Improving Force Fields by Identifying and Characterizing Small Molecules with Parameter Inconsistencies

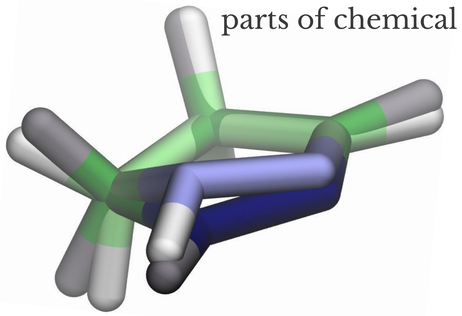
Jordan N. Ehrman, Caitlin C. Bannan, Victoria T. Lim, Nam Thi, Daisy Y. Kyu, David L. Mobley
University of California, Irvine

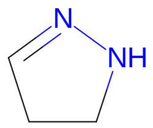
Abstract

Computer-aided drug design utilizes force fields to simulate chemical structures. Force fields are sets of functions and parameters which return the potential energy of a chemical system. Force fields are widely used, but 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 molecules minimized with a variety of force fields. We apply this pipeline to the eMolecules database, and highlight molecules that appear to be parameterized inconsistently across different force fields. We aim to identify molecules that are informative for future force field development, and therefore display these inconsistencies between force fields. We then characterize these sets by identifying overrepresented functional groups. This project is a subset of the Open Force Field Initiative, which is working to automate force field parameterization. Molecules identified by our pipeline will be used to parameterize future force fields.

Background

Force fields aren't consistent with each other for all parts of chemical space

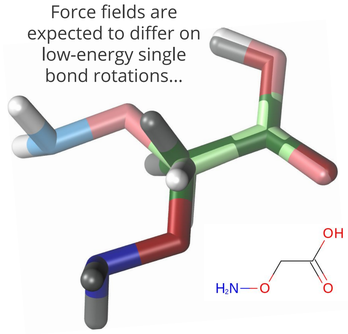


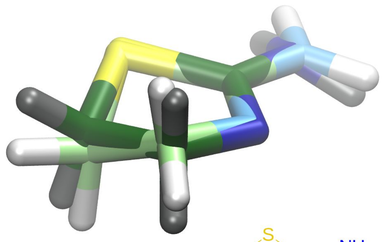


Overlaid conformers display differences in optimized geometry

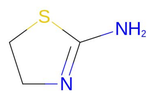
Some differences between force fields are expected, but others display gross differences in parameterization

Force fields are expected to differ on low-energy single bond rotations...





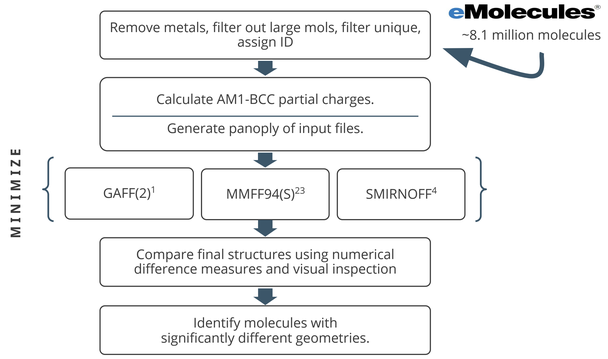
...but not on high-energy hybridization differences



Our goal is to identify sets of molecules that are abundant in these parameterization differences.

Methods

Millions of molecules were generated and minimized for use in this project

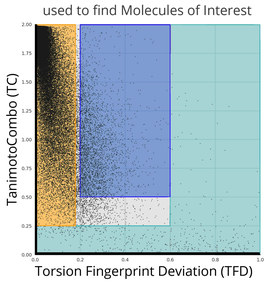


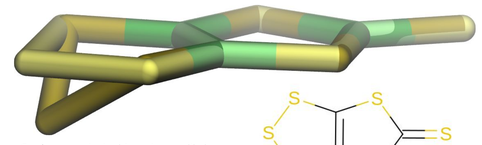
1. Wang, J.; et al. *J. Comp. Chem.* 2004. 2. Halgren, T. *J. Comp. Chem.* 1999. 3. Halgren, T. *J. Comp. Chem.* 1996. 4. Mobley, D.; et al. *BioRxiv*. 2018.

Development

A combination of torsional and coordinate-based evaluation methods can identify molecules with likely parameterization differences

TanimotoCombo[5] and TFD[6] can be used to find Molecules of Interest





Conformer pairs in this region are likely to contain a parameter difference, and are assigned a difference flag.

Conformer pairs in this region are sufficiently similar in shape, and are assigned a similarity flag.

