Improving Small Molecule Force Fields by Identifying and Characterizing Small Molecules with Inconsistent Parameters

Jordan N. Ehrman (ORCID: 0000-0002-0150-082X)[1], Victoria T. Lim (ORCID: 0000-0003-4030-9312)[2], Caitlin C. Bannan (ORCID: 0000-0003-2777-1174)[2], Nam Thi[1], Daisy Y. Kyu[1], David L. Mobley (ORCID: 0000-0002-1083-5533)[1,3]

[1]Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, California 92697, United States; [2]Department of Chemistry, University of California, Irvine, Irvine, California 92697, United States

***For correspondence:**
dmobley@mobleylab.org (DM)

Abstract Many molecular simulation methods use force fields to help model and simulate molecules and their behavior in various environments. Force fields are sets of functions and parameters used to calculate the potential energy of a chemical system as a function of the atomic coordinates. Despite the widespread use of force fields, their inadequacies are often thought to contribute to systematic errors in molecular simulations. Furthermore, different force fields tend to give varying results on the same systems with the same simulation settings. Here, we present a pipeline for comparing the geometries of small molecule conformers. We aimed to identify molecules or chemistries that are particularly informative for future force field development because they display inconsistencies between force fields. We applied our pipeline to a subset of the eMolecules database, and highlighted molecules that appear to be parameterized inconsistently across different force fields. We then identified over-represented functional groups in these molecule sets. The molecules and moieties identified by this pipeline may be particularly helpful for future force field parameterization.

0.1 Keywords
Molecular Mechanics simulations · Force Fields · Geometry Optimization · Molecular Modeling · Conformer Comparison

0.2 Abbreviations
FF Force field
QM Quantum Mechanical
TFD Torsion Fingerprint Deviation
RMSD Root-Mean-Square Deviation
MMFF Merck Molecular Force Field
GAFF General AMBER Force Field
SMIRNOFF SMIRKS Native Open Force Field; here, also typically used as shorthand for the SMIRNOFF99Frosst force field version 1.0.8

1 Introduction
Molecular simulations are widely used in drug design, materials design, and in the study of biophysical processes. Large systems, like biomolecules or even small molecules in solution, prove to be computationally difficult to simulate at the quantum mechanical (QM) level of theory. For this reason, classical empirical potential energy functions known as force fields are often

used in place of quantum mechanics in order to efficiently simulate chemical and biological systems. General small force fields, such as the general AMBER force fields GAFF and GAFF2 [39–41], OPLS [17, 23], CGenFF [36, 37], and the Merck molecular force fields MMFF94 and MMFF94S [10–16], were built to model a wide variety of small organic molecules. These force fields are often fit to attempt to reproduce energies and geometries observed in QM calculations. However, when applied to new molecules, they have been observed to differ from both quantum mechanical calculations and from each other in predicted energies and optimized geometries for important areas of chemical space [3, 7, 27, 33].

In the present study, we aimed to identify regions of chemical space where parameterization differences between force fields lead to different optimized geometries for small drug-like molecules in the gas phase. Geometric differences between force fields for some molecules would indicate that the underlying force fields describe the molecule differently, and thus are indicative of force field differences. Here, a subset of molecules from the eMolecules database [5] was used as a broad sample of small molecule chemical space. Five energy minimizations were performed on each molecule using one of five force fields: GAFF, GAFF2, MMFF94, MMFF94S, and the Open Force Field Initiative's SMIRNOFF99Frosst [27]. Two geometric measurements, Torsion Fingerprint Deviation [31] (TFD) and TanimotoCombo [18], were used to better identify meaningful geometric differences that may suggest parameterization inconsistencies.

One key assumption in our work is that large geometric differences in optimized geometries tend, overall, to be indicative of substantial differences in the underlying force fields. In other words, we operate with the belief that differences in force fields which are substantial enough to result in large differences in optimized geometries are interesting to force field developers. This assumption does not mean that such force field differences are necessarily large; indeed, small force field differences can result in large differences in optimized geometries [6, 27, 33]. This is because many organic molecules have a large number of conformational minima often separated by relatively small barriers, so small force field differences may cause a molecule to optimize into different minima. Rather, we assume that force field differences which are large enough to substantially alter optimized geometries are of interest, even if the force field differences themselves are relatively small. All minimizations were performed with the same starting structure to ensure that differences observed are as attributable as possible to differences in force fields.

In part, our work is motivated by the Open Force Field Initiative (OpenFF), which seeks to develop open data sets and infrastructure which can be used to produce new force fields which improved accuracy. It recently released an initial prototype force field, SMIRNOFF99Frosst [27] and, given our connection with OpenFF, SMIRNOFF99Frosst is one focus of our testing in the present study.

By identifying particular functional groups or substructures that lead to drastically different geometrically optimized conformers, we will have identified a portion of chemical space that is inconsistently parameterized by the gamut of force fields studied, and thus is likely to be inaccurately described by at least some of these force fields. In the future, these molecules could be prioritized when training new force fields through inclusion in QM reference calculations or searches for new experimental data.

2 Results and Discussion

In this study, we aimed to identify portions of small molecule chemical space which are particularly informative for force field development. After filtering eMolecules as described in Section 3.3, we were left with 2.7 million molecules. We optimized each of these molecules with each of the five force fields considered – GAFF, GAFF2, MMFF94, MMFF94S, and SMIRNOFF99Frosst [10–16, 27]. For any given molecule, we performed pairwise comparisons of these five minimized conformers, yielding ten comparisons that we here call "molecule pairs" (though each member of a molecule pair is actually the same molecule in different conformations). Each of the molecule pairs was evaluated for geometric differences using Torsion Fingerprint Deviation (TFD) [32] and TanimotoCombo [18]. We limited our analysis to molecules having 25 or fewer heavy atoms. Furthermore, we restricted our analysis to molecule pairs which yielded a TFD value less than 0.60 and a TanimotoCombo value between 0.25 and 2.0. These cutoffs were chosen based on visual inspection, as explained in detail in Section 3. Last, we sort molecules into different sets, which were then characterized using the Checkmol [8, 9] functional group identification tool.

Here, we chose TFD and TanimotoCombo, rather than the more common RMSD, as key metrics for this analysis. The primary trouble with RMSD is that it is highly dependent on molecular size. For example, a value of 1.0 Åmight correspond to a very large geometric difference for an extremely small molecule (e.g. butane) but a trivial geometric difference for a large, drug-like molecule (e.g. lipitor). Both TFD and TanimotoCombo are dimensionless numbers covering a well defined scale (TFD from 0 to 1; TanimotoCombo from 0 to 2) allowing us to define similarity and difference flags which are independent of molecular size. As

described above, these metrics also track well with the qualitative structural differences we hope to identify in molecule pairs. While RMSD also captured some of these differences, its size dependence makes it impractical for surveying a wide variety of molecules.

