water	0.109 M NaCl
VII[78]	2	1.8	1.08: 1.15: 1	—	7.2	Water	0.05 M Tris-HCl

For a given MW, (1°: 2°: 3°) ratio and PR, the numbers of primary, secondary, tertiary, and

protonated amines were determined (see SI **Section S7**). With these descriptors, the number of

possible chemical structures was still large. In this work, three branched structures, semi-linear

(SL), moderately-branched (MB), and hyper-branched (HB), were proposed for 2 and 25 kDa CG

PEIs (**Figure 8**). They were named SL2, MB2, HB2, SL25, MB25, and HB25, denoting their DB

and M_w of the PEI they are modeling (**Table 2**). When designing these structures, we used

repeating units of t(sspq) and t(spq) as well as terminals of pq, p, and sspq such that all bonded

parameters were available from coarse-graining the 2kDa PEI in **Section 3.1**. As shown in **Figure**

8, SL structures had the smallest side chain and longest main chain; MB structures had a side chain between SL and HB; and HB structures had comparable main and side chains. Typically, the main chain was composed of tertiary beads, to which side chains of similar lengths were attached. Some side chains in the MB and HB structures were small (spq) to match the experimental descriptors. Due to the significant difference among the SL, MB, and HB structures, they are well representative of branched PEIs with $M_n \leqslant 10$ kDa, and the properties of other branched structures would likely lie between those of SL and HB PEIs.
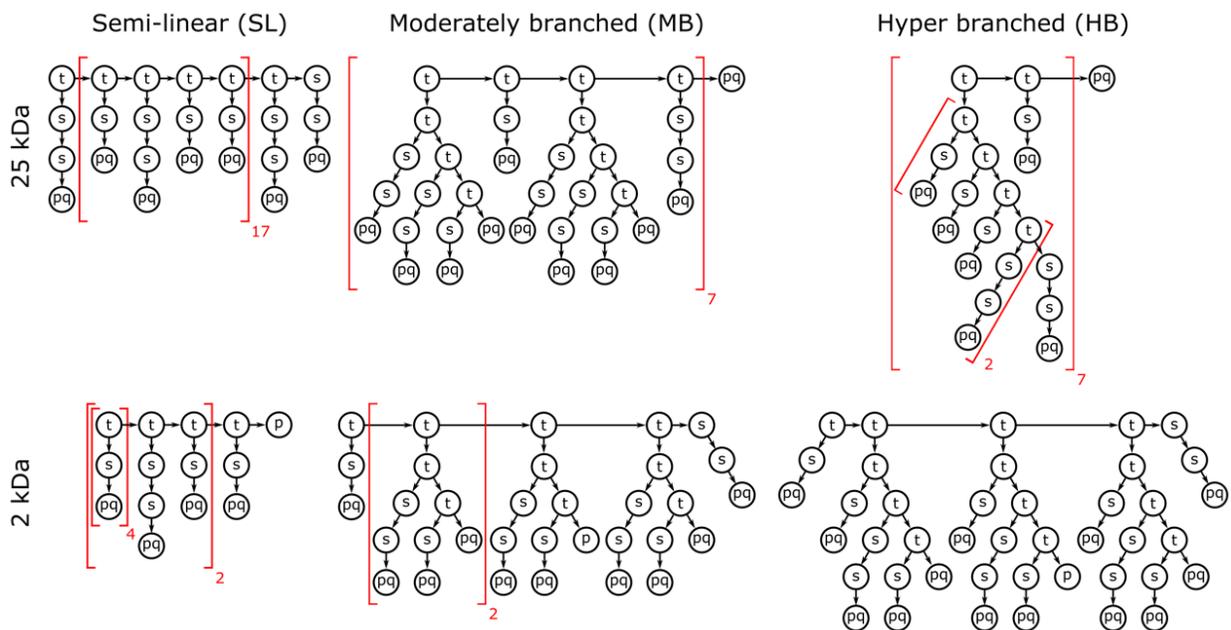
Figure 8. Proposed CG structures for branched 25 kDa and 2 kDa Millipore-Sigma PEIs. The molecular weight of each structure is consistent with the M_n of the Millipore-Sigma PEI it represents.

The diffusion coefficient D_0 (**Figure 9a**) and radius of gyration R_0 (**Figure 9c**) at infinite dilution and the intrinsic viscosity $[\eta]$ (**Figure 9c**) were evaluated for the CG PEIs. A comparison between 2 and 25 kDa CG PEIs showed that D_0 was higher for 2 kDa and R_0 was higher for 25 kDa (**Figure 9a, b**), which is consistent with the understanding that D_0 and the Stokes-Einstein radius are inversely proportional.[78] The differences in $[\eta]$ between 2 and 25 kDa CG PEIs were insignificant (**Figure 9c**). For 25 kDa PEIs, D_0 increased and R_0 decreased with the increase in DB,

while the properties of 2 kDa PEIs were not significantly affected by DB. The dependence of $[\eta]$

