PROJECT SUMMARY/ABSTRACT

The study of biomolecular interactions and design of new therapeutics requires accurate physical models of the atomistic interactions between small molecules and biological macromolecules. Over the least few decades, molecular mechanics force fields have demonstrated the potential that physical models hold for quantitative biophysical modeling and predictive molecular design. However, a significant technology gap exists in our ability to build force fields that achieve high accuracy, can be systematically improved in a statistically robust manner, be extended to new areas of chemistry, can model post-translational and covalent modifications, are able to quantify systematic errors in predictions, and can be broadly applied across a high-performance software packages.

In this project, we aim to bridge this technology gap to enable new generations of accurate quantitative biomolecular modeling and (bio)molecular design for chemical biology and drug discovery. In **Aim 1**, we will produce a modern, open infrastructure to enable practitioners to rapidly and conveniently construct and employ accurate and statistically robust physical force fields via automated machine learning methods. In **Aim 2**, we will construct open, machine-readable experimental and quantum chemical datasets that will accelerate next-generation force field development. In **Aim 3**, we will develop statistically robust Bayesian inference techniques to enable the automated construction of type assignment schemes that avoid overfitting and selection of physical functional forms statistically justfied by the data. This approach will also provide an estimate of the *systematic error* in predicted properties arising from uncertainty in parameters or functional form choices—generally the dominant source of error—to be quantified with little added expense. In **Aim 4**, we will integrate and apply this infrastructure to produce open, transferable, self-consistent force fields that achieve high accuracy and broad coverage for modeling small molecule interactions with biomolecules (including unnatural amino or nucleic acids and covalent modifications by organic molecules), with the ultimate goal of covering all major biomolecules.

This research is significant in that the technology developed in this project has the potential to radically transform the study of biomolecular phenomena by providing highly accurate force fields with exceptionally broad chemical coverage via fully consistent parameterization of organic (bio)molecules. In addition, we will produce new tools to automate force field creation and tailoring to specific problem domains, quantify the systematic error in predictions, and identify new data for improving force field accuracy. This will greatly improve our ability to study diverse biophysical processes at the molecular level, and to rationally design new small-molecule, protein, and nucleic acid therapeutics. This approach will bring statistical rigor to the field of force field construction and application by providing a means to make data-driven decisions, while enhancing reproducibility by enabling it to become a rigorous and reproducible science using a fully open infrastructure and datasets.

SPECIFIC AIMS

Molecular simulation is a powerful tool to predict the properties of biomolecules, interpret biophysical experiments, and design new small molecules or biomolecules with therapeutic utility. In recent years, biomolecular simulation software and applications have benefited from massive increases in computational power. However, improvements in the force fields at the heart of simulations have proceeded at a slower pace. As a consequence, the utility of molecular simulations is still significantly limited by the insufficient accuracy and limited domain of applicability of current force fields. The overall goal of this project, therefore, is to improve the accuracy of simulations by driving the development of new and improved force fields. To do this, we will develop and apply new technology, comprising an open, extensible, and shared software and data infrastructure, implementing statistically robust methods of parameterizing force fields and of choosing the structure of force fields in a statistically sound manner. This infrastructure will enable modelers to rapidly create new force fields that incorporate new data, choose and utilize new functional forms, and extend to new chemistries, significantly enabling computational biophysical and biomedical investigation. This work will create not just a new generation of force fields, but an open technology to continue advancing force field science. This will be accomplished via the following Aims:

Aim 1: Create a modern, open software infrastructure for automatically generating and validating force fields and utilizing them broadly in modeling packages. We will develop an automated, easy-to-use infrastructure for generating, optimizing, and validating force fields using specified experimental and quantum chemical datasets. These tools will allow practitioners to develop, refine, validate, and apply highly accurate force fields with either broad biomolecular coverage or tailored to application domains of interest, in a manner that enables their use in all major modeling packages. In doing so, we will reformulate legacy atom type based force fields to use *direct chemical perception* via modern cheminformatic tools. This reformulation will resolve a decades-old design flaw, allowing us to simplify and streamline force field development, especially for small molecules. We also propose new ways to accelerate the parameterization process, making it orders of magnitude faster by using hierarchical surrogate models of simulation properties.

Aim 2: Construct open datasets and databases for next-generation force field development. Constructing and validating high-accuracy force fields requires access to large quantum chemical datasets and curated experimental measurement datasets. We will create and provide rapid, facile access to large, high-quality, open, and revision-controlled datasets. First, we will construct a quantum chemical database to provide high-accuracy, comprehensive, and easily queried large-scale *ab initio* datasets useful for parameterizing small molecules with significantly expanded coverage of drug-like space, as well as chemical fragments of a broad range of biomolecules and biological macromolecules. Second, we will generate high-quality physical property and biophysical datasets to inform the training and validation of parameters we are fitting. Third, we will automate the synthesis of chemically diverse host molecules, and the measurement of their binding thermodynamics with varied guest molecules, generating a large new dataset to drive creation of force fields well-suited for modeling noncovalent binding.

Aim 3: Develop Bayesian inference techniques to address key questions in force field physical modeling and predict systematic error. Traditional approaches to force field development based on minimizing an error function provide little information regarding which models of physical interactions, differing in either functional form or number of parameters, are most appropriate, which new data would most rapidly improve prediction accuracy, and when predictions will be unreliable. We will embed the force field parameterization problem in a Bayesian context, providing practical, statistically robust tools to address these questions. The same Bayesian formalism will allow us to identify minimal sets of experiments that maximally reduce force field uncertainty.