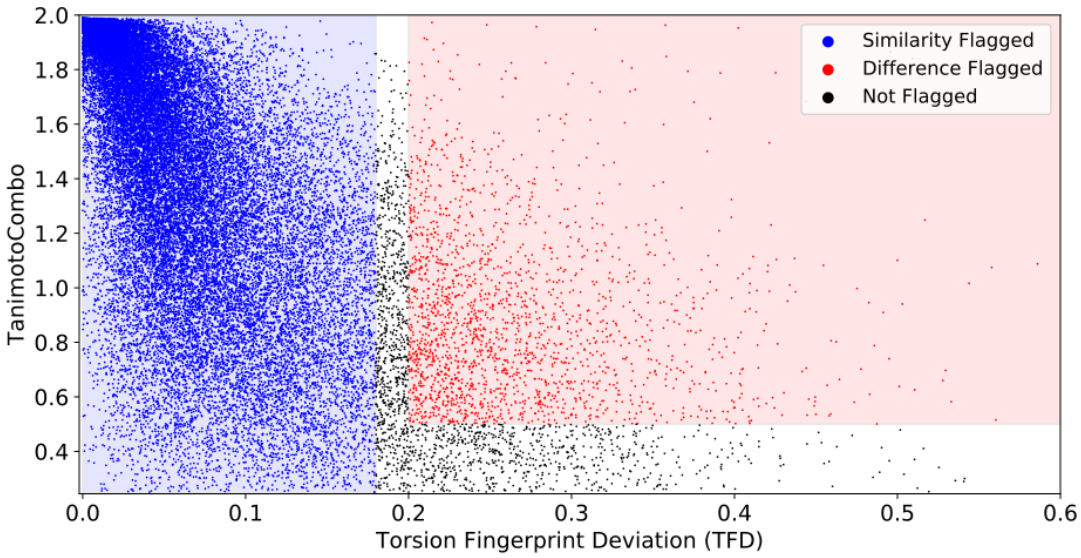

Figure 1. The vast majority of molecule pairs are geometrically similar by TanimotoCombo and TFD. Each point on this graph represents a molecule pair, i.e., a pair of structures of the same molecule, geometrically optimized with different force fields. The points are plotted at the resulting TFD and TanimotoCombo scores of the molecule pair. If the two minimized structures are identical, we would expect a Tanimoto Combo score of 2.0 and a TFD score of 0.0. For the purposes of this project, we flag molecule pairs yielding a TFD score above 0.2 and a TanimotoCombo score above 0.5 as being informatively different. This region is shaded red on the graph. Pairs judged as similar are shaded blue; the white region is included in neither category to avoid extreme sensitivity to choice of cutoff. This graph displays a random sample of 38,880 molecule pairs out of the total of 26,984,560 molecule pairs analyzed in this project.

2.1 Molecule pairs were flagged as similar or different based on TFD and TanimotoCombo

We used TanimotoCombo and TFD to identify molecules with dissimilar geometries in order to potentially identify molecules with parameter inconsistencies. We assign a “difference flag” to a molecule pair (in a “molecule pair”, the comparison is made across force fields) when it yields a TFD value over 0.20 and a TanimotoCombo value over 0.50. These pairs visually exhibit different minimized geometries that may be indicative of parameterization differences. Out of 26,984,560 possible molecule pairs involving any pair of force fields, the combination of the SMIRNOFF99Frosst and GAFF2 force fields yielded the largest number of difference flags (305,582, Table 1)). This indicates that these force fields are quite different. In contrast, the combination of MMFF94 and MMFF94S yielded the smallest number of difference flags at 10,048 difference flags, indicating that these two force fields are the most similar among those being compared. These numbers are sensible given the history of these force fields – GAFF2 has undergone considerable recent reparameterization [39] and SMIRNOFF99Frosst inherits parameters from parm@Frosst [1], a sibling force field of GAFF, while reducing the number of parameters with an entirely different form of chemical perception [26, 27]. In contrast, MMFF94 and MMFF94S are identical aside from their treatment of some nitrogen atoms [15] consequently their optimized conformers should be rather similar, as reflected in our scores. Thus, these results match what would be expected from the parameterization history of these force fields.

Table 1. Number of difference flags in analysis for each FF pair (out of 2,698,456 molecules). Shown are the number of difference flags obtained when comparing each FF pair, with each difference flag representing a molecule with a substantially different geometry after minimization with those two force fields.

FF	GAFF	GAFF2	MMFF94	MMFF94S	SMIRNOFF
GAFF	-	87,829	153,244	142,369	268,830
GAFF2	-	-	138,716	131,528	305,582
MMFF94	-	-	-	10,048	267,131
MMFF94S	-	-	-	-	246,894
SMIRNOFF	-	-	-	-	-

We also label molecule pairs with highly similar geometries. To do this, we assign "similarity flags" to molecule pairs that yielded TFD values under 0.18, indicative of similar geometries (Table 2). In order to visualize the number of molecule pairs with each flag, we plot TFD versus TanimotoCombo for all molecule pairs. We highlight regions flagged as similar and different along with regions outside the interest of this analysis (Figure 1). Figure 1 likewise shows that the vast majority of molecule pairs were rated similar by both TFD and TanimotoCombo.

Table 2. Number of similarity flags in analysis for each FF pair (out of 2,698,456 molecules). Shown are the number of similarity flags obtained when comparing each FF pair, with each similarity flag representing a molecule with a similar geometry after minimization with those two force fields.

FF	GAFF	GAFF2	MMFF94	MMFF94S	SMIRNOFF
GAFF	-	2,577,081	2,467,654	2,481,084	2,324,408
GAFF2	-	-	2,483,650	2,493,171	2,277,081
MMFF94	-	-	-	2,678,568	2,294,096
MMFF94S	-	-	-	-	2,319,197
SMIRNOFF	-	-	-	-	-

2.2 Sets of molecules were created based on their similarity and difference flags

We then sort the molecules into sets of interest by their patterns of difference and similarity flags. As molecule pairs were formed from a set of five conformers, each resulting from optimization with a different force field, each molecule results in ten different molecule pairs which can be assigned either a difference or similarity flag. All molecules that yielded five or more difference flags out of ten were added to the set named "FivePlus." We also categorized molecules of particular interest for each force field. For each force field, we identified molecules in which two conditions held: (1) all molecule pairs involving that force field were flagged as different, and (2) the molecule pairs not including that force field were flagged as similar. Accordingly, molecules in these sets must result in four difference flags and six similarity flags; molecules in these sets can not also be in the FivePlus set. This allows us to highlight molecules which were treated differently by only one force field, potentially indicating problems in the force field's parameters for the represented chemistries of the molecule. We called this set the "Individually Different" set for that force field. For example, the molecules identified in this scheme for SMIRNOFF99Frosst were added to the "Individually Different SMIRNOFF" ($ID_{SMIRNOFF}$) set. This latter analysis is probably most relevant to the SMIRNOFF force field, as GAFF/GAFF2 and MMFF94/MMFF94S come in families which would reduce the number of cases meeting these criteria if intra-family similarity is high – specifically, if both family members treat a molecule consistently, it will not be flagged as "individually different" for that force field.

Our results after categorizing put 111,162 molecules into the FivePlus and 93,859 molecules in the $ID_{SMIRNOFF}$ set out of a total of 2,698,456 molecules. The $ID_{SMIRNOFF}$ set was the largest of the individually different force field sets, as is displayed in Table 3. As noted, we had some expectation SMIRNOFF might be relatively distinct from the other force fields considered.

Table 3. Number of molecules in each set of interest. Shown are the number of molecules in each of six sets of interest (described in Section 2.1); briefly, the FivePlus set contains molecules with substantially different geometries across multiple force fields, whereas the other sets contain molecules in which only the indicated force field yields a substantially different geometry from other force fields. The set with the largest number of molecules, the FivePlus set, contains 111,162 molecules out of the 2,698,457 molecules analyzed. No molecule can appear in more than one set of interest.

Set of Interest	Number of Molecules
FivePlus	111,162
Individually Different SMIRNOFF	93,859
Individually Different GAFF2	13,689
Individually Different GAFF	813
Individually Different MMFF94S	718
Individually Different MMFF94	72

2.3 Certain functional groups are more likely to be responsible for geometric differences

We characterized molecules with five or more difference flags

Molecules which yielded five or more out of ten possible difference flags were separated into what we call our FivePlus set. This set contained 111,162 total molecules, comprising 4.62% of all molecules included in this analysis. Visualizations of selected molecule pairs from the FivePlus set displaying significant geometric differences are provided in Figure 2.

