A Knowledge-Based Approach to the Derivation of Bonded Parameters for Small Molecules

www.ccdc.cam.ac.uk

Florian D. Roessler[a], Oliver Korb[b], Robert Glen[a], Peter J. Bond[c]
[a] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge,CB2 1EW, UK
[b] The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK
[c]Bioinformatics Institute (A*STAR), Singapore; Department of Biological Sciences, National University of Singapore, Singapore

Introduction

One of the central assumptions made in force-field (FF) development is the transferability of parameters. Based on this principle, most common FFs supply parameters for e.g. bond lengths or valence angles. These are represented by atom-type combinations derived from model compounds that can readily be applied to novel, not yet parameterized molecules. This study assesses the impact of modifying the general parameterisation protocol at two essential steps: atom-typing, and the replacement of *ab initio* fitted parameters by experimentally-refined parameters. In both cases this is done by comparing results based on information derived from the Cambridge Structural Database (CSD) with data from the AMBER general force-field (GAFF) and the CHARMM general force-field (CGenFF).

Representation of Chemical Environments and Parameter Set Size across Approaches

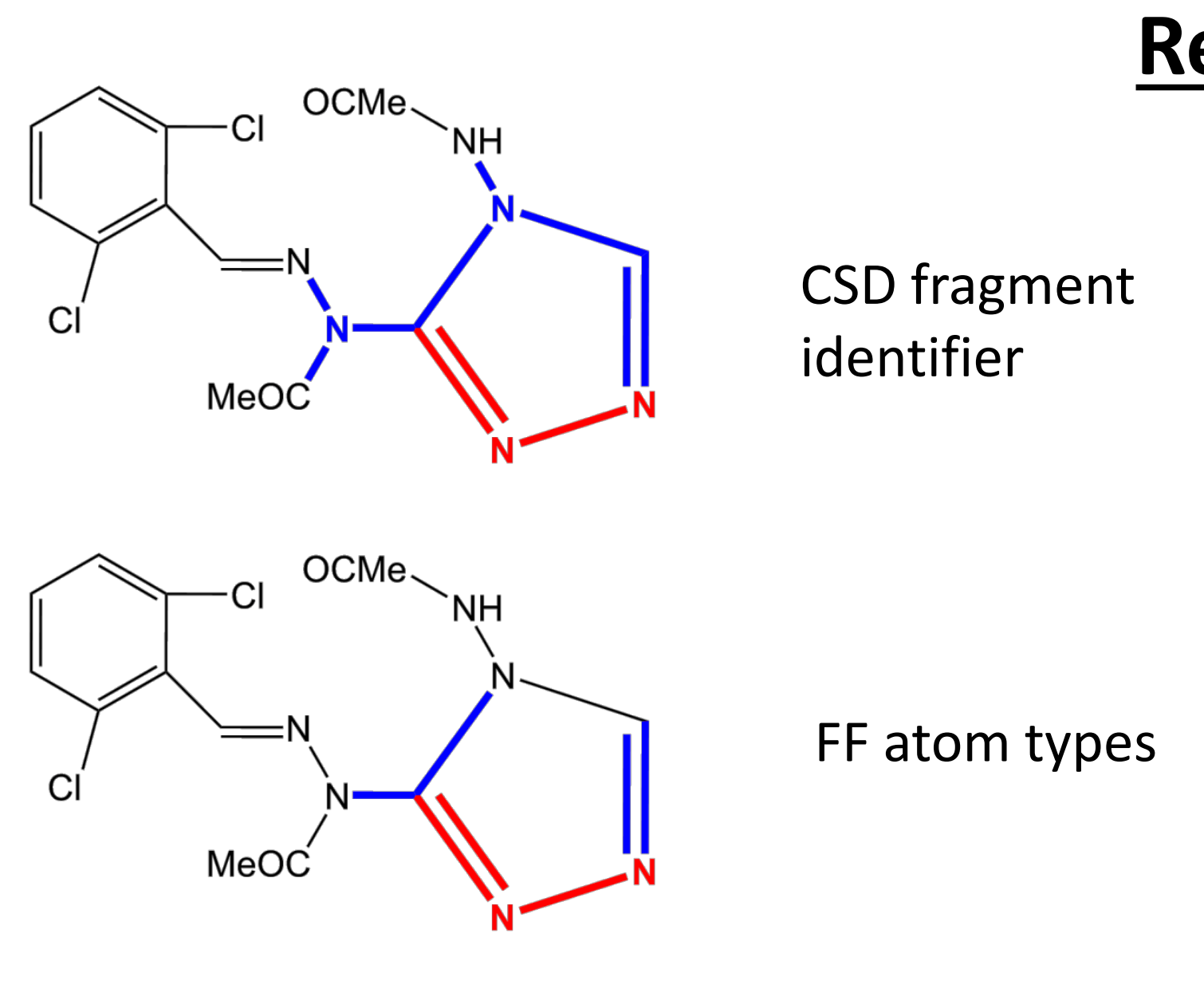

Figure 1: The chemical environment of the valence angles (red) is dependent on the atom and bond types highlighted in blue.