Aim 4: Integrate the results of Aims 1–3 to construct and validate open source, transferable, and self-consistent force fields for small molecule interactions with complex biomolecular systems. We will utilize the infrastructure developed here to produce an open set of versioned force fields applicable to , along with versioned datasets used to generate the force fields so that practitioners can utilize or further refine them with in-house data. From the same datasets, we will generate consistent sets of force fields of varying functional complexity (such as fixed-charge and polarizable electrostatics) to enable systematic study of accuracy-generalizability of more complex models. These force fields will be validated by comparison to experimental NMR, X-ray, HDX, and temperature-dependent measurements, using open and automated simulation workflows to ensure reproducibility.

To accomplish these Aims, we have established the Open Force Field Initiative, a multidisciplinary team with extensive individual and collaborative experience in force field development and related areas to integrate and apply the diverse technologies involved. The resulting infrastructure and force fields will have a transformative effect on the rapidly growing field of biomolecular simulation and modeling through increased accuracy, expanded domains of applicability, and facile extensibility to meet modern challenges in biology and drug discovery.

SIGNIFICANCE

Molecular simulations are powerful but have not reached their full potential. Atomistic molecular simulations can now address significant biomolecular questions encountered in human health and disease and provide predictions of small molecule binding affinity, selectivity, and bioavailability.[7–18] While we now possess the techniques, algorithms and computing power to investigate many details of biomolecular function and guide small molecule and biomolecular design, these techniques could still be significantly improved in both predictive accuracy and domain of applicability in order to broadly probe disease mechanisms and enable truly *de novo* molecular engineering.

More accurate force fields are needed to overcome current limitations in small molecule design. Advances in computing power have enabled massive increases in conformational sampling, with μs-timescale simulations now routine.[19] Although computing power is not yet sufficient for many problems involving very long time scales, it has become clear the accuracy of current force fields limits the utility of atomistic modeling where available computational power is much less of a limitation.[16;20;21] With the specialized supercomputer Anton 2[22] that has enabled simulation to longer time scales, the Shaw group has also found it necessary to address force field inaccuracy in protein structural dynamics.[23;24] In the cases where sampling challenges are easily overcome, such as the prediction of binding thermodynamics of druglike small molecules to macromolecular hosts, or small molecule transfer free energies, force field inaccuracy is consistently shown to limit predictive utility.[11;12;25–31] Force fields are still insufficiently accurate to deliver on the promise of eliminating trial-and-error cycles in real-world molecular design projects.[7;8] While recent small molecule force field efforts such as GAFF,[32;33] GAFF2,[34] CGenFF,[35–37] and ATB[38] represent notable progress, problematic legacy assumptions and errors are retained[1] and modern models are still lacking in predictive accuracy.[35;36;39] While the proprietary OPLS3[40;41] reports somewhat improved accuracy within a legacy force field framework,[9;42;43] its license terms prohibit independent accuracy benchmarks or the decryption of its parameters, effectively prohibiting improvements or examination of failures. It is clear from users in the pharmaceutical industry that OPLS3 provides insufficient accuracy and coverage of chemical and biomolecular space (see letters from Bayer, BI, BMS, GSK, Merck, Pfizer, and Roche).

Significant changes in force field parameterization approaches are needed to accelerate progress in force field development. The core of most present-day force fields—pairwise-additive nonbonded interactions using partial charges and Lennard-Jones interactions—dates primarily from decades-old work.[44;45] Over the past decade, new force field development has primarily been limited to small, incremental improvements on the existing paradigm;[37;40;41;46–50] Even the recent COMPASS II[51] and OPLS3,[40;41] with significant commercial effort, are primarily partial parameter refits. The slow pace of progress stems from several causes that can and should be changed: *(1)* Current force fields are built via decades of human decisions guided by chemical intuition rather than reproducible methodologies, making it difficult to backtrack on decisions made long ago that prove limiting for accuracy or extensibility, or to refit the entire force using new datasets. *(2)* Parameterization procedures lack a statistical framework to prevent overfitting, resulting in a proliferation of atom types and hence valence parameters. *(3)* Legacy software infrastructure—in particular, custom atom-based typing models—frustrates the development of force fields that can generalize to the breadth of chemical functionalities of interest. *(4)* Academic efforts to build new force fields, such as the AMOEBA and Drude approaches, focus on wholly new physical models such as multipoles and polarizability that can carry significant performance penalties[52–54] without consistent validation strategies to specifically assess the impact of these functional forms. Evidence suggests the accuracy of faster fixed-charge force fields may still be substantially improved,[12;37;41;50;55–57] so we must quantify the predictive gain of more detailed and expensive models in order to properly weigh accuracy/expense tradeoffs.

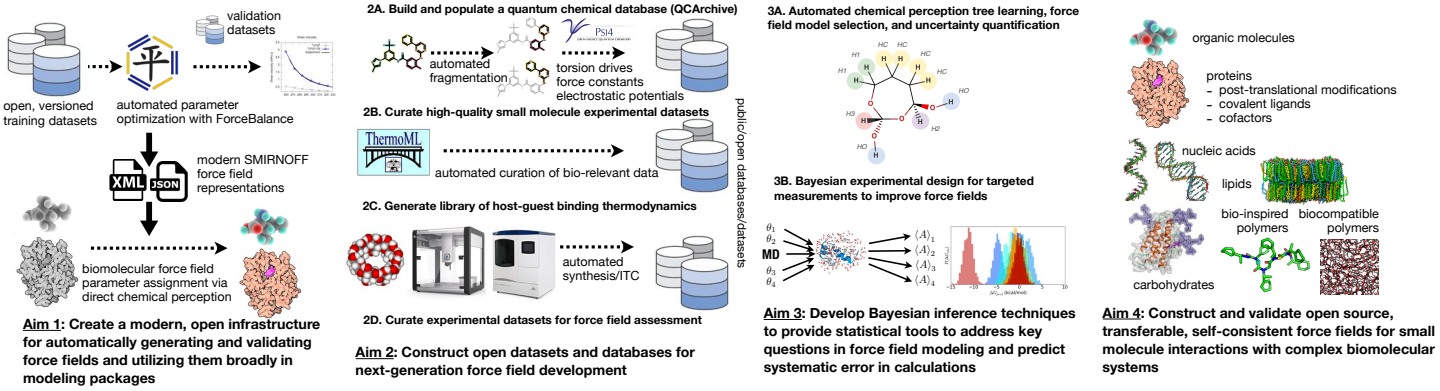

Figure 1. **The proposed work will develop technologies for producing significantly improved force fields for molecular simulation.**

