# Supporting Information: Escaping atom types in force fields using direct chemical perception

David L. Mobley,*,† Caitlin C. Bannan,‡ Andrea Rizzi,¶ Christopher I. Bayly,§ John D. Chodera,‖
Victoria T. Lim,‡ Nathan M. Lim,⊥ Kyle A. Beauchamp,# David R. Slochower,@ Michael R. Shirts,△
Michael K. Gilson,@ and Peter K. Eastman▽

†*Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine*
‡*Department of Chemistry, University of California, Irvine*
¶*Tri-Institutional Program in Computational Biology and Medicine, New York NY 10065*
§*OpenEye Scientific Software, Santa Fe, NM 87507*
‖*Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York NY 10065*
⊥*Department of Pharmaceutical Sciences, University of California, Irvine*
#*Counsyl, South San Francisco, CA 94080*
@*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego*
△*Department of Chemical Engineering, University of Colorado, Boulder*
▽*Department of Chemistry, Stanford University, Stanford, CA 94305*

E-mail: dmobley@mobleylab.org
Phone: 949-824-6383

## Abstract

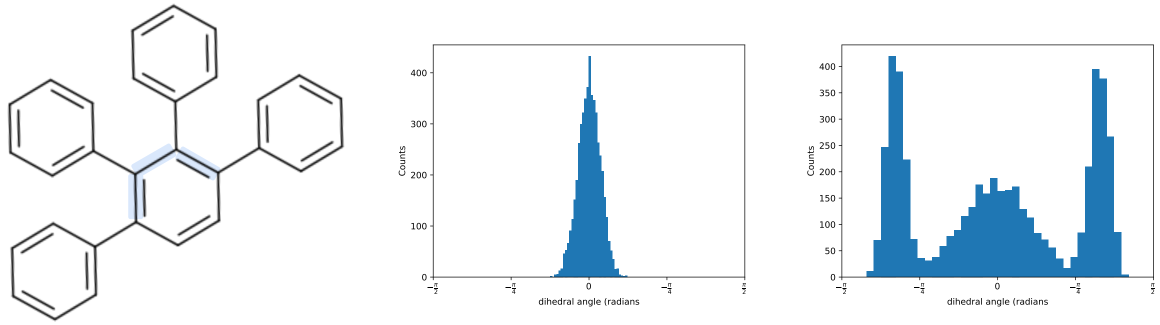
Supporting information

For the torsion distributions shown in these plots, simulations were relatively short gas-phase simulations with OpenMM, as detailed in our online supporting files on GitHub (`https://github.com/mobleylab/smirnoff_paper_code`, specifically in the `selected_molecules` directory and the python scripts therein; they consisted of 5 ns in the gas phase at 300K with full details given in the scripts in the aforementioned directory.

## 1 Aromaticity models and the MDL aromaticity model employed here

The choice of aromaticity model is an important part of chemical perception, and is particularly important for perceiving fragments using SMIRKS patterns. In particular, a SMIRKS pattern for an aromatic bond between two aromatic atoms can only match a bond which is determined to be "aromatic" and not a formal double or single bond. While this is desired for many kinds of aromatic six-membered rings such as benzene and pyridine, it is a liability for five-membered heteroaromatic rings such as imidazole or oxazole. In six-membered rings, such as benzene and pyridine, there is no significant difference between the formally drawn single and double bonds. For those molecules all six bonds should often be treated identically. But this is not necessarily the case for five-membered rings. Consider imidazole as a specific example: Empirical evidence shows that the bonds between carbon and nitrogen are not all identical; some have more single bond character while others more closely resemble double bonds. The Supporting Information provides more detailed analysis of both of these possibilities. If the bonds in imidazole were all perceived as aromatic, the SMIRKS required to correctly describe the differences in these valence terms would be significantly more complex, most likely incorporating all atoms in the ring. If the bonds in the five-membered rings are identified as formal single or double bonds then the SMIRKS only require information about the atoms and bonds directly involved in the valence term.