- Fragment identifiers (CSD) take more local chemistry into account than current FF atom type combinations (Figure 1)
- The number of parameters associated with a bonded interaction limits the level of detail with which an approach can describe the chemical space (table on the right)

# of parameters	GAFF[1]	CGenFF[2]	CSD (2004)[3]
Atom types	57	156	-
Bonds	52	492	~6500
Angles	-	1518	~9700
Dihedrals	200	3136	~4000

Comparison of CSD-derived and Force-Field Parameters

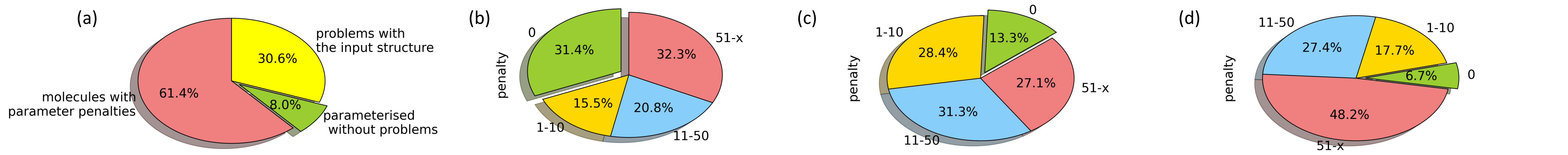

Figure 2: (a) CGenFFs performance in parameterising 10,000 random CSD molecules. (b, c, d) Penalties[2] assigned to bond, angle and dihedral parameters, respectively. Parameters with no penalty can be used straight away, penalties between 1-10 need caution while penalties higher than 10 need small or extensive refinement.

- 10,000 randomly selected CSD molecules were parameterised using CGenFF and GAFF
- Unknown chemistries are parameterised based on their chemical analogy to known chemistries. CGenFF quantifies the degree of analogy through a penalty score (Figure 2)
- Over 50% of bonded parameters assigned by analogy have high penalties and need refinement either through an knowledge or QM based approach
- Comparing CSD and FF data (Figure 3) shows that valence angles (and bond lengths) can be refined based on equilibrium values derived from CSD structures using the MOGUL[3] software