on DB was negligible considering the standard deviation.
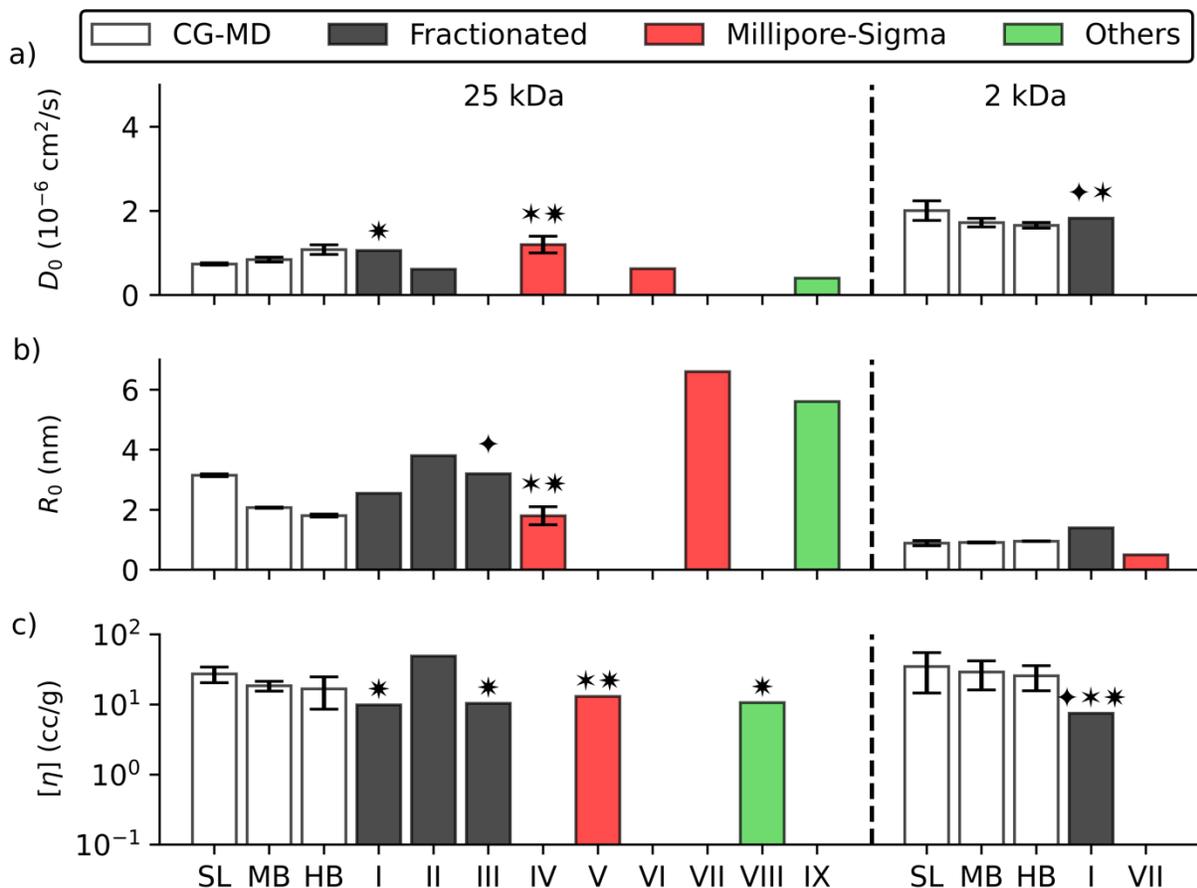
Figure 9. Cross-validation of the CG PEI forcefield with experimental results. Comparison of (a) diffusion coefficient and (b) radius of gyration at infinite dilution, and (c) intrinsic viscosity. CG PEI results were calculated for three branched structures: semi-linear (SL), moderately-branched (MB), and hyper-branched (HB) based on the characterization of branched M_w = 25 kDa and 2 kDa Millipore-Sigma PEIs. The experimental results are numbered I-IX. Experimental data statistically similar within a 95% confidence level to the data of SL, MB, and HB CG PEIs were respectively marked with ♦, ✶, and ✳. Data from (I) Hostetler and Swanson,[73] (II) Lindquist and Stratton,[74] (III) Park and Choi,[75] (IV) Clamme et al.,[76] (V) Ghriga et al.,[77] (VI) Zhou et al.,[20] (VII) Andersson et al.,[78] (VIII) Chen et al.,[79] and (IX) Nguyen et al.[80]

Previously published experimental data for PEI had diverse branched structures and solvent conditions (**Table 2**). These PEIs were either polydisperse with an average $M_w \sim 2$ or 25 kDa, or fractionated with $M_n \sim 1.8$ kDa or 10 kDa. Available experimental values for D_0, R_0 (Stokes-Einstein radius), and $[\eta]$ are shown in **Figure 9**. Overall, the experimental data compared well with CG simulations, which was confirmed quantitatively via the statistical z-test (see SI **Section**

S8). For each experimental data, the CG PEI model(s) that provided predictions within a 95% confidence interval from the experiment are indicated by ✦, ✶, and ✳ for SL, MB, and HB, respectively (**Figure 9**). Based on the comparison in **Figure 9** and the statistical z-test data (SI **Section S8**), we can discern that HB25 is a good representation of the 25 kDa PEI in I[73], III[75], IV[76], V[77], and VIII[79], and SL2 can reasonably represent the structure of the 2 kDa PEI in I[73]. It is important to note that there are likely other branched structures that can produce results similar to those of HB25 or SL2. However, they should also possess physical properties that are comparable to HB25 or SL2. In other words, one should have confidence in modeling the PEIs in these experiments using HB25 or SL2 and performing CG simulations to investigate phenomena that cannot be observed at experimental resolution.