Conformer pairs in this region are often false-positives for geometric differences due to indexing and chirality issues, and so are removed from our analysis.

TFD	.417
TC	1.95

This data can be used to find interesting sets of molecules

	MMFF 94	MMFF 94S	GAFF	GAFF2	SMIRN OFF
MMFF 94					
MMFF 94S					
GAFF					
GAFF2					
SMIRN OFF					

Molecules where one force field is different, and all others are in consensus

Shown: All combinations including SMIRNOFF yield a difference flag, while all other combinations yield a similarity flag.

	MMFF 94	MMFF 94S	GAFF	GAFF2	SMIRN OFF
MMFF 94					
MMFF 94S					
GAFF					
GAFF2					
SMIRN OFF					

Molecules where one family of force fields is different, and all others are in consensus

Shown: MMFF94 and MMFF94S yield a similarity flag together, but difference flags with any other force field.

	MMFF 94	MMFF 94S	GAFF	GAFF2	SMIRN OFF
MMFF 94					
MMFF 94S					
GAFF					
GAFF2					
SMIRN OFF					

Molecules with many difference flags

Shown: Molecules with more than five difference flags, regardless of origin, are useful for future force field development.

5. Hawkins, P.; et al. *J. Med. Chem.* 2006. 6. Schulz-Gasch, T.; et al. *JCIM.* 2012.

Results

This pipeline was applied to a 5.1 million molecule subset

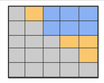
1.56 million conformer pairs yielded difference flags...

	MMFF94	MMFF94S	GAFF	GAFF2	SMIRNOFF
MMFF94		7180	133556	114407	244213
MMFF94S			124287	106923	222494
GAFF				70512	251937
GAFF2					294324
SMIRNOFF					

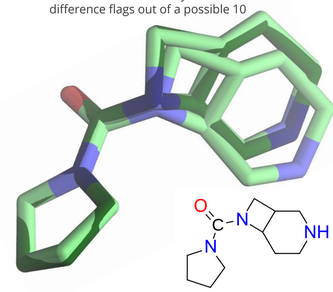
And 26.78 million conformer pairs yielded similarity flags

	MMFF94	MMFF94S	GAFF	GAFF2	SMIRNOFF
MMFF94		2927918	2715722	2742077	2518282
MMFF94S			2727395	2750756	2547336
GAFF				2835968	2543741
GAFF2					2487888
SMIRNOFF					

These molecules can then be sorted into sets of interest

Description	Diagram	Molecule Count
Molecules where SMIRNOFF is different and all other force fields are in consensus		38137
Molecules where the MMFF94 family is different than all other force fields, which are in consensus		6228
Molecules with five or more difference flags, regardless of the force fields of origin	*Various*	43637

55 molecules were found that yielded a total of 9 difference flags out of a possible 10



Analysis

These sets of molecules can be characterized by the frequencies of structural descriptors within them

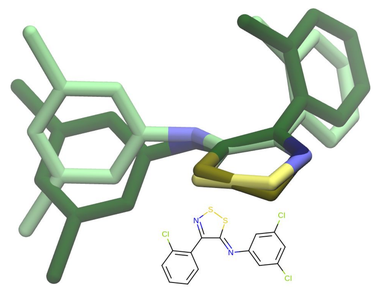
Specific structural descriptors were generated from pairs of checkmol[7] functional groups

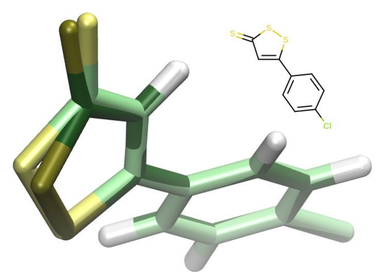
The most overrepresented structural descriptor in the 5 or more difference flag set was iminohetarene / disulfide molecules

Proportion of Molecules with Descriptor in 5+ Dif. Flag Set	Proportion of Molecules with Descriptor in Total Set	Overrepresentation Factor
$2.18 * 10^{-3}$	$9.97 * 10^{-4}$	21.8

The second most overrepresented structural descriptor in the 5 or more difference flag set was arylchloride / disulfide molecules

Proportion of Molecules with Descriptor in 5+ Dif. Flag Set	Proportion of Molecules with Descriptor in Total Set	Overrepresentation Factor
$6.25 * 10^{-4}$	$5.26 * 10^{-5}$	11.9





7. Haider N, The checkmol/matchmol homepage. http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html

Acknowledgements

NIH Grant R01GM108889
UCI's Greenplanet Cluster is supported in part by NSF Grant CHE-0840513