We observed 150 Checkmol functional group descriptors with at least two occurrences within the FivePlus set. For each descriptor, we compared the proportion of FivePlus molecules with this descriptor to the proportion of molecules with this descriptor in the total set (Eq. 1), to assess whether any particular chemistries/functional groups tend to increase the likelihood of force fields treating molecules differently (and thus it ending up in the FivePlus set). We then identified the descriptors that are over-represented within the FivePlus set. For each of the descriptors we include in this section, we will provide an inline SMILES pattern for that descriptor along with the number of molecules with that descriptor in the current set of interest and the total set in the form (`SMILES`, number of molecules with the descriptor in the set of interest, number of molecules in total). For example, disulfides (`[R1]SS[R2]`, 51, 302) yield an over-representation factor of 4.04 in the FivePlus set.

The most over-represented descriptor within the FivePlus set was the thiocarbonic acid monoester (`OC(O[R])=S, 5, 26`), which were over-represented in the FivePlus set by a factor of 4.67. Three other descriptors were over-represented in the FivePlus set by a factor greater than 4:

1. Thiocarbamic acid halides (`[F,Cl,Br,I]C(N([R])[R])=S`, 3, 17) were over-represented in the FivePlus set by a factor of 4.28.
2. Phosphoric acid amides (`[R]P(N([R])[R])([R])=O`, 51, 302) were over-represented in the FivePlus set by a factor of 4.10.
3. Disulfides (`[R]SS[R]`, 149, 895) were over-represented in the FivePlus set by a factor of 4.04.

The most under-represented descriptor in the FivePlus set was the ketene (`[R]C([R])=C=O`, 9, 2124), with an over-representation factor of 0.11. This suggests that most force fields describe geometries of ketenes consistently, possibly due to the ketene function group's simple linear structure.

We repeated this process with pairs of Checkmol descriptors to see whether particular combinations of descriptors are especially indicative of discrepancies. We observed 6,500 descriptor pairs occurring in at least two cases in the FivePlus set. As with singular descriptors, we compared the proportion of molecules displaying a descriptor pair in the FivePlus set to the proportion of molecules displaying a descriptor pair in the total set (we applied the same expression, Eqn. 1, but for A+B descriptor pairs). The most over-represented descriptor pair in the FivePlus set were imidoyl halides paired with oxime molecules (`[R]/C([F,Cl,Br,I])=N\[R]` & `[R]/C([R])=N \O`, 3, 3), which was over-represented in the FivePlus set by a factor of 24.28, but the number of molecules with this particular combination is so low it makes it hard to know how much weight to give this observation. We determined by visual inspection that the imidoyl halide and oxime functional groups were in close proximity in these molecules, such that they may form a conjugated system. The force fields inconsistently predicted planar groups within this larger system. Two other descriptor pairs were over-represented in the FivePlus set by a factor greater than 19:

1. Quaternary ammonium salts paired with secondary aromatic amine molecules (`[R][N+]([R])([R])[R]` & `[R]N[R]`, 11, 12) were over-represented in the FivePlus set by a factor of 22.25.

2. Secondary aliphatic amines paired with disulfide molecules (`[R]N[R]` & `[R]SS[R]`, 11, 12) were over-represented in the FivePlus set by a factor of 19.90.

Again, these combinations are rare, so conclusions must be tentative at best.

Table 4. Selected Over-Represented Checkmol Descriptors and Descriptor Pairs in the FivePlus Set. Shown are the over-representation factors corresponding to selected descriptors or descriptor pairs, calculated using Equation 1. Descriptor pairs are denoted with an ampersand, e.g. "Descriptor One & Descriptor Two". The four descriptors and three descriptor pairs shown are the most over-represented descriptors and descriptor pairs of the FivePlus set.

Descriptor or Descriptor Pair	Over-Representation Factor
Thiocarbonic Acid Monoester	4.67
Thiocarbamic Acid Halide	4.28
Phosphoric Acid Amide	4.10
Disulfide	4.04
Imidohalide & Oxime	24.28
Quaternary Ammonium Salt & Secondary Aromatic Amine	22.25
Secondary Aliphatic Amine & Disulfide	19.90

Some pairs of descriptors are more likely to appear in the set of interest together more often than they are apart. We quantify this dependence by our pair enrichment factor (PEF) measurement (Eq. 2). The descriptor pair that showed the greatest degree of this dependence is quaternary ammonium salts paired with secondary aromatic amines (`[R][N+]([R])([R])[R]` & `[R]N[R]`, 11, 12), which yielded a pair enrichment factor of 2,807. Two other descriptor pairs yielded pair enrichment factors greater than 1,000:

1. Imines paired with thioxohetarenes (`[R]/C([R])=N\[R]` & `[R]N1C=CC=CC1=S`, 13, 24) yielded a PEF of 1,967.
2. 1,2-amino alcohols paired with carboxylic acid hydrazides (`[R]C(N([R])O)=O` & `[R]C(N([R])N)=O`, 2, 3) yielded a PEF of 1,188.

These findings display that heteroatoms, especially in delocalized pi-systems, are likely to lead to inconsistent optimized geometries. In particular, nitrogen, phosphorus, and sulfur atoms were found in all of the most over-represented descriptors and descriptor pairs. This is in line with our expectations, as QM treatments of sulfur and phosphorus are computationally expensive. Early force field development may have prioritized parameters for only the most common functional groups that involve sulfur and phosphorus. Our procedure has identified molecular fragments that yielded inconsistent geometries, and therefore can be improved upon in future force fields. Furthermore, nitrogen planarity errors are a known issue across force fields [15, 27]. We therefore believe that the descriptors identified by this procedure may be informative for the creation/training of higher accuracy small molecule force fields. Molecules containing these fragments should be included in future force field training sets in order to create more accurate and general small molecule force fields.

Table 5. Selected Pair Enriched Checkmol Descriptor Pairs in the FivePlus Set. Shown are the pair enrichment factors corresponding to selected descriptor pairs, calculated using Equation 2. The three descriptor pairs shown are the three pairs that yielded the highest pair enrichment factor in the FivePlus set.

Descriptor or Descriptor Pair	Pair Enrichment Factor
Quaternary Ammonium Salt & Secondary Aromatic Amine	2807
Imine & Thioxohetarene	1967
1,2-Amino Alcohol & Carboxylic Acid Hydrazide	1188