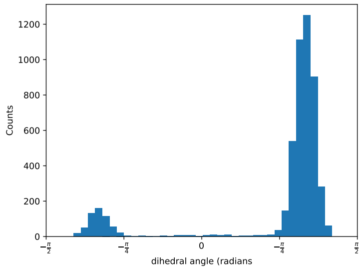
Because of the simplification possible if we treat many five-membered rings as alternating single and double bonds, we selected the MDL aromaticity model for use with SMIRNOFF99Frosst — specifically, OpenEye Scientific's OEChem Toolkit implementation of the



(a) 1,2,3,4-tetraphenylbenzene torsion

(b) smirnoff99frosst

(c) GAFF



(d) GAFF2

Figure 1: **Torsional errors in 1,2,3,4-tetraphenylbenzene.** As discussed in the main text (and shown in Figure 1), 1,2,3,4-tetraphenylbenzene has incorrect torsional barriers in GAFF and GAFF2, resulting in buckling of the central ring in gas phase dynamics simulations. Shown are torsional plots for the central torsion of the bonds highlighted in blue in (a), from 10 ns of dynamics in the gas phase. (b), (c), and (d) show the distribution of torsional angles observed in smirnoff99Frosst, GAFF, and GAFF2, respectively. In GAFF and GAFF2, the torsion in question is effectively rotatable, whereas in smirnoff99Frosst it is not.

MDL model (here, "OpenEye MDL" for short) [a]. The OpenEye MDL model assigns aromaticity based on a limited definition of Hückel's $4n+2$ rule; the only pi-electrons counted are those coming from double bonds in the ring system. This means the electrons on hetero-atoms or anions in five-membered rings are not counted as contributing to the ring system. In other words, the only atoms and bonds perceived as aromatic with this model are those in ring systems with an odd number of alternating double bonds. This includes six membered rings with alternating single bonds such as benzene or pyridine and also fused systems such as azulene (a fused seven-membered and five-membered ring system). The ForceField class (and the SMIRNOFF format) can optionally perceive aromaticity (via specification of an aromaticity model) or use a molecule as provided, but SMIRNOFF99Frosst was designed assuming the use of OpenEye's MDL aromaticity model and specifies this in its XML. RDKit has this same OpenEye MDL aromaticity model available in versions 2017.09.3 and later, and an RDKit-based implementation of the ForceField class is under development.

## 2 Reproducing parm@Frosst energies on AlkEthOH for problematic molecules

As discussed in the main text, we had problems reproducing parm@Frosst energies on molecules containing the H1, H2, and H3 atom types in specific environments due to human error in parm@Frosst torsional parameters. This resulted in us needing to introduce new, more specialized SMIRKS patterns for torsions to reproduce bugs in parm@Frosst, as highlighted in the main text.

---

[a] OpenEye?s OEChem Toolkit implementation is a modified version of the original MDL model. Originally, the MDL model considered two types of aromatic groups: six-membered rings with alternating single/double bonds including the perimeter bonds in azulene, and five-membered rings with two double bonds and hetero atoms or a carbon anion at the ring apex. The OpenEye implementation uses a more limited definition of aromaticity including only the former group. That is to say, only six-membered rings with alternating single and double bonds and azulene are considered aromatic. The five-membered rings then have formally assigned single or double bonds.

In Figure 7(a), we had to introduce a specialized SMIRKS pattern `[#1:1]-[#6X4:2](-[#8])-[#6X4:3]-[#6X4:4]` to reproduce parm@Frosst energies on this set because parm@Frosst applies the generic `X -CT-CT-X` torsion (barrier height $1.4/9$ kcal/mol) to all `H*-CT-CT-CT` torsions *except* `HC-CT-CT-CT` (parm@Frosst barrier height $0.16/3$ kcal/mol). We also had to introduce a specialized SMIRKS pattern `[#1:1]-[#6X4:2]-[#6X4:3](-[#8])-[#1:4]` and associated parameters to reproduce parm@Frosst energies because parm@Frosst applies the generic `X -CT-CT-X` torsion (parm@Frosst barrier height $1.4/3$ kcal/mol) to all `H*-CT-CT-H*` torsions *except* `HC-CT-CT-HC` (parm@Frosst barrier height $0.05$ kcal/mol). In Figure 7(b), the same two patterns also occur, but we additionally had to introduce the specialized SMIRKS pattern `[#1:1]-[#6X4:2](-[#8])(-[#8])-[#6X4:3]-[OX2:4]` to reproduce parm@Frosst energies where the `X -CT-CT-X` torsion is applied to all `H2,H3-CT-CT-OH,OS` torsions (where the comma denotes an "or"). In this case, `HC,H1-CT-OH,OS` torsions get use two terms, with a periodicity 3 barrier height of exactly zero and a periodicity 1 barrier height of $0.25$ kcal/mol, whereas the generic has a single term with a periodicity 3 barrier height of $1.4/9$ kcal/mol. These three problematic torsions occur for many molecules in our set, and all three issues persist in GAFF and in beta versions of GAFF2 (such as the AmberTools 16 and 17 versions examined here). summarizes the conclusions about GAFF – no need to address specifically since it's exactly the same bugs, just (in some cases) very slightly different barrier heights. These issues appear to be a case of human error, where introduction of new derivative atom types (`H2` and `H3` are derivatives of `H1`) was not accompanied by reproducing or refitting some of the relevant torsional potentials. Thus, to reproduce parm@Frosst energies we actually had to create two versions of our AlkEthOH force field – a minimal one which reproduces parm@Frosst as it was intended for these molecules, and a more extensive one which also reproduces these three bugs in parm@Frosst.