4. Discussion

In this work, the CG forcefield for PEI with an arbitrary MW and DB was determined from the AA trajectory and topology using graph analysis and adaptive gradient descent, consistent with the Martini methodology.[53] CG simulations with the identified optimal bonded parameters could accurately predict AA local properties such as RDF (**Figure 5**) and bonded distributions (SI **Section S3**), as well as global properties such as D, R_e, and R_g (**Figure 6**). Compared to our previous study,[28] the present work: (i) automates the AA to CG mapping and CG topology generation with the consideration of bead asymmetry; (ii) introduces new cost functions that allow fast convergence (8 hours for a 2 kDa PEI vs. a few months for eight 600 Da PEIs) and accurate predictions of bonded parameters; (iii) demonstrates that bonded parameters can be partially (**Section 3.1**) or fully (**Section 3.2**) repurposed for PEIs with different MW, DB, PR, mass concentration, and in different Martini water models; and (iv) makes direct comparison with experimental data to predict the probable chemical structures of synthesized PEIs.

The CG forcefield generated from different replicas had similar accuracy in reproducing AA local and global properties (**Figure 5**, **Figure 6**, SI **Section S3**). The performance of the parameterization was not significantly affected by small changes in the settings of the adaptive gradient descent algorithm (e.g., initial learning rate and parameter constraints, see **SI Section S2**). While the CG parameterization in **Section 3.1** was demonstrated for a single hydrated PEI, the algorithm and the package *coarsen* can parameterize AA trajectories containing multiple PEIs with diverse MW, DB, and PR in the presence or absence of solvent(s).

During the NPT equilibration of system CG2k, the volume of the simulation box changed, increasing the mass concentration of CG2k PEI from 5.7 g/L (same as the AA simulation) to 8.4 g/L. Using the van der Waal's radius of 0.12 nm when adding refined polarizable water,[65] the number of water molecules added to the system was insufficient, leading to volume reduction during equilibration. An additional simulation was performed with the same optimal bonded parameters as replica 5 and van der Waal's radius, but the number of water molecules was increased (along with a larger initial box) to achieve an equilibrium mass concentration of PEI comparable to that in the AA simulation (6.0 g/L). The comparison of AA and CG RDFs improved after the mass concentration correction (**Figure 5** vs. **Figure S11**), whereas the quality of bonded distributions and global properties remained the same (**SI Section S3** and **S9**). The mass concentration correction was applied to the CG HB2k c37 system discussed in **Section 3.2**, thereby producing more accurate RDF curves. The comparison between CG PEI and experimental data did not need mass concentration correction because global properties were determined at infinite dilution using linear regression.

Potter et al.[36] recently developed an automated mapping algorithm for Martini by modifying the algorithm of Webb et al.[35] In their algorithm, coarse-graining a simple linear PEI

such as (C-C-N)$_4$ produced dissimilar beads (-(C-C-N)-, -(C-C-N)-, -(C-C-N-C)-, and -(C-N)-), which was inconsistent with the Martini methodology[53]. Furthermore, the algorithm cannot capture the asymmetry of -(C-C-N)- beads and the resulting asymmetric bonded distributions (as discussed in Methods). These issues were addressed by our work.

There are differences in design philosophies between our iterative parameterization algorithm and that of Empereur-Mot et al.[48] Here we elaborate on the performance of the two algorithms by comparing the parameterization of our 2 kDa PEI in **Figure 2b** with the parametrization of C3-symmetric benzotrithiophene and naphthalene diimide in Empereur-Mot et al.[48] The 2 kDa PEI contained 27 new bonded distributions and 180 new parameters. C3-symmetric benzotrithiophene had a comparable number of distributions (25) but a lower number of parameters (50), and naphthalene diimide had a larger number of distributions (49) but half the number of parameters (98). First, we examine computational efficiency. Empereur-Mot et al.[48] obtained optimal parameters for C3-symmetric benzotrithiophene and naphthalene diimide after 240 and 360 iterations, respectively, which is significantly higher than the number of iterations needed by our algorithm (at most 84 including the sub-iterations for Jacobian calculation). This is expected because particle swarm optimization searches more parameter space than adaptive gradient descent. In contrast, gradient descent is known to converge to the nearest minimum. Using initial guesses obtained from direct Boltzmann inversion, our optimal CG bonded parameters could reproduce local and global AA properties, which suggests that gradient descent likely converged to a global minimum. Therefore, our algorithm provides an efficient way to optimize parameters, especially when the number of parameters is high.

Second, we examine the quality of reproducing the local structural properties. By minimizing the earth mover's distance,[60,61] Empereur-Mot et al.[48] could match most of the

reference bonded distributions for C3-symmetric benzotrithiophene and naphthalene diimide, although in some cases they struggled to match the standard deviation. Our algorithm achieved a similar level of matching with the reference bonded distributions. As discussed in the Results, some reference N-type bond angle distributions were difficult to model due to strong repulsion between beads at small angles. It was also challenging to perfectly model certain reference dihedral angle distributions (SI **Section S3**). However, this is not a flaw associated with the adaptive gradient descent algorithm. Instead, we had demonstrated in a previous work that different reference bonded distributions are possible for the same dihedral angle due to local steric hindrance.[28] Among the dihedral angles parameterized in the present work, the CG distribution did not perfectly match the reference bonded distribution for tttt, Nsqttpq, Nsttt, and Npqttsq (see SI **Section S3**); these dihedral angles have at least two tertiary beads, signifying the importance of local steric hindrance.

