NREL
NATIONAL RENEWABLE ENERGY LABORATORY

NREL is a national laboratory of the U. S. Department of Energy, Office of Energy Efficiency and Renewable Energy, operated by the Alliance for Sustainable Energy, LLC.

Systematic parameterization of lignin for the CHARMM force field

Josh Vermaas[1], Loukas Petridis[2], Gregg Beckham[1], and Michael Crowley[1]
[1]National Renewable Energy Laboratory
[2] Oak Ridge National Laboratory

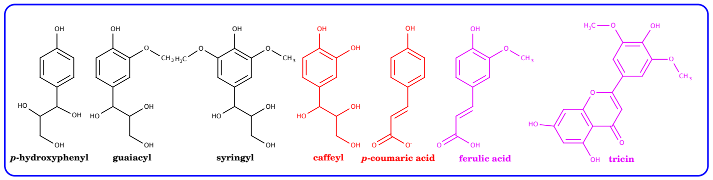
Why Lignin?

Plant lignin is the largest source of aromatics in the biosphere, composing approximately 10-30% of plant biomass dry weight, making it an attractive carbon source for industrial processes to make fuels and chemicals

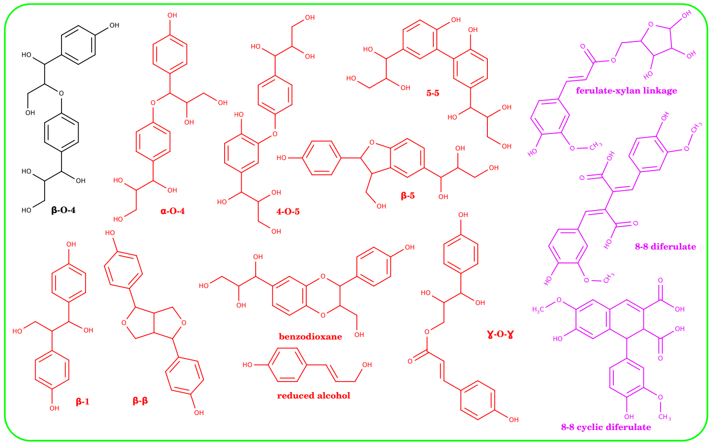
Lignin is a branched heteropolymer formed through radical chemistry, and as a result, the structure on the atomic level is not well understood due to the combinatorics of assembling the diverse complement of linkages seen in nature

By developing a complete force field for lignin that incorparates all the linkages both within lignin and to hemicellulose, we enable subsequent investigations of biomass structure including natural lignin biodiversity

Monomers



Linkages and Modifications



*Examples of lignin monomers and linkages covered in the new force field, highlighting **monomers** and **dimers** that were covered by previous force fields (**black**), are new lignin linkages not previously included (**red**), and are linkages to hemicellulose (**violet**).*

Parameterization Approach

$$U_{non-bonded} = U_{VDW} + U_{electrostatic}$$

$$= \sum_{i,j \in pairlist} \boxed{\epsilon_{ij}} \left(\left(\frac{\boxed{R_{ij}^{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{\boxed{R_{ij}^{min}}}{r_{ij}} \right)^{6} \right) + \frac{\textcircled{q_i}\textcircled{q_j}}{4\pi\epsilon_0 r_{ij}}$$

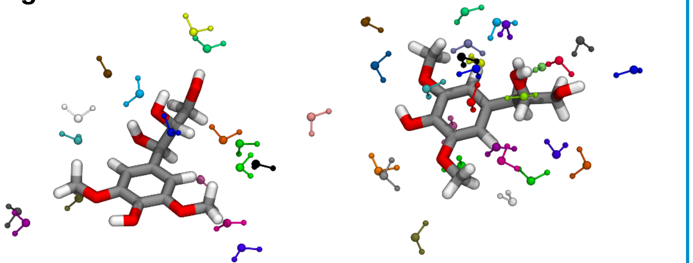
$$U_{bonded} = U_{bonds} + U_{angles} + U_{dihedrals} + U_{impropers}$$

$$= \sum_{i \in bonds} \textcircled{k_i} \left(b_i - \textcircled{b_0}\right)^2 + \sum_{j \in angles} \textcircled{k_j} \left(a_j - \textcircled{a_0}\right)^2$$

$$+ \sum_{k \in dihedralterms} \textcircled{k_k} \left(1 + \cos\left(\textcircled{n_k}\chi_k + \textcircled{\delta_k}\right)\right) + \sum_{l \in impropers} \boxed{k_l} \left(\chi_l - \boxed{\chi_0}\right)^2$$