## 3 Hydroxyl radius adjustments

As part of developing SMIRNOFF99Frosst, we wanted to add hydroxyl radii to polar hydrogens, in deviation from AMBER force fields, partly because we had previously observed that having multiple hydroxyls in close proximity to one another and other polar atoms (such as oxygens) could result in crashes in some situations. For example, both DLM and CIB had previously encountered systems for which simulations which inevitably crashed due to the lack of nonzero Lennard-Jones parameters for polar hydrogens.

Our first iteration of SMIRNOFF99Frosst did not include these parameters, and kept the zero LJ parameters for polar hydrogens that are used in AMBER-family force fields. However, the extensive benchmarking for this paper seemed to provide an ideal opportunity to validate a solution to this known issue, so we added nonzero hydrogen LJ parameters for all hydroxyl hydrogens for our benchmarking for this study.

The hydroxyl radius selected was chosen to ensure that Lennard-Jones interactions involving polar hydroxyl hydrogen atoms (SMIRKS `[#8X2H1+0:1]`) are perturbed as small as possible; full details of our analysis for this are given in our `SMIRNOFF_paper_code` GitHub repository (`https://github.com/mobleylab/SMIRNOFF_paper_code`) and especially in the `polar_hydrogens` directory. But briefly, we initially chose a $r_{min}/2$ for hydroxyl hydrogen of 0.3 angstroms, half of that used for amide nitrogens and for `HX` in parm@Frosst. We then chose the $\epsilon$ parameter such that combined interactions between hydrogen and iodine (chosen since it is one of the atoms with the largest $\epsilon$ parameter) would be less than 0.1% of $k_BT$ outside the radius of the oxygen atom which protects the hydrogen, yielding a final proposed value of $0.00527$ kcal/mol for the well depth. To test whether this would significantly perturb interactions, we considered an artificially extreme example and simulated TIP3P water as normal, and compared with a comparable three-point water model but with these parameters artificially applied to the hydrogens (though we have no intention of applying these parameters to water, this seemed to be a good overly pessimistic test of how these parameters might perturb a system). With these parameters, we found that the density of water would run about 0.2% lower than it otherwise would. We concluded this was too significant a perturbation, and thus reduced the $\epsilon$ value by another order of magnitude, to $5.27e-4$ kcal/mol, and repeated our density calculation, now finding that the perturbation to density was roughly $0.04\%$, which seemed tolerable (since it is far less than the variation across water models).

We then turned to check that this difference was substantial enough to alleviate any problematic interactions, so we looked at the association of neutral acetic in the gas phase (see `Dimer energetics.ipynb` in supporting code). Consider an acetic acid dimer without nonzero LJ parameters on the polar hydrogens, and consider a hypothetical configuration where the carboxylic acid protons rest on top of the center of the oxygen atom in the adjacent molecule of the dimer. In such a configuration, the steric repulsion between the hydroxyl oxygen and the oxygen in the alternate member of the dimer is quite large, but finite, whereas the electrostatic energy is $-\infty$ because a positive charge sits directly on top of a negative charge. While it might be difficult to *reach* this configuration due to steric repulsion between oxygen atoms as the hydroxyl hydrogen approaches the other oxygen, this must be the true minimum energy configuration for such a dimer. And, in fact, that is exactly what we find here – a large unfavorable energy tends to resist forming of such a complex, but once the separation between atoms becomes small enough, the attractive electrostatic energy diverges and the complex forms. In contrast, with our new proposed hydroxyl LJ parameters, the energy of the complex grows steadily more unfavorable as the separation distance decreases, as it should. Data is shown in Figure 3.

