Electronic Supplementary Material (ESI) for Green Chemistry.

Supporting Information: Systematic Parameterization of Lignin for the CHARMM Force Field

Josh V. Vermaas,[†] Loukas Petridis,[‡] John Ralph,[¶] Michael F. Crowley,[*,†] and Gregg T. Beckham[*,§]

†*Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401*

‡*UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37831*

¶*Department of Biochemistry, University of Wisconsin, Madison, WI 53726, USA*

§*National Bioenergy Center, National Renewable Energy Laboratory, Golden, CO 80401*

E-mail: michael.crowley@nrel.gov; gregg.beckham@nrel.gov

Supporting Archives

Four archives are provided alongside this document to assist the reader in reproducing the results and using the force field, all accessible via https://cscdata.nrel.gov/#/datasets/61b10985-8ed4-412c-9353-1889f200778f. One, parameters.zip, contains the output topology and parameter files in CHARMM format, suitable for use in constructing simulation systems. Another, crystal.zip, contains the directory and minimal output related to building and simulating crystalline lignin. The remaining two hold the logic of the parameterization, featuring scripts that do the optimization (including the source for the GPU-accelerated objective function evaluation) and target data creation (scripts.zip), as well as minimal outputs needed to recreate the work (datafiles.zip). Due to licensing restrictions, Gaussian log files that created the target data are not provided here. Instead, only binary-formatted parsed target data as well as the input decks used to create the target data are provided in datafiles.zip.

Introduction

In addition to the typical ancillary tables and figures, the Supporting Information contains significant discussion related to the different optimization protocols tried, and how we select from among the different resulting parameter sets generated to determine the optimum.

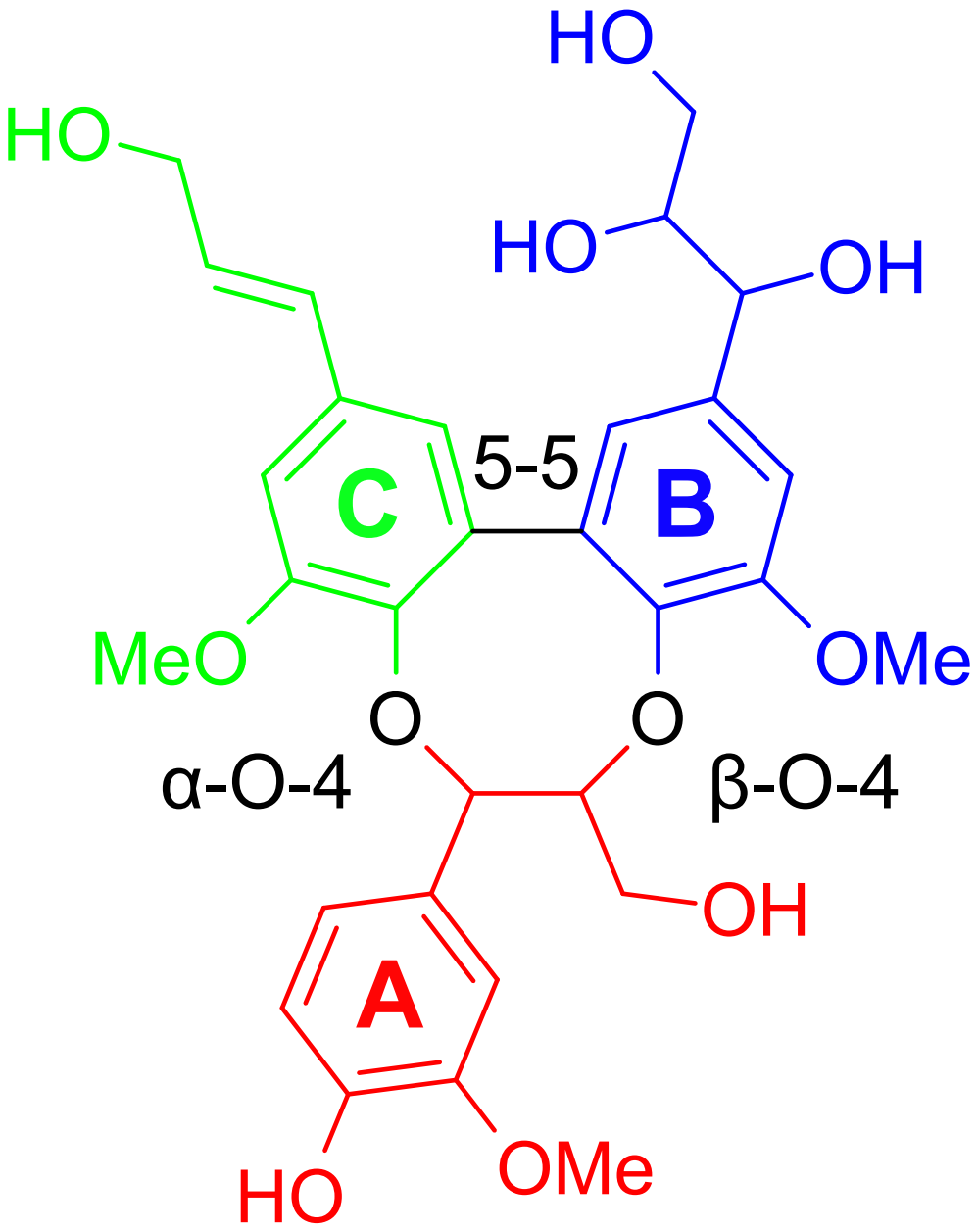

Figure S1: A simple example of a lignin trimer demonstrating the combination possibilities enabled by the forcefield. Dibenzodioxocin-like structures have been isolated from native lignins. In this example, monomer A (red), B (blue) & C (green) are all G-type lignin. The original α and β alcohols from the base monomer A are used to specify the stereochemistry of the α-O-4 and β-O-4 linkages with the C & B monomers, respectively. These alcohols are removed entirely in monomer C by adding a double bond through the typical CHARMM "patching" process. Similarly, the three linkages shown (black) are applied as patches applied to form the linkages of the trimer.

Extended Parameterization Theory

General CHARMM Parameterization Scheme

CHARMM has a two-stage approach to parameterization,[S1,S2] following the demarcation in Eq. 1 between bonded an non-bonded terms. In the first stage, non-bonded parameters alone are adjusted. Typically, the Lennard-Jones parameters (red boxed terms in Eq. 1) are taken by analogy from other compounds with a similar structure,[S1,S2] as high-quality target data are not readily available, and it has been shown to only minimally change the outcome of observables such as solvation free energy.[S3] What sets CHARMM apart from other force fields is how it handles the atomic charge assignment, representing the red circles within Eq. 1. Rather than fitting the charges to the results of a restrained electrostatic potential (RESP) fit, as is done for the AMBER,[S4] GROMOS,[S5] and OPLS[S6] force fields, in CHARMM, quantum calculations of the interaction energy between TIP3 water[S7] and the newly parameterized compound and the quantum electrical dipole moment, are used as the target data to determine the atomic charge distribution across each molecule.[S1,S2]

The charges obtained in the first stage are then used as part of the input for parameterization in the second stage to determine the bonded parameters. For describing the bond and angle terms, in principle a single Hessian calculation can be used as input, with the bond and angle force constants determined from a scaled vibrational analysis.[S1,S8,S9] However, given the increased availability of computing power, relaxed energy scans, where one degree of freedom is systematically changed and the remainder of the molecule is allowed to relax, become tractable for all the required bond and angle terms, and can alternatively be used to determine these parameters.[S10] This approach mirrors what is done for dihedral torsions, where relaxed potential energy scans are used to determine the dihedral parameters that best fit the underlying potential energy surface.[S1,S2,S9] Together, these two successive steps determine all of the circled parameters from Eq. 1. When combined with the free parameters with squares around them in Eq. 1, this parameterization approach creates a complete descrip-

tion of the forces between individual atoms within the molecules of interest. To maximize compatibility with other CHARMM parameter sets for carbohydrates[S11–S14] or proteins,[S15] following the whole parameterization scheme, including the water interaction-based charge determination, is required.

Additional Considerations for Branched Polymers

For small, discrete molecules, many tools exist to assist in the parameterization process, like the force field toolkit (ffTK),[S1] ForceBalance,[S16] the Visual Force Field Derivation Toolkit (VFFDT),[S17] ForceFit,[S18] or the general automated atomic model parameterization (GAAMP).[S19] However, there are specific required features that arise in parameterizing branched polymers, and it is worth reviewing what those features are before further explaining why new software was required. Chiefly, a mechanism like the atom-typer in CGenFF[S20] is required to recognize equivalent chemical environments for an atom. In lignin, there are many structural elements (such a methoxy group) that repeat in many molecules. Conceptually, the charges for these repeated structural elements should be consistent, otherwise the parameter set is overly specific to the tested geometries and not broadly transferable. In principle a single lignin monomer can be connected to up to five other monomers (e.g. participating in both an α-O-4 and a β-O-4 linkage off of the α and β carbons of the C1 branch, a γ ester, and a 5-5 and β-O-4 linkage on C5 and C4). This means that the charge changes induced by these linkages should be local enough so as to be largely independent of other surrounding linkages. If this is not the case, the charge distributions would need to be separately parameterized for all possible combinations of monomers, a task that is currently intractable due to the combinatorial explosion of possible lignin linkage permutations, even granting that only a subset of these permutations are found in native lignins.

Furthermore, existing publicly available tools do not easily integrate target data from multiple molecules into a single objective function for simplified simultaneous optimization of the parameters in a CHARMM-compatible scheme. This is a critical feature for parame-

terizing lignin, since the same parameter set must simultaneously describe each linkage type, ideally without creating overfitting artifacts that could arise from creating many new atom types for each compound. An additional feature required by the aromatic lignin monomers is that the equilibrium angle parameters around each of the aromatic carbons should sum to 360 °. If the sum is not exactly 360, a persistent bias develops towards puckering the aromatic ring, which can cause distorted geometries during simulation away from a typical planar aromatic ring. The combination of these features is not found in another CHARMM-compatible tool similar to how ForceBalance can be used for Amber,[S21] and demands custom code to incorporate these features into our parameterization workflow, which we describe in detail within the methods.

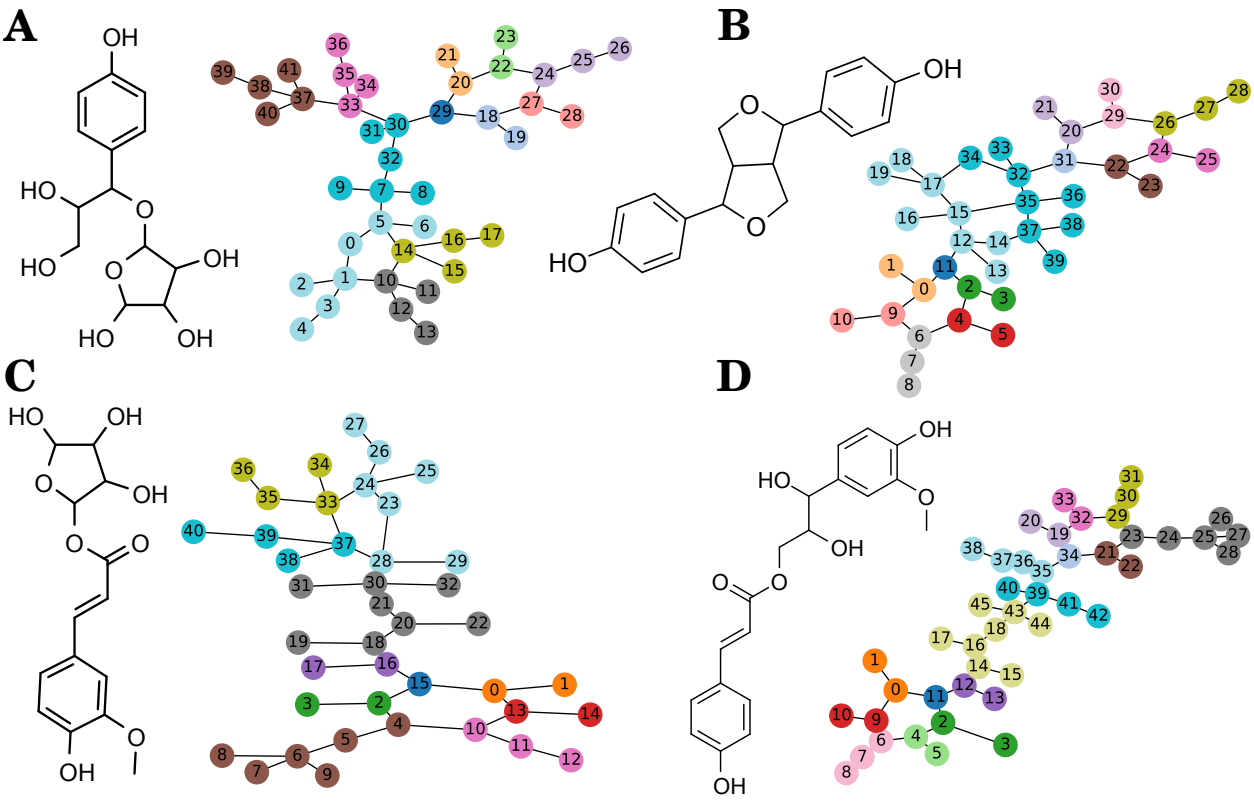

Figure S2: Examples of how structural elements are divided into near-integer charge groups of atoms for four non-trivial examples, with the chemical structure shown on the left and the graph representation of the charge groups on the right. As in Fig. 2 previously, each charge group is assigned its own unique color within the molecular graph, and the numbers labeling each node of the graph indicate the atomic index of the assembled molecular structure. The specific molecules chosen here represent dimeric structures testing different linkers. (A) An α-6 linkage from lignin to sugar. (B) A β-β linkage between lignins. (C) A ferulate-xylan linker. (D) A γ-O-γ linker between lignins. The algorithm consistently picks out similar chemical functionalities across the diverse molecules tested, such as the groups shared between the linkages in (A) and (C) or (C) and (D).

Extended Optimization Methods

Charge Optimization

To reoptimize charges in CHARMM, three different elements need to be considered; the interaction energy (E^{int}) between water and the compound at the quantum optimum geometry after scaling and shifting the value to better represent liquid conditions, the distance between

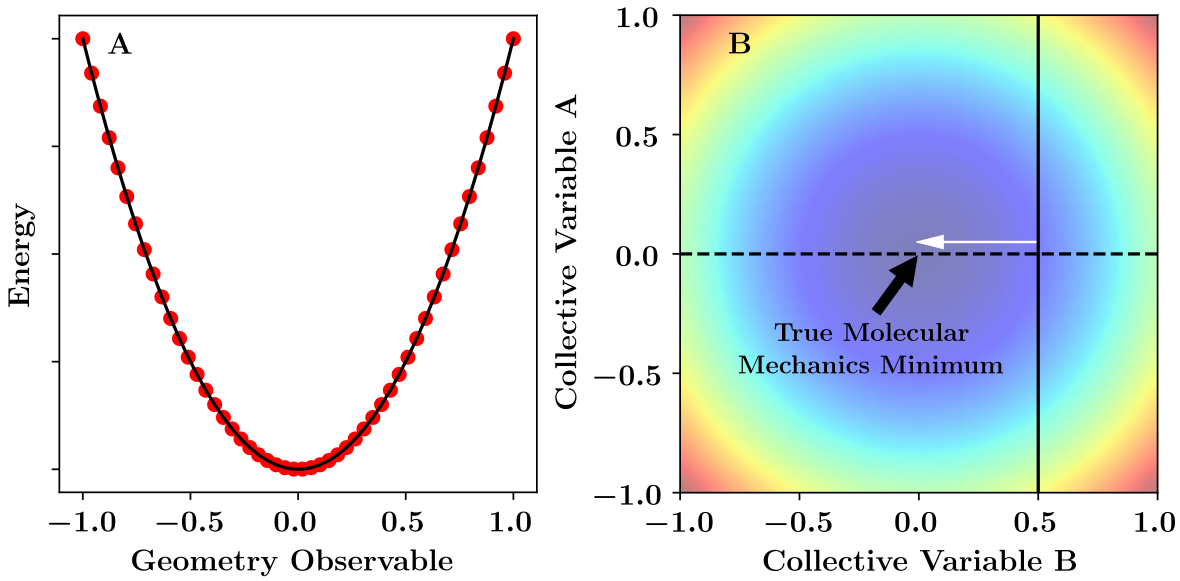

Figure S3: Example potential energy surface dilemma. Suppose, as in (A), that our observed quantum potential energy surface (red points) is perfectly fit by a series of parameters that result in a harmonic potential energy surface along that particular geometrical perturbation or collective variable. When viewed in a multidimensional space (B), this 1D optimized potential energy surface can either overlay on the minimum of the local classical potential energy surface (horizontal dashed line, where the geometry observable is collective variable B), or be offset somewhat (vertical solid line, where the geometry observable is collective variable A). The offset case represents an instance where a classical molecular dynamics minimizer will change the geometry away from the starting geometry, and bring the structure to the local minimum of its own multidimensional potential energy surface by changing a degree of freedom orthogonal to the scanned potential energy surface, as represented by the white arrow.