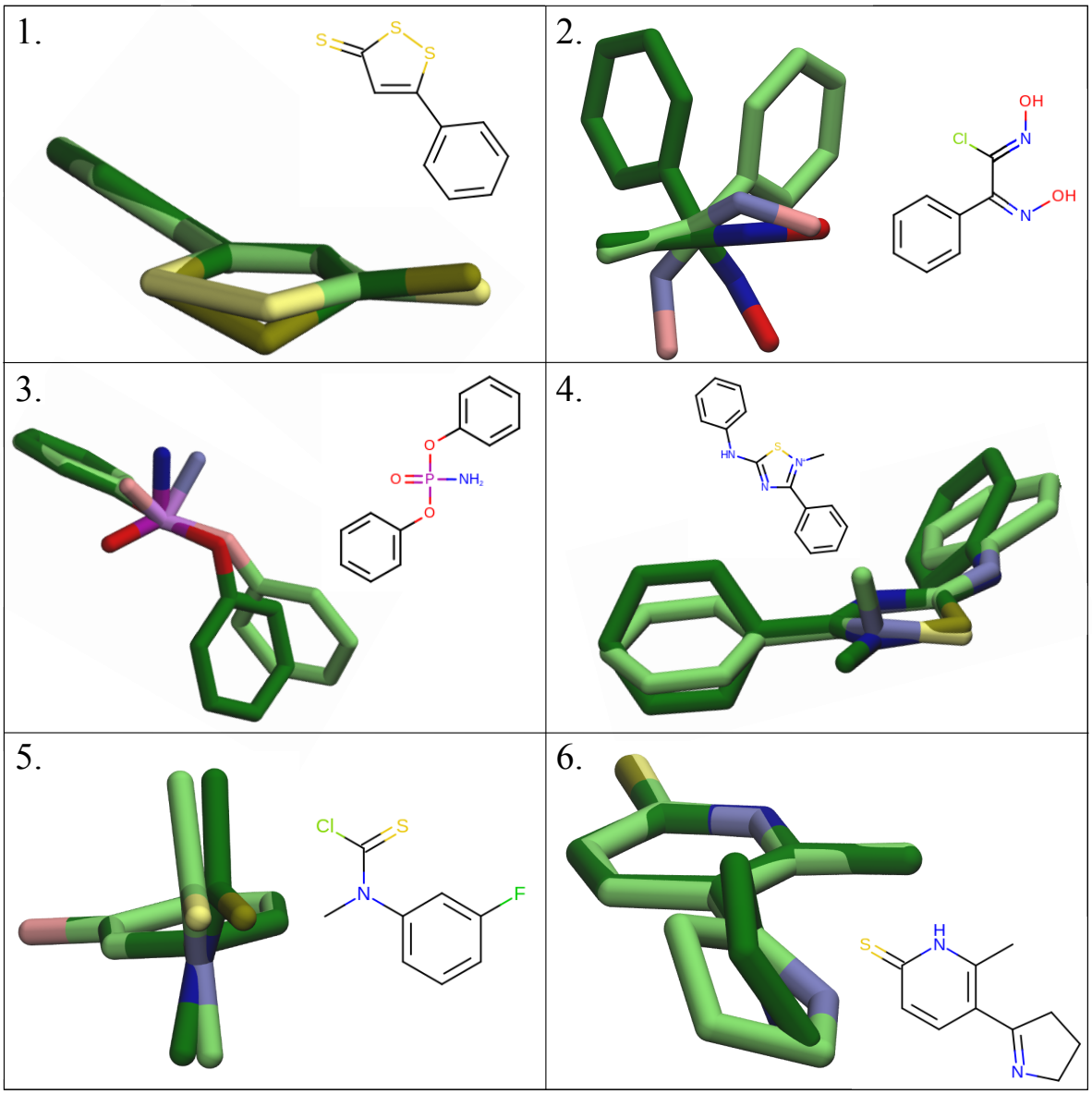

Figure 2. Molecule pairs from the FivePlus set display visual geometric differences. The six molecules displayed here were identified from the FivePlus set using the over-represented descriptor and descriptor pair method described in Section 3.1 and thus are molecules where geometries differ substantially across force fields. Each panel shows a molecule (with the 2D structure shown as inset) and a pair of minimized conformers resulting from optimization with different force fields. These highlight geometric differences between minimized structures. While many structure pairs yield difference flags for molecules in the FivePlus set, only one structure pair is displayed for each molecule here. The lightly colored structure was optimized with GAFF, while the darkly colored structure was optimized with SMIRNOFF. (1) While GAFF predicts a planar structure of the ring system, SMIRNOFF predicts a buckled ring for this molecule with the disulfide descriptor. (2) GAFF predicts the imidoyl halide group to be nonplanar in this molecule with the imidoyl halide and oxime descriptors, while SMIRNOFF predicts it to be planar. (3) SMIRNOFF predicts a larger bond angle between the amine and non-bridging oxygen than does GAFF in this molecule displaying the phosphoric acid amide descriptor. (4) This molecule displays both the quaternary ammonium cation and the secondary aromatic amine descriptors. While SMIRNOFF predicts a planar thiadiazolium ring, GAFF predicts it to be nonplanar. (5) While GAFF predicts the thiocarbamic acid halide fragment to be planar and perpendicular to the aromatic ring, SMIRNOFF predicts it to be nonplanar and off-perpendicular to the aromatic ring. (6) This molecule displays both the thioxohetarene and imine descriptors. While GAFF predicts a planar pyrroline ring, SMIRNOFF predicts this ring to be buckled.

We characterized molecules where SMIRNOFF was individually different

The OpenFF Initiative seeks to improve force fields via a series of progressive improvements, thus we focus on the SMIRNOFF force field in particular in order to help our work with OpenFF. Specifically, we identify molecules where parameterization dif-

ferences in SMIRNOFF relative to other force fields lead to geometry differences. Molecules that yielded four difference flags from combinations including the SMIRNOFF-minimized conformer, and six similarity flags from combinations not including the SMIRNOFF-minimized conformer, were likewise grouped into a set of interest. We refer to this set as the Individually Different SMIRNOFF ($ID_{SMIRNOFF}$) set. This set contained 93,859 molecules in total, or 3.48% of all molecules included in this analysis. Visualizations of molecule pairs from the $ID_{SMIRNOFF}$ set displaying geometric differences are provided in Figure 3.

We observed 139 Checkmol descriptors in at least two molecules in the $ID_{SMIRNOFF}$ set. We compared the proportion of molecules exhibiting some descriptor within the $ID_{SMIRNOFF}$ set to the proportion of molecules exhibiting the descriptor in the total set (Equation 1). We can then identify descriptors that are over-represented or underrepresented within the $ID_{SMIRNOFF}$ set. The most over-represented descriptor within the $ID_{SMIRNOFF}$ set was the azo compound descriptor (`[R]/N=N/[R]`, 717, 1500) which was over-represented in the $ID_{SMIRNOFF}$ set by a factor of 13.74. Such compounds have been a focus of reparameterization efforts in more recent versions of SMIRNOFF-based force fields, in particular in OpenFF 1.1. [21, 38], consistent with our observation here that these may be poorly treated. We discuss later OpenFF releases further below. Four other descriptors were over-represented in the $ID_{SMIRNOFF}$ set by a factor greater than 4:

1. Carbodiimides (`[R]N=C=N[R]`, 4, 19) were over-represented by a factor of 6.05.
2. Acylcyanides (`[R]C(C#N)=O`, 5, 30) were over-represented by a factor of 4.79.
3. Hydrazones (`[R]/C([R])=N/N([R])[R]`, 2962, 20,025) were over-represented by a factor of 4.25.
4. Thioaldehydes (`[R]C([H])=S`, 23, 165) were over-represented by a factor of 4.01.

The most underrepresented descriptor in the $ID_{SMIRNOFF}$ set was the 1,2-amino alcohol (`[R]N([R])CCO`, 159, 24,344), with an over-representation factor of 0.19.

We observed 5,805 descriptor pairs in at least two molecules in the $ID_{SMIRNOFF}$ set. As with singular descriptors, we compared the proportion of molecules displaying a descriptor pair in the $ID_{SMIRNOFF}$ set to the proportion of molecules displaying a descriptor pair in the total set (Equation 1). These descriptor pairs and their over-representation factors are likewise included in Table 4. Six different descriptor pairs were tied as most over-represented in the $ID_{SMIRNOFF}$ set. For these, all molecules displaying these pairs in the total set were also included in the $ID_{SMIRNOFF}$ set. For example, there were five molecules characterized as both ketene acetal derivatives and oximes (`[R]/C([R])=C([R])\[R]` & `[R]/C([R])=N \O`, 5, 5), and all five of these molecules were also present in the $ID_{SMIRNOFF}$ set. We observed two other descriptor pairs which occurred in greater than 10 molecules in the $ID_{SMIRNOFF}$ set and had an over-representation factor greater than 20:

1. Azo compounds paired with Aldehydes (`[R]/N=N/[R]` & `[R]C([H])=O`, 41, 49) were over-represented by a factor of 24.06.
2. Hydrazones and Hydroxamic Acids (`[R]/C([R])=N/N([R])[R]` & `[R]C(N([R])O)=O`, 14, 18) were over-represented by a factor of 22.36.

Table 6. Selected Over-Represented Checkmol Descriptors and Descriptor Pairs in the $ID_{SMIRNOFF}$ Set. Shown are the over-representation factors corresponding to selected descriptors or descriptor pairs, calculated using Equation 1. Descriptor pairs are denoted with a "&" symbol, e.g. “Hydrazone & Hydroxamic Acid”. These are some of the most over-represented descriptors and descriptor pairs of the $ID_{SMIRNOFF}$ set. Note that the “Ketene Acetal Derivative & Oxime” pair has a very high over-representation factor because all 5 molecules displaying this descriptor pair are in the $ID_{SMIRNOFF}$ set.