Following this analysis, we concluded that our hydroxyl LJ parameters were sufficient to remove the problem, so we proceed to validate the full SMIRNOFF99Frosst on density and dielectric constants and hydration free energies as discussed in the main text.

We also ran all hydration free energy calculations with the original hydroxyl hydrogen Lennard-Jones parameters and cross-compared results. Plots are shown below; no significant perturbations to hydration free energy were observed with the new hydroxyl parameters, so these were formally incorporated into SMIRNOFF99Frosst. We also ran (data not shown) a full set of density and dielectric constant calculations with the earlier version of SMIRNOFF99Frosst and found no significant perturbations to the density/dielectric constant based on introduction of the hydroxyl Lennard-Jones parameters.

## 4 Parameters imported from other force fields to improve coverage

After initially generating smirnoff99Frosst, we found that its coverage of general chemistry was still less than ideal in tests on diverse chemistry such as DrugBank. Therefore we adjusted some SMIRKS patterns to improve generality, and also imported several parameters from GAFF2. These changes included:

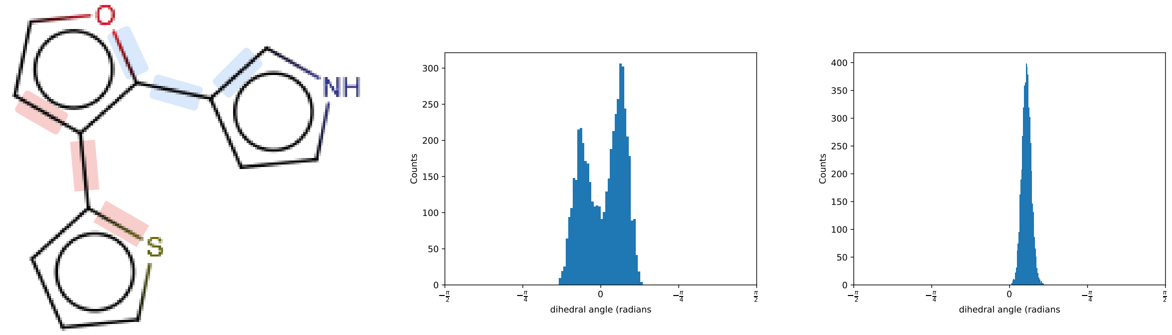
- **Bonds**
  - Remove neutral requirement in the bonds to hypervalent sulfur by adjusting the SMIRKS pattern
  - Add a `C=S` bond from parm@frosst which had been dropped
  - Remove the `X1` requirement for nitrogens in nitriles
  - Remove the `X4` requirement for the most generic `S-C` bond
  - Add `C-halogen` bonds to cover all carbons
  - Remove `X2` requirement for sulfur in `S-H` bonds
  - Make some bonds more generic to address other corner cases, such as the `X3` requirement for the most generic `N-N` bonds to cover cases involving positively charged nitrogens; change `[#6X4:1]-[#8X2:2]` to `[#6:1]-[#8:2]` to cover corner cases; remove `X2` on carbon in nitriles to cover charged corner cases
  - Change order of carbonyl bonds to properly type carbonyl groups
  - Change the first carbonyl group to be more general for groups of the type `X=C=O`
- **Angles**
  - Add a slightly more generic `#7X2` angle for linear bonds
  - Make the `X-N-O` angle in nitro groups generic enough that `X` could be carbon, oxygen or nitrogen
  - Update the most generic divalent sulfur to accommodate aromatic cases
- **Tosions**
  - Update aromatic nitrogen to cover all generic aromatic bonds involving nitrogens connected to carbons
  - Remove neutral requirement in bonds to hypervalent sulfur
  - Add torsion for nitro group tetrahedral carbons based on parm@Frosst and GAFF2
- **Parameters estimated from GAFF2**
  - `S-P` bond stretching parameters
  - `H-P` bond stretching parameters
  - `N-halogen` bond stretching parameters