Third, we compare the quality of reproducing global structural properties. Empereur-Mot et al.[48] accurately reproduced the mean of R_g but not its standard deviation. In comparison, our optimal bonded parameters correctly predicted the mean and standard deviation of R_g and R_e. The higher accuracy in our results is most likely due to the definition of the cost function, which minimizes the difference in means and standard deviations between CG and reference bonded distributions. That is, by modeling the fluctuations in local structural properties, the fluctuations in global structural properties are predicted with higher precision. Other considerations in our algorithm, such as the asymmetry of beads and multiple periodic functions for dihedral angles, can also contribute to higher accuracy.

The recent work of Empereur-Mot et al.[47] demonstrated the parameterization of bonded and nonbonded interactions of several lipids in implicit/explicit water. A multi-objective

optimization approach was taken to reproduce AA bonded distributions and RDFs as well as experimentally measured area per lipid and lipid bilayer thickness. The methodology presents several advantages, including automated parameterization of bonded and nonbonded interactions and good accuracy due to the application of multiple objective functions. On the other hand, simultaneously parameterizing bonded and nonbonded interactions can lead to unintended artifacts. One artifact could be the over- or under-stretching of the bond lengths to match the experimental lipid bilayer thickness instead of seeking the optimal nonbonded interactions. For example, large discrepancies in the reference and CG bond length distributions were observed for POPC in explicit water.[47] Whether or not this has a strong influence on the transferability of the parameters remains to be explored. In comparison, in this work, nonbonded and bonded interactions are parameterized separately following the Martini methodology, which avoids such an artifact.[53] Nonbonded parameters are determined from solvation and partitioning free energies, which have proven to be robust and reproduce the potential of mean force (PMF) profiles from AA simulations.[28,31]

For the first time, chemical structures of CG PEIs were proposed based on a commercial product (Millipore-Sigma PEIs), and their physical properties were compared with experiments. The results suggest the feasibility of using simulations to predict probable chemical structures of monodisperse PEI that can represent PEIs in experiments. Simulating these chemical structures can provide unique insights into the structure-function relationships that govern biological or physical phenomena that may not be accessible by experiments. It should be noted that the reported experimental data were obtained under varied conditions (pH, solvent, buffer, and salt concentration). Furthermore, synthesized PEI solutions are polydisperse even after fractionation. Replicating all experimental conditions and performing corresponding CG simulations is beyond

27

the scope of this work. While our simulation systems were designed to match experimental characterization and solution conditions as closely as possible, a better comparison can be achieved by developing Martini models for buffers and modeling polydisperse PEI systems. Additional experimental data, such as multidimensional NMR and the distribution of PEI's MW, will benefit the construction of such polydisperse systems.

Additional discussions are warranted for the PEI properties being compared. In experiments, the diffusion coefficient is calculated from the Stokes-Einstein radius (and vice versa) at a specified mass concentration or infinite dilution. Since Stokes-Einstein radius and R_g are by definition different, a comparison of D_0 (or D) between CG PEI and experiments is more appropriate. In **Figure 9**, the standard deviation in $[\eta]$ is much higher than that in D_0 and R_0. Attempts to decrease the standard deviation by increasing the number of random walks, interior samples, and PEI configurations in the ZENO program[81,82] were unsuccessful. Calculating η and $[\eta]$ using Green-Kubo relations generated unphysical values, probably because of the large timesteps used in CG simulations, which increase the integration error of the time autocorrelation function. For this work, the standard deviation in $[\eta]$ appears acceptable. However, further testing with different salt concentrations and validation with corresponding experiments are recommended for future studies.

Currently, the software is limited to PEIs. Due to the large diversity of PEI structures and number of unparameterized bonded distributions, we expect this software would greatly speed up their CG modeling and simulation. The same algorithm can be adapted to automate the CG parameterization of other polymers, which requires only three changes in the open-source software. First, different repeating units should be searched, e.g., -(C-C(-C))- for branched or linear polypropylene (PPL) and -(O-C(=O)-C(-C))- for linear poly(lactide) (PLA). To handle block

polymers or polymers with large repeating units, different units can be searched one after another. For example, AA to CG mapping for PEI-g-polyethylene glycol (PEG) can be determined by searching all -(C-C-N)- units first, followed by searching all -(C-O-C)- units. For polymethyl methacrylate (PMMA), the units of -(C-O-C(=O))- and -(C-C(-C))- can be searched sequentially. Second, the criteria for identifying the "head node" might need modification. For example, the "head node" in PLL is a carbon atom with three hydrogens (the same as in PEI), while in PLA it is an oxygen atom with one hydrogen. Although PLL does not have a unique "head node", starting from any "head node" and searching for the repeating units would produce the same AA to CG mapping. Finally, to handle CG molecules with only symmetric repeating units, the CG bonds should be represented by undirected edges. The iterative algorithm to determine bonded parameters remains the same. The automated structure generation from SMILE strings, AA to CG mapping, and CG parameterization scheme are particularly beneficial for modeling branched polymers and would allow high throughput simulations of complex polydisperse solutions. In our upcoming work, a general mapping algorithm that can handle molecules consisting of both symmetric and asymmetric units and fractional AA to CG mapping will be discussed, which is beyond the scope of this study.