There is an unmet need for force fields that are more chemically complete. Beyond the force field issues in protein-ligand interactions described above, key classes of biomolecules cannot be accurately simulated with current force fields. Nucleic acid force fields have lagged behind protein force fields despite recent progress,[46;54;58] Greater accuracy is needed to model or design protein-DNA (e.g., transcription factors[59]) or protein-RNA-DNA (e.g., CRISPR[60;61]) interactions. Lipid force fields suffer from accuracy issues when simulating membranes[62–64] while carbohydrates and glycoproteins are difficult to simulate with current force fields. Models such as GLYCAM[65] have made some headway but development has not kept pace with proteins. Protein force fields still require improvement,[21;24] with accurate modeling of intrinsically disordered proteins being particularly difficult.[24;24;66;67] Heterogeneous systems, combining standard amino acid or nucleic acid residues with nonstandard/modified ones, pose particular problems. Examples include proteins with post-translational modifications, unnatural amino acids, and biomolecules bound to covalent inhibitors. In principle, a force field should assign consistent parameters to such systems, but in practice, the parameters for components of a heterogeneous system (if available) are often developed by different research groups following different strategies, resulting in models that risk incompatibility.

Force field development must move from manual decisions to reproducible, statistically-guided processes. There is a clear need for software tools and resources to automate the process of building physical force fields, tools which also draw on statistically sound approaches to avoid overfitting and ensure generalizability. Current force fields include hundreds or thousands of adjustable parameters fit to small and inadequate datasets in stages, or using trial and error, guided by human expertise and intervention. Existing force field efforts generally target one class of biomolecules, are executed independently by different groups, share little or no common infrastructure or datasets, and lack clear ways to subsequently improve the resulting force fields. The field needs better mechanism to systematically, automatically, and reproducibly parameterize general, accurate atomistic force fields from disparate experimental datasets with minimal human intervention, hindering our ability to learn from and correct failures. Current-generation force fields also lack the ability to quantify their systematic error in predicted properties, leaving practitioners without tools to assess the dominant source of error in predictions.

To resolve these challenges, we will automate fully consistent parameterization of small molecule and biopolymer force fields using a modern approach free from past limitations on extensibility. Our proposal focuses on building tools and datasets to drive the next generations of force field development, resolving these issues and providing a new generation of force fields. Our open infrastructure and data will also enable existing force field developers to advance their own efforts using our innovations, bringing broad benefits to the community.

INNOVATION

This project will yield the first open, scalable data and software infrastructure designed to reproducibly build, apply, and validate statistically robust force fields, and introduces significant innovations at multiple levels:

Fully automated force field parameterization breaks from tradition. We will implement a new infrastructure that will no long require force field creators to decide what atom types are needed *a priori*, or which parameters should be optimized; instead, creators will assemble training and validation datasets, select physical functional forms to consider (e.g., fixed partial charges with the point polarizable dipoles or multipoles where needed), and choose a model representation (e.g., constrained hydrogens, virtual-site point charges for lone pairs, etc.). The infrastructure will automatically evaluate the evidence supporting each model, defined by both choice of parameters and if desired, choice of functional form. This approach is a complete break from traditional incremental, “artisanal” parameterization. To make this possible, we will integrate modern, scalable software infrastructures to coordinate automated parameterization efforts across available computational resources while keeping the software modular, so that researchers can easily integrate new experimental or quantum chemical data sources.

We will develop an open collection of high-quality data sets designed to enable reproducible biomolecular force field construction and validation. Our effort will collect, curate, and organize high-quality open datasets for small molecule and biomolecular force field parameterization, coordinating with existing force field development communities (see letters from Cheatham, Ollila, Roux, MacKerell, Riniker, Roitberg, Simmerling, and Wang) to maximize impact. Working with the Molecular Software Sciences Institute (MolSSI), we will populate a new public quantum chemical database (QCArchive[68])—developed to publicly aggregate and share quantum chemical data—with key datasets for biomolecular force field parameterization using automated workflows, allowing the community to submit molecules to greatly expand coverage of and accuracy within chemical space. To generate experimental datasets necessary to build and validate general small molecule and biomolecular force fields, we will curate high-quality experimental physical property datasets in collaboration with NIST (see Letter), utilize automated synthesis and calorimetry to generate large new thermodynamic datasets for small molecule binding, and compile high-quality experimental biophysical datasets for force field parameterization and validation.