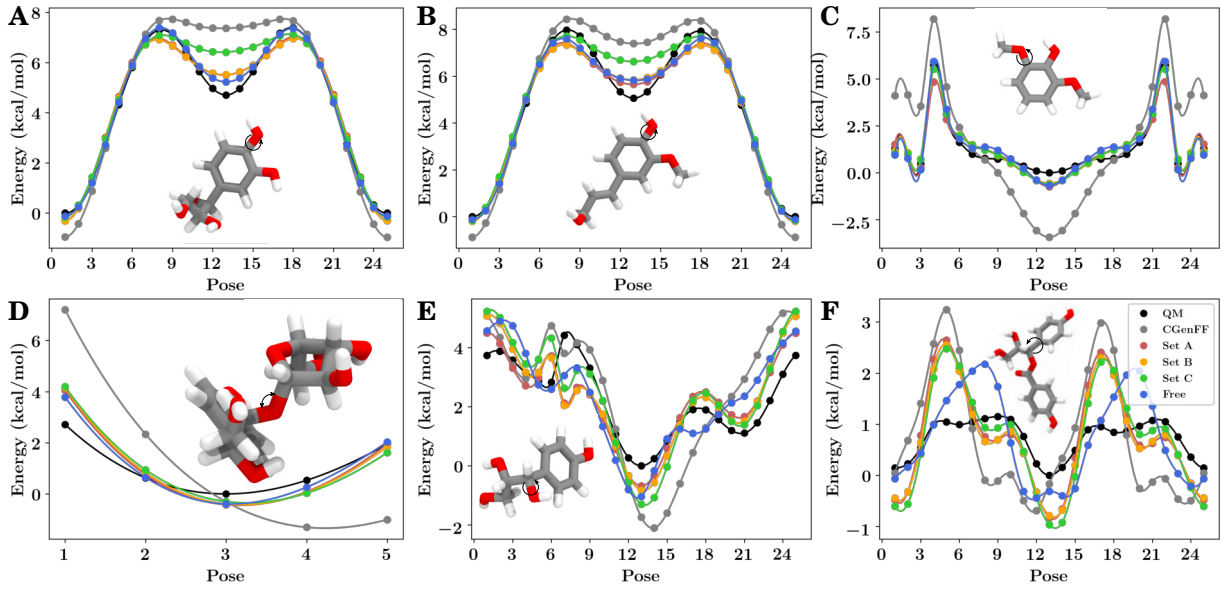

Figure S4: Examples of the quantum mechanical and classical potential energy surface for a limited subset of the 2574 bond, angle or dihedral scans used as target data. To simplify the visualization, $v = 0$ during the optimizations that lead to this set of parameters. The bounds enforced on the dihedral terms during the optimization for each set are listed in Table S1. Each set is drawn in a consistent color, as indicated in the in-figure legend, with the quantum mechanical target data from the scan shown in black. A molecular image of the compound being scanned in its central pose can be found within each panel, with a black arrow indicating what degree of freedom is being probed by the scan. (A) and (B) highlight the energy change when an unbalanced hydroxyl group rotates, (C) shows a typical methoxy rotation, (D) demonstrates an angular change between a lignin and sugar monomer, (E) shows an α-hydroxyl rotation, and (F) shows a rotation around an ester-adjacent bond.

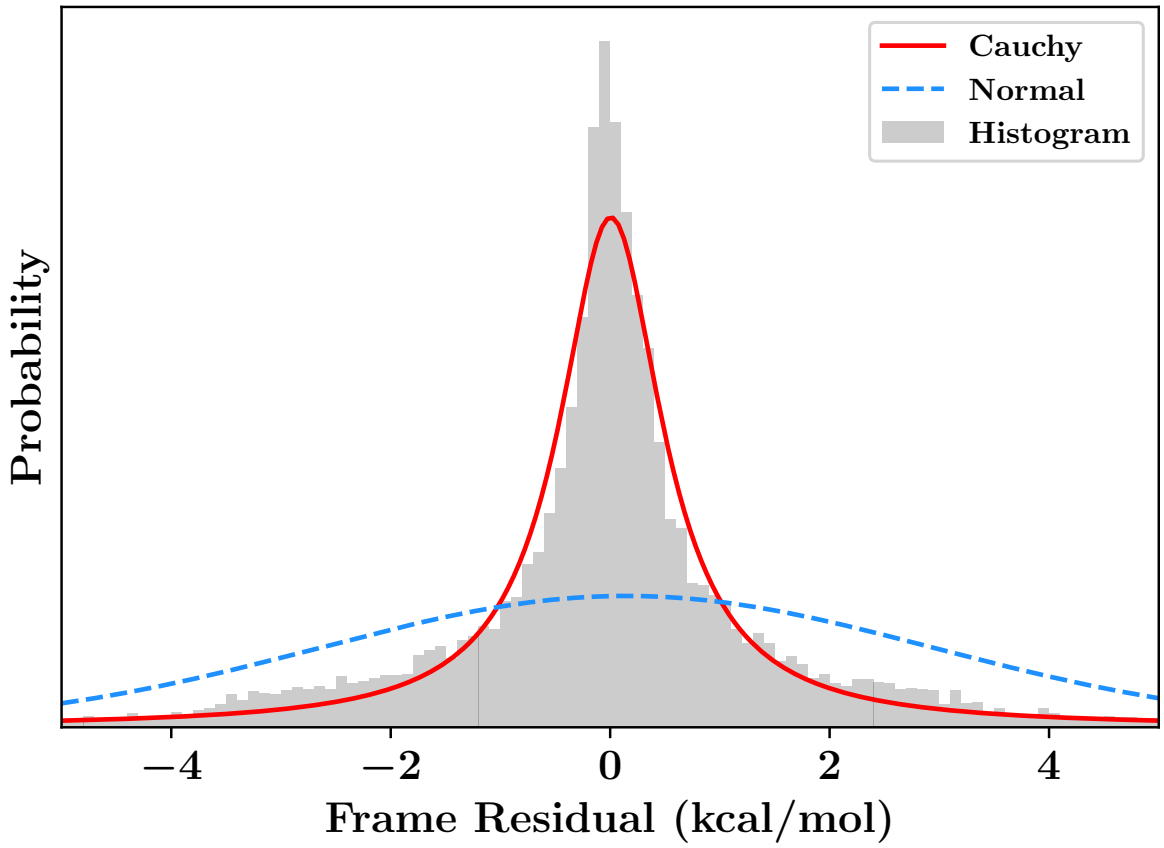

Figure S5: The actual distribution of residuals (grey histogram) is not fit by a normal distribution (blue dashed line), as the width is not related to the standard deviation of the distribution. Instead, a cauchy distribution centered around zero (red solid line) describes the distribution much better.

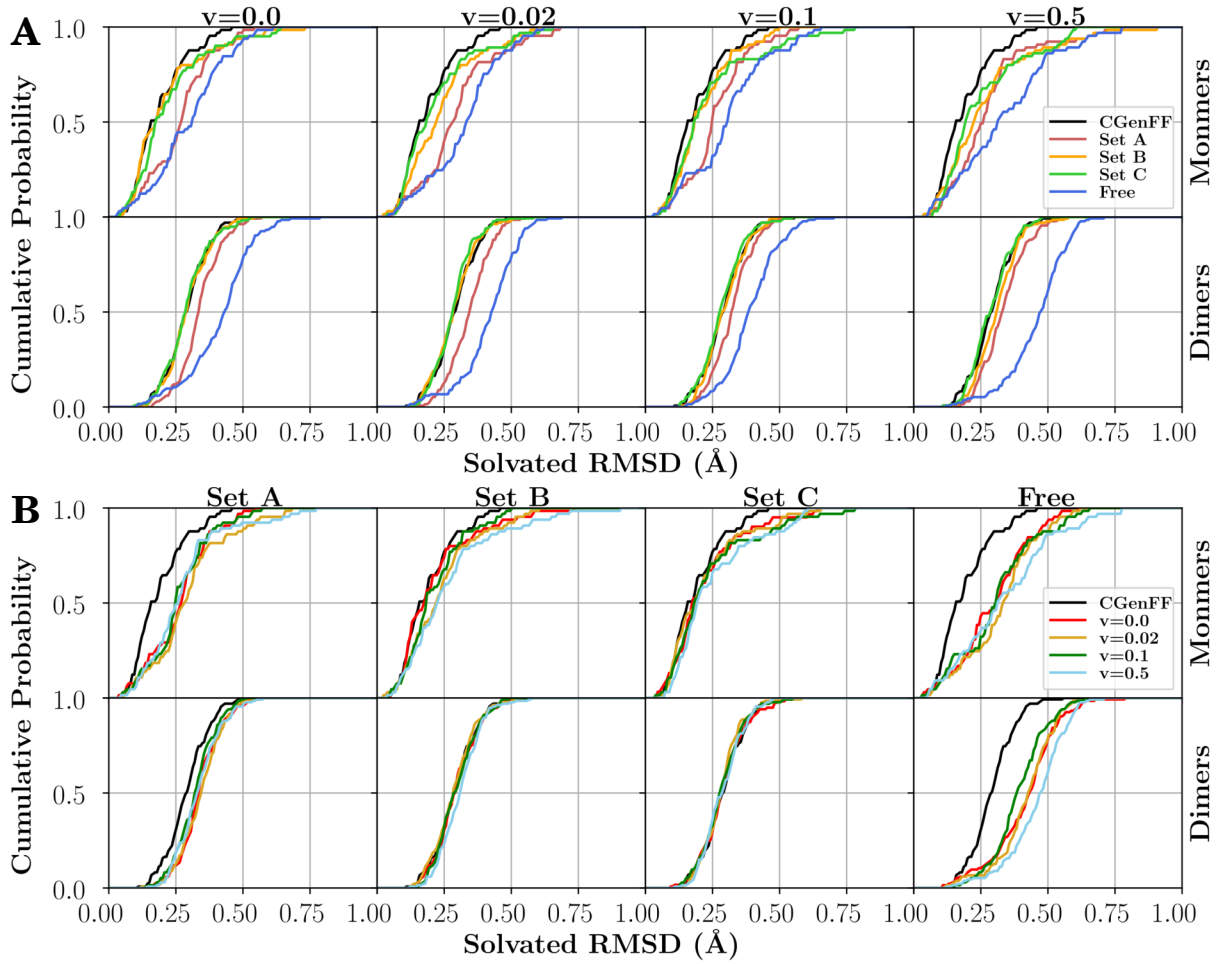

Figure S6: Root mean square deviation (RMSD) distribution of the resulting structures for both monomers and dimers after minimization in solution relative to the gas-phase minimum energy structures determined quantum mechanically. As in Fig. S32, the parameter v from Eq. S2 is held fixed in (A), highlighting the effect of different choices for dihedral bounds, with the effect of force inclusive optimization demonstrated in (B). Within each subpanel, the results from the CGenFF starting point are also shown in black as a benchmark. Monomers and dimers are separated here due to their disparate sizes, which tends to make the RMSD for dimers larger overall. The mean and standard deviations are reported in Table S6. For a similar analysis in vacuum, see Fig. S33 and Table S8.

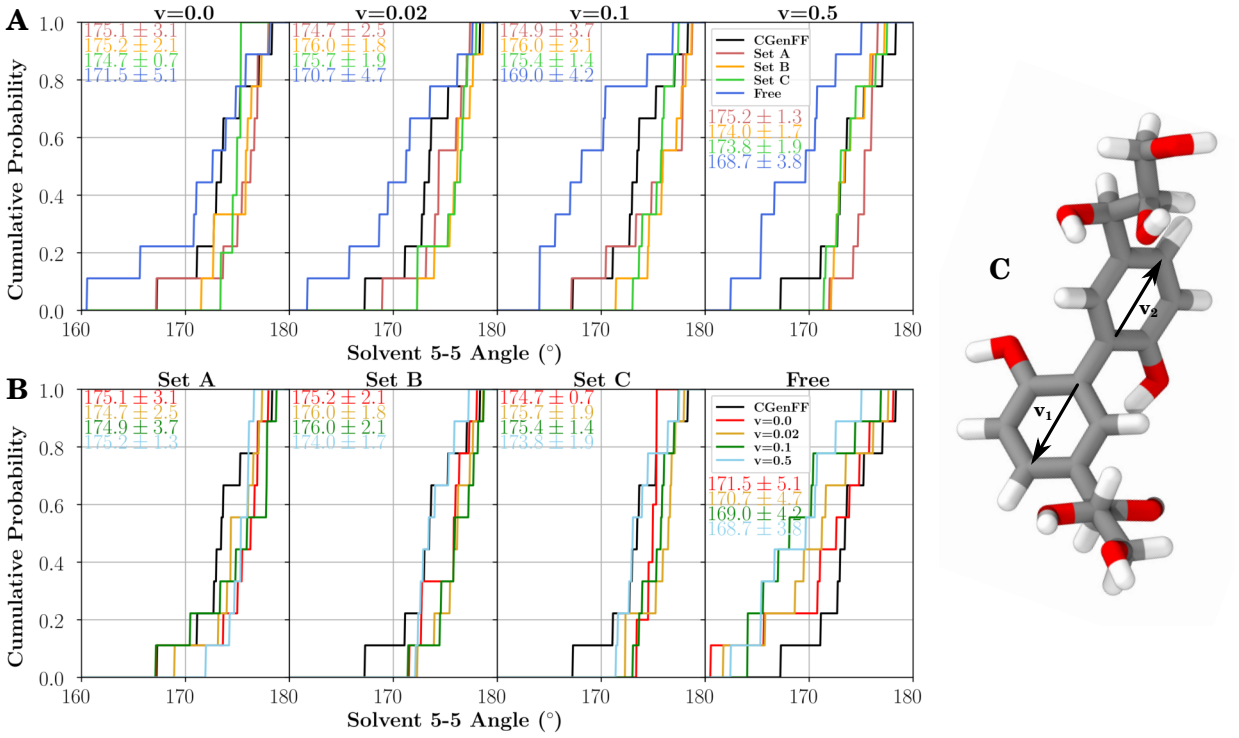

Figure S7: Distribution of the angle between vectors in lignin dimers linked through a 5-5 linkage after minimization in solution. The angle itself is defined through the vectors shown in (C), which go from C5 to C2 of each monomer involved in the linkage. The angle between these vectors is the measured quantity. At the quantum level, this angle is almost exactly 180°; however, relatively large deviations from the perfect line between the vectors were noted during minimization. The distributions of the observed angles are reported in panels (A) and (B). The parameter v from Eq. S2 is held fixed in (A), highlighting the effect of different choices for dihedral bounds, with the effect of force inclusive optimization demonstrated in (B). Mean values and their standard deviation over the distribution are reported in-figure, using the same color as those given in the in-figure legend.

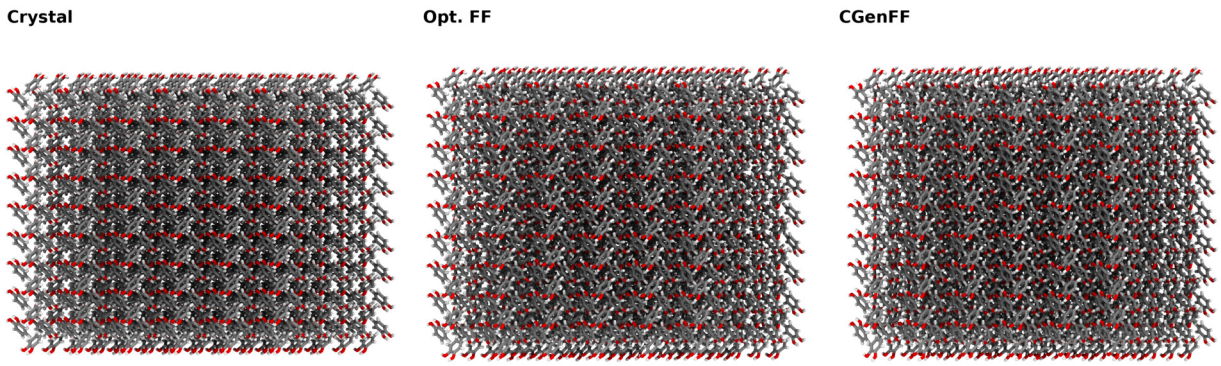

Figure S8: Crystal comparison for catechol (CATCOL13) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.977 ± 0.008) and CGenFF (right, mean RMSD 1.336 ± 0.008). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

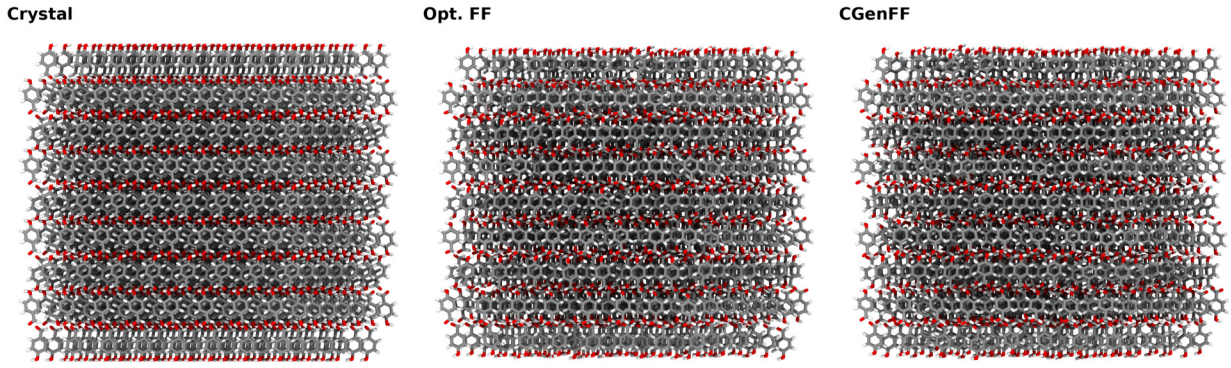

Figure S9: Crystal comparison for *p*-hydroxybenzaldehyde (PHBALD11) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.09 ± 0.02) and CGenFF (right, mean RMSD 1.26 ± 0.02). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

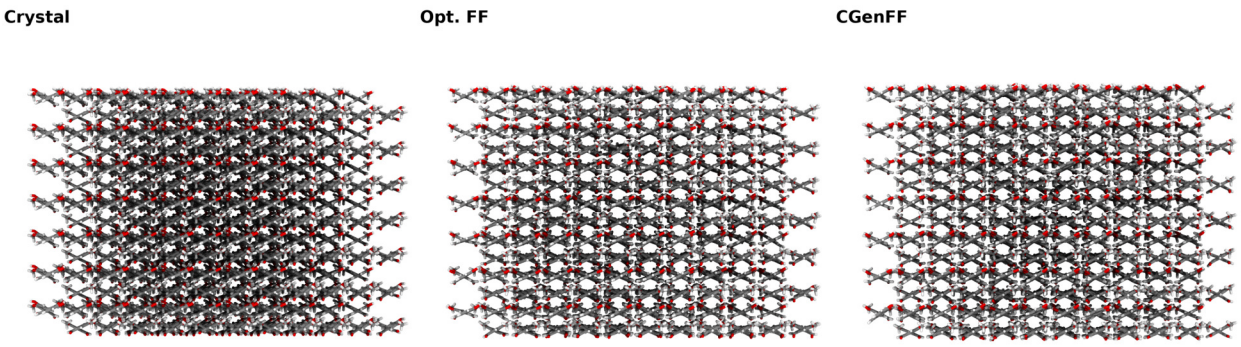

Figure S10: Crystal comparison for vanillin (YUHTEA01) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.331 ± 0.008) and CGenFF (right, mean RMSD 1.426 ± 0.009). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

