Escaping atom types in force fields using direct chemical perception

David L. Mobley,*,† Christopher I. Bayly,‡ John D. Chodera,¶ Caitlin C. Bannan,§ Nathan M. Lim,§ Michael R. Shirts,‖ and Michael K. Gilson⊥

Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine, OpenEye Scientific Software, Memorial Sloan Kettering Cancer Center, Department of Chemistry, University of California, Irvine, Department of Chemical Engineering, University of Colorado, Boulder, and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego

E-mail: dmobley@mobleylab.org

Abstract

Molecular mechanics force fields have typically relied on indirect chemical perception to assign parameters, typically by way of atom types. Particularly, first a molecule is processed by a machinery which recognizes chemical environments and assign atom types, and then the atom-typed system is processed to assign parameters. Thus, assignment of atom types either relies on specialized chemical perception code or manually assigned via a template that matches an overall chemical unit. While this approach

*To whom correspondence should be addressed
†University of California, Irvine
‡OpenEye Scientific Software
¶Memorial Sloan Kettering Cancer Center
§University of California, Irvine
‖University of Colorado, Boulder
⊥University of California, San Diego

has served well, it impairs flexibility, as atom types must encode all of the information necessary for assigning parameters. This can require considerable additional complexity in atom typing when additional parameters are needed, when the underlying chemistry does not necessarily merit such additional complexity. Here, we describe a new approach to defining molecular mechanics force fields based on the standard SMARTS chemical perception language (with extensions to identify specific atoms available in SMIRKS), where each force field term (atoms, bonds, angles, and torsions) features separate definitions assigned in a hierarchical manner that is free of atom types. Here, we accomplish this using direct chemical perception, where parameters are assigned directly based on queries operating on the molecule being parameterized, thereby avoiding the intermediate step of assigning atom types. This allows for substantial simplification of force fields, as well as additional generality. Additional flexibility can also be gained by allowing force field terms to be interpolated based on the assignment of fractional bond orders via the same procedure used to assign partial charges. This approach is flexible and applicable to a wide variety of (bio)molecular systems, and can greatly simplify the number of parameters needed to create a complete force field. As an example of its utility, we provide a minimalist small molecule force field derived from Merck's parm@Frosst (a parm99 descendant), in which a parameter definition file only 300 lines long can parameterize a large and diverse spectrum of pharmaceutically relevant small molecule chemical space.

Introduction

Classical all-atom force fields see widespread use in molecular simulations in diverse areas in chemistry, biochemistry, biology, drug discovery, and materials [references]. Often, these are two-body additive fixed-charge force fields with the relatively simple Lennard-Jones functional form for nonpolar interactions. However, despite this simplicity, force fields have achieved remarkable successes predicting a wide range of properties and behaviors far beyond the simple condensed phase [refs] or biomolecular [refs] properties they have been pa-

rameterized for. For example, successful predictions include protein-ligand binding affinities [refs], hydration free energies [refs], partitioning coefficients [refs], dielectric constants [refs], ligand binding modes [refs], and many others. It seems safe to say these relatively simple models have succeeded far beyond original expectations, likely in part because of their strong physical basis, careful parameterization, and a reasonable balance of speed versus accuracy for any of these applications.

One key concern in developing force fields is the balance of accuracy versus generality. Since the underlying functional form is certainly approximate, it is always possible to improve accuracy by adding more parameters which are tuned to particular use cases. Of course, in some cases this is warranted; for example, a tetrahedral carbon obviously requires a different bonding geometry than a planar one and any model missing this will result in major structural errors. However, in some cases specialization may be unwarranted or even lead to problems of transferability due to overfitting. Typically, the major, general-purpose force field families used for molecular simulations have tried to achieve the right level of balance in this regard, with sufficient chemical sophistication to ensure adequate coverage of major distinct chemical functionalities, while also not adding additional unnecessary parameters. Perhaps, force fields have often managed to hit the presumed sweet spot balancing accuracy with generality, thus enabling the remarkable breadth of applications seen in the field.

One major challenge in force field development is the amount of human time and expertise involved. Development of a new general force field (covering all or almost all of normal organic chemistry and biomolecules) from scratch typically takes many years, judging based on historical precedent. For a concrete example, consider the AMOEBA polarizable force field, for which the first publications appeared in [year, ref] and a truly general force field is still forthcoming, in that applying to new small molecules still requires a considerable amount of expert attention and parameterization [refs]. Thus, while there have been numerous adjustments to biomolecular force fields over the years, especially terms relating to proteins and nucleic acids [refs], the core of most of our present-day force fields, at least aside from

the torsions and charges, still seems to typically date to the 1980s and early 1990s. Building a new fixed-charge force field from scratch would simply require too much human time, and too large an investment of effort, over too long a time for academic groups to tackle the problem, not to mention the fact that funding is difficult to impossible. Small molecule force fields have thus received much less attention than biomolecular force fields, and have typically been developed at least in part by generalizing biomolecular force fields to cover more chemical space [ref GAFF, GAFF2?]. This is in part because small molecule force fields necessarily introduce vastly more chemical complexity – and when human expertise and time plays a key role in force field development, it means they require vastly more time to develop.