Descriptor or Descriptor Pair	Over-Representation Factor
Azo Compound	13.74
Carbodiimide	6.05
Acylcyanide	4.79
Hydrazone	4.25
Thioaldehyde	4.01
Ketene Acetal Derivative & Oxime	287.50
Azo Compound & Aldehyde	24.06
Hydrazone & Hydroxamic Acid	22.36

We also calculated pair enrichment factors (PEFs), as described in Equation 2, for the $ID_{SMIRNOFF}$ set of molecules. The descriptor pair that showed the greatest degree of this dependence in the $ID_{SMIRNOFF}$ set is the iminohetarene & secondary

alcohol pair (`[R]/N=C1C=CC=CN/1[R]` & `[R]C(O)[R]`, 3, 10), which yielded a PEF of 2,308, relative to a mean PEF of 49.83 for the $ID_{SMIRNOFF}$ set. Two other descriptor pairs yielded PEFs greater than 2,000:

1. Iminohetarenes paired with tertiary alcohols(`[R]/N=C1C=CC=CN/1[R]` & `[R]C(O)([R])[R]`, 4, 6) yielded a PEF of 2,187.
2. Thiocarboxylic acid amides paired with primary aliphatic amines (`[R]C(N([R])[R])=S` & `[R]N`, 2, 3) yielded a PEF of 2155.

Descriptor pairs with high pair enrichment factors may suggest unique chemistries that lead to geometric inconsistencies that were not accurately described by single descriptors.

Nitrogen atoms in conjugated systems make up a large portion of molecules that were optimized to unique structures by SMIRNOFF. While other force fields have likewise had problems with nitrogen planarity, our results display two checkmol descriptors, azo compound and hydrazone, that are especially informative for SMIRNOFF. By visual inspection, molecules with one of these descriptors in between two aromatic rings are especially prominent, as can be seen in boxes 2, 3, and 4 of Figure 3. QM calculations are necessary to determine if SMIRNOFF's minimized conformers were more or less accurate than other force fields. Still, molecules like these will be useful in training sets of future force fields. In other cases, such as those displayed in boxes 5 and 6 of Figure 3, SMIRNOFF disagrees with other force fields on the geometry of secondary carbon atoms in certain environments. SMIRNOFF assigns parameters to molecules separately by type (i.e. bonds, angles, and torsions are treated independently) with explicit treatment for bond order which differs from the atom-type approach used by the other force fields in this study [26]. It is possible this change in chemical perception can help account for the change in treatment of these systems. QM data on these molecules will be useful for future iterations of the SMIRNOFF force field, which are already in development.[2, 21, 24, 29]

Table 7. Selected Pair Enriched Checkmol Descriptor Pairs in the $ID_{SMIRNOFF}$ Set. Shown are the pair enrichment factors corresponding to selected descriptor pairs, calculated using Equation 2. The three descriptor pairs shown are the three pairs that yielded the highest pair enrichment factor in the $ID_{SMIRNOFF}$ set.

Descriptor Pair	Pair Enrichment Factor
Iminohetarene & Secondary Alcohol	2,308
Iminohetarene & Tertiary Alcohol	2,187
Thiocarboxylic Acid Amide & Primary Aliphatic Amine	2155

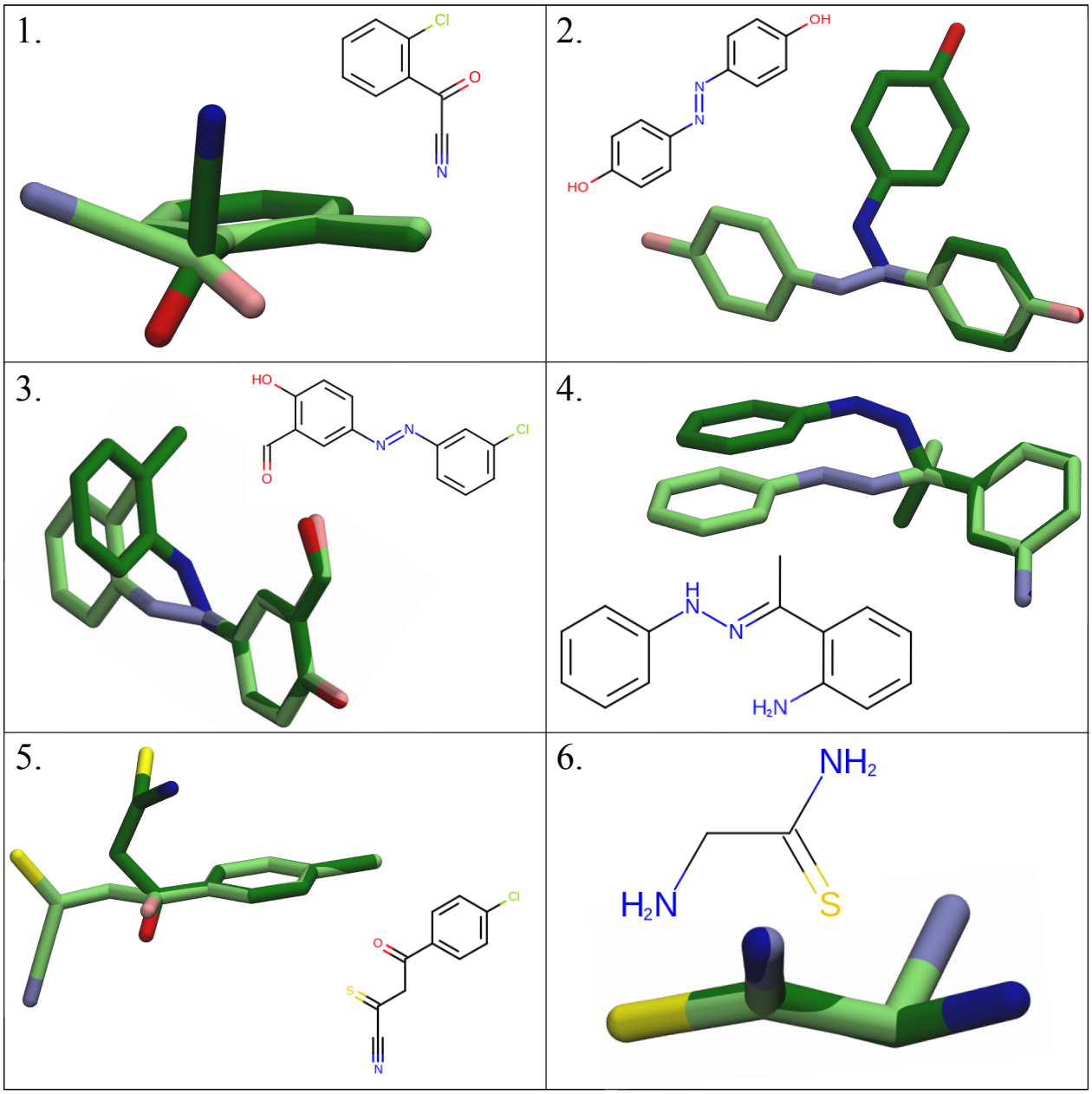

Figure 3. Molecule pairs from the Individually Different SMIRNOFF set display visual geometric differences. The displayed molecules were identified from the $ID_{SMIRNOFF}$ set using the descriptor method in Section 3.1. For each molecule, shown in a 2D inset, we visualize the minimized conformer with SMIRNOFF (darker colors) and GAFF - representing all non-SMIRNOFF force fields (lighter colors). (1) SMIRNOFF predicts the acylcyanide group to be near perpendicular to the aromatic ring, while GAFF makes it near planar. (2) The two force fields disagree on the appropriate torsion angle for the C-N=N-C bond in the azo group. (3) Again, the SMIRNOFF and GAFF force fields disagree on the planarity of the azo group where the orientation of the aldehyde group agrees between the two force fields. (4) The substituted hydrazone's change in planarity dominates this molecule with the hydrazone descriptor. (5) In this molecule with a thiocarbonyl, GAFF keeps all carbons planar while SMIRNOFF allow the carbon-carbon single bonds to rotate. (6) This molecule pairs thiocarboxylic acid amide and primary amine descriptors where GAFF predicts the primary amine to bend out of plane, while SMIRNOFF predicts all heavy atoms to be planar.