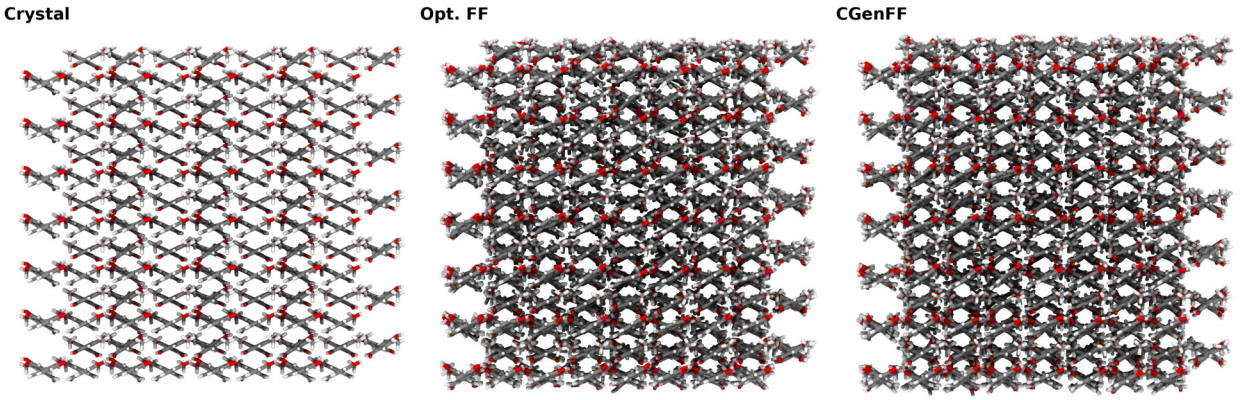

Figure S11: Crystal comparison for vanillin (YUHTEA03) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.52 ± 0.03) and CGenFF (right, mean RMSD 1.61 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

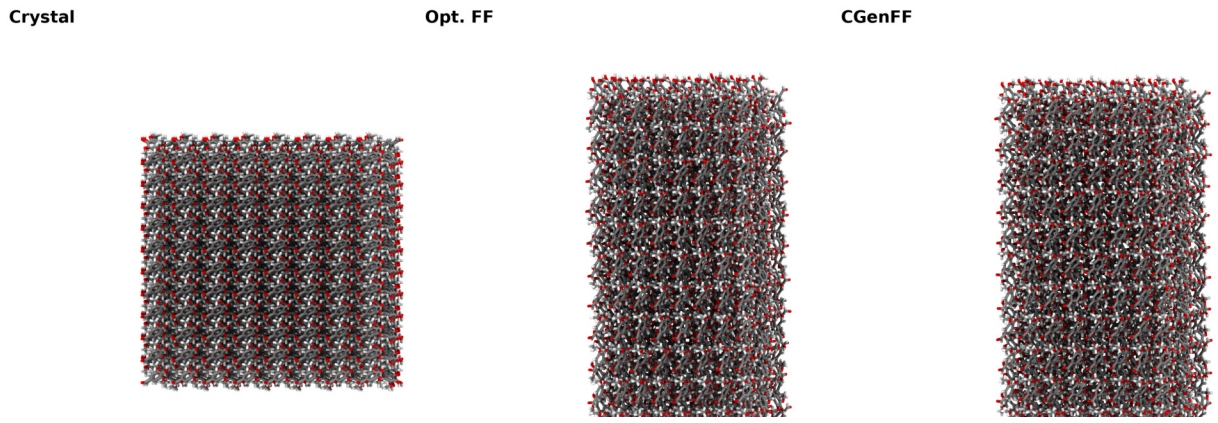

Figure S12: Crystal comparison for syringaldehyde (IZALAW) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 8.92 ± 0.01) and CGenFF (right, mean RMSD 8.40 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

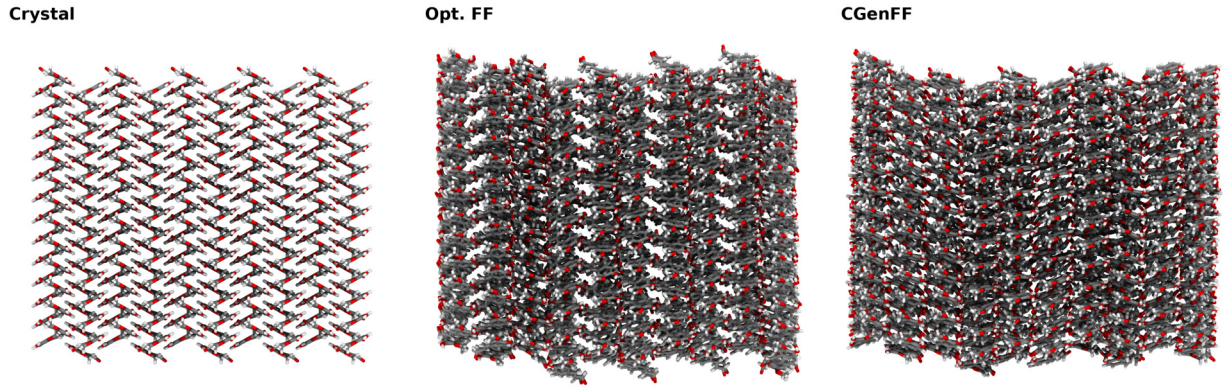

Figure S13: Crystal comparison for coniferaldehyde (SIPKEH) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 3.16 ± 0.02) and CGenFF (right, mean RMSD 3.14 ± 0.02). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

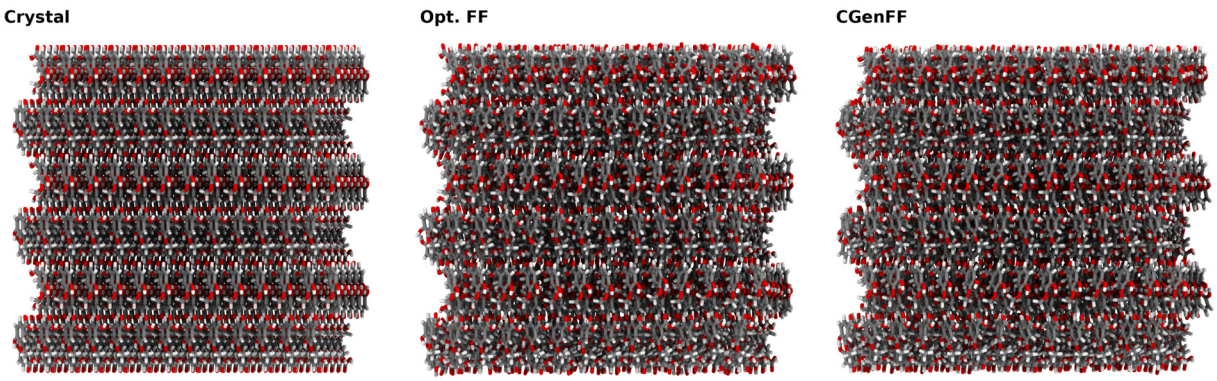

Figure S14: Crystal comparison for vanillic acid (CEHGUS) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.01 ± 0.01) and CGenFF (right, mean RMSD 1.06 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

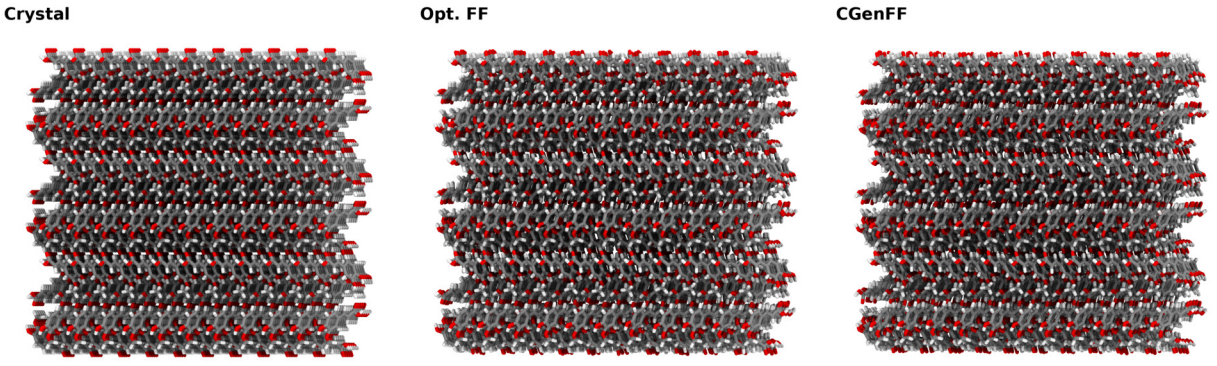

Figure S15: Crystal comparison for ferulic acid (GASVOL01) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.837 ± 0.006) and CGenFF (right, mean RMSD 1.352 ± 0.007). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

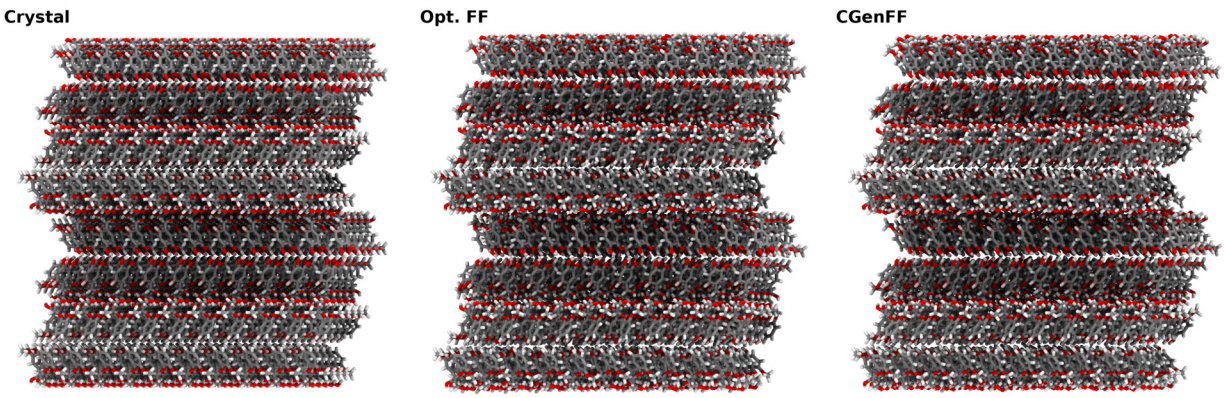

Figure S16: Crystal comparison for G-βO4-G (RABWUM) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.592 ± 0.007) and CGenFF (right, mean RMSD 0.631 ± 0.006). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

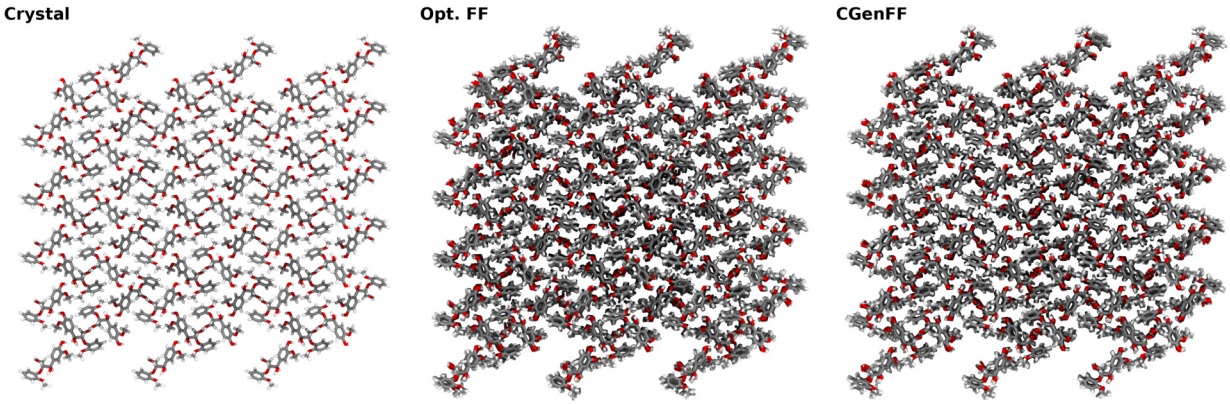

Figure S17: Crystal comparison for G-βO4-G (SIPPEM) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.93 ± 0.01) and CGenFF (right, mean RMSD 1.100 ± 0.009). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

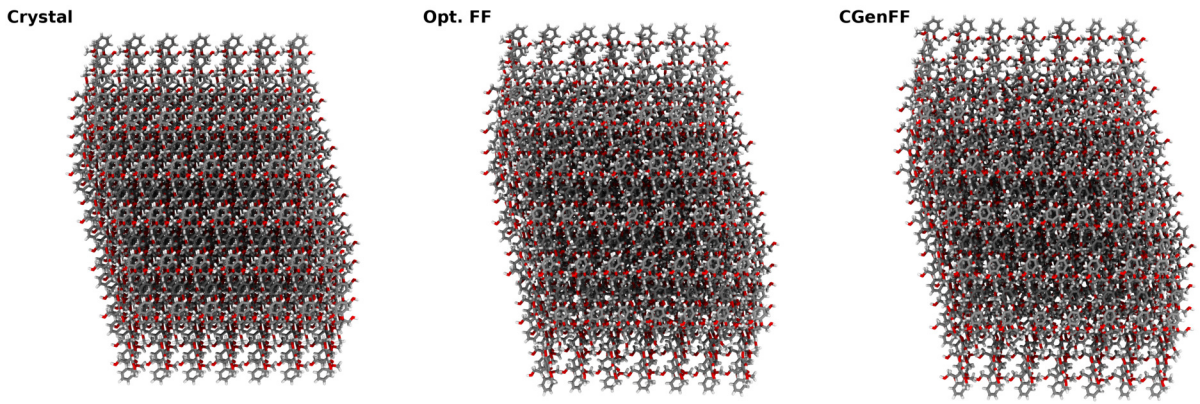

Figure S18: Crystal comparison for S-βO4-G (VADDOT) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.15 ± 0.01) and CGenFF (right, mean RMSD 1.72 ± 0.02). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

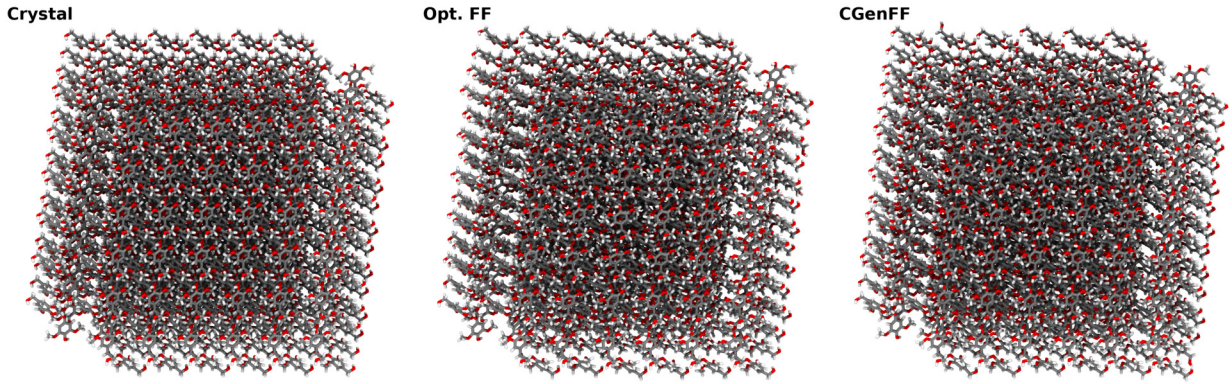

Figure S19: Crystal comparison for S-βO4-S (SAZHEG) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.95 ± 0.01) and CGenFF (right, mean RMSD 0.84 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

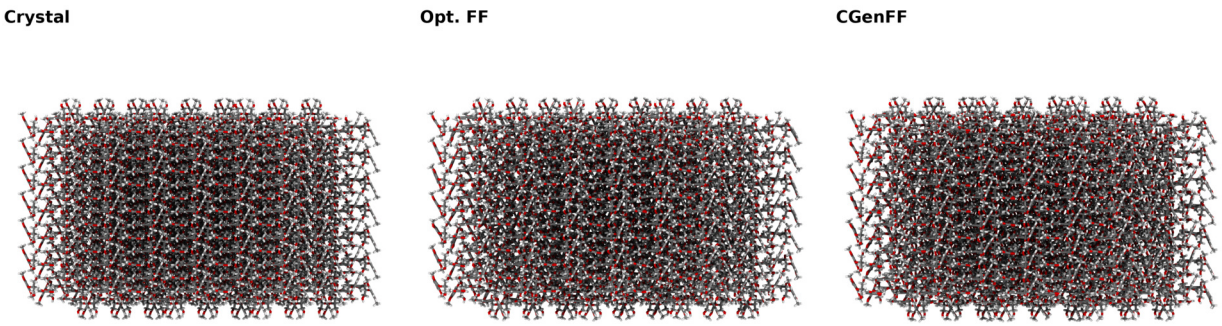

Figure S20: Crystal comparison for S-βO4-S (FOCGUA) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.917 ± 0.007) and CGenFF (right, mean RMSD 0.84 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

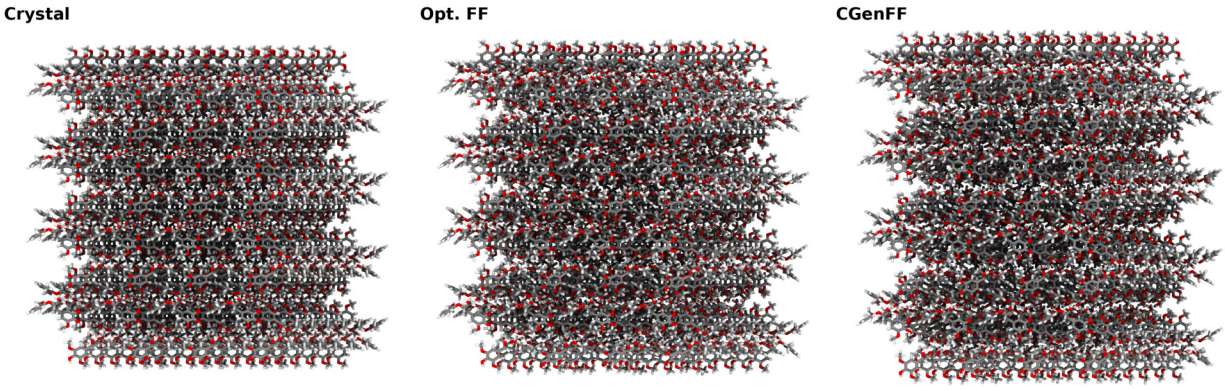

Figure S21: Crystal comparison for S-βO4-S (IDIKIP) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.84 ± 0.01) and CGenFF (right, mean RMSD 1.83 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