All tunable parameters in the CHARMM force field are either boxed (taken directly from CGenFF) or circled (reparameterized based on fitting the molecular mechanics potential energy function to generated quantum mechanical target data)

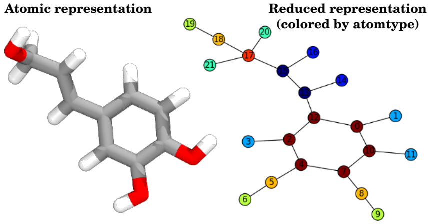
Charge Parameterization

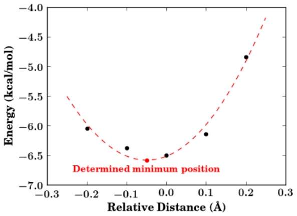


In CHARMM, charge parameterization starts with computing the interaction energy with water to every potential donor or acceptor on the molecule (in this case a single syringyl unit). The water can act either as an proton acceptor (left) or a donor (right), and its position is optimized to determine the optimal interaction distance. Additionally, the dipole of each molecule at the QM level is used to compare with the dipole that results from our MM results. Across all species, this involved setting up 5319 calculations in Gaussian 09.

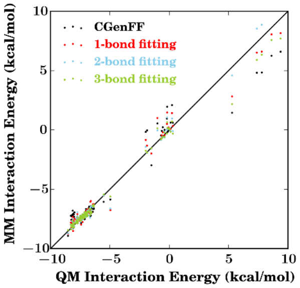
$$\underset{\text{Minimal charge vector}}{f(\bar{q})} = \sum_{species} \left(\underbrace{\frac{1}{N_{species}} \left(D_{QM} - D_{MM}(\bar{q})\right)^2}_{\text{Dipole term}} + \sum_{sites} \left[\underbrace{\left(\frac{E_{QM}^{int} - E_{MM}^{int}(\bar{q})}{0.2\,\text{kcal/mol}} \right)^2}_{\text{Energy term}} + \underbrace{\left(\frac{d_{QM}^{int} - d_{MM}^{int}(\bar{q})}{0.1\,\text{Å}} \right)^2}_{\text{Distance term}} \right] \right)$$

The objective function that we are minimizing has three terms. The dipole term is scaled by the number of atoms in the molecule, and takes into account both the magnitude and direction of the dipole vector. The energy term compares the interaction energies computed via QM to the scaled molecular mechanics interaction energies based on the current distribution of charge. In order to obtain a gradient for the distance term, slightly offset energy vs. distance plots were fit to a quadratic to determine the apparent molecular mechanics minimum energy interaction distance.





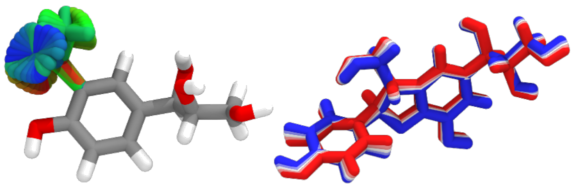
Charges need to be consistent across molecules. To enforce this, charges were constrained to be equal if their chemical environment is the same, as judged by the atomtypes within a 1, 2, or 3 bond neighborhood. In this example with a reduced caffeyl alcohol, atoms 1 and 3 have identical neighborhoods within 1 or 2 bonds, and would be forced to have the same charge. 3 bonds away, the topologies are different, and so the two atoms could have different charges. These charge neighborhoods were assigned and compared across all molecules.