2.4 This work has been used to improve training datasets for the OpenFF Parsley series

In the present work, discrepancies between optimized geometries from different force fields highlight potential issues, but we have no ground truth or point of reference for sorting out which geometries are correct and which are not. This data simply helps us select molecules/chemistries which may be informative, and prioritize them for further study. Particularly, ideally one might generate optimized geometries for these same molecules with QM calculations and then use these to help assess which

force fields produce the best results, or use these in force field training sets to improve force field quality.

Indeed, informative molecules from the present study are being used for precisely that purpose. Particularly, a subset of the FivePlus set was used as the basis for the "coverage" set used for the first OpenFF Parsley release, OpenFF 1.0 [29]. A larger portion was used in benchmarking OpenFF 1.0. Then, for OpenFF 1.2, training data was completely redesigned, in part drawing from what was called the "eMolecules Discrepancies Set" [22, 25], corresponding to the first portion of the FivePlus set generated here. This training data redesign resulted in improved performance on a variety of benchmarks [21, 24]. The relevant optimized geometries are freely available in QCArchive [Smith et al.] as part of the OpenFF 1.2 training and benchmarking datasets.

While subsequent OpenFF work building on the data generated here is not formally part of this study, it does appear that molecules identified as potentially informative by this approach do serve well as input for QM calculations and force field training, at least when coupled with additional data selection/curation steps.

3 Methods

In order to help improve force fields, we sought to to identify where current force fields differ from one another. Here, we compared results of force fields (particularly, optimized geometries) after energy minimizing a large subset of the eMolecules database to identify sets of molecules for use in future force field parameterization.

Multiple force fields were used to minimize conformers

We created input files for multiple force fields from a filtered eMolecules set (filtering described in Section 3.3). Partial charges were assigned to molecules before minimization using the OpenEye implementation of AM1-BCC [19, 20]. The input generation process yields one Tripos MOL2 file to be minimized directly with SMIRNOFF99frosst, MMFF94, and MMFF94S, as well as individual input coordinate and parameter topography files for use by GAFF (1.8) and GAFF2 (2.1). These force fields were chosen because they are widely-used, easily available, and compatible with our workflow. Other force fields were either incompatible with our toolchain without substantial additional work, or were commercial and proprietary. For example, comparisons with CGenFF [36, 37], OPLS-AA [23], or the Schrödinger OPLS series [17, 30] would be of considerable interest, but these require substantially different toolchains, and the most recent Schrödinger force fields are also proprietary and require paying for a license.

We minimized each molecule using the parameters from each of the five aforementioned force fields, making sure to start all five minimizations from the same conformer. Minimizations with force fields other than MMFF were performed with OpenMM [4] 7.0.1 using the L-BFGS algorithm [28] with an energy tolerance of 5.0e-9 kJ/mol and a maximum of 1500 iterations. MMFF minimizations were performed with OpenEye's Szybki Toolkit [35, 42]. Sample run files can be found in the Supporting Information. Molecules that did not successfully result in five minimized structures (one from each force field), were removed from analysis. For each molecule with five minimized structures, pairwise comparisons yielded a total of ten molecule pairs for geometric evaluation. We call these pairs of minimized conformers generated by different force fields "molecule pairs."

Molecule pairs were assessed using Torsion Fingerprint Deviation and TanimotoCombo

We then assessed each molecule pair for geometric differences. Molecule pairs were evaluated using two distinct measurements: Torsion Fingerprint Deviation (TFD) and TanimotoCombo.

TFD is a method of measuring geometric differences between two conformers of the same molecule based on torsion angles. The TFD score between two structures represents a weighted sum of torsional differences as defined by Schulz-Gasch et al. in 2012 [31]. Torsions central to the molecule are given more weight than torsions on the periphery of the molecule. Similarly with RMSD, geometric similarity is inversely correlated with TFD score. TFD scores range from 0 to 1, with 0 being most similar and 1 being most different. The authors of TFD consider scores over 0.2 to represent significantly different geometries. In contrast to RMSD, TFD is bounded and less sensitive to molecular size, making it particularly helpful here.

TanimotoCombo, from OpenEye Scientific, is a normalized method of measuring geometric similarity between molecules. It is the sum of ShapeTanimoto, a measure of overall spatial overlap between two molecules, and ColorTanimoto, a measure of spatial overlap of specific functional groups between two molecules, both of which are also metrics from OpenEye. TanimotoCombo values between two conformers range between 0 and 2 (it is the sum of two values running from 0 to 1), with 2 being the most similar and 0 being the most different.

By visual inspection, we determined that TanimotoCombo is useful for recognizing cases where geometric differences are caused by particularly flexible moieties, such as single bond rotations in an alkyl chain. These differences can often be attributed to minor differences between force fields leading to flexible bond rotations, not to larger differences in force fields that result

in more substantial geometric differences. Thus, here, we find that TanimotoCombo alone does not serve to help us isolate geometry differences that are likely due to substantial force field differences; instead, low TanimotoCombo values can result from simple bond rotations that result from molecules energy minimizing to different local minima that we do not consider particularly interesting by visual inspection. However, TanimotoCombo in conjunction with TFD can be used to identify geometric differences that suggest underlying inconsistencies in parameterization.

Molecule pairs were flagged as similar or different based on TFD and TanimotoCombo

We identified molecule pairs displaying parameterization differences which led to different geometries using TFD and TanimotoCombo. TFD is sensitive to ring deformations, torsional differences, and atom planarity changes, which makes it useful for recognizing differences in parameterization. TanimotoCombo, with greater sensitivity to coordinate differences caused by conformational flexibility in a molecule, is more useful for removing cases that are less likely to be caused by parameterization differences, or which may simply be due to minor differences in which rotamer a molecule minimizes to.

We chose cutoffs to flag molecule pairs displaying parameterization differences (flagged "different") and pairs displaying no parameterization differences (flagged "similar"). TFD values below 0.20 are believed to be pharmacologically similar [31], so we chose a TFD value greater than 0.20 to label molecule pairs as different. After visual inspection of a variety of molecules, we observed that molecule pairs with a TanimotoCombo under 0.5 typically had changes due to single bond rotations. Because such bond rotations can arise from a variety of reasons aside from substantial differences in parameterization, we did not wish to focus on such cases. Thus, molecule pairs with a TFD value greater than 0.20 as well as a TanimotoCombo value greater than 0.50 were flagged as different – allowing us to focus on cases with substantial torsional differences which were NOT simply due to rotations around highly flexible bonds. We used a substantial amount of manual inspection of these thresholds to help us make these choices. As a result of these choices, any pair of molecules with a TFD value of 0.18 or less was assigned a similarity flag, as it will display geometrically similar structures. We left a small buffer region between 0.18 and 0.2 when defining similarity flags in order to avoid an extreme sensitivity to small changes around the 0.20 cutoff.