One final comment is that in the present work, the bonded parameters for PEI have been determined and validated for dilute hydrated systems (mass concentrations 6-37 g/L) and could be different for PEIs with high mass concentrations, in non-aqueous solvent(s), or PEI melts. In such cases, the bonded parameters derived here should not be repurposed without care. If an AA forcefield is able to accurately simulate such conditions, our automated algorithm can derive new bonded parameters from the reference CG trajectories. The nonbonded parameters are expected to remain the same because they were derived from partitioning free energy data between water and

different organic solvents; however, additional validation is recommended. The ability of our CG model to predict complex phenomena such as entanglements also needs further exploration.

5. Conclusion

An automated algorithm was developed to determine Martini CG forcefield parameters for PEI with arbitrary molecular weight, chemical structure, and protonation state. The parameterization of a 2 kDa PEI was demonstrated as an example, where the optimization converged within 30 iterations. Four additional replica parameterizations were performed to demonstrate the uniqueness of the optimized parameters. The derived CG PEI forcefield reproduced the structural (radius of gyration, end-to-end distance of the longest linear chain, radial distribution function, and bonded distributions) and dynamic (diffusion coefficient) properties of AA PEI.

Experimental characterization data were used to propose PEI structures and model them using the newly derived parameters. Specifically, three branched structures were proposed for 2 and 25 kDa Millipore-Sigma PEIs. Multiple CG simulations were performed for each PEI to calculate the diffusion coefficient and radius of gyration at infinite dilution as well as the intrinsic viscosity. The properties of proposed CG PEI structures agreed well with experimental data, thus providing a strategy to infer and validate probable chemical structures in an experimental system.

6. Supporting Information

Supporting methods; discussions on parameter constraints and learning rate in adaptive gradient descent; comparing reference and CG bonded distributions; z-score between CG and AA data; CG time scaling factor; determining protonation ratio from buffering capacity; determining the number of primary, secondary, tertiary, and protonated beads in PEI; z-score between CG and

experimental data; transferability of CG parameters to new PEIs; and effects of equilibrium mass concentration of PEI and water model on CG PEI properties.

7. Conflict of Interest

Authors declare no competing financial interests.

8. Acknowledgements

9. Data and Code Availability

The open-source software *coarsen* is available in Github, https://github.com/subhamoy mahajan/Martini_PEI_ff.

10. References

(1)   Luo, H.; Peng, H.; Zhao, Q. High Flux Mg2+/Li+ Nanofiltration Membranes Prepared by Surface Modification of Polyethylenimine Thin Film Composite Membranes. *Appl. Sufrace Sci.* **2022**, *579*, 152161.
(2)   Ding, J.; Wu, H.; Wu, P. Preparation of Highly Permeable Loose Nanofiltration Membranes Using Sulfonated Polyethylenimine for Effective Dye/Salt Fractionation. *Chem. Eng. J.* **2020**, *396*, 125199.
(3)   Hack, J.; Frazzetto, S.; Evers, L.; Maeda, N.; Meier, D. M. Branched Versus Linear Structure: Lowering the CO2 Desorption Temperature of Polyethylenimine-Functionalized Silica Adsorbents. *Energies* **2022**, *15*, 1075.
(4)   Meng, Y.; Ju, T.; Meng, F.; Han, S.; Song, M.; Jiang, J. Insights into the Critical Role of Abundant-Porosity Supports in Polyethylenimine Functionalization as Efficient and Stable CO2 Adsorbents. *ACS Appl. Mater. Interfaces* **2021**, *13*, 54018–54031.
(5)   Xue, Q.; Bai, J.; Han, C.; Chen, P.; Jiang, J.-X.; Chen, Y. Au Nanowires@Pd-Polyethylenimine Nanohybrids as Highly Active and Methanol-Tolerant Electrocatalysts toward Oxygen Reduction Reaction in Alkaline Media. *ACS Catal.* **2018**, *8*, 11287–11295.
(6)   Wang, X.; Ma, X.; Song, C. Deep Removal of SO2 from Cathode Air over Polyethylenimine-Modified SBA-15 Sorbents for Fuel Cells. *Catal. Today* **2021**, *371*, 240–246.
(7)   Li, P.; Jiang, Z.; Huang, H.; Tian, S.; Hu, X.; Li, W.; Bai, X.; Wang, Y. Improving Electron Transport by Using a NaCl/Polyethylenimine Ethoxylated Double Layer for High-Efficiency Polymer Solar Cells. *Mater. Sci. Semicond. Process.* **2020**, *116*, 105150.
(8)   Sadeghianlemraski, M.; Aziz, H. Reducing Ultraviolet-Induced Open-Circuit Voltage Loss in Inverted Organic Solar Cells by Maintaining Charge Selectivity of the Electron Collection Contact Using Polyethylenimine. *Sol. Energy* **2020**, *198*, 427–433.
(9)   Li, J.; Wang, X.; Shen, M.; Shi, X. Polyethylenimine-Assisted Generation of Optical Nanoprobes for Biosensing Applications. *ACS Appl. Bio Mater.* **2020**, *3*, 3935–3955.
(10)   Calamak, S.; Ulubayram, K. Polyethylenimine-Mediated Gold Nanoparticle Arrays with Tunable Electric Field Enhancement for Plasmonic Applications. *J. Mater. Sci. Mater. Electron.* **2019**, *30*, 10013–10023.