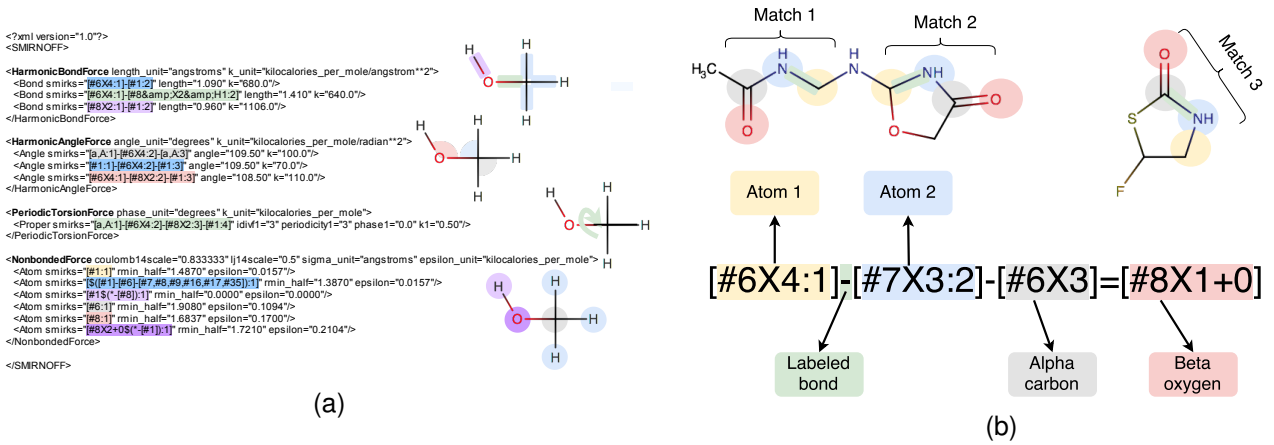

Figure 2. **The SMIRKS Native Open Force Field (SMIRNOFF) approach to direct chemical perception.** (a) Excerpt of the SMIRNOFF 0.1 format representation of a force field covering alkanes, ethers, and alcohols, highlighting terms that match force types in methanol.[1] (b) SMIRKS allows for direct chemical perception via chemical substructure matching; here, a single SMIRKS pattern matches three different substructures in two different molecules, with colors denoting corresponding components of matching SMIRKS string.[1]

Direct chemical perception breaks free from limitations of atom types. Typically, atom types, crafted by human experts, are used to assign force field parameters. These atom types must encode the entire range of distinct chemical environments the force field requires, which hampers extension of these force fields into new areas of chemical space. Assigning bond, angle, and torsion parameters using atom types leads to needless proliferation of parameters, as well as a multitude of opportunities for human error when attempts a new type is introduced without rebuilding the entire force field.[1] We plan to avoid atom typing altogether, and instead assign parameters directly based on the chemical graph.[1] Recently, we introduced a new force field specification format called the **SMIR**KS **N**ative **O**pen **F**orce **F**ield (SMIRNOFF) specification which implements this concept (Figure 2b),[1] which uses the widely supported SMARTS chemical perception standard[69] and the atom-tagging features of SMIRKS[70] to directly assign van der Waals, bond, angle, and torsion parameters based on the local chemical environments. This approach solves numerous problems with legacy atom typing, and easily accommodates new van der Waals types without causing an explosion in additional bonded types. The format is *hierarchical*, with more specific child parameters overriding general parent parameters only as needed, enormously reducing complexity and aiding extensibility (Figure 2a).[1] By adopting SMIRNOFF, our optimization approach can automate determination of *both* the number and definition of types as well as the more traditional numerical parameters (equilibrium values, force constants, Lennard-Jones parameters, etc.).

Bayesian model construction blends the best aspects of physical modeling and data-driven machine learning, automating physical model selection, avoiding overfitting, and estimating error. This project takes a novel approach to automate learning physical force fields from experimental and quantum chemical data, leveraging our knowledge of fundamental physical interactions to construct predictive models that generalize beyond their training data. This approach differs fundamentally from recent work that uses pure machine learning methods to learn physical interaction laws and invariants using general function approximators.[6;71–73] Instead, we embrace physics-based force field functional forms, which have already demonstrated remarkable predictive power and transferability even for properties far outside their expected domain of applicability,[74] have widespread support in high-performance simulation codes, and show clear opportunities for improvement.[12;37;41;50;55–57] Despite the primary focus on physical models, there may be opportunities to use deep learning models[6;71–73] to describe valence interactions, which are not always physically well described by harmonic and Fourier terms, and we will enable the construction and assessment of these models from physical property and quantum chemical data.

Our Bayesian approach uses inferential machine learning to automate the statistically-guided selection of physical models and their parameters in a manner that penalizes complexity,[75–77] ensuring the resulting models generalize broadly. Specifically, we will exploit Bayesian inference to determine the SMARTS typing hierarchy and associated parameters (described above), ensuring that the number of types is statistically grounded in the data.[78;79] Furthermore, the process will generate not just a *single* best-fit force field, but also *ensembles* of parameter sets that can be used to rapidly estimate the *systematic error* in a prediction after a simulation has been conducted, by considering correlated parameter (and typing) uncertainties and their impact on predicted properties.[3;80–82] To make this possible, we employ a set of hierarchical parameter estimation techniques to greatly reduce the cost of evaluating properties at large numbers of of force field parameters. This approach uses reweighting techniques developed by the PIs[83–88] to adaptively construct inexpensive surrogate models that learn how physical properties depend on parameter changes.[3;89]

APPROACH

We will build a modern, open infrastructure for the development of *general* force fields that model organic small molecules, biomolecules, and chemical adducts self-consistently, recasting force field parameterization as a primarily automated machine learning problem that fully exploits known physics. By "modern" and "open", we mean modular, extensible, open source, using and documenting open standards and best practices (see Data Sharing Plan). This infrastructure is needed for transferable, general force fields that easily model new molecular systems.

Aim 1: Create a modern, open software infrastructure for automatically generating and validating force fields and utilizing them broadly in modeling packages.