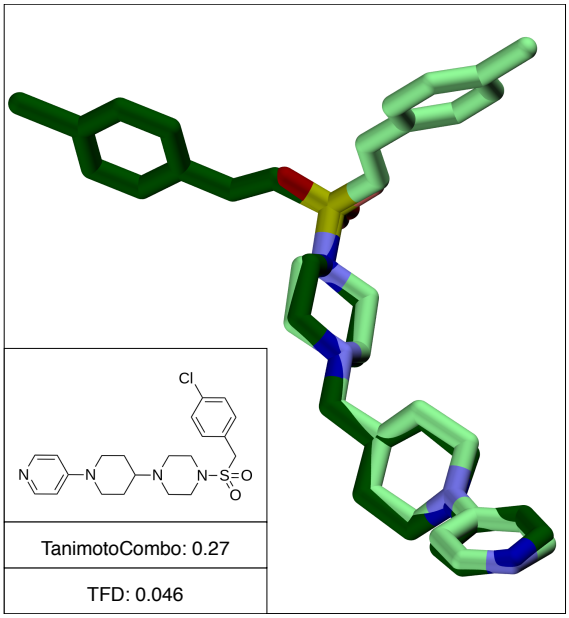

Figure 4. Molecule pairs with low TanimotoCombo and low TFD scores are often uninformative. Here, we show an example of a molecule pair that we did not deem informative for force field parameterization. The lightly colored molecule was minimized with GAFF, while the darker molecule was minimized with SMIRNOFF. The two minimized structures display little geometric differences outside of the placement of substituents around the sulfonamide group; most of the geometric difference appears due to rotation of a single torsion. The low TFD value of 0.046 implies that these structures are highly similar by TFD, while the low TanimotoCombo value of 0.27 implies that these structures are starkly different by TanimotoCombo. By visual inspection of this molecule and others, we determined that molecule pairs with low Tanimoto Combo and low TFD scores were often not as informative, at least with respect to our goals in this project.

Molecule pairs that yielded very high TFD or very low TanimotoCombo values were determined to often be uninformative. Tagging these molecule pairs as 'different' would be unhelpful because the differences are not due to substantial changes in

force field parameters. Most molecule pairs in this category displayed cases of what might be called "conformer chirality" – where an achiral molecule was minimized to two similar but non-superimposable structures – essentially, collapsing down to two minima which are equivalent to the force field but not geometrically identical. A number of molecule pairs, most with small flexible ring systems, yielded TFD values greater than 1 (which should not be possible). While this behavior was unexpected, we continued to use RDKit's implementation without modifications. Considering these results, we removed molecule pairs with a TFD greater than 0.60 or a TanimotoCombo less than 0.25.

3.1 We created and characterized sets of interest

Molecules can be sorted into sets of interest by considering the combinations of their difference and similarity flags. One molecule in this pipeline will yield five minimized structures. Taking pairwise combinations of these structures can yield ten molecule pairs, and thus up to ten flags. Molecules that yielded a large number of difference flags, regardless of the force fields of origin, are of particular interest for force field parameterization. Specifically, we set aside molecules with five or more difference flags for further analysis, we call this our *FivePlus* set.

The other sets of interest are based on the origin of the difference flags with the goal of identifying molecules which behave differently with one force field than all the others. For a molecule to be considered different with that one force field, all four molecule pairs including that force field should be flagged as different and all other molecule pairs need to be flagged as similar. We call these *Individually Different* Sets for each force field, i.e. for SMIRNOFF we create the SMIRNOFF Individually Different set with the label $ID_{SMIRNOFF}$. A molecule in the $ID_{SMIRNOFF}$ set would have 4 difference flags, one for each pair including SMIRNOFF, and six similarity flags for all other force field combinations.

3.2 Sets of interest were analyzed by the frequencies of the functional groups

Identifying functional groups which are more prevalent in our sets of interest could be informative for future force field parameterization. To this end, we used Checkmol [9] to describe the combination of functional groups in each molecule. When given a molecule, Checkmol provides a list of descriptors for the functional groups it has. We count the number of molecules overall and in each set of interest that display each descriptor. From there, we can determine the most over-represented descriptors in each set of interest. We only considered descriptors and descriptor pairs that appeared at least twice in our full molecule set.

We compute the *over-representation factor* describing how over-represented a particular descriptor is in a given set by dividing the frequency of the descriptor in the set by the frequency of the descriptor in all molecule. Mathematically, we can write

$$f_A = \frac{N_{A,\text{set}}/N_{\text{mols,set}}}{N_{A,\text{total}}/N_{\text{mols,total}}} \tag{1}$$

where $N_{A,\text{set}}$ is the number of molecules containing descriptor A in a particular set, $N_{\text{mols,set}}$ is the number of molecules in that particular set. $N_{A,\text{total}}$ is the number of molecules in total with descriptor A, and $N_{\text{mols,total}}$ is the number of molecules in total.

Force field behavior could change with combinations of functional groups, thus we repeated this calculation with pairs of Checkmol descriptors. We can apply Equation 1 to analyze pairs of descriptors by replacing A with $A + B$ to represent molecules containing both descriptors. However, with pairs of descriptors, we are more concerned about if the combination of the descriptors is important. For example, if both descriptors A and B are highly probably in a set of molecules, then finding the combination in that set at a higher frequency is not particularly interesting. Thus, we try to determine if the descriptor pair is more likely to show up in a set of interest than the individual descriptors separately. To that end we calculate an *enrichement* factor given by

$$\frac{f_{A+B}}{f_A \cdot f_B} \tag{2}$$

where f_{A+B} denotes the frequency of the combined A and B descriptors in molecules in the set of interest, and f_A and f_B denote the individual frequencies of descriptors A and B in the set of interest. A larger enrichment factor indicates the combination of descriptors A and B is more likely to occur in a set of interest than those descriptors individually. Thus, descriptor pairs with a larger enrichment factor should be considered as important for future parameterization because the combination of functional groups changes a force field's behavior.

3.3 Molecules were sourced from the eMolecules online database

Approximately 8.1 million molecules were initially sourced from the eMolecules database as SDF files (September 2016 version) [5]. Molecules from this set were then filtered by several criteria. We removed all molecules that contained any metal or

metalloid atoms, were over 200 heavy atoms, or had a nonphysical valence (such as a pentavalent carbon atom). Molecules which failed at any step of the process were also removed, i.e. could not be parameterized by one of the force fields. While we minimized all these molecules with each force fields, very large molecules are impractical for visual inspection or future QM calculations. Thus, we filtered the molecules for analysis here to remove molecules with more than 25 heavy atoms.

4 Conclusions

Here, we sought to determine informative molecules for force field parameterization. We assume that conformational differences in molecules minimized with different force fields indicates those molecules ought to receive additional attention in future force field parameterization.

Thus, we energy minimized a large portion of eMolecules with various force fields, and cross-compared the resulting optimized geometries based on TFD and TanimotoCombo metrics. We chose cutoffs for each of these metrics in order to prioritize conformational differences likely due to changes in force field parameters.

Our analysis flags molecules for further analysis in several ways. First, we single out molecules that differ in treatment across many force fields as molecules which are likely to be particularly informative in general. Second, we can separate out molecules which are treated differently by only one force field as perhaps indicative of problems with that force field in particular. We can further break down informative molecules by looking at representation of functional groups, and pairs of functional groups, to identify those that are over-represented among informative molecules, perhaps indicating these functional groups require additional attention in force field parameterization.

The descriptors which were over-represented in the FivePlus set could be informative for understanding the limitations of current force field parameterization procedure. All general small molecule force fields currently available depend on human determined typing rules – atom types in most force fields and the `SMARTS` patterns used in SMIRNOFF-based force fields. The differences in geometries around heteroatoms, especially sulfur and phosphorous, point to the potential bias of the scientists parameterizing each force field. Most of the time new parameter typing rules are added to force fields out of necessity and each group is going to prioritize different chemistry. Including typing rules in automatic force field parameterization should help reduce this bias since typing rules would be driven by training data rather than human choices.

Finding the more accurate conformation in each molecule pair would require performing a quantum mechanical optimization(QM). QM calculations are significantly more expensive than a simple force field optimization. Our protocol allowed us to compare 26,984,560 molecule pairs. Our approach has identified regions of chemical space where force field parameterization is currently inconsistent. Our approach and results have identified descriptor and descriptor pairs which are different for each individual force field. Molecules with these descriptors could be prioritized for future parameterization leading to more accurate force fields over all. Some work along these lines is already in progress [21, 22, 25, 29].