(11) Hao, F.; Li, Y.; Zhu, J.; Sun, J.; Marshall, B.; Lee, R. J.; Teng, L.; Yang, Z.; Xie, J. Polyethylenimine-Based Formulations for Delivery of Oligonucleotides. *Curr. Med. Chem.* **2019**, *26*, 2264–2284.

(12) Zou, Y.; Li, D.; Shen, M.; Shi, X. Polyethylenimine-Based Nanogels for Biomedical Applications. *Macromol. Biosci.* **2019**, *19*, 1900272.

(13) Wei, X.; Hong, J.; Zhu, S.; Chen, J.; Lv, B. Structure–Performance Study of Polyamide Composite Nanofiltration Membranes Prepared with Polyethyleneimine. *J. Mater. Sci.* **2017**, *52*, 11701–11714.

(14) Werth, S.; Urban-Klein, B.; Dai, L.; Höbel, S.; Grzelinski, M.; Bakowsky, U.; Czubayko, F.; Aigner, A. A Low Molecular Weight Fraction of Polyethylenimine (PEI) Displays Increased Transfection Efficiency of DNA and SiRNA in Fresh or Lyophilized Complexes. *J. Control. Release* **2006**, *112*, 257–270.

(15) Bieber, T.; Elsässer, H. P. Preparation of a Low Molecular Weight Polyethylenimine for Efficient Cell Transfection. *Biotechniques* **2001**, *30*, 74–77, 80–81.

(16) Krämer, M.; Stumbé, J.-F.; Grimm, G.; Kaufmann, B.; Krüger, U.; Weber, M.; Haag, R. Dendritic Polyamines: Simple Access to New Materials With Defined Treelike Structures for Application in Nonviral Gene Delivery. *ChemBioChem* **2004**, *5*, 1081–1087.

(17) Brissault, B.; Kichler, A.; Guis, C.; Leborgne, C.; Danos, O.; Cheradame, H. Synthesis of Linear Polyethylenimine Derivatives for DNA Transfection. *Bioconjug. Chem.* **2003**, *14*, 581–587.

(18) Dick, C. R.; Ham, G. E. Characterization of Polyethylenimine. *J. Macromol. Sci. Part A- Chem.* **1970**, *4*, 1301–1314.

(19) Mattice, W. L.; Porter, L. J. Molecular Weight Averages and 13C NMR Intensities Provide Evidence for Branching in Proanthocyanidin Polymers. *Phytochemistry* **1984**, *23*, 1309–1311.

(20) Zhou, B.; Shen, M.; Bányai, I.; Shi, X. Structural Characterization of PEGylated Polyethylenimine-Entrapped Gold Nanoparticles: An NMR Study. *Analyst* **2016**, *141*, 5390–5397.

(21) Alexson, D. E.; Blake, S. L. Multinuclear (13C, 15N) NMR Study of Polyethyleneimine-Based Polymers. *J. Polym. Sci.* **1985**, *23*, 2507–2525.

(22) Holycross, D. R.; Chai, M. Comprehensive NMR Studies of the Structures and Properties of PEI Polymers. *Macromolecules* **2013**, *46*, 6891–6897.

(23) Suh, J.; Paik, H. J.; Hwang, B. K. Ionization of Poly(Ethylenimine) and Poly(Allylamine) at Various PH′s. *Bioorg. Chem.* **1994**, *22*, 318–327.

(24) Kondinskaia, D. A.; Kostritskii, A. Y.; Nesterenko, A. M.; Antipina, A. Y.; Gurtovenko, A. A. Atomic-Scale Molecular Dynamics Simulations of DNA-Polycation Complexes: Two Distinct Binding Patterns. *J. Phys. Chem. B* **2016**, *120*, 6546–6554.

(25) Ziebarth, J.; Wang, Y. Molecular Dynamics Simulations of DNA-Polycation Complex Formation. *Biophys. J.* **2009**, *97*, 1971–1983.

(26) Beu, T. A.; Ailenei, A.-E.; Farcaş, A. CHARMM Force Field for Protonated Polyethyleneimine. *J. Comput. Chem.* **2018**, *39*, 2564–2575.

(27) Sun, C.; Tang, T.; Uludağ, H.; Cuervo, J. E. Molecular Dynamics Simulations of DNA/PEI Complexes: Effect of PEI Branching and Protonation State. *Biophys. J.* **2011**, *100*, 2754–2763.

(28) Mahajan, S.; Tang, T. Martini Coarse-Grained Model for Polyethylenimine. *J. Comput. Chem.* **2019**, *40*, 607–618.