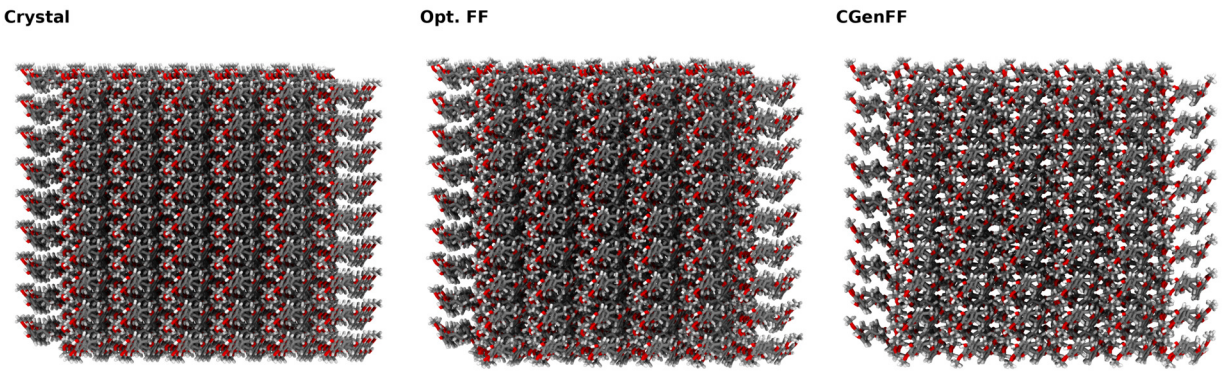

Figure S22: Crystal comparison for G-$\beta\beta$-G (INELIW) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.786 ± 0.009) and CGenFF (right, mean RMSD 1.42 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

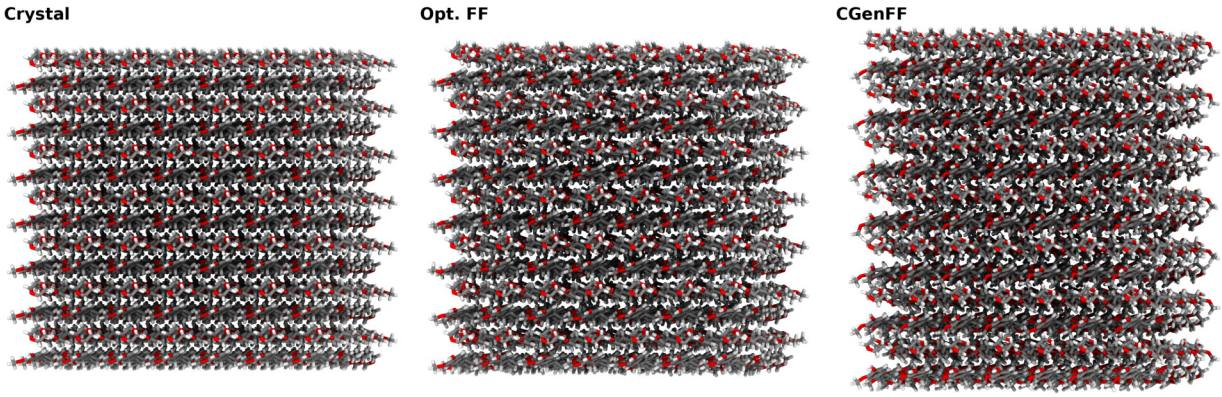

Figure S23: Crystal comparison for G-$\beta\beta$-G (INELIW01) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.82 ± 0.02) and CGenFF (right, mean RMSD 2.322 ± 0.008). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

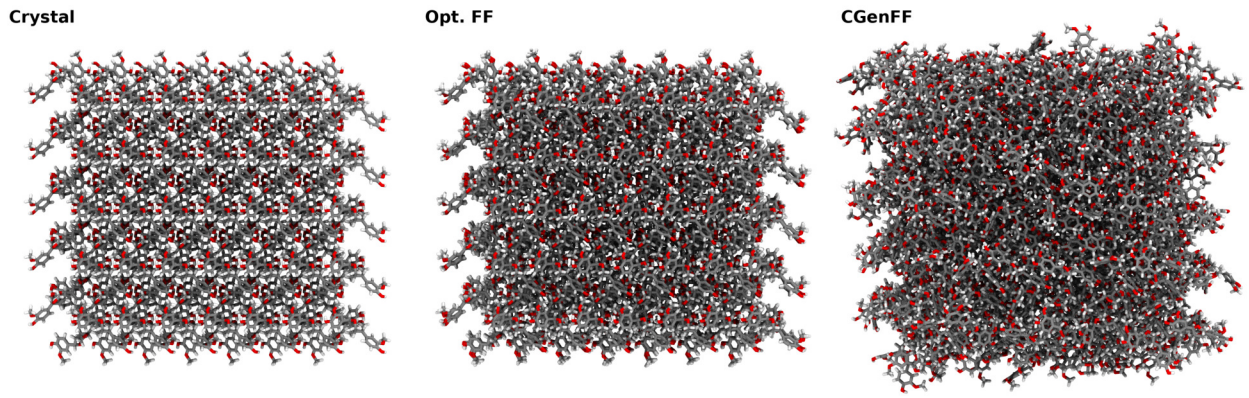

Figure S24: Crystal comparison for G-$\beta\beta$-G (FAFXUF) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.99 ± 0.01) and CGenFF (right, mean RMSD 3.97 ± 0.02). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

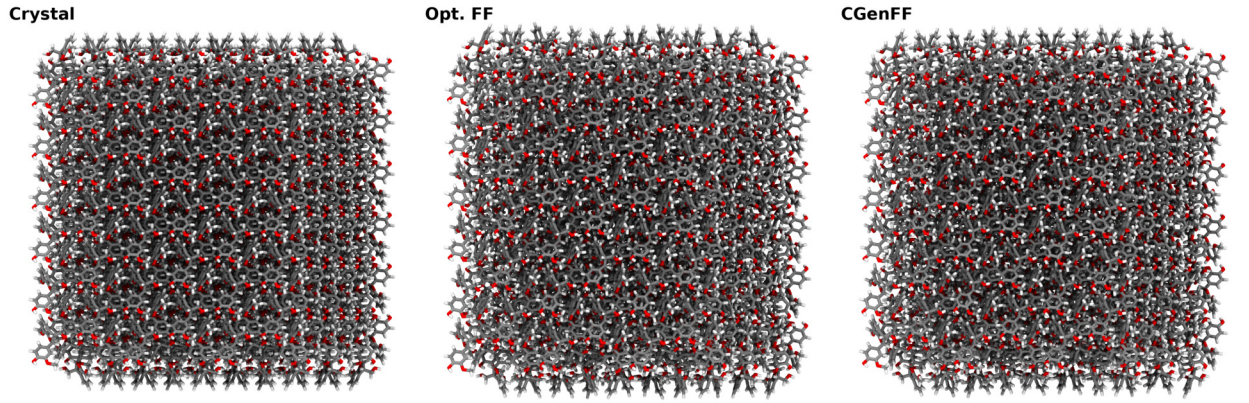

Figure S25: Crystal comparison for G-β5-G (FUMVUE) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 1.01 ± 0.01) and CGenFF (right, mean RMSD 1.07 ± 0.01). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

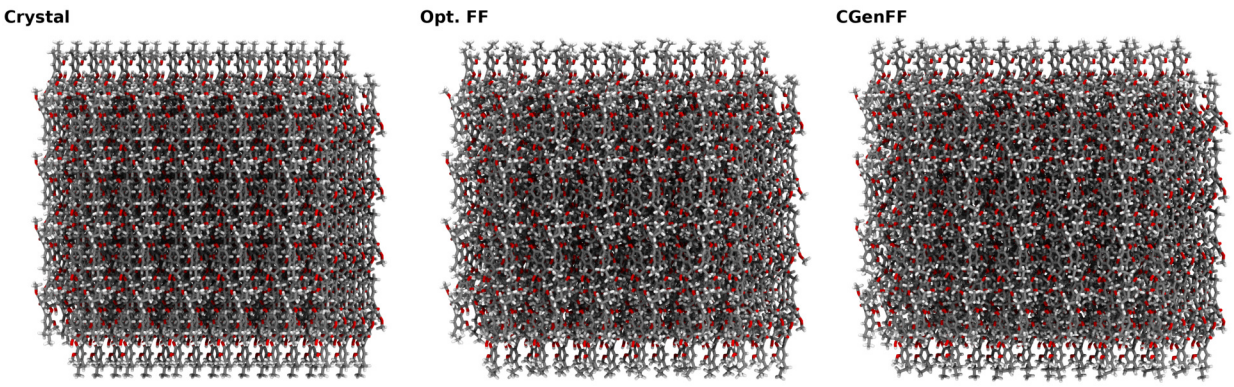

Figure S26: Crystal comparison for dibenzodioxocin (TUGWAT) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 0.88 ± 0.02) and CGenFF (right, mean RMSD 1.23 ± 0.02). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

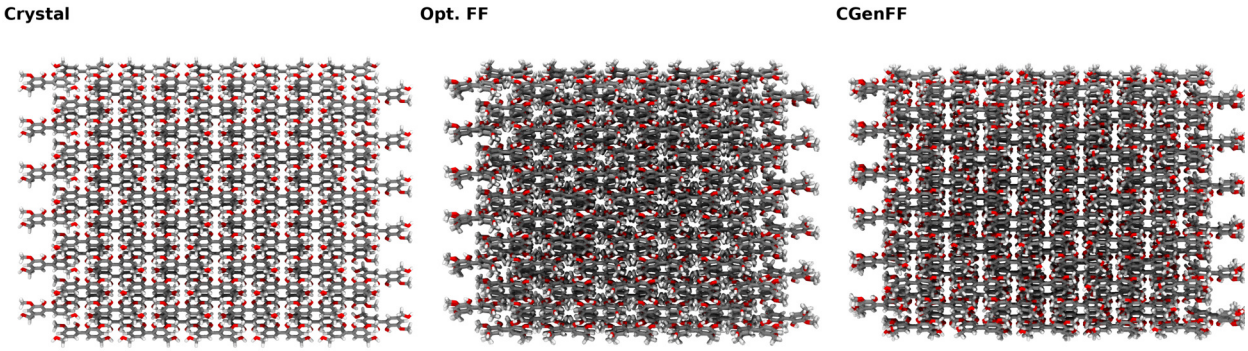

Figure S27: Crystal comparison for G-55-G (UJOGIK) between the original crystal (left), and the eventual structure after 20 ns of simulation with the lignin-optimized force field (middle, mean RMSD 2.521 ± 0.004) and CGenFF (right, mean RMSD 1.77 ± 0.04). Each molecule is shown in a stick representation, with carbons shown in gray, oxygens in red, and hydrogens in white.

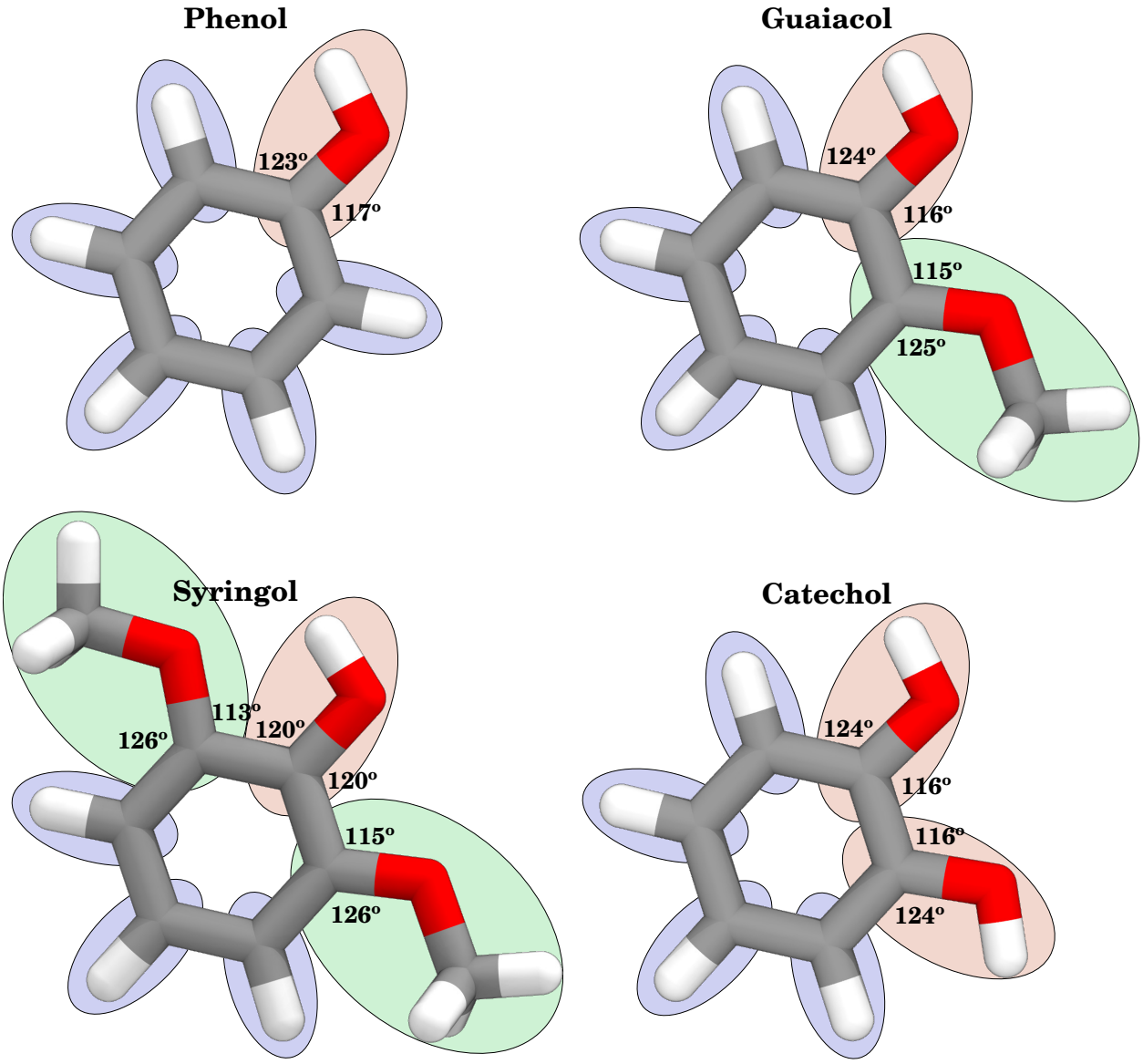

Figure S28: Angles (rounded to the nearest integer) around aromatic carbons bonded to oxygen in quantum mechanically optimized structures. Each monomer parameterized is shown after truncating the group attached to C1 (Fig. 1), with the common name for each resulting compound shown above the molecular structure. Within the molecular representation, carbons are grey, oxygens are red, and hydrogens are white. The oval underlays represent the near-integer charge groups for these compounds, first laid out in Fig. 2. There are clear differences between angles depending on surrounding functional groups, indicating that a single atom type for all aromatic carbons, as CGenFF initially created, cannot reproduce the diversity of angle values seen here.

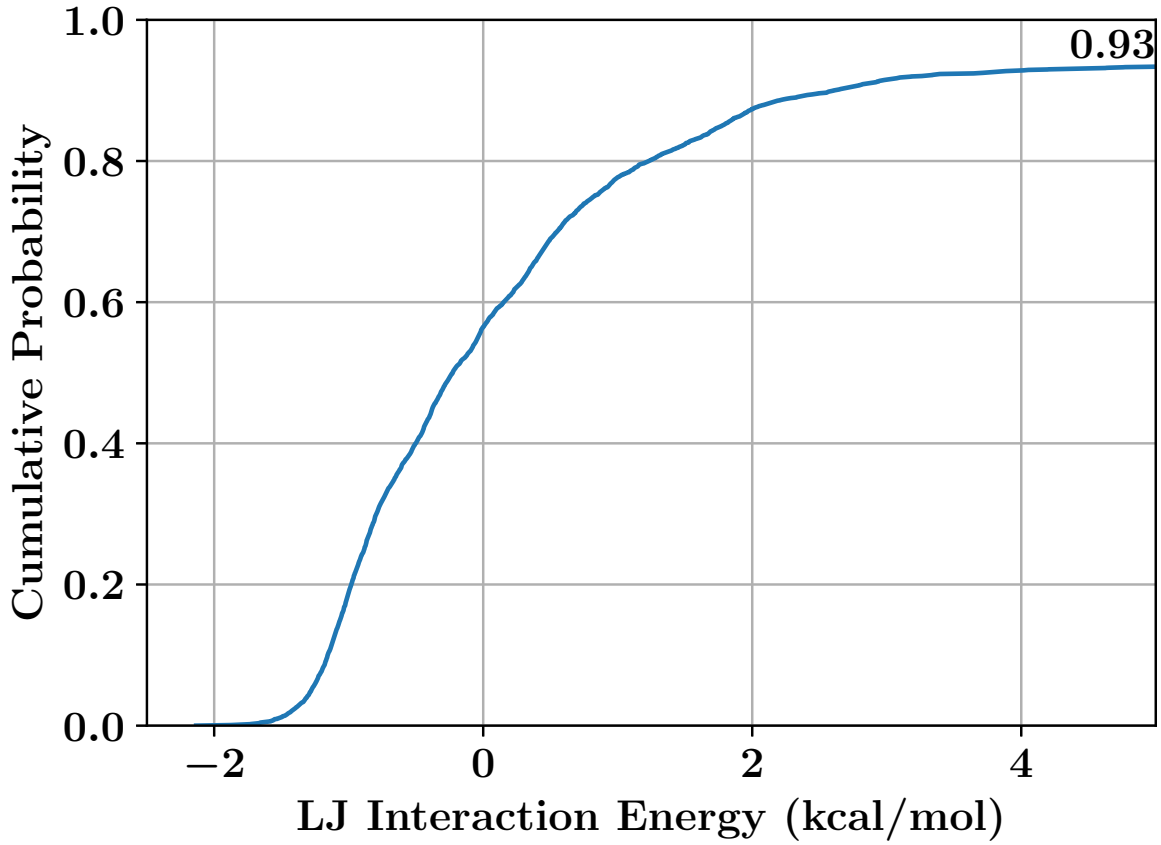

Figure S29: Distribution of the Lennard Jones (LJ, equivalent to U_{VDW} from Eq. 1) contributions to the water interaction energy. Since CHARMM requires that water distances be shifted inward to better match solution conditions based on gas phase quantum calculations, sometimes the LJ contribution at this new geometry can be very large, with LJ interactions in the hundreds of kcal mol^{-1} observed. If these large LJ contributors were not eliminated, they would dominate the objective function, as their individual residuals from the quantum-determined interaction energy would be very large. Thus, we use the 5 kcal mol^{-1} threshold for exclusion from the optimization target data, which eliminates less than 7% of the initial target data set, as indicated by the cumulative sum at the cutoff reported in the upper right.