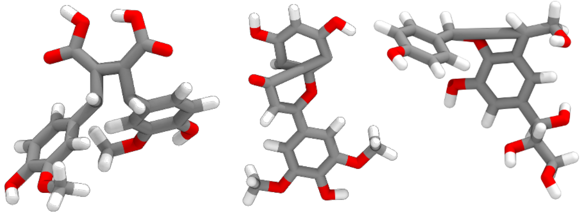
	Small E ($E^{int} < 0\,\text{kcal/mol}$)		Larger E ($E^{int} < 20\,\text{kcal/mol}$)	
Charge Scheme	$\left\langle \left(E_{QM}^{int} - E_{MM}^{int}\right)^2 \right\rangle^{\frac{1}{2}}$	R^2	$\left\langle \left(E_{QM}^{int} - E_{MM}^{int}\right)^2 \right\rangle^{\frac{1}{2}}$	R^2
CGenFF	0.5394 kcal/mol	0.95230	0.7768 kcal/mol	0.95714
Revised 1-bond	0.3906 kcal/mol	0.97376	0.4712 kcal/mol	0.98309
Revised 2-bond	0.2907 kcal/mol	0.98289	0.3583 kcal/mol	0.99238
Revised 3-bond	0.2077 kcal/mol	0.99134	0.4144 kcal/mol	0.98938

No matter how many bonds we look around to group charges, we improve the fit between QM and MM interaction energies relative to CGenFF by doing the parameterization ourselves. These fits could be further improved with additional charge groups at the cost of increased complexity in assembling complete lignin polymers. Due to this complexity, we advocate using the 1-bond fits.

Bonded Optimization



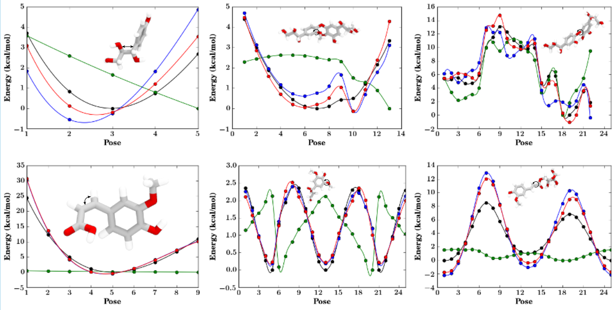
Bonded term parameterization depends on reproducing in a molecular mechanics setting the results of individual quantum mechanical bond, angle or dihedral scans around the initial minimum energy geometry, with examples shown above. There are 16637 optimized single point geometries calculated this way, not including scan results that needed to be excluded due to geometric irregularities (primarily topological rearrangment), such as those shown below.



Programmatically, these errant structures are identified by testing for non-hydrogens that are within 1.65Å of other atoms but are not bonded to them. Particularly long bonds whose length is greater than 1.65Å are also excluded.

$$f(\bar{p}) = \sum_{scans} \sum_{poses} w_{pose} \left(E_{pose}^{QM} - E_{pose}^{MM}(\bar{p}) - C_{scan}\right)^2$$

The objective function we are minimizing is very similar to that for the charges. However, the number of degrees of freedom is much larger (101 independent charges for the 1-bond fit, vs 3824 unknowns across all parameters), and many terms need to be recomputed as the parameters are adjusted. To deal with the slow function and gradient evaluations, this was written into a CUDA-aware python library, speeding up the function and gradient evaluations by up to 100x



*Using the gradient to the objective function, we can try to reproduce the **QM energy profiles** across the scans by adding bonded terms to the **electrostatic and van der Waals energy profile** for the same poses. The **unrestricted fits** are slightly better than **fits where we use the same terms as found in CGenFF**. The next stages are to see which terms can be safely eliminated without impacting the quality of the fit, and which atomtypes need to be split due to conflicting scans.*

This work was supported by the Director's Fellowship Program, which is part of the Laboratory Directed Research and Development (LDRD) Program at the National Renewable Energy Laboratory. NREL is a national laboratory of the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy operated by the Alliance for Sustainable Energy, LLC..

Biophysical Society 61st Annual Meeting, New Orleans, Louisiana
February 11, 2017
NREL/PO-2700-67929