We will enable reproducible force field parameterization with a new open software infrastructure that combines multiple advanced force field technologies. Practitioners of virtual screening are increasingly screening large libraries, with recent work reporting screens as large as 170M compounds[90]—which will continue to grow with purchasable compound libraries such as eMolecules (22M compounds, `emolecules.com`), Enamine REAL (500M compounds, `enamine.net`), and ZINC15[91] (750M compounds). Parameter assignment must therefore be designed to be fast (seconds or less per molecule). This precludes the use of expensive *ab initio* quantum calculations to assign force field parameters to individual molecules;[92;93] instead, a *general* force field is needed.

We will therefore employ our new SMIRNOFF format[1] (Figure 2b), which can not only make legacy force fields portable, but enables hierarchical atom-type-free force fields that are dramatically more compact (with many fewer adjustable parameters) to cover the same chemical space, providing an excellent, extensible foundation for next-generation force fields. We will use our ForceBalance tool[94–98] to automate force field parameter optimization, and develop an adaptive reweighting and surrogate model construction approach[3;84–86;89] to accelerate physical property estimation for use in parameter optimization and assessment. To benchmark this technology, we will reparameterize the SMIRNOFF99Frosst small molecule force field and evaluate its performance on partition coefficients (logD/logP),[99] relative solubilities,[99] and host-guest binding thermodynamics.[100]

1A: Reformulate legacy atom type based force fields to use direct chemical perception via modern cheminformatic tools, resolving a decades-long design flaw to further force field development.

Preliminary results. A key issue impairing the accuracy, transferability, and extensibility of existing force fields is the use of atom typing. Traditional approaches to defining molecular mechanics force fields use a discrete set of atom types to encode all information about the chemical environments of atoms. Parameters are then assigned by looking up combinations of these atom types in tables. This approach leads to a variety of problems, making it difficult to expand the set of atom types, and leading to unnecessary proliferation of parameters.[1]

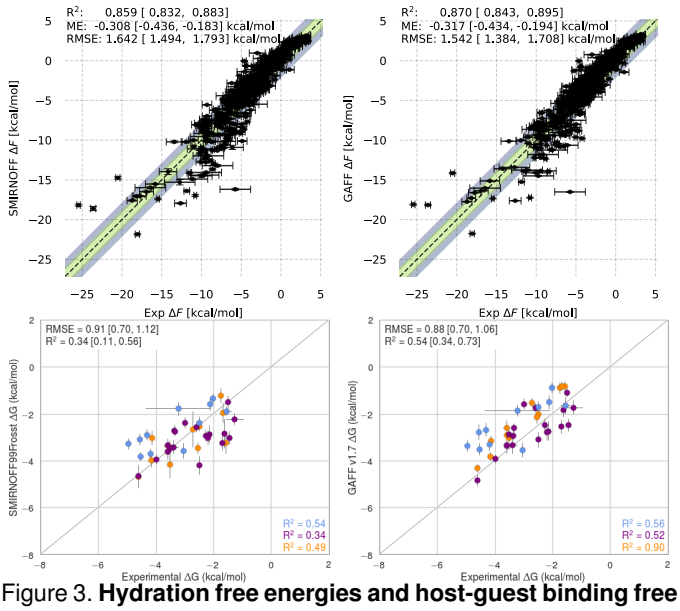

Figure 3. **Hydration free energies and host-guest binding free energies for GAFF and SMIRNOFF99Frosst.** The 300-line unoptimized SMIRNOFF99Frosst force field[1] shows comparable accuracy to the widely used ∼7000-line GAFF[101] for both FreeSolv hydration free energies[102] (top) and host-guest binding free energies[103] (bottom). Densities and dielectric constants show comparable performance.[1]

We have developed a new approach to assigning force field parameters based on *direct chemical perception* using the industry standard SMARTS chemical perception language, with extensions to identify specific atoms available in SMIRKS. In this approach, each force field term (bonds, angles, torsions, and nonbonded interactions) uses separate chemical typing to assign parameters in a hierarchical manner, instead of using atom types. This approach can greatly reduce the number of parameters needed to describe small molecules, biopolymers, lipids, carbohydrates, and small molecules. As a demonstration, we developed a minimalist small molecule force field[1] derived from Merck's parm@Frosst[104] (an Amber *parm99* descendant), in which a parameter definition file only ∼300 lines long can parameterize a large and diverse set of pharmaceutically relevant small molecules, providing better coverage of chemical space than the ∼7000-line GAFF and GAFF2 force fields.[1] Hydration free energies, host-guest binding affinities, densities, and dielectric constants are reproduced with comparable accuracy despite the drastic reduction in complexity and increase in coverage (Fig. 3).[1] Existing tools allow SMIRNOFF-parameterized molecules to be used in all major simulation engines.[1]