water and the compound at the minimum (d), and the compound dipole moment (D).[S1,S2] Each of these components is incorporated into an objective function that is minimized to yield an optimized charge distribution such that the new molecular mechanics (MM) model matches the values observed from the quantum mechanical target data (QM).

$$f(\bar{q}) = \sum_{compounds} \sum_{stereoisomers} \left[\left(\sum_{sites} w^{-2} \left(E^{int}_{QM} - E^{int}_{MM}(\bar{q}) \right)^2 \right. \right. \\ \left. \left. + \, w_d^{-2} \left(d_{QM} - d_{MM}(\bar{q}) \right)^2 \right) + w_D \left(D_{QM} - D_{MM}(\bar{q}) \right)^2 \right] \quad \text{(S1)}$$

It should be emphasized that Eq. S1 is a restatement of the objective function used by default in ffTK ($w = 0.2\,\mathrm{kcal\,mol^{-1}}$, $w_d = 0.1\,\text{Å}$, w_D proportional to the number of atoms in the compound),[S1] with the exception that Eq. S1 explicitly incorporates target data from multiple compounds with multiple chiralities and discards interaction energies that are unphysically high after shifting and scaling. In keeping with standard CHARMM methodology, E^{int}_{QM} is scaled up by 1.16 for uncharged compounds from the computed value from Gaussian and d_{QM} is shifted inward by 0.2 Å to better reproduce liquid phase properties from gas phase calculations.[S1,S2] Occasionally, the inward shift greatly increases the non-bonded Lennard-Jones contribution to the energy (U_{VDW}), which makes fitting the quantum mechanical interaction impossible, as no possible combination of charges would be able to counterbalance the high U_{VDW} that is observed for a subset of the interactions (Fig. S29). To ameliorate this, both w and w_d are additionally multiplied by 0 if $U_{VDW} > 5\,\mathrm{kcal\,mol^{-1}}$, and by $\min\left(1, \exp\left(-0.2E^{int}_{QM}\right)\right)^{-1}$ otherwise, which weights repulsive interactions more weakly than attractive interactions, and completely discounts contributions from interactions that are too strongly repulsive after the shift is applied. These highly repulsive interactions account for 7 % of the total interactions observed (Fig. S29). The target magnitude of the dipole moment was chosen such that the molecular dipole (D_{MM}) is between 1.2 and 1.5 times the magnitude of the quantum dipole (D_{QM}), and that harmonic penalties were not

applied if the direction of D_{MM} was within 20° of D_{QM}, consistent with standard CHARMM methodology.[S1,S2]

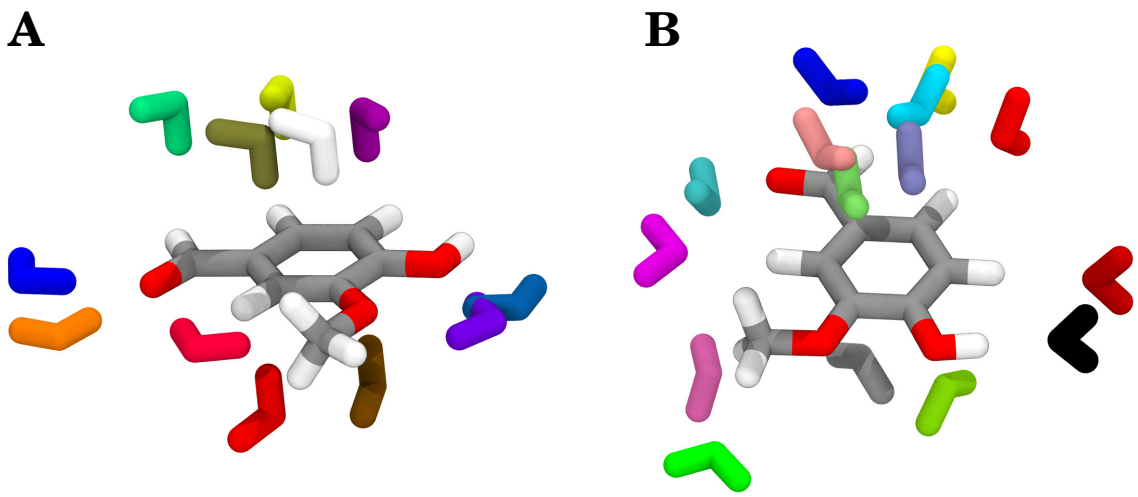

Figure S30: Examples of the optimized positions of water relative to a test compound, in this case the combination of the G-monomer with an aldehyde substituent on C1. In (A), water is acting as a proton donor to accessible sites on the test compound, whereas in (B), water is acting as a proton acceptor relative to potential proton donor sites on the compound. Each calculation is of the target molecule with a single water molecule, and are carried out separately. However here all water molecule positions are shown simultaneously with a distinct solid color.

The target data fed into this objective function are created through the same process as in ffTK,[S1] using the MP2-optimized geometries computed earlier as the basis for calculation. To determine the interactions with water, water molecules are algorithmically placed as either potential hydrogen bond donors or acceptors (Fig. S30), and then optimized at an HF/6-31G* level of theory[S22] to determine the optimal distance for interaction between the single water and the target compound. This required 4836 individual optimizations in Gaussian 09.[S23] The dipole is determined at the MP2/6-31G* level of theory[S24] from the optimized geometry computed previously.

With the quantum mechanical target data needed for Eq. S1 in hand, it remains to determine the exact charge vector $\bar{q}$ that optimizes the objective function. However, unrestrained

optimization can result in aphysical results (such as positive partial charges for oxygens) if not carefully bounded. The bounds chosen restrict hydrogens to their typical CHARMM charges of 0.09 for non-polar hydrogens, 0.115 for aromatic hydrogens, 0.15 for sp2 hydrogens, and 0.42 for polar hydrogens, whose units are with respect to the charge of a proton. Heavy atoms were allowed to change their charges by ± 0.25 charge units relative to the CGenFF charges.

Equivalent atoms on different compounds should have equivalent charges, which must be imposed within the optimizer. Since no specific prescription exists for determining equivalent charges, we tried two different schemes (Fig. 2). In one scheme, a neighborhood is determined for each atom of every compound, and atoms with identical neighborhoods are grouped together and are forced to carry the same charge (Fig. 2). A neighborhood in this instance is defined to be a subgraph representation of the molecule centered around a specific atom, where the nodes are the atoms labeled by atom type that are within 1, 2, or 3 bonds of the original atom, and the edges represent the bonded topology of the molecule within that subset of atoms. Neighborhoods can then be grouped together by checking whether the subgraphs are isomorphic relative to one another when the atom type is used as a key.[S25,S26] This reduces the number of independent charges significantly, while still reducing the residual between quantum and classical interaction energies relative to CGenFF. However, this scheme does have the unfortunate side-effect of making the patches that would link together monomers dependent on the identity of the monomers being linked. Additionally, monomers with multiple adjacent linkages (such as a 5-5 linkage and an additional linkage on C4) cannot always have their charges determined by this method, since the two dimers that we use to describe each linkage can result in conflicting charge assignments for the same atoms.

To rectify this, we also present an alternative scheme, where compounds are broken up into integer charge groups based on the charge assignments from CGenFF (Fig. 2). In this setup, equivalent atoms within equivalent charge groups are assigned equivalent charges, keeping charges consistent for similar functional groups across all target compounds. The

charge groups are determined algorithmically, starting from atoms at extreme points on the molecule (typically hydrogen), and growing charge groups until the net charge is within 0.05 charge units of an integer. This is continued until all atoms are assigned to a group, with the caveat that large groups are split if possible by checking alternative starting sites for group assignment. If the absolute value of the net charge of the last assigned charge group exceeds 0.05, adjacent charge groups are removed, and new seeds are chosen for the charge groups. This algorithm results in compact charge groups within 100 assignment attempts for all the compounds studied.

For both schemes, the objective function was minimized using the L-BFGS-B algorithm[S27] as implemented in SciPy. The L-BFGS-B algorithm is a modification to the conventional L-BFGS algorithm that has been used in previous parameterization studies,[S19,S21] which can handle bounded and constrained optimization simultaneously. Since L-BFGS-B requires derivatives, derivatives for the charge objective function were computed for the energy and distance terms analytically. The dipole derivative was estimated numerically by taking steps of 0.0001 for all elements of the charge vector. The results were written to a topology file, where the output charges were rounded to the nearest thousandth of a charge unit using integer programming to arrive at a solution that minimally changed the output charges while making sure that the charge sums remained unchanged for groups and molecules. The topology file was used to generate the molecular topologies required for the subsequent bonded term optimization.

Bonded Term Optimization

Similar to the charge optimization, determining the bonded term parameters of Eq. 1 depends on creating quantum mechanical target data and using those data to inform an objective function that is minimized. The target data in this case are optimized bond, angle, and dihedral scans[S10] performed in Gaussian 09 at the MP2/6-31G* level of theory.[S23,S24] Due to their increased computational cost, only non-redundant dimer scans were performed, while

for monomers, all possible scans were performed. All bonds were stretched and compressed from their optimized geometry values by 0.1 Å in two steps, for a total of 5 molecular poses for each stretched bond. All angles were increased and decreased from their optimized geometry values by 10° in two steps. Dihedrals centered around sp^2 centers were scanned in 5° increments for 30° in both directions around the geometry optimum. Similarly, other dihedrals were scanned in 15° increments for 180° around the geometry optimum to generate a potential energy surface for a complete rotation of the bond. In sum, these 2574 scans generated 28473 valid poses where the energy of the pose is known quantum mechanically, and the bonding topology remained unchanged from the input structure, as judged by no bonds within a scan geometry exceeding 1.65 Å in length and no new unbonded atom pairs come within 1.65 Å of each other.

These target data are passed along to the objective function, which aims to match the energy changes between the individual poses from a single scan with our classical forcefield. The objective function contains terms not found in any prior parameterization efforts, incorporating information about the forces at quantum minima as well as biasing the sum of the angles around sp^2 centers to be 360°.

$$f(\bar{p}) = \sum_{poses} w \left(E_{QM} - E_{VDW+Elec} - E_{Bonded}(\bar{p})\right)^2 + \sum_{minima} v f_{MM}(\bar{p})^2 + \sum_{sp^2 sites} 10 \left(\left(\sum_{angles} \right) - 360 \right)^2 \quad \text{(S2)}$$

The first term is the typical energy term found in other parameterization schemes, which treats the bonded energy as a correction to the nonbonded energy that comes from the completed charge parameterization, which is precalculated prior to optimization. This term is weighted by $w = \exp(0.25 E_{QM})$, where E_{QM} has been shifted such that the minimum energy for a single scan is defined to be 0, under the principle that it is more important for the forcefield to accurately model near the minima of energy landscapes to obtain the correct

statistical distribution between minima rather than exactly recapitulating barrier heights of slow transitions. The second term was added based on the knowledge that the force on each atom at a quantum mechanical minima should be exactly zero. Different values for the weighting term v were tried, including 0, 0.02, 0.1, and 0.5, to see how this term improves the structures seen during simulation relative to computed quantum structures. The final term in Eq. S2 is an additional bias to make the angle terms around sp^2 centers sum to 360°. If this term is not added, these centers will tend to pucker and move the central atom out of the plane formed by the surrounding three atoms without an improper dihedral term added as well. Since we choose not to fit improper terms at all, we elect to make the aromatic rings flat through this modest bias, which is less than 1 % of the energy contribution to the objective function.

As with the charge optimization, we place bounds on the parameters during the bonded optimization process. The equilibrium bond lengths and Urey-Bradley lengths are allowed to drift only by 10 % from the initial value provided by CGenFF, whose initial values did not always reflect the geometry optimum observed in quantum calculations. Similarly, equilibrium angles are allowed to drift by at most 5 ° from the starting point. These bounds were found to be needed to prevent the optimizer from moving the terms far away from their observed geometry to improve fits in unrelated scans. Additionally, the force constants were restrained with a low bound of half the CGenFF estimate, and a high bound of 1000 $\mathrm{kcal\,mol^{-1}}$ $\mathrm{Å^{-2}}$ for bonds, and 500 $\mathrm{kcal\,mol^{-1}}$ $\mathrm{unit^{-2}}$ for both angles and Urey-Bradley terms. In practice, these force constant bounds were largely superfluous, as in most cases the force constants from CGenFF were found to be largely retained in the final fit. In the rare instances where this was not the case, the lower bound on the force constants prevented them from becoming zero, which would ruin the fit for the equilibrium values, in addition to permitting unphysical geometries.

Four different sets of dihedral parameter bounds were tried (Table S1), with results shown later. In dihedral set A, the only nonzero dihedral force constant (k_k in Eq. 1) occurs

Table S1: Bounds imposed on k_k during dihedral optimization for the four parameter sets generated depending on the periodicity. Note that for the case of sets A, B, and C, the bounds are implicitly $[0, 0]$ unless the force constant is specified in CGenFF (k_C) or was one of the select few terms added afterward to improve the fits, as detailed in the methods.

	Dihedral Set			
N	A	B	C	Free
1	$[-5, 3]$	$\begin{cases} [-5, 0] & k_C < 0 \\ [0, 3] & k_C > 0 \\ [-5, 3] & k_C = 0 \end{cases}$	$\begin{cases} [\max(-5, -2k_C), 0] & k_C < 0 \\ [0, \min(3, 2k_C)] & k_C > 0 \\ [-5, 3] & k_C = 0 \end{cases}$	$[-5, 3]$
2	$[-10, 5]$	$\begin{cases} [-10, 0] & k_C < 0 \\ [0, 5] & k_C > 0 \\ [-10, 5] & k_C = 0 \end{cases}$	$\begin{cases} [\max(-10, -2k_C), 0] & k_C < 0 \\ [0, \min(5, 2k_C)] & k_C > 0 \\ [-10, 5] & k_C = 0 \end{cases}$	$[-10, 5]$
3	$[-5, 4]$	$\begin{cases} [-5, 0] & k_C < 0 \\ [0, 4] & k_C > 0 \\ [-5, 4] & k_C = 0 \end{cases}$	$\begin{cases} [\max(-5, -2k_C), 0] & k_C < 0 \\ [0, \min(4, 2k_C)] & k_C > 0 \\ [-5, 4] & k_C = 0 \end{cases}$	$[-5, 4]$
4	$[-1.5, 2.5]$	$\begin{cases} [-1.5, 0] & k_C < 0 \\ [0, 2.5] & k_C > 0 \\ [-1.5, 2.5] & k_C = 0 \end{cases}$	$\begin{cases} [\max(-1.5, -2k_C), 0] & k_C < 0 \\ [0, \min(2.5, 2k_C)] & k_C > 0 \\ [-1.5, 2.5] & k_C = 0 \end{cases}$	$[-1.5, 2.5]$
6	$[-1, 1]$	$\begin{cases} [-1, 0] & k_C < 0 \\ [0, 1] & k_C > 0 \\ [-1, 1] & k_C = 0 \end{cases}$	$\begin{cases} [\max(-1, -2k_C), 0] & k_C < 0 \\ [0, \min(1, 2k_C)] & k_C > 0 \\ [-1, 1] & k_C = 0 \end{cases}$	$[-1, 1]$

when those terms are defined in the parameter set from CGenFF, including if the terms are zero, leveraging heuristics included in CGenFF about which dihedral terms are essential to describing the potential energy surface (e.g. $n_k = 2$ for terms within aromatic rings, or $n_k = 3$ for most rotateable bonds). The reliance on the heuristics is expanded in dihedral set B, where k_k is restricted to have the same phase as what was determined through CGenFF, which heavily penalizes situations where two related dihedral terms nearly cancel by being opposite in sign. As a result, the force constants in set B are in general smaller than they are in set A. Dihedral set C goes even further by restricting the magnitude of the force constants. Whereas sets A and B have force constant bounds determined by the range of values observed throughout the CHARMM force field (Table S1), set C places bounds based on the force constants determined through CGenFF. The effect of this approach is to fine tune the allowable range such that the dihedral force constants respond to different chemistries. For instance, $n_k = 3$ terms involving four heavy atoms tend to have larger force constants than they would if a hydrogen is involved, reflecting the higher cost of eclipsed conformations for larger species. The bounds imposed in set C (Table S1) would reflect this reality better than either set A or B, where such force constants all share the same bounds. In all three letter dihedral sets tested, additional $n_k = 1$ terms not found in CGenFF were permitted to be nonzero in order to reflect the preference of specific alcohols to have their hydrogen pointed away from nearby large functional groups.

The final set of bounds tested is the "free" parameter set, where all $n_k = 1, 2, 3, 4, 6$ dihedral terms were allowed to be nonzero, regardless of whether they appear in the initial parameter set provided by CGenFF. This approach follows other fully automatic parameterization schemes,[S19,S28] while still retaining the bounds from set A (Table S1). Since this set has the most free parameters, it naturally will have the lowest the objective function value as described in Eq. S2 relative to the other dihedral sets. However, this free set loses much of the chemical intuition of the lettered sets, and is primarily included as a reference point for how far from optimal the other sets are, as it was shown to produce inferior geometries

during simulation.

The bonded optimization uses the same SciPy L-BFGS-B routine[S27] as was used for charges. However, rather than simultaneously optimizing bond, angle, and dihedral terms, we cycle between only optimizing bond and angle terms and optimizing dihedral terms separately five times, and then optimize once more with all terms allowed to float. To bound our time to solution, each optimization stage is limited to 8000 L-BFGS-B steps or 16000 objective function evaluations. This staged approach prevents the dihedral terms from rapidly adjusting to eliminate the initial residual before the bond and angle terms can respond. Due to the large number of free parameters, evaluation of the energy contribution to the objective function and its gradients would take between three and four seconds when written purely in python, whereas each step was a fraction of a second for the charge objective function. The slow evaluations became a significant bottleneck in the parameterization pipeline, and so the energy and force contributions to the objective function were reimplemented within a GPU-accelerated library. Specifically, since optimization of bonded terms simplifies to the population and reduction of a N by M matrix (N being the total number of poses, M being the number of free parameters),[S10,S28] we use a combination of CUDA[S29]-kernels and Thrust[S30] to fill the matrix and cuBLAS to perform the reduction required to compute the gradients. This GPU implementation reduces the runtime overall by approximately two orders of magnitude relative to the original python implementation, allowing many more optimization steps to be taken, and is provided as Supporting Information for others to take advantage of. Accelerating the evaluation of the objective function allows us to take many more steps in this very high dimensional space, thereby converging on the optimum parameter set given the bounds placed upon the optimizer.