5 Code and Data Availability

We provide the code used in this project in our GitHub repository (https://github.com/mobleylab/off-ffcompare and with a DOI at https://dx.doi.org/10.5281/zenodo.3995606). Additionally, at https://dx.doi.org/10.5281/zenodo.3995059 we provide a supporting data package. This includes a .csv file which has TanimotoCombo and TFD scores, SMILES strings, and eMolecules identifiers for all 2,698,456 molecules analyzed. Additionally, we provide optimized geometries of 265,847 molecules with five or more difference flags. An archived copy of the GitHub repository is provided in the electronic Supporting Information associated with this paper.

6 Acknowledgments

JNE appreciates financial support from the National Institute of Health (R01GM108889). VTL appreciates funding from the National Science Foundation Graduate Research Fellowship Program. CCB appreciates financial support from The Molecular Sciences Software Institute under NSF grant ACI-154758. DLM appreciates financial support from the National Institutes of Health (R01GM108889 and R01GM132386) and the National Science Foundation (CHE 1352608). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

7 Disclaimers

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8 Disclosures

DLM is a member of the Scientific Advisory Board of OpenEye Scientific Software and an Open Science Fellow with Silicon Therapeutics.

References

[1] Bayly, C., McKay, D., and Truchon, J. (2010). An Informal AMBER Small Molecule Force Field: Parm@ Frosst. http://www.ccl.net/cca/data/parm_at_Frosst/.

[2] Chodera, J., Qiu, Y., Boothroyd, S., Wang, L.-P., and Mobley, D. (2019). The Open Force Field 1.0 small molecule force field, our first optimized force field (codename "Parsley").

[3] Dauber-Osguthorpe, P. and Hagler, A. T. (2018). Biomolecular force fields: Where have we been, where are we now, where do we need to go and how do we get there? *J Comput Aided Mol Des*.

[4] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput Biol*, 13(7):e1005659.

[5] eMolecules (2015). eMolecules Database Download. *https://www.emolecules.com/info/plus/download-database*.

[6] Fennell, C. J., Wymer, K. L., and Mobley, D. L. (2014). A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. *The Journal of Physical Chemistry B*, 118(24):6438–6446. Publisher: American Chemical Society.

[7] Hagler, A. T. (2018). Force field development phase II: Relaxation of physics-based criteria… or inclusion of more rigorous physics into the representation of molecular energetics. *J Comput Aided Mol Des*.

[8] Haider, N. (2010). Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: An Open-Source Approach. *Molecules*, 15(8):5079–5092.

[9] Haider, N. (2020). Checkmol/Matchmol Homepage. http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.

[10] Halgren, T. A. (1992). The representation of van der Waals (vdW) interactions in molecular mechanics force fields: Potential form, combination rules, and vdW parameters. *Journal of the American Chemical Society*, 114(20):7827–7843.

[11] Halgren, T. A. (1996a). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519.

[12] Halgren, T. A. (1996b). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.*, 17(5-6):520–552.

[13] Halgren, T. A. (1996c). Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.*, 17(5-6):553–586.

[14] Halgren, T. A. (1996d). Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.*, 17(5-6):616–641.

[15] Halgren, T. A. (1999). MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry*, 20(7):720–729.

[16] Halgren, T. A. and Nachbar, R. B. (1996). Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J. Comput. Chem.*, 17(5-6):587–615.

[17] 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., and Friesner, R. A. (2016). OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.*, 12(1):281–296.

[18] Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010). Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.*, 50(4):572–584.

[19] Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, 21(2):132–146.

[20] Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641.

[21] Jang, H. (2020). Update on Parsley minor releases (openff-1.1.0, 1.2.0).

[22] Jang, H., Maat, J., Qiu, Y., Smith, D. G., Boothroyd, S., Wagner, J., Bannan, C. C., Gokey, T., Lim, V. T., Lucas, X., Tjanaka, B., Shirts, M. R., Gilson, M. K., Chodera, J. D., Bayly, C. I., Mobley, D. L., and Wang, L.-P. (2020). Openforcefield/openforcefields: Version 1.2.0 "Parsley" Update. Zenodo.

[23] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935.

[24] Lim, V. T., Hahn, D. F., Tresadern, G., Bayly, C. I., and Mobley, D. (2020). Benchmark Assessment of Molecular Geometries and Energies from Small Molecule Force Fields. *chemRxiv*.

[25] Maat, J. (2020). Training dataset selection.

[26] Mobley, D. L., Bannan, C. C., Rizzi, A., Bayly, C. I., Chodera, J. D., Lim, V. T., Lim, N. M., Beauchamp, K. A., Shirts, M. R., Gilson, M. K., and Eastman, P. K. (2018a). Open Force Field Consortium: Escaping atom types using direct chemical perception with SMIRNOFF v0.1. *bioRxiv*, page 286542.

[27] Mobley, D. L., Bannan, C. C., Rizzi, A., Bayly, C. I., Chodera, J. D., Lim, V. T., Lim, N. M., Beauchamp, K. A., Slochower, D. R., Shirts, M. R., Gilson, M. K., and Eastman, P. K. (2018b). Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.*

[28] Nash, S. G. and Nocedal, J. (1991). A Numerical Study of the Limited Memory BFGS Method and the Truncated-Newton Method for Large Scale Optimization. *SIAM J. Optim.*, 1(3):358–372.

[29] Qiu, Y., Smith, D. G. A., Boothroyd, S., Wagner, J., Bannan, C. C., Gokey, T., Jang, H., Lim, V. T., Stern, C. D., Rizzi, A., Lucas, X., Tjanaka, B., Shirts, M. R., Gilson, M. K., Chodera, J. D., Bayly, C. I., Mobley, D. L., and Wang, L.-P. (2019). Introducing the first optimized Open Force Field 1.0.0 (codename "Parsley").

[30] Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., Dahlgren, M. K., Mondal, S., Chen, W., Wang, L., Abel, R., Friesner, R. A., and Harder, E. D. (2019). OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.*, 15(3):1863–1874.

[31] Schulz-Gasch, T., Schärfer, C., Guba, W., and Rarey, M. (2012a). TFD: Torsion Fingerprints As a New Measure To Compare Small Molecule Conformations. *J. Chem. Inf. Model.*, 52(6):1499–1512.

[32] Schulz-Gasch, T., Schärfer, C., Guba, W., and Rarey, M. (2012b). TFD: Torsion Fingerprints as a new measure to compare small molecule conformations. *J Chem Inf Model*, 52(6):1499–1512.

[33] Sellers, B. D., James, N. C., and Gobbi, A. (2017). A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.*, 57(6):1265–1275.

[Smith et al.] Smith, D. G. A., Altarawy, D., Burns, L. A., Welborn, M., Naden, L. N., Ward, L., Ellis, S., Pritchard, B. P., and Crawford, T. D. The MolSSI QCArchive project: An open-source platform to compute, organize, and share quantum chemistry data. *WIREs Computational Molecular Science*, n/a(n/a):e1491.

[35] Szybki ToolKit (2015). Version 1.9.0. OpenEye Scientific Software Inc. *Santa Fe, NM*.

[36] Vanommeslaeghe, K. and MacKerell, A. D. (2012). Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.*, 52(12):3144–3154.

[37] Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D. (2012). Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.*, 52(12):3155–3168.

[38] Wagner, J. (2020). Openforcefield/openforcefields: Version 1.1.0 "Parsley" Update. Zenodo.

[39] Wang, J. (2017). A Snapshot of GAFF2 Development.

[40] Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260.

[41] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174.

[42] Wlodek, S., Skillman, A., and Nicholls, A. (2010). Ligand entropy in gas-phase, upon solvation and protein complexation. fast estimation with quasi-newton hessian. *Journal of chemical theory and computation*, 6(7):2140–2152.