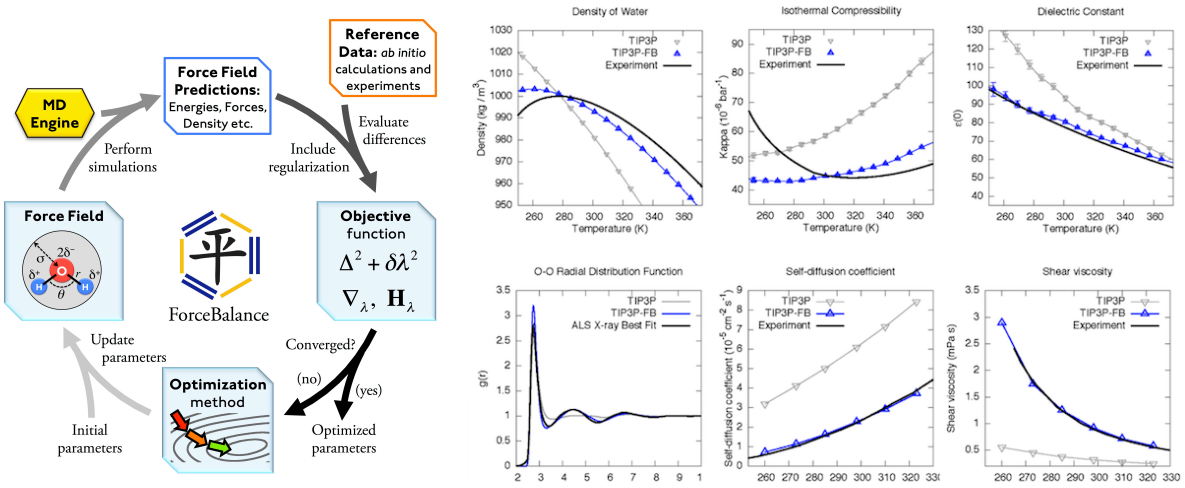

Figure 4. **ForceBalance allows for automated parameterization.** (Left) Workflow: Starting with an input force field, observables are computed and compared with training data. The objective function is calculated and minimized iteratively to improve agreement with training data. This procedure enables fully automated fitting of force fields of diverse functional forms. (Right) TIP3P-FB[95] is a substantially better water model with the same number of adjustable parameters, developed to fit six properties relative to experiment (three of which are shown at top) leading to overall better performance than TIP3P, including properties not fitted shown at bottom.

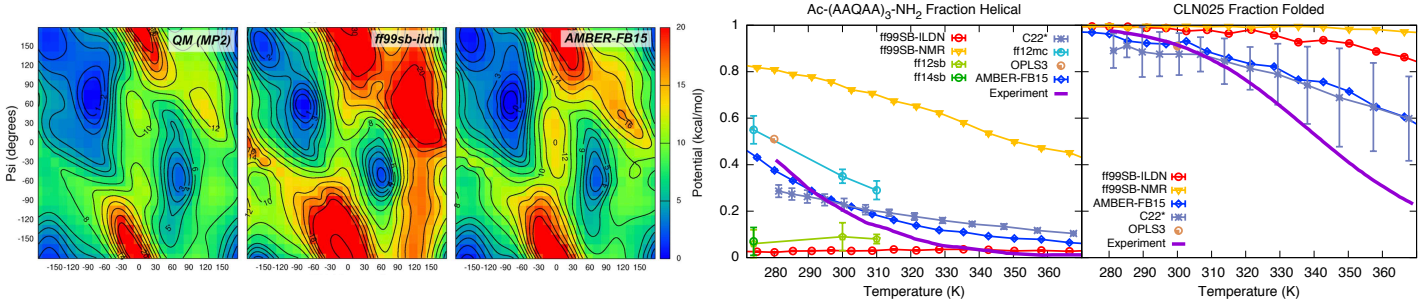

Figure 5. (Left): **ForceBalance yields more accurate potential energy surfaces for proteins.** Shown here are potential energy scans for alanine dipeptide, with quantum mechanics (MP2/aug-cc-pVTZ), AMBER ff99sb-ildn[23] and the new ForceBalance[94–98] protein force field, AMBER-FB15.[98] Color indicates the relative potential energy with respect to the minimum; AMBER-FB15 accuracy relative to QM is substantially improved here and for other protein properties. (Right): **Temperature dependence validation studies of ForceBalance protein force fields.** Melting curves of a small helical peptide, ACE-$(AAQAA)_3$-NME, and a small beta hairpin, CLN025 are shown. The experimental data is obtained from established circular dichroism and NMR methods. The AMBER-FB15 model developed using ForceBalance (blue curve) is one of the best performing models, despite not being optimized to any folding data.

Improved automated tools for force field parameter optimization are needed, because the process involves many interconnected tasks that are traditionally arduous to carry out and difficult to reproduce, such as collecting the training data set of properties, computing these properties for trial force fields, and tuning parameters to maximize agreement. Semi-automated approaches to fitting have been in use since the late 1990s,[105–107] but the diversity of functional forms, simulation packages, and training datasets has limited their broader applicability.

Recently, ForceBalance,[94–98] developed by co-I Wang, has been very successful for automatically fitting force fields to specific target properties, including quantum mechanical data, condensed-phase properties, etc. (Figure 4 left). Recent uses of ForceBalance include development of a new three-point water model with dramatically improved properties[95] (Figure 4 right), a new AMBER-family protein force field with improved accuracy[98] (Figure 5), and polarizable models of nanoporous graphene for desalination.[108] ForceBalance is designed to fit force fields of essentially any type, using an interface that does not assume a particular functional form, and contains interfaces to several molecular simulation packages (e.g., AMBER, GROMACS, and OpenMM). It uses trust-radius Newton-Raphson to minimize an error function that penalizes deviations from experimental and quantum chemical targets, in contrast to the Bayesian approach we explore in Aim 3, making its application independent of that Aim.

It is especially important to note that SMIRNOFF force fields, which use direct chemical perception (DCP), provide a dramatically better starting point for optimization. Traditional force fields, due to limitations of atom typing, include hundreds or thousands of potentially redundant parameters. For example, SMIRNOFF99Frosst's parent force field, parm@Frosst, has hundreds of duplicated angle parameters and thousands of duplicated torsional parameters; similar issues occur in GAFF and GAFF2. By virtue of DCP, SMIRNOFF99Frosst uses 10–20 times fewer adjustable parameters than parm@Frosst and GAFF/GAFF2,[1] and will ultimately be far easier to optimize.