Extended Results and Discussion

We discuss here the rationale for choosing a group-based charge assignment paradigm and letting v=0 in Eq. S1. This discussion is rather technical, as the differences between individual parameter sets are rather small, and is not required to show that the developed force field strikes a good balance between improved accuracy and simplicity in model construction.

Charge Optimization

As briefly described in Methods (Fig. 2), we attempted two alternative approaches to determine equivalent charge environments for the compounds under study. In one approach, a neighborhood around an atom was defined through a molecular subgraph that extended 1, 2, or 3 bonds away from the originator atom. If two neighborhoods were identical, with the same internal topology between atoms of the same type, those charges were forced to be equivalent during optimization. The other approach was to define charge groups based on the initial assigned charges from CGenFF and use those groups as the subgraphs to be compared. As will be shown in the subsequent discussion, the two approaches are equally good at fitting the target data, however the group-based method produces modular lignin models, and is therefore our preferred method for determining equivalent charges.

Under either scenario, optimization reduces the interaction energy residual between the quantum interaction energy (E^{int}_{QM}) and the classical interaction energy (E^{int}_{MM}), with varying magnitudes of success, depending on the scheme used (Fig. S31). From Fig. S31B, we see that after optimization, approximately 50 % of the calculated water interactions are within 0.5 kcal mol^{-1} of their quantum targets, a significant improvement on the 40 % from the CGenFF starting point. For more problematic water interactions where the residual remains large, as in the extreme 5 % of the residual distribution, optimization can reduce the residual by up to 1 kcal mol^{-1}. The optimization significantly reduces the residual for most interactions by somewhat increasing the residual for others (Fig. S31A), improving the

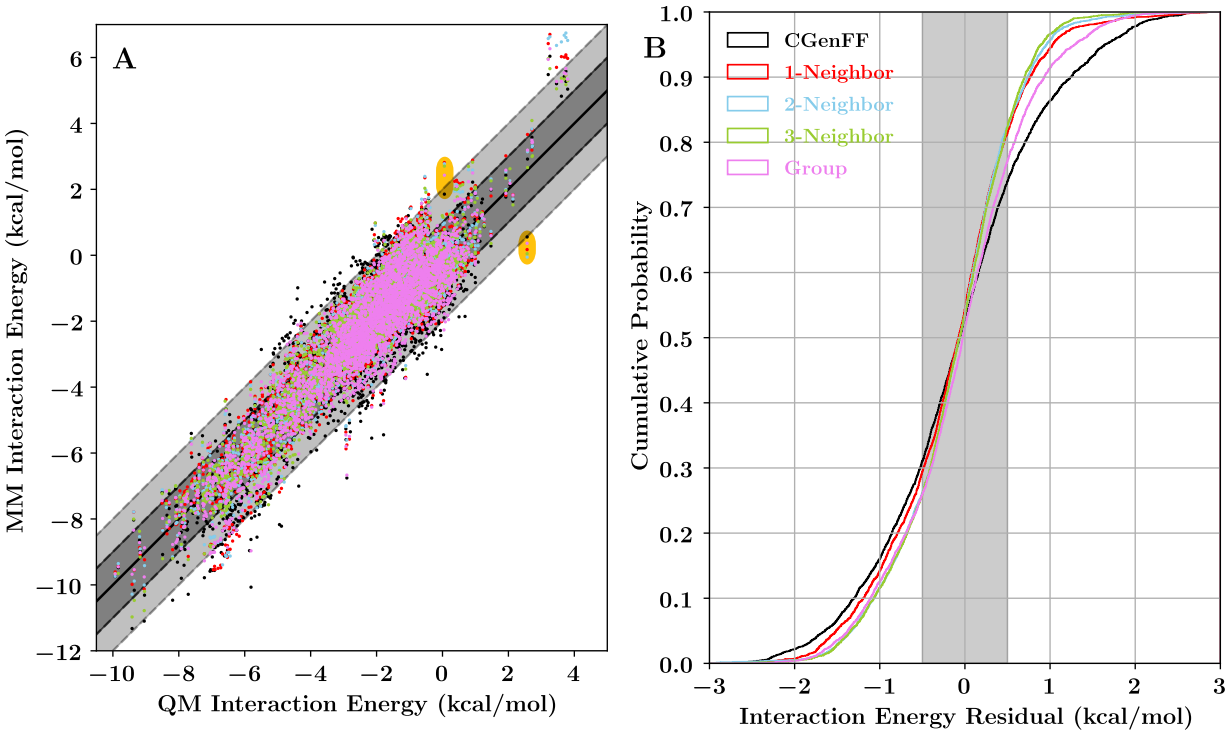

Figure S31: Comparison of water interaction energies determined through quantum calculations and the parameterized point charges in our molecular mechanics framework. (A) Scatter diagram comparing the adjusted quantum (QM) and classical (MM) interaction energies for the low interaction energy poses ($E^{int}_{QM} < 5\,\text{kcal mol}^{-1}$ and $E^{int}_{VDW} < 1\,\text{kcal mol}^{-1}$) under different parameterization schemes. These cutoffs reduce the number of points plotted, which improves the visual clarity of the plot. The solid black diagonal line indicates the line where $E^{int}_{QM} = E^{int}_{MM}$, which is surrounded by darker and lighter bands indicating deviations of $1\,\text{kcal mol}^{-1}$ and $2\,\text{kcal mol}^{-1}$. The scattering points for two specific isolated poses have been highlighted with an orange underlay, indicating that not all points have been improved by the fitting procedure. In (B), the scatter plot is transformed into a cumulative distribution of the interaction energy residuals ($E^{int}_{QM} - E^{int}_{MM}$), with a highlighted grey region representing residuals less than $0.5\,\text{kcal mol}^{-1}$. Both plots use the same colors to discriminate between parameterization schemes. Black is used for the charges taken directly from CGenFF. Red, blue, and green are used for successively larger neighborhood schemes, and violet is used to denote the group-based parameterization scheme.

overall fit.

Table S2: Mean absolute charge shift, in charge units, comparing the difference in optimized heavy atom charges between different parameter sets.

Charge Type	Group	1-Neighbor	2-Neighbor	3-Neighbor
CGenFF	0.025	0.045	0.049	0.039
Group	–	0.049	0.047	0.044
1-Neighbor	–	–	0.055	0.048
2-Neighbor	–	–	–	0.033

Table S3: Charge optimization comparison statistics. Each of the four parameterization schemes is compared against CGenFF, principally through their root mean squared error for the water interaction energies. This is done under two conditions, once for the favorable interaction energies that contribute the most to the optimization due to the weighting applied in Eq. S1, and again for higher energy interactions that are largely excluded from the optimization. As in Fig. S31, only if $E^{int}_{VDW} < 1\,\mathbf{kcal\,mol^{-1}}$ is the datapoint included in the reported statistics, so that the reported $\mathbf{R^2}$ matches the scatter shown in Fig. S31A.

	Small E ($E^{int} < 0$ kcal/mol)			Larger E ($E^{int} < 20$ kcal/mol)	
Charge Scheme	$\left\langle \left(E^{int}_{QM} - E^{int}_{MM}\right)^2 \right\rangle^{\frac{1}{2}}$	$\left\langle E^{int}_{QM} - E^{int}_{MM} \right\rangle$	R^2	$\left\langle \left(E^{int}_{QM} - E^{int}_{MM}\right)^2 \right\rangle^{\frac{1}{2}}$	R^2
CGenFF	0.98 kcal/mol	-0.08 kcal/mol	0.80	0.98 kcal/mol	0.81
1-Neighbor	0.79 kcal/mol	-0.16 kcal/mol	0.87	0.79 kcal/mol	0.87
2-Neighbor	0.70 kcal/mol	-0.13 kcal/mol	0.89	0.72 kcal/mol	0.89
3-Neighbor	0.68 kcal/mol	-0.13 kcal/mol	0.89	0.69 kcal/mol	0.90
Group	0.78 kcal/mol	-0.06 kcal/mol	0.86	0.79 kcal/mol	0.87

The degree of success in improving the fit is alternatively quantified in Table S3 through the root mean squared error (RMSE). Through this lens, the group scheme is effectively equivalent to the neighbor scheme when the subgraph only considers direct neighbors (1-Neighbor), with very similar RMSE values and correlation coefficients. The RMSE improves further when the neighbor scheme uses a larger neighborhood to determine which charges should be equivalent, thereby increasing the number of individual charges allowed and the parameter space the optimizer can explore. The number of free parameters roughly doubles (from 259 to 450) in going from the 1-Neighbor to the 2-Neighbor scheme, with the 3-

Neighbor scheme nearly doubling the total of free parameters again to 798. This makes the relatively small 10 % decrease in the RMSE in going to a more expansive charge scheme much less impressive than it otherwise might be, and suggests that the 2- and 3-Neighbor approaches may overfit the target data. The ultimate conclusion is that the number of free charge parameters is the primary determinant for the observed RMSE, and neatly explains why the 1-Neighbor (259 independent charges) and group bond (290 independent charges) are so similar, and are effectively interchangeable from an RMSE standpoint, even if there are modest changes in charges between the two sets (Table S2).

Since the quality of the fits are largely unchanged between the two approaches (Fig. S31, Table S3), other considerations can drive the final choice. In this case, we choose the group-based approach for the practical reason that it makes the topology file simpler. To understand why, it helps to consider why the CG2R61 atom type, which is used in CGenFF for most aromatic carbons, was split into three atom types in the current lignin force field. The atom type split was required by monomer-specific angle changes near oxygenated functional groups found throughout the quantum optimized geometries (Fig. S28). Given this diversity in the observed angles, there is no way to reproduce these geometries if the aromatic carbon atom types were all equal. Instead, the aromatic carbon atom type is split based on the functional group that is attached to the carbon, representing the only case where atom types in the lignin force field are mapped surjectively onto the CGenFF atom types.

The Neighbor-1 variant of charge optimization therefore assigns different charges on the aromatic carbons depending on the monomer type, which in turn means that the topological patches that describe individual lignin linkages are monomer-specific. The Neighbor-1 approach quadruples the number of possible topological patches to choose from when linking lignin monomers together. Different linkages may also both modify the charges on a specific atom, such as if a monomer was involved in both a β-O-4 and a 5-5 linkage to carbons 4 and 5 (Fig. 1), which could lead to a non-integer charge after both linkages are applied to the system.

By contrast the charge groups assigned as in Fig. 2 logically subdivide the ring by functional group, as shown by the underlaying ovals in Fig. S28. The rational division of lignin functionality is replicated for more complicated dimeric systems as well (Fig. S2). Since the group method isolates neighboring functional groups from their neighbors, the topological patches that link together individual monomers are largely independent of monomer identity, simplifying the lignin construction process. The isolated charge groups also mean that multiple linkages can be applied to the same monomer without two linkages trying to apply different charges to the same atom. Thus from this point onward, we only consider charges that are determined through this group scheme, which is just as accurate as the other scheme tested (Fig. S31, Table S3), and comes with fewer side effects when constructing systems that use the force field.

Bonded Optimization

With the charges established in the previous section, determining bonded term parameters is the final choice prior to a full parameter set. At the heart of the bonded optimization are the potential energy scans, with a small subset of the 2574 scans performed shown in Fig. S4, which are chosen to be representative of different behaviors observed over the whole set. Figs. S4A and S4B led to the introduction of extra n=1 terms to adjust the existing n=2 dihedral to allow for the energies of the two states to be unequal, rather than just being dictated by the non-bonded terms as it was originally in CGenFF (grey). This was the only clear case in surveying the individual scans where CGenFF had obviously missed a dihedral term required to correctly fit the potential energy scans. In Fig. S4C, we see a different kind of limitation of the general force field, where CGenFF already contained the exact multiplicities needed to recapitulate the scan, but did not have the right weighting between them. This is even clearer in Fig. S4D, where the optimum angle at the bridging oxygen was not originally recapitulated by CGenFF, but is improved in all of our new optimizations. However, not every scan is perfect, as evidenced by Figs. S4E and S4F, where the relatively

well-behaved quantum potential energy surface is not perfectly fit by any of the molecular mechanics parameter sets. Sometimes, as in Fig. S4E, the overall shape is preserved, whereas in others like S4F, the shape of the potential energy surface is not broadly reproduced even when all possible dihedral parameters are allowed to be nonzero, as in the free dihedral set.

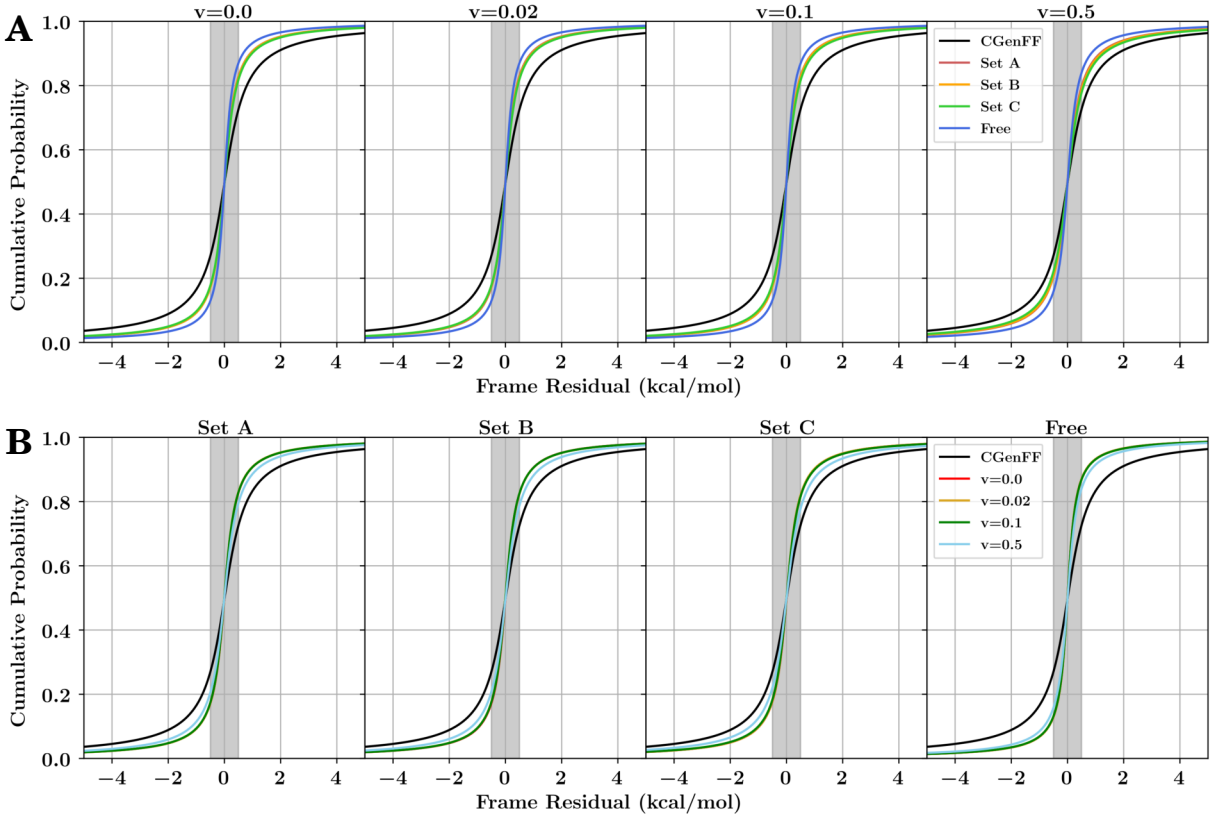

Figure S32: Cumulative distribution of the residual between the optimized potential energy surface and the target potential energy surface under different optimization conditions. In (A), the parameter v in Eq. S2 is held fixed in each subpanel, and the dihedral optimization set (Table S1) changes. The reverse is true in (B), where the impact of the v parameter is directly probed. In each graph, the black line is the original CGenFF distribution, the other colored lines represent the newly optimized parameter sets, and a gray background for the region where the residual is less than 0.5 kcal mol^{-1}.

Individual scans, such as those in Fig. S4, do not provide a holistic view of parameter quality. Instead, we aggregate the residuals within each scan under the different optimization conditions and monitor their distribution, as in Fig. S32. In all cases, the newly optimized parameters outperform the original CGenFF parameter set, with 10-20 % more of the pose population near zero residuals, although they do so to differing degrees. In general, large val-

ues for the v parameter within the optimization objective function (Eq. S2) are detrimental to the fit, with less of the population being near zero residual (Fig. S32B). Likewise, the energetic fit becomes worse with increasing constraints on the dihedral parameters (Fig. S32A). Both of these phenomena are consequences of disfavoring recapitulating exactly the potential energy surface by adding in a competing term that is force-dependent and by bounding the solution space.

Table S4: γ parameters (in kcal mol^{-1}) for the Cauchy distributions centered on zero that best fit the remaining residuals for all tested combinations of v parameters and dihedral sets, in addition to the original CGenFF fits.

Dihedral Set	v=0	v=0.02	v=0.1	v=0.5
Set A	0.3	0.3	0.3	0.38
Set B	0.3	0.31	0.31	0.39
Set C	0.32	0.32	0.33	0.42
Free	0.22	0.21	0.22	0.27
CGenFF	0.57			

A more typical quantification metric to measure the energy deviation would be the root square mean error (Table S7). However, the residuals observed are not normally distributed, as the standard deviation of the residuals would predict a much broader distribution than what is actually observed (Fig. S5). Instead, we find that a Cauchy distribution centered around zero, whose probability density has the form:

$$P(x) = \left(\gamma\pi\left(1 + x^2\gamma^{-2}\right)\right)^{-1} \tag{S3}$$

fits the observed residuals much better (Fig. S5). In the formalism of Eq. S3, x would be the energy residual, and γ would be a scale parameter that determines the probability at the peak ($P(0) = (\pi\gamma)^{-1}$). These γ parameters are reported in Table S4, with smaller values for γ indicating a distribution with a higher peak and less population in the long tails of the distribution. The γ parameters reinforce the findings from Fig. S32, in that the best fits to the energy scans come from ignoring forces at minimum energy structures and giving the

optimizer as many free parameters as possible.

Based on the data presented so far, one would then clearly choose the optimization parameter combination of no force contribution to the optimization criteria (v=0) and to allow as many dihedral terms as possible to perfectly recapitulate the target potential energy surfaces. This is indeed what many automated parameter optimization techniques do when reoptimizing dihedral parameters,[S19,S28] and numerically reduces energy residuals by exploiting these other degrees of freedom. However, having good energy fits does not always imply that the derived structures are accurate, as was shown in Fig. S3 for a hypothetical potential energy surface. Thus before making a final assessment of what parameter set to choose, we examine structural information after optimization to see what extra impact and possibly overfit degrees of freedom might have on the observed structures to be expected in simulation.