  - `P-halogen` bond stretching parameters
  - `S-halogen` bond stretching parameters
  - Torsion where central atoms are two singly bonded oxygens
  - Sulfur-carbon torsion where the central bond is double or aromatic
  - `P-N` torsions

A GitHub pull request which discusses all of these changes (and made the changes) and links to more detailed discussion of them on the `smarty` repository is available at `https://github.com/openforcefield/smarty/pull/232`.

## 5 Other material

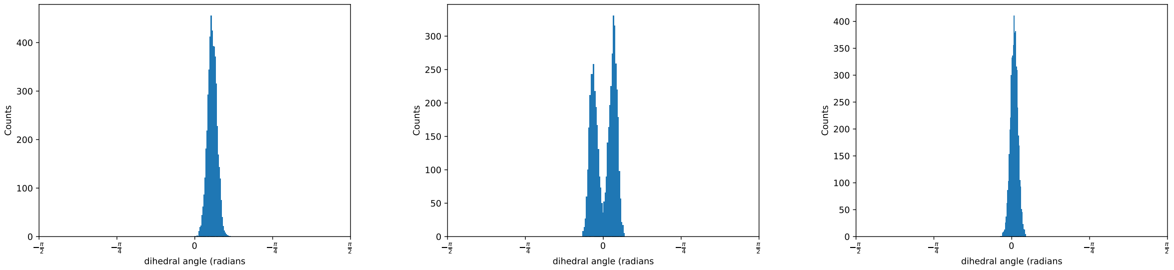
PDF files containing a visualization of the full 1500 molecule AlkEthOH set are available in the additional electronic supporting information, but are not reproduced in this document for space reasons.



(b) smirnoff99frosst, blue

(c) GAFF, blue

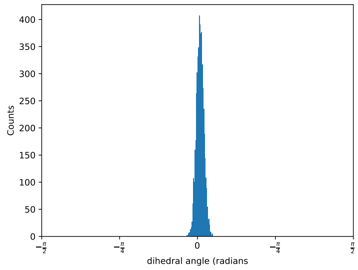
(a) bridgehead torsions



(d) GAFF2, blue

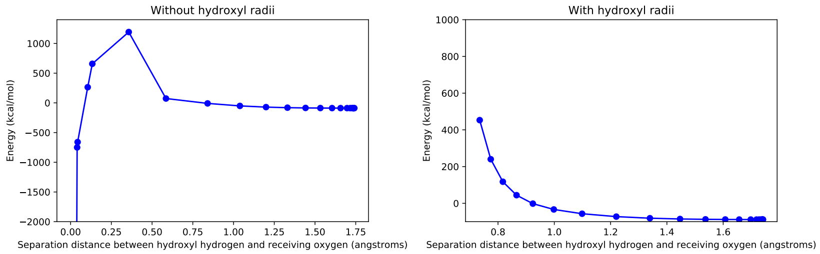
(e) smirnoff99frosst, red

(f) GAFF, red



(g) GAFF2, red

Figure 2: **Torsional errors in the bridgehead problem case from the main text.** As discussed in the main text (and shown in Figure 2e), our sample bridgehead problem case has incorrect torsions for the bonds connecting bridgehead atoms in GAFF and GAFF2, with the central torsions for the bonds in blue and red bonds being treated as non-rotatable in GAFF and GAFF2. Here we show torsional distributions measured from 10 ns of dynamics in the gas phase in smirnoff99Frosst, GAFF, and GAFF2 for the highlighted torsions. (b), (c), and (d) show distributions for the torsion highlighted in blue; in GAFF and GAFF2, this has to be slightly out of plane due to a steric clash with the planar sulfur-containing ring, but the ring is never able to rotate and flip, whereas in smirnoff99Frosst (b), the ring flips back and forth between orientations. (e), (f), and (g) show distributions for the torsion highlighted in orange; in GAFF and GAFF2, the sulfur-containing ring is co-planar with the oxygen-containing ring, whereas in smirnoff99Frosst it is tilted slightly out of plane and the bond is rotatable so the distribution is bimodal.