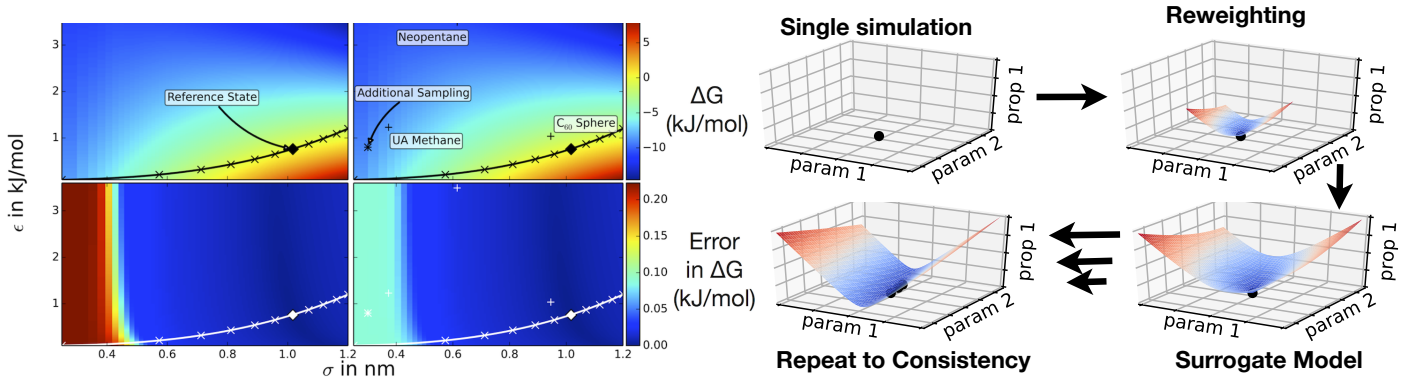

Figure 6. **Reweighting allows physical properties to be extrapolated in parameter space.** (Far Left): Estimated solvation free energies can be extrapolated using reweighting over a large range of Lennard-Jones $\{\epsilon, \sigma\}$ parameter space (upper panel) with low statistical error (lower panel), using only data from 12 original simulations (crosses). (Near Left): Addition of a single new equilibrium simulation extends the region of parameter space over which the extrapolation has low statistical error (Naden et al.[87]). (Right): **Surrogate models can be used to rapidly estimate properties as a function of parameters**. The posterior distributions after each round of (1) simulation of a single state point, (2) multistate reweighting (MBAR) from all previously sampled points calculates the local response of observables to change in parameters, and (3) A Gaussian process surrogate model describes the local response to parameter changes. MCMC is carried out on the surrogate model, rather than by peforming thousands of simulations. At each step, a new simulation is carried out at the projected optimum. The process is continued until the model converges and no new simulations are required.

Plan: Link ForceBalance to SMIRNOFF infrastructure and property databases to automatically optimize force fields. ForceBalance[94–98] provides a framework to automate optimization of force field parameters to fit target physical and quantum chemical properties. We will extend ForceBalance to utilize the new SMIRNOFF approach to direct chemical perception, and build a general framework for the efficient calculation of experimental and quantum chemical physical properties necessary for building high-accuracy biomolecular force fields: This includes properties abundant in the ThermoML Archive, including heat capacities, mixture densities, and dielectrics, as well as host-guest binding thermodynamics from BindingDB (Aim 2). Our SMIRNOFF99Frosst force field[1] will provide a starting point for initial refitting. ForceBalance already provides a framework to run the requisite simulations, manage jobs, etc., so the main task is to extend it to handle SMIRNOFF and additional properties.

1B: Develop methods to accelerate and automate force field assessment and optimization using hierarchical surrogate functions of observables rather than direct simulations.

Fitting force fields to condensed-phase measurements is enormously expensive, generally precluding its use in simultaneous optimization of parameters (e.g.[37]). A naïve approach to optimization involves new simulations at least several nanoseconds in length (minutes to hours of wall clock time) to evaluate target properties for each proposed parameter set. To overcome these limitations and permit co-optimization of all parameters simultaneously, we are developing a hierarchical approach that increases efficiency by orders of magnitude.

Preliminary results. We have developed fast multistate reweighting techniques such as MBAR[83] that predict properties in the neighborhood of parameters for which simulation data has already been collected, using only energy evaluations in the new parameter sets with the previously sampled configurations. **This allows physical properties for new parameter sets to be reliably estimated using up to three orders of magnitude less computational effort** than the naïve approach, allowing rapid searches through parameter space[3;84–86;89]. These reweighting techniques provide estimates of their own statistical uncertainties[85] (Figure 6, lower left), allowing us to trigger additional simulations when parameters move outside the high-confidence region. With each new simulation dataset, this high-confidence region is greatly expanded (Figure 6, lower left).

Even with the efficiency gains through reweighting, further enhancements are both necessary and feasible. We have developed a framework to employ *surrogate models*—cheap approximations of condensed phase properties given the simulation parameters—that can be fit on-the-fly to reweighted expectations and their uncertainties, further reducing computational cost to optimize or explore parameter space in well-explored regions (Figure 6, right). These models approximate the parameter-property relationship over previously-explored regions of parameter space, and can predict when the uncertainty in estimated properties becomes sufficiently large to require new reweighting calculations or simulations. Our prototype uses Gaussian processes[109;110] to converge on optimal regions of parameter space with a small number of simulations. Thus, an approximate, locally correct, model is used to drive sampling towards the areas of parameter space with the highest data evidence, falling back to reweighting, or again to simulations, when the confidence region of the inexpensive approach is exceeded. New simulations will be performed and the process repeated until the optimization converges.