This structural information is added through examining how far the optimized structures from gas-phase quantum calculations will drift either when placed in solution, where water molecules are present to screen intra-molecular electrostatic interactions (Fig. S6), or when the molecule is isolated in vacuum (Fig. S33). In either case, using the free dihedral set results in larger deviations from the starting point than any of the other dihedral sets tested, suggesting that indeed the added complexity from the additional terms is creating new local minima away from the quantum minima along the orthogonal degrees of freedom within a single scan, as was exemplified in Fig. S3. In our view, this is compelling evidence that allowing any periodicity to contribute during minimization creates overfitting artifacts, as the free dihedral set has consistently poor fits (Fig. S6A). For this reason, we exclude the free dihedral set as a candidate for the final lignin parameter set, despite having the lowest energy residuals (Fig. S32, Table S4).

We examine the three CGenFF-based dihedral sets as candidates for the final parameter set. Based on the mean RMSDs after minimization, as presented in Tables S6 and S8, dihedral sets B or C, where the phases of the dihedral terms are fixed to their CGenFF values,

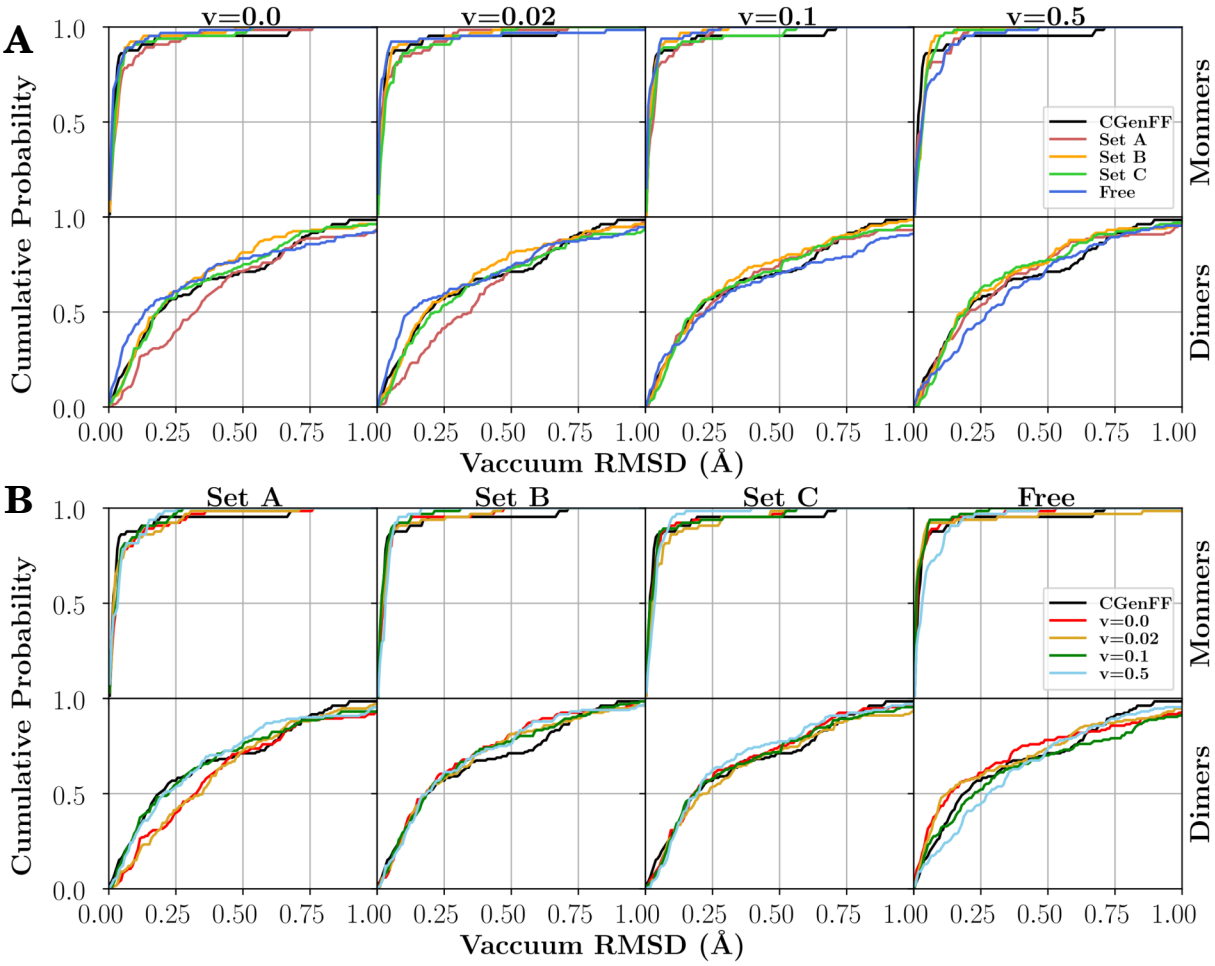

Figure S33: Root mean square deviation (RMSD) distribution of the resulting structures for both monomers and dimers after minimization in vacuum (no other molecules in the simulation system) relative to the gas-phase minimum energy structures determined quantum mechanically. In (A), the parameter v from Eq. S2 is held fixed, highlighting the effect of different choices for dihedral bounds, with the effect of force inclusive optimization demonstrated in (B). Within each subpanel, the results from the CGenFF starting point are also shown in black as a benchmark. The mean and standard deviation for these distributions are reported in Table S8. For a similar analysis in solution, see Fig. S6.

Table S5: Periodic unit cell dimensions a, b, c, α, β, and γ for the original crystal (black), as well as averaged over the last 10 ns of simulation with both the optimized lignin force field (red), and CGenFF (gray). The standard deviation is reported as an uncertainty in the last digit in parentheses.

CSD Code	a			b			c		
CATCOL13	58.39	58.259(6)	57.186(4)	50.58	52.589(4)	52.352(4)	51.66	50.227(6)	50.477(5)
PHBALD11	53.59	52.78(1)	52.964(9)	54.22	53.17(2)	53.04(2)	50.01	52.39(1)	52.73(1)
YUHTEA01	56.40	53.975(5)	54.120(4)	53.75	54.928(8)	55.690(9)	59.41	62.584(4)	62.715(6)
YUHTEA03	56.20	53.91(2)	53.95(1)	55.05	57.37(5)	58.12(2)	60.05	56.61(3)	56.69(1)
IZALAW	55.26	37.81(1)	39.88(1)	53.85	77.55(2)	76.68(1)	50.43	55.10(1)	54.615(9)
SIPKEH	62.31	67.56(3)	68.84(5)	51.51	54.87(3)	53.21(2)	55.62	48.57(4)	49.39(3)
CEHGUS	50.83	53.77(1)	53.83(1)	52.17	52.37(1)	51.89(1)	56.64	56.86(1)	57.39(1)
GASVOL01	50.58	52.982(6)	51.072(5)	50.26	49.399(5)	49.059(5)	59.02	59.996(6)	63.300(9)
RABWUM	50.80	51.630(7)	50.731(6)	60.22	61.658(5)	61.212(5)	52.02	51.900(4)	53.512(4)
SIPPEM	70.95	71.36(1)	72.16(1)	60.72	63.33(2)	63.63(1)	50.62	50.59(2)	49.759(8)
VADDOT	50.54	50.587(8)	51.12(1)	53.74	56.63(1)	58.22(1)	60.22	61.17(1)	60.87(1)
SAZHEG	52.23	53.64(1)	53.66(1)	53.36	53.09(1)	53.40(2)	52.84	55.08(1)	54.454(9)
FOCGUA	72.19	75.161(6)	73.696(9)	51.42	51.626(3)	52.513(4)	67.34	67.301(8)	69.240(8)
IDIKIP	51.62	52.748(5)	52.061(6)	62.49	64.33(1)	68.51(1)	59.48	60.34(1)	58.568(8)
INELIW	58.99	60.556(6)	61.512(6)	54.59	56.136(9)	55.78(1)	55.48	54.565(6)	55.014(6)
INELIW01	54.85	53.93(1)	50.930(7)	54.61	56.12(2)	60.892(9)	53.77	54.96(1)	54.808(5)
FAFXUF	57.75	58.84(1)	60.5(1)	54.12	53.87(1)	58.0(1)	56.91	59.90(1)	60.3(2)
FUMVUE	52.54	51.991(8)	51.783(7)	58.69	61.00(1)	60.12(1)	54.90	56.83(1)	57.88(1)
TUGWAT	59.08	59.76(4)	62.00(2)	52.16	52.546(9)	52.853(9)	62.60	64.89(8)	62.08(1)
UJOGIK	62.46	58.538(4)	62.80(4)	51.06	49.816(4)	48.81(5)	53.15	59.972(3)	58.3(1)
CSD Code	α			β			γ		
CATCOL13	90.00	90.000(4)	90.000(4)	114.24	111.049(9)	108.495(7)	90.00	90.000(4)	90.000(4)
PHBALD11	90.00	90.00(1)	90.00(1)	112.87	111.01(2)	109.83(2)	90.00	90.00(1)	90.00(1)
YUHTEA01	90.00	90.00(1)	90.00(1)	115.02	9(3)e+01	9(3)e+01	90.00	90.000(6)	90.000(8)
YUHTEA03	90.00	90.00(2)	90.00(2)	90.00	90.00(2)	90.00(2)	90.00	90.00(2)	90.00(1)
IZALAW	90.00	90.00(1)	90.00(1)	91.35	101.95(2)	102.90(2)	90.00	90.00(1)	90.00(1)
SIPKEH	90.00	89.9(1)	90.04(6)	90.00	90.13(4)	89.96(1)	90.00	92.7(2)	89.69(4)
CEHGUS	90.00	90.00(2)	90.00(1)	95.31	93.67(2)	93.61(2)	90.00	90.00(3)	90.00(2)
GASVOL01	90.00	90.000(9)	90.000(6)	91.72	92.117(8)	92.92(1)	90.00	90.000(8)	90.000(4)
RABWUM	90.00	90.000(6)	90.000(6)	96.16	96.142(7)	96.622(6)	90.00	90.000(6)	90.000(5)
SIPPEM	90.00	90.00(2)	90.000(8)	90.00	90.00(1)	90.000(9)	90.00	90.00(1)	90.000(8)
VADDOT	109.41	111.44(1)	113.23(1)	90.57	92.04(2)	92.24(2)	103.28	103.19(2)	103.92(1)
SAZHEG	106.23	107.61(1)	106.16(2)	93.71	94.91(2)	93.38(1)	82.97	83.83(1)	84.53(1)
FOCGUA	90.00	90.000(6)	90.000(7)	114.33	114.162(5)	115.180(5)	90.00	90.000(5)	90.000(7)
IDIKIP	90.00	90.000(8)	90.00(1)	109.84	111.578(9)	111.54(2)	90.00	90.000(9)	90.000(7)
INELIW	90.00	90.000(8)	89.999(7)	92.89	91.949(9)	87.75(1)	90.00	90.000(7)	90.001(6)
INELIW01	90.00	90.02(9)	90.01(1)	95.48	93.42(1)	98.733(9)	90.00	89.999(9)	90.00(1)
FAFXUF	90.00	90.000(9)	88.8(1)	90.00	90.00(1)	89.9(2)	90.00	90.000(9)	89.3(1)
FUMVUE	90.00	90.00(2)	90.00(1)	109.81	109.21(2)	109.30(1)	90.00	90.00(1)	90.000(8)
TUGWAT	90.00	90.02(2)	90.01(3)	100.26	101.82(4)	103.67(3)	90.00	90.00(2)	89.99(8)
UJOGIK	90.00	90.000(4)	90.011(8)	90.00	90.000(3)	90.041(6)	90.00	90.000(4)	89.78(2)

Table S6: The RMSD mean and standard deviation (in Å) for monomers and dimers minimized in solution relative to the gas-phase minimum energy structures determined quantum mechanically.

	Monomers			
Dihedral Set	v=0	v=0.02	v=0.1	v=0.5
Dihedral Set A	0.27 ± 0.12	0.30 ± 0.14	0.26 ± 0.12	0.28 ± 0.15
Dihedral Set B	0.22 ± 0.14	0.24 ± 0.13	0.22 ± 0.11	0.27 ± 0.17
Dihedral Set C	0.22 ± 0.14	0.22 ± 0.14	0.24 ± 0.16	0.25 ± 0.15
Free Dihedral	0.30 ± 0.13	0.33 ± 0.14	0.31 ± 0.15	0.34 ± 0.18
CGenFF	0.19 ± 0.10			
	Dimers			
Dihedral Set A	0.34 ± 0.08	0.35 ± 0.08	0.33 ± 0.08	0.34 ± 0.08
Dihedral Set B	0.30 ± 0.08	0.29 ± 0.08	0.30 ± 0.08	0.31 ± 0.08
Dihedral Set C	0.29 ± 0.09	0.29 ± 0.07	0.29 ± 0.08	0.29 ± 0.08
Free Dihedral	0.42 ± 0.12	0.42 ± 0.11	0.39 ± 0.10	0.46 ± 0.11
CGenFF	0.29 ± 0.08			

Table S7: Root mean square error, in kcal mol^{-1}, comparing the difference between quantum and classical energies across all potential energy scans. The tabulated numbers are heavily influenced by the poses with the poorest fits, and do not reflect typical residuals.

Dihedral Set	v=0	v=0.02	v=0.1	v=0.5
Set A	2.3	2.3	2.3	2.5
Set B	2.3	2.3	2.3	2.5
Set C	2.3	2.3	2.3	2.4
Free	2.1	2.1	2.1	2.3
CGenFF	2.8			

Table S8: The RMSD mean and standard deviation (in Å) for monomers and dimers minimized in vacuum relative to the gas-phase minimum energy structures determined quantum mechanically.

	Monomers			
Dihedral Set	v=0	v=0.02	v=0.1	v=0.5
Dihedral Set A	0.07 ± 0.11	0.06 ± 0.11	0.05 ± 0.07	0.05 ± 0.06
Dihedral Set B	0.05 ± 0.09	0.05 ± 0.09	0.04 ± 0.05	0.04 ± 0.03
Dihedral Set C	0.06 ± 0.11	0.07 ± 0.11	0.06 ± 0.11	0.05 ± 0.05
Free Dihedral	0.06 ± 0.19	0.06 ± 0.19	0.03 ± 0.05	0.06 ± 0.08
CGenFF	0.06 ± 0.14			
	Dimers			
Dihedral Set A	0.40 ± 0.33	0.39 ± 0.30	0.36 ± 0.36	0.36 ± 0.37
Dihedral Set B	0.29 ± 0.29	0.31 ± 0.30	0.31 ± 0.30	0.32 ± 0.33
Dihedral Set C	0.32 ± 0.31	0.35 ± 0.33	0.35 ± 0.36	0.31 ± 0.30
Free Dihedral	0.32 ± 0.37	0.33 ± 0.38	0.37 ± 0.37	0.38 ± 0.34
CGenFF	0.34 ± 0.32			

are superior to A, where the phase was allowed to change for a term with a given periodicity. We suspect that in the case of set A, the optimizer chased particularly large residuals, and corrected them by flipping around the phase for specific dihedral terms. This can have deleterious effects on the structure, since flipping the phase will reverse the positions of minima and maxima within a single term. Since these phases encode within them specific chemical intuition, such as n=2, $\delta = 180°$ forcing a flat interaction suitable for the core of an aromatic ring, or n=3, $\delta = 0°$ correctly emphasizing staggered rather than eclipsed configurations around sp^3 centers, flipping the phase can push structures in unnatural directions.

It should be emphasized that these minimized structure results were the primary motivation for using the CGenFF parameter set as a basis for the non-zero dihedral terms. The parameter set determined by CGenFF already properly encodes the required chemical intuition to minimize overfitting problems that we see with the free parameter set. Thus, rather than develop a novel and untested algorithm to assign nonzero dihedrals and phases, we leverage prior general force field development efforts[S2] to guide us towards a good starting point. Since few potential energy scans initially saw results where the free dihedral set was decidedly better than the restricted dihedral sets (Figs S4A and S4B), we think that

CGenFF indeed was an excellent starting point.

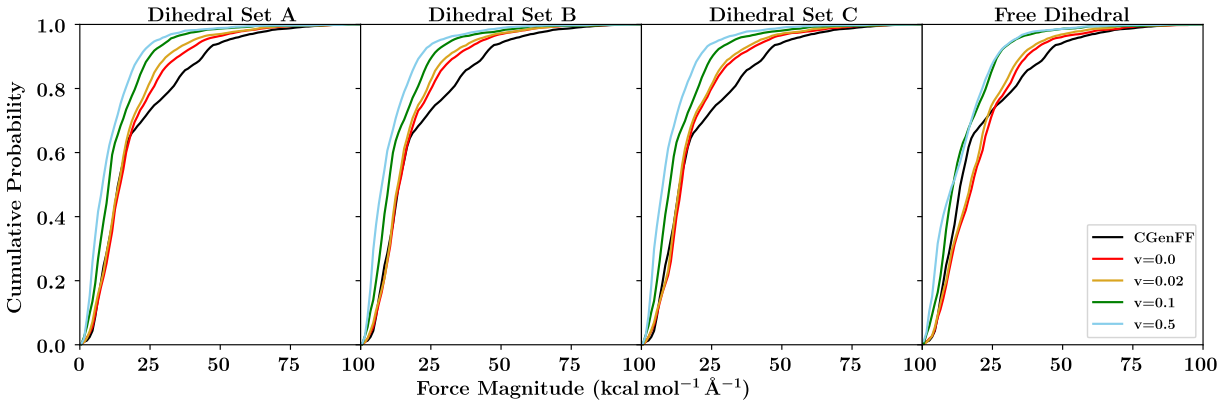

Figure S34: Distribution of the force magnitudes on each atom from minimized structures depending on the different v parameters applied while optimizing with a specific dihedral set.

Having settled on either dihedral set B or C, the parameter v in Eq. S2 needs to be addressed. In principle, having v be nonzero could reduce the type of problem exemplified in Fig. S3, since if the computed force is near zero at the quantum minimum, that implies that the molecular mechanics minimum energy structure coincides with the quantum structure. However, since force information at minimum energy configurations has not been used to inform CHARMM-style parameterization before, there is a high burden of proof to show that this is indeed the case. Based on the evidence shown so far (Figs. S32B and S6B), there is no compelling evidence that a nonzero v actually improves the quality of the parameterization, although it does clearly shift the distribution of the forces experienced by individual atoms towards zero (Fig. S34).