(29) Ziebarth, J. D.; Wang, Y. Understanding the Protonation Behavior of Linear Polyethylenimine in Solutions Through Monte Carlo Simulations. *Biomacromolecules* **2010**, *11*, 29–38.

(30) Gallops, C. E.; Ziebarth, J. D.; Wang, Y. Coarse-Grained Simulations of the Impact of Chain Length and Stiffness on the Formation and Aggregation of Polyelectrolyte Complexes. *Macromol. Theory Simul.* **2020**, *29*, 2000015.

(31) Mahajan, S.; Tang, T. Erratum : "Martini Coarse-Grained Model for Polyethylenimine" [J. Comput. Chem. 2019, 40, 607-618, DOI:10.1002/Jcc.25747]. *J. Comput. Chem.* **2020**, *41*, 1730–1734.

(32) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure* **2006**, *14*, 1767–1777.

(33) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073–5083.

(34) Bereau, T.; Kremer, K. Automated Parametrization of the Coarse-Grained Martini Force Field for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 2783–2791.

(35) Webb, M. A.; Delannoy, J. Y.; De Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *J. Chem. Theory Comput.* **2019**, *15*, 1199–1208.

(36) Potter, T. D.; Barrett, E. L.; Miller, M. A. Automated Coarse-Grained Mapping Algorithm for the Martini Force Field and Benchmarks for Membrane-Water Partitioning. *J. Chem. Theory Comput.* **2021**.

(37) Noid, W. G. Perspective: Coarse-Grained Models for Biomolecular Systems. *J. Chem. Phys.* **2013**, *139*, 090901.

(38) Jin, J.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. *J. Chem. Theory Comput.* **2022**, *18*, 5759–5791.

(39) Schmid, F. Toy Amphiphiles on the Computer: What Can We Learn from Generic Models? *Macromol. Rapid Commun.* **2009**, *30*, 741–751.

(40) Soper, A. K. Empirical Potential Monte Carlo Simulation of Fluid Structure. *J. Chem. Phys.* **1996**, *202*, 295–306.

(41) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of Coarse-Grained Potentials via Multistate Iterative Boltzmann Inversion. *J. Chem. Phys.* **2014**, *140*, 224104.

(42) Ganguly, P.; Mukherji, D.; Junghans, C.; Vegt, N. F. A. van der. Kirkwood−Buff Coarse-Grained Force Fields for Aqueous Solutions. *J. Chem. Theory Comput.* **2012**, *8*, 1802–1807.

(43) Oliveira, T. E. de; Netz, P. A.; Kremer, K.; Junghans, C.; Mukherji, D. C-IBI: Targeting Cumulative Coordination within an Iterative Protocol to Derive Coarse-Grained Models of (Multi-Component) Complex Fluids. *J. Chem. Phys.* **2016**, *144*, 174106.

(44) Lyubartsev, A. P.; Naômé, A.; Vercauteren, D. P.; Laaksonen, A. Systematic Hierarchical Coarse-Graining with the Inverse Monte Carlo Method. *J. Chem. Phys.* **2015**, *143*, 243120.

(45) Shahidi, N.; Chazirakis, A.; Harmandaris, V.; Doxastakis, M. Coarse-Graining of Polyisoprene Melts Using Inverse Monte Carlo and Local Density Potentials. *J. Chem. Phys.* **202AD**, *152*, 124902.

(46) Christensen, R.; Sørensen, S. S.; Liu, H.; Li, K.; Bauchy, M.; Smedskjaer, M. M. Interatomic Potential Parameterization Using Particle Swarm Optimization: Case Study of Glassy Silica. *J. Chem. Phys.* **2021**, *154*, 134505.

(47) Empereur-Mot, C.; Capelli, R.; Perrone, M.; Caruso, C.; Doni, G.; Pavan, G. M. Automatic Multi-Objective Optimization of Coarse-Grained Lipid Force Fields Using SwarmCG. *J. Chem. Phys.* **2022**, *156*.

(48) Empereur-Mot, C.; Pesce, L.; Doni, G.; Bochicchio, D.; Capelli, R.; Perego, C.; Pavan, G. M. Swarm-CG: Automatic Parametrization of Bonded Terms in MARTINI-Based Coarse-Grained Models of Simple to Complex Molecules via Fuzzy Self-Tuning Particle Swarm Optimization. *ACS Omega* **2020**, *5*, 32823–32843.

(49) 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.

(50) Ercolessi, F.; Adams, J. B. Interatomic Potentials from First-Principles Calculations: The Force-Matching Method. *Europhys. Lett.* **1994**, *26*, 583–588.

(51) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 1. General Principles. *J. Chem. Theory Comput.* **2013**, *9*, 2466–2480.

(52) Shell, M. S. The Relative Entropy Is Fundamental to Multiscale and Inverse Thermodynamic Problems. *J. Chem. Phys.* **2008**, *129*, 144108.

(53) 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.

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

(55) Uusitalo, J. J.; Ingólfsson, H. I.; Marrink, S. J.; Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophys. J.* **2017**, *113*, 246–256.

(56) Uusitalo, J. J.; Ingólfsson, H. I.; Akhshi, P.; Tieleman, D. P.; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11*, 3932–3945.

(57) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, 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.