At the same time, fixed charge force fields show clear room for improvement. Certainly not all accuracy problems are due to force field problems, but it seems increasingly clear that results of calculations often are quite sensitive to force field parameters [Rocklin sensitivity; Gilson host-guest stuff, something from Merz], and that force field issues do result in significant systematic errors in a fairly wide range of cases [refs] Indeed, in some cases, systematic errors can be traced back to problems with force field parameters for particular functional groups [ref hydration papers; papers citing them] and follow-up work can in some cases fix these issues. For example, GAFF parameters yielded systematic errors for alkenes which could be fixed by a minor adjustment to Lennard-Jones parameters [ref] and larger errors for alcohols in general due to issues with underpolarization of the hydroxyl group which were fixed by a focused effort [ref]. But these isolated efforts serve as band-aids rather than a general fix, and are themselves human-intensive.

The evidence seems clear that a new generation of fixed-charge force fields developed from scratch could do dramatically better than our current force fields, but the investment of human time and expertise required is so large that no general effort has gone forward. In our view, the solution to this problem is to dramatically reduce the amount of *human* effort required for force field development. This can be achieved by automating the force field

development process so that human expertise is used to select the input data for parameterization and the functional form, but then a completely automated machinery produces the force field itself.

While a reasonable amount of effort has gone into improving fitting of parameters for a particular force field given input data, such as in the Force Balance effort [ref], this approach still requires a great deal of human expertise deciding which parameters need to be fitted. Particularly, a human expert must decide how many atom types (and thus how many bond, angle, torsional, Lennard-Jones and charge parameters) are needed to represent all of the relevant chemistry, and then, given these choices and others, automated machinery can go to work. To automate this process, we need to reduce the human expertise required even in this early stage of the process.

Atom typing can be thought of as a type of *indirect chemical perception*, where a molecule or molecules are processed via some machinery to assign labels to atoms (atom types) and then these labels are subsequently processed to assign parameters. Thus, key for success is ensuring that the atom types encode all of the relevant information but no unnecessary information, as once parameterization is done, the atom typing is considered fixed. *Direct chemical perception*, in contrast, would assign parameters based on processing of a molecule itself via a chemically-aware engine. To see the distinction, note that force fields in the AMBER force field family do not retain bond order when assigning parameters, so if any bond order information is necessary, this must be encoded in the labels or atom types themselves, as we discuss further below. In contrast, a tool doing direct chemical perception can use information about a molecule such as bond order, as it operates directly on the molecule itself rather than simply on a labeled graph representing the molecule.

In our view, indirect chemical perception based on atom typing results in at least three subsequent challenges in force field development. First, indirect chemical perception results in unnecessary parameters and thus overly complex force fields, because introduction of a new atom type because it is needed for one parameter type (such as a torsion) results in

a need for new parameters across all parameter types (such as van der Waals, bond, and angle parameters). Second, it results in unnecessary atom types because the atom types must encode all necessary information about the relevant chemistry. Third, atom typing is typically hard-coded, difficult to change, and requires great human insight in deciding which information to encode in atom types and how, adding to the expertise required in force field development.

Here, we introduce an alternative to atom typing, direct chemical perception, where parameters are assigned directly based on processing a molecule rather than based on processing a molecular graph labeled with predefined atom types. Since the parameterization engine has access to the molecule itself, this provides a variety of benefits, including that bond order information is available, as well as as much information as needed about the chemical environment, and a variety of tools can be applied in parameterization, including (if needed) electronic structure calculations. This also allows the chemical perception to be easily changed on the fly rather than hard-wired and, potentially, learned via a computer (as we will explore in a separate study).

In this work, we introduce a specific implementation of direct chemical perception, based on the chemical query language SMARTS [refs] and its extension in SMIRKS, use it to develop a new, SMIRKS native open force field (SMIRNOFF) format, and implement some AMBER-family force fields covering a small region of chemical space in this format. We show here how SMIRKS, and the SMIRNOFF format, can dramatically reduce the complexity (in terms of number of apparently independent parameters) in existing force fields while still yielding the same energies, while at the same time allowing a variety of new innovations which would be quite difficult in typical force fields. We also introduce a new force field, SMIRNOFF99Frosst, which is a prototype general small molecule force field in the SMIRNOFF format, and an AMBER-family descendant of Merck's Merck-Frosst force field.

Background

Theory and methodology

Usage examples and prototypes

Discussion and conclusions

Acknowledgement

DLM and CCB appreciate the financial support from the National Science Foundation (CHE 1352608) and the National Institutes of Health (1R01GM108889-01) and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513. JDC appreciates support from the Sloan Kettering Institute and NIH grant P30 758 CA008748. We appreciate helpful discussions with Christopher Fennell (Oklahoma State), Bryce Manubay (University of Colorado), and Patrick Grinaway (MSKCC).

Supporting Information Available

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Graphical TOC Entry