If we look further at specific structural features that we would like to recapitulate, such as the alignment of aromatic rings coupled by 5-5 linkages (Fig. S7C), we can further probe the effect of v, as well as provide a practical example of how large structural deviations with the new force field compare with CGenFF. For dihedral sets B or C, both improve on CGenFF uniformly, shifting the distribution to the right and closer to the 180° We observe that changing the dihedral set (Fig. S7A) has a larger impact than increasing v (Fig. S7B), although an appropriate v of intermediate size can better recapitulate the 180° geometries

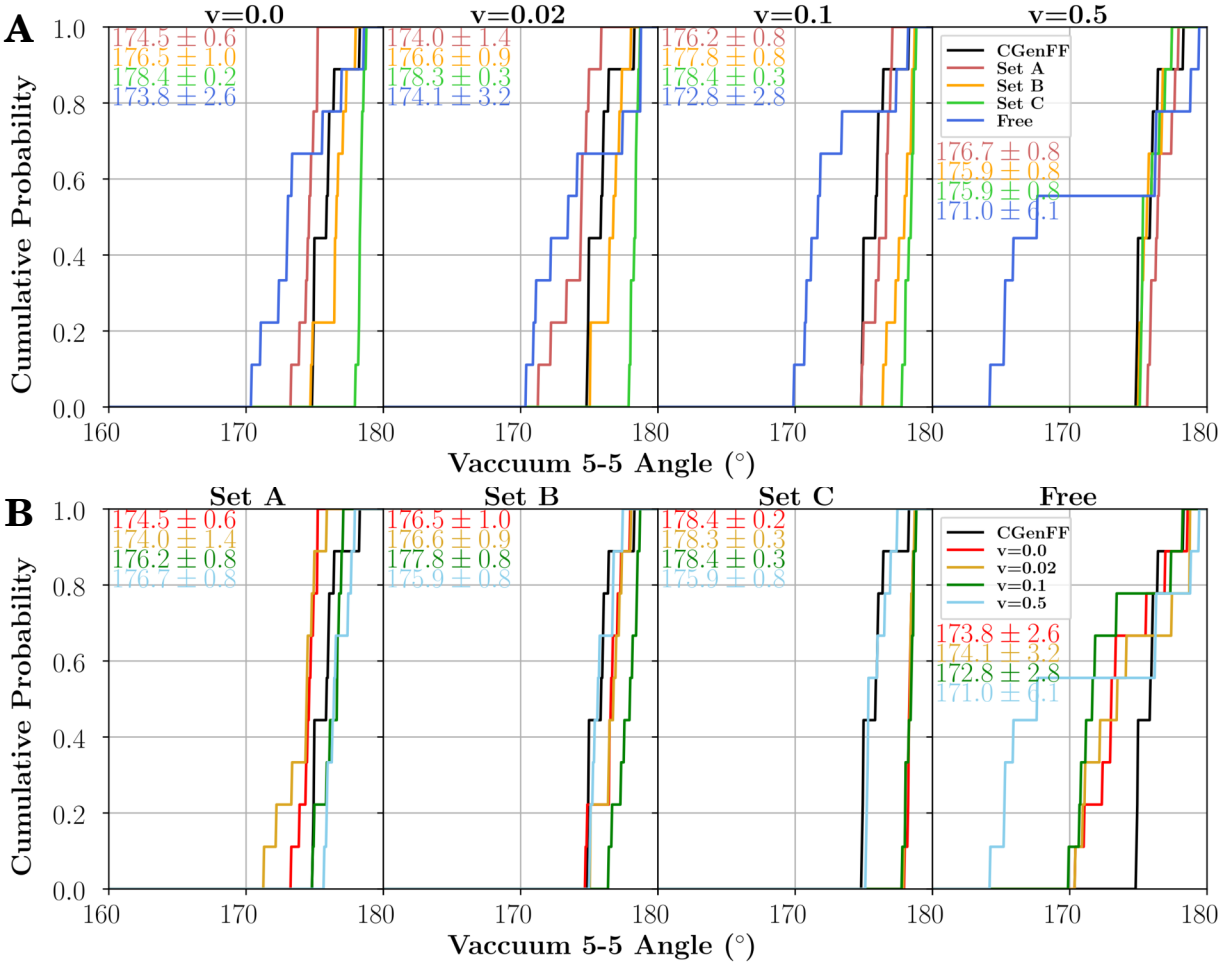

Figure S35: Distribution of the 5-5 angle, as defined in Fig. S7C, when dimers containing a 5-5 linkage are optimized in vacuum. In (A), the parameter v from Eq. S2 is held fixed, highlighting the effect of different choices for dihedral bounds, with the effect of force inclusive optimization demonstrated in (B). Mean values and their standard deviation over the distribution are reported in-figure, using the same color as those given in the in-figure legend.

expected from quantum calculations. These results also hold in vacuum (Fig. S35). While this does eliminate the v=0.5 case, which had already performed poorly on prior metrics, it does not provide a compelling reason on its own to adopt either v=0.02 or v=0.1 because of how small both the effect size as well as the sample size, since we only are considering the 5-5 linkage. However, given what we know so far, v=0.02 is slightly better than v=0.1, and that is what we will use in the final comparisons.

At this stage, four candidate parameter sets remain, a combination of v=0 or v=0.02, and dihedral sets B or C. Deciding between these candidates is done in part by directly comparing how different each of the parameter sets are, by assessing their correlation coefficients with respect to one another (Fig. S36). What we observe is that the parameter sets that satisfy our criteria are highly correlated with one another, with a few minor visible changes. For instance, the equilibrium values for the bonds and the Urey-Bradley terms show a checkerboard pattern, suggesting that these changed depending on the value for v. By contrast, the force constants for these terms are effectively unchanging between all the combinations shown. The angle term changes significantly from CGenFF, a consequence of both the angle sum constraint around sp^2 centers (Eq. S2) and some instances where CGenFF did not fit particularly well due to an incorrect minimum position (Fig. S4D). Finally, the correlation between the dihedral sets that emerge as the best overall fit (B, C, v=0 or 0.02) is actually quite high, suggesting that again we can make the final determination between these options based on other considerations.

We use this freedom to select our final parameter set based on arguments of simplicity. The rest of the CHARMM force field was parameterized using v=0, since the forces at the minimum were not explicitly considered. Since we find no compelling reason to use a different value for v, our final parameter set should also use v=0. We then choose to report only the parameters from set B because of the slightly narrower distribution of energy residuals (Table S4), generally running to the left of set C in Fig. S6A, and the fact that the 5-5 angle distribution is slightly better when v=0 (Fig. S7A). These changes are minute, as is

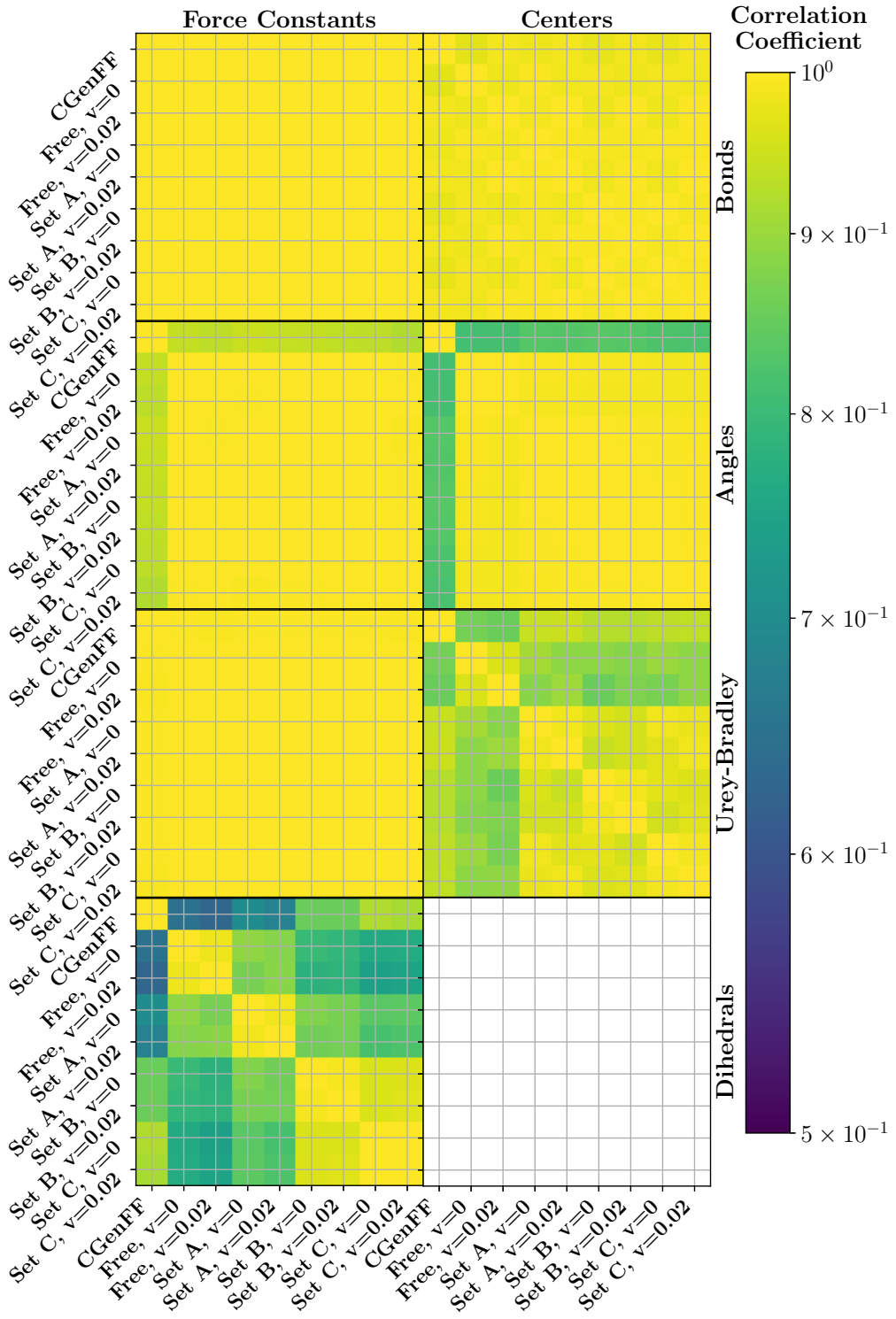

Figure S36: Correlation coefficients between specific parameter sets tried, as labeled on the axes. The individual parameter vectors that describe specific bonded terms from Eq. 1 are separated out, such that the measured correlations only pertain to either the force constants or the equilibrium position (center of each harmonic potential) for each term in the force field. Higher correlation coefficients between parameter sets are shown in yellow, and lower correlation coefficients are bluer, with a minimum correlation of 0.5.

expected given the differences in the bounds imposed during optimization (Table S1) between sets B and C. The final complete parameter set is provided as a separate download in the Supporting Information.

Implications and Applications for Polymeric Force Field Development

While the force field determined here is strictly applicable to lignin, the process presented has important implications for parameterization efforts of other polymers. The first implication is that we demonstrate that general force fields are often perfectly adequate to recreate structures of the polymer at the atomic level, as evidenced by the low RMSD when the structures are minimized (Table S6). Thus, significant progress can be made even without a tailored force field, and for lignin systems that we have not parameterized, would be a perfectly acceptable starting point for simulation. However, there are significant improvements in the energetic description of structural changes that come from explicit parameterization (Tables S3 and S4), which means that the significant cost to parameterization should only be borne when structural changes are expected. Since lignin structures are thought to be amorphous, implying that significant structural changes will occur regardless of starting structure, accurate energetics can guide modeled lignin towards native-like states. For crystalline polymer simulations, parameterization may not be required if structural changes are not the desired result.

On the parameterization front, the exhaustive balancing performed with a tunable objective function (Eq. S2) has significant implications on future parameterization efforts. While it is true that the energy residual always improves with more free parameters (Fig. S32A), this comes at the cost of frequently moving the minimum of the classical molecular mechanics surface along orthogonal degrees of freedom away from the true minimum energy structures (Fig. S6A). This means that automated processes that fit dihedrals run the risk of perturbing molecular structures unless specific steps are taken to reduce this risk, such

as not including higher order terms such as the n=4 or n=6 terms by default,[S15,S31] or even better using chemical intuition to determine what the logical periodicities should be.[S32]

Though not explored in depth here, bounds also play an important role in keeping the optimizer confined to an acceptably small search space,[S1,S15] which can be important for correctly describing interactions between different molecules. This is particularly true for the charges, where the optimizer can find creative solutions that are very dissimilar from other species in CHARMM if not appropriately guided through placing upper and lower limits on the optimization. Likewise, optimizers can override chemical intuition for dihedral terms if given the opportunity, similarly distorting structures during simulation.

What was surprising to us is the lack of improvement in the quality of the parameters if forces at minimum energy geometries were considered. Despite trying a number of different levels of strength for the parameter v in Eq. S2, there was little to no gain in quantifiable metrics of parameter performance. We suspect that rather than forcing the minimum energy structures to coincide, as we had hoped given the discussion around Fig. S3, the optimizer just scaled down the forces overall, including away from the minima. Reducing force magnitudes thereby perturbs the energetics, and does not actually cause the molecular mechanical and quantum mechanical potential energy minima to coincide. If forces were also considered when the structure is perturbed, perhaps adding in force information would be more successful. However, since the CHARMM corrections between quantum and molecular mechanical forces is not always straightforward, as we see in the scaling of water interaction energies,[S1,S2] this was not attempted here.

Finally, we see elements of our workflow being useful in other parameterization efforts, and is provided as two separate archives provided as Supporting Information. The code we use to generate the target data exposes to a wider research community the data-generation utilities of ffTK,[S1] enabling other researchers to better automate the tedious data acquisition required to start optimization. The optimization routines themselves also may find a use by others, particularly those who would like a monolithic optimization process to determine the

best fit across all of their compounds. This is doubly true for the GPU-accelerated bonded optimization function, which can be directly used or extended with minimal reconfiguration within other parameterization tools.

References

(S1) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *Journal of Computational Chemistry* **2013**, *34*, 2757–2770.

(S2) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *Journal of Computational Chemistry* **2010**, *31*, 671–690.

(S3) Knight, J. L.; Yesselman, J. D.; Brooks, C. L. Assessing the Quality of Absolute Hydration Free Energies Among CHARMM-compatible Ligand Parameterization Schemes. *Journal of Computational Chemistry* **2013**, *34*, 893–903.

(S4) Jämbeck, J. P. M.; Lyubartsev, A. P. Update to the General Amber Force Field for Small Solutes With an Emphasis on Free Energies of Hydration. *The Journal of Physical Chemistry B* **2014**, *118*, 3793–3804.

(S5) Lemkul, J. A.; Allen, W. J.; Bevan, D. R. Practical Considerations for Building GROMOS-Compatible Small-Molecule Topologies. *Journal of Chemical Information and Modeling* **2010**, *50*, 2221–2235.

(S6) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad

Coverage of Drug-Like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016**, *12*, 281–296.

(S7) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* **1983**, *79*, 926.

(S8) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *The Journal of Physical Chemistry* **1996**, *100*, 16502–16513.

(S9) Xu, Y.; Vanommeslaeghe, K.; Aleksandrov, A.; MacKerell, A. D.; Nilsson, L. Additive CHARMM Force Field for Naturally Occurring Modified Ribonucleotides. *Journal of Computational Chemistry* **2016**, *37*, 896–912.

(S10) Vanommeslaeghe, K.; Yang, M.; Mackerell, A. D. Robustness in the Fitting of Molecular Mechanics Parameters. *Journal of Computational Chemistry* **2015**, *36*, 1083–1101.

(S11) Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; Mackerell, A. D. Additive Empirical Force Field for Hexopyranose Monosaccharides. *Journal of Computational Chemistry* **2008**, *29*, 2543–2564.

(S12) Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages Between Hexopyranoses. *Journal of Chemical Theory and Computation* **2009**, *5*, 2353–2370.

(S13) Hatcher, E. R.; Guvench, O.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Aldopentofuranoses, Methyl-Aldopentofuranosides, and Fructofuranose. *The Journal of Physical Chemistry B* **2009**, *113*, 12466–12476.

(S14) Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate-Protein Modeling. *Journal of Chemical Theory and Computation* **2011**, *7*, 3162–3180.

(S15) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. *Journal of Chemical Theory and Computation* **2012**, *8*, 3257–3273.

(S16) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters* **2014**, *5*, 1885–1891.

(S17) Zheng, S.; Tang, Q.; He, J.; Du, S.; Xu, S.; Wang, C.; Xu, Y.; Lin, F. VFFDT: A New Software for Preparing AMBER Force Field Parameters for Metal-Containing Molecular Systems. *Journal of Chemical Information and Modeling* **2016**, *56*, 811–818.

(S18) Waldher, B.; Kuta, J.; Chen, S.; Henson, N.; Clark, A. E. ForceFit: A Code to Fit Classical Force Fields to Quantum Mechanical Potential Energy Surfaces. *Journal of Computational Chemistry* **2010**, *31*, 2307–2316.

(S19) Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *Journal of Chemical Theory and Computation* **2013**, *9*, 3543–3556.

(S20) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *Journal of Chemical Information and Modeling* **2012**, *52*, 3144–3154.

(S21) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *The Journal of Physical Chemistry B* **2017**, *121*, 4023–4039.

(S22) Slater, J. C. A Simplification of the Hartree-Fock Method. *Physical Review* **1951**, *81*, 385–390.

(S23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; J. A. Montgomery, J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01. 2013.

(S24) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **1934**, *46*, 618–622.

(S25) Cordella, L.; Foggia, P.; Sansone, C.; Vento, M. A (Sub)graph Isomorphism Algorithm for Matching Large Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2004**, *26*, 1367–1372.

(S26) Hagberg, A.; Swart, P.; Chult, D. Exploring Network Structure, Dynamics, and

Function Using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008). Pasedena, CA USA, 2008; pp 11–15.

(S27) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software* **1997**, *23*, 550–560.

(S28) Hopkins, C. W.; Roitberg, A. E. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *Journal of Chemical Information and Modeling* **2014**, *54*, 1978–1986.

(S29) Nickolls, J.; Buck, I.; Garland, M.; Skadron, K. Scalable Parallel Programming With CUDA. *ACM Queue* **2008**, *6*, 40–53.

(S30) Bell, N.; Hoberock, J. *GPU Computing Gems Jade Edition*; Morgan Kauffmann Publishers: Burlington, MA, 2011; Chapter 26, pp 359–371.

(S31) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics With the OPLS-AA Force Field. *Journal of Chemical Theory and Computation* **2015**, *11*, 3499–3509.

(S32) Bartell, L. S. Representations of Molecular Force Fields. 3. Gauche Conformational Energy. *Journal of the American Chemical Society* **1977**, *99*, 3279–3282.