(58) Lee, H.; de Vries, A. H.; Marrink, S. J.; Pastor, R. W. A Coarse-Grained Model for Polyethylene Oxide and Polyethylene Glycol: Conformation and Hydrodynamics. *J. Phys. Chem. B* **2009**, *113*, 13186–13194.

(59) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(60) Rubner, Y.; Tomasi, C.; Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121.

(61) Pele, O.; Werman, M. Fast and Robust Earth Mover's Distances. *IEEE 12th Int. Conf. Comput. Vis.* **2019**, 460–467.

(62) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. Prediction of N-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**,

41, 1407–1421.

(63)   2014., V. 1. *RDKit: Open-source cheminformatics,*. Online: http://www.rdkit.org (accessed: July 1, 2014).

(64)   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.

(65)   Michalowsky, J.; Schäfer, L. V.; Holm, C.; Smiatek, J. A Refined Polarizable Water Model for the Coarse-Grained MARTINI Force Field With Long-Range Electrostatic Interactions. *J. Chem. Phys.* **2017**, *146*, 054501.

(66)   Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(67)   Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; A, D. M. J. CHARMM General Force Field (CGenFF): A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(68)   Sun, C.; Tang, T.; Uludağ, H. Molecular Dynamics Simulations for Complexation of DNA with 2 KDa PEI Reveal Profound Effect of PEI Architecture on Complexation. *J. Phys. Chem. B* **2012**, *116*, 2405–2413.

(69)   Song, J.; Ma, M.; Dong, Y.; Wan, M.; Fang, W.; Gao, L. Coarse-Grained Force Field for Polyethylene Oxide and Polyethylene Glycol Aqueous Solutions Based on a Polarizable Water Model. *J. Chem. Theory Comput.* **2023**, *19*, 1864–1874.

(70)   Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(71)   Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Comput. Biol.* **2010**, *6*, e1000810.

(72)   von Harpe, A.; Petersen, H.; Li, Y.; Kissel, T. Characterization of Commercially Available and Synthesized Polyethylenimines for Gene Delivery. *J. Control. Release* **2000**, *69*, 309–322.

(73)   Hostetler, R. E.; Swanson, J. W. Diffusion Into and Adsorption of Polyethylenimine on Porous Silica Gel. *J Polym Sci* **1974**, *12*, 29–43.

(74)   Lindquist, G. M.; Stratton, R. A. The Role of Polyelectrolyte Charge Density and Molecular Weight on the Adsorption and Flocculation of Colloidal Silica with Polyethylenimine. *J. Colloid Interface Sci.* **1976**, *55*, 45–59.

(75)   Park, I. H.; Choi, E. J. Characterization of Branched Polyethyleneimine by Laser Light Scattering and Viscometry. *Polymer (Guildf).* **1996**, *37*, 313–319.

(76)   Clamme, J. P.; Azoulay, J.; Mély, Y. Monitoring of the Formation and Dissociation of Polyethylenimine/DNA Complexes by Two Photon Fluorescence Correlation Spectroscopy. *Biophys. J.* **2003**, *84*, 1960–1968.

(77)   Ghriga, M. A.; Gareche, M.; Khodja, M.; Andreu, N.; Lebouachera, S. E. I.; Khoukh, A.; Drouiche, N.; Grassl, B. Structure–Property Relationships of the Thermal Gelation of Partially Hydrolyzed Polyacrylamide/Polyethylenimine Mixtures in a Semidilute Regime. *Polym. Bull.* **2020**, *77*, 1465–1488.

(78)   Andersson, M. M.; Hatti-Kaul, R.; Brown, W. Dynamic and Static Light Scattering and Fluorescence Studies of the Interactions between Lactate Dehydrogenase and Poly(Ethyleneimine). *J. Phys. Chem. B* **2000**, *104*, 3660–3667.

(79)   Chen, Y.; Shen, Z.; Frey, H.; Pérez-Prieto, J.; Stiriba, S. E. Synergistic Assembly of Hyperbranched Polyethylenimine and Fatty Acids Leading to Unsual Supramolecular Nanocapsules. *Chem. Commun.* **2005**, No. 6, 755–757.

(80)   Nguyen, J.; Xie, X.; Neu, M.; Dumitrascu, R.; Reul, R.; Sitterberg, J.; Bakowsky, U.; Schermuly, R.; Fink, L.; Schmehl, T.; Gessler, T.; Seeger, W.; Kissel, T. Effects of Cell-Penetrating Peptides and Pegylation on Transfection Efficiency of Polyethylenimine in Mouse Lungs. *J. Gene Med.* **2008**, *10*, 1236–1246.

(81)   Juba, D.; Audus, D. J.; Mascagni, M.; Douglas, J. F.; Keyrouz, W. ZENO: Software for Calculating Hydrodynamic, Electrical, and Shape Properties of Polymer and Particle Suspensions. *J. Res. Natl. Inst. Stand. Technol.* **2017**, *122*, 1–2.

(82)   Mansfield, M. L.; Douglas, J. F. Improved Path Integration Method for Estimating the Intrinsic Viscosity of Arbitrarily Shaped Particles. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2008**, *78*, 1–11.

For Table of Contents Only

t: tertiary   s: secondary   p: primary   q: charged

→ bead asymmetry