(a) Without hydroxyl LJ parameters (b) With hydroxyl LJ parameters

Figure 3: **Dimerization of acetic acid** As discussed, we examined dimerization of neutral acetic acid; shown are energies of minimum energy configurations for conformations of acetic acid with the hydrogen-oxygen distances restrained to zero with steadily increasing force constants (resulting in smaller and smaller separations), with the energy of the restraints turned off. This allows us to find minimum energy configurations at varying separations. (a) shows energy for acetic acid dimers without nonzero hydroxyl hydrogen LJ parameters, and (b) shows the energy for the same system with the new proposed LJ parameters. Both tests use the same restraints; in (a), once the separation becomes small enough, the hydroxyl hydrogen is pulled on top of the oxygen atom in the neighboring molecule in the dimer due to strong electrostatic interactions, and the energy becomes negative and infinite, whereas in (b) the energy instead grows steadily less favorable and the restraints are not able to make these atoms overlap.

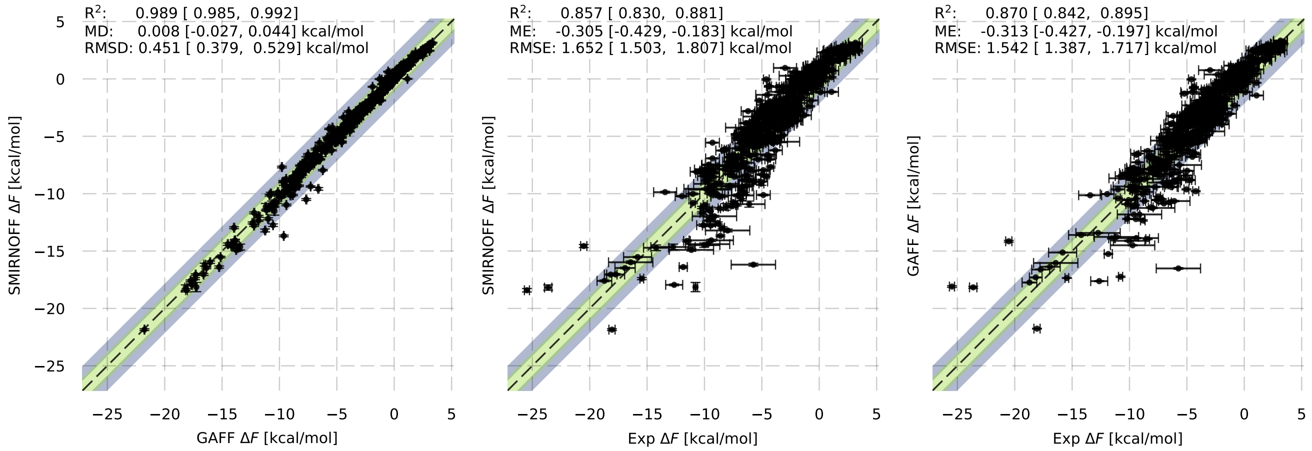


Figure 4: **Hydration free energies for FreeSolv from GAFF and SMIRNOFF99Frosst before applying nonzero LJ parameters.** Shown are hydration free energies computed for the FreeSolv set with GAFF (from previous work[?] ) and with SMIRNOFF prior to introduction of nonzero Lennard-Jones parameters for hydroxyl hydrogens, in this work. This is analogous to Figure X of the main text but with an older version of SMIRNOFF99Frosst. No statistically significant differences in performance are observed here relative to Figure X. The **left panel** shows SMIRNOFF99Frosst versus GAFF (left), the **middle panel** shows SMIRNOFF99Frosst versus experiment, and the **right panel** shows GAFF versus experiment. Statistics, with bootstrapped uncertainties representing 95% confidence intervals, are shown at the top of each panel. Here, the mean difference between SMIRNOFF99Frosst and GAFF is statistically indistinguishable from zero (left panel) though there is a significant discrepancy based on the RMS difference. However, compared to experimental values, the coefficient of determination $R^2$, mean error, and RMS error for GAFF and SMIRNOFF are within confidence intervals of one another (middle and right panels) indicating that the performance of SMIRNOFF99Frosst is essentially comparable on this dataset.