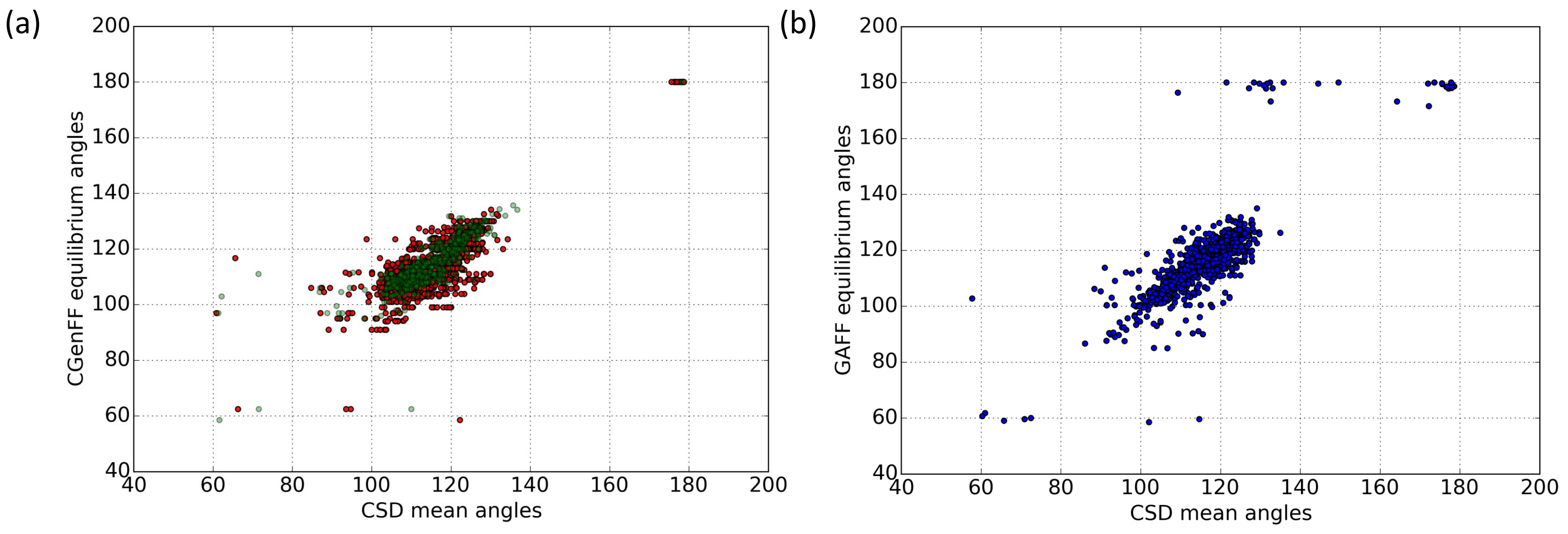

Figure 3: Comparison of FF equilibrium values against CSD derived mean values for (a) CSD plotted against CGenFF. Values in green have no penalties while values in red have penalties higher than one. (b) CSD plotted against GAFF, for GAFF no penalties are provided

- Outliers in the comparison (Figure 3) can identify potential errors that arise during the parameterisation procedure

Comparison of CSD-derived Statistical Potentials and Quantum Mechanical Potentials for Dihedrals

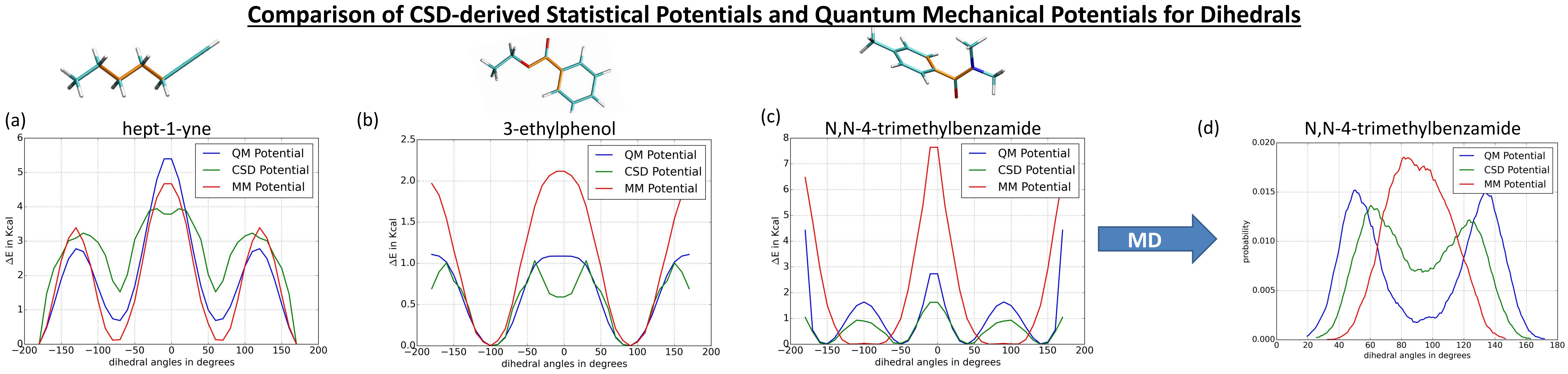

Figure 4: (a-c)Different target energy potentials for dihedrals highlighted in orange. (d) Probability distributions of the dihedral highlighted in (c) when using the individual potentials during an MD simulation

- Statistical potentials were generated via an inverse Boltzmann approach $P(x) = -RT \log\left(\frac{x_i}{x_{ref}}\right)$
- QM potentials were calculated in 10° intervals using B3LYP 6-31G*
- Molecular Mechanics (MM) potentials were calculated from minimisation using geometries at the same intervals as in QM before application of any correction potential

- Good agreement between QM potentials and statistical CSD potentials
- Speed-ups up to factor of 50 can be achieved using the knowledge based approach

Conclusions and Future Work

- Fragment identifiers result in parameters with a higher level of detail
- A statistical approach using crystal databases can provide dihedral target data faster than QM approaches while maintaining a similar level of accuracy
- This approach may also be used to derive parameters for other FF types, such as those employed in coarse-grained models

References

1) Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. a. Development and Testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–74 (2004).
2) Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31,** 671–90 (2010).
3) Bruno, I. J. *et al.* Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44,** 2133–44 (2004).

Acknowledgments:

Cambridge Crystallographic Data Centre, Department of Chemistry Cambridge, High Performance Computing Centre Cambridge, Archer National Supercomputing Service