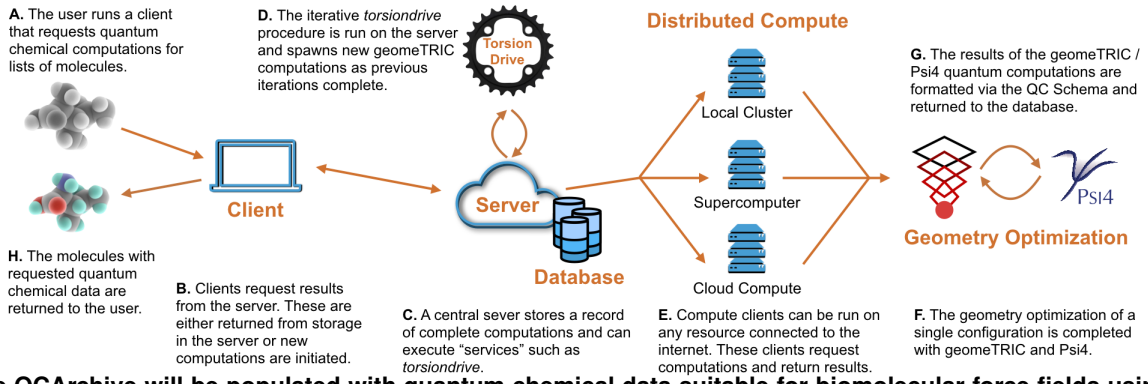

Figure 7. **The QCArchive will be populated with quantum chemical data suitable for biomolecular force fields using automated workflows.** The QCArchive,[68] designed and built in partnership with MolSSI, stores data in the open MolSSI QC JSON schema.[116] When molecules representing novel chemistries of interest are submitted, a number of connected workflows are triggered, such as automated fragmentation, capping, and multidimensional torsion drives, Hessians, or electrostatic potentials. Portable containerized workflows are run on pooled available connected computational resources (investigator clusters, cloud computing, and supercomputers). The QCPortal[117] Python tool provides a convenient Python object model that can be used to build various front ends for interacting with the database. While the QCArchive is intended to be general, force field specific workflows and APIs will be developed with MolSSI as part of this proposal.

Plan: Extend our automated property estimation framework to utilize hierarchical surrogate models. We will develop a scalable property estimation framework—which computes target physical properties and parameter gradients for force field assessment and optimization—to enable efficient calculation of condensed phase properties. Physical property computation protocols will be incorporated via a plug-in framework for easy extensibility to new physical measurements, with initial properties including densities, dielectric constants, and enthalpies of mixing of organic molecular liquids and mixtures, as well as host-guest binding thermodynamics for druglike guests. As we explore higher dimensional spaces using surrogate models, we will also examine polynomial chaos expansions[111–115]. Hierarchical modeling approaches may fail for sufficiently high-dimensional parameter optimizations. In these cases, We can perform principal coordinate analysis to identify parameters to collectively optimize, and those that are relatively uncoupled for independent, iterative optimization. We will work with collaborator Nathan Baker at PNNL on alternative surrogate model approaches (see letter).

Aim 2: Construct open datasets and databases for next-generation force field development.

Force field parameterization and assessment requires high-quality data. We will therefore curate high-quality, open datasets that can be used for automated parameterization and validation of force fields.

2A: Construct an open, comprehensive quantum chemical database of high-accuracy, *ab initio* datasets for parameterizing small molecules and biomolecules.

Force fields generally rely on quantum chemical calculations for determining torsional potentials, valence terms, and partial charges. While it has recently become popular to determine some of these terms (such as torsional potentials and partial charges) in a bespoke manner for molecules of interest in low-throughput applications such as free energy calculations—where seconds to minutes can be spent to derive parameters for a much longer calculation—generalizable small molecule force fields that can be applied instantly require large datasets of quantum chemical calculations from which generalizable parameters can be derived. Typically, these datasets have been produced at great cost, used once, and then discarded—despite their considerable value to the community for force field improvements, new force field efforts, and machine learning research .[6;71–73]

Preliminary data. Working with the NSF-sponsored Molecular Sciences Software Institute (see letter from MolSSI and Senior Person Daniel Smith), we have built automated workflows to populate the open quantum chemical database QCArchive[68] with data from a small molecule fragmentation, torsion drive, and constrained optimization energy workflow (Figure 7). All components of this pipeline are built using open source tools whenever possible, such as MongoDB (high-performance database), FireWorks (workflow engine), Psi4 (high-performance quantum chemical calculations), and RDKit (cheminformatics). While database instances can be hosted locally (and privately) by users, MolSSI will provide a public instance of this database for use by both our force field effort and the broader community, which can utilize large quantum chemical datasets in myriad ways.

Plan: We will populate the open QCArchive with data necessary for building biomolecular force fields, storing configurations and associated quantum chemical properties (energies, gradients, Hessians, fractional bond orders, and electrostatic potentials) in the widely-supported MolSSI QC JSON standard.[116] We will populate the database with data relevant to biomolecular force field development using automated workflows (Figure 7), including: *(1)* an automated torsion drive workflow[98;118] that uses geomeTRIC[119;120] for constrained geometry

optimization and Psi4 for QC single-point calculations; *(2)* a conformation enumeration and optimization scheme to compute Hessians for fitting valence forces; *(3)* an electrostatic potential fitting scheme for deriving improved charge models, supported by a novel open-source implementation[121] that will enable improvements over the widely-used two-stage RESP.[122] To populate the QCArchive, we will process fragment-size small molecules generated from druglike small molecules, biopolymers, and other biomolecules using a new chemical fragmentation algorithm[123] to minimize disruption of conjugated systems during fragmentation and capping.

The MolSSI QCArchive[68] will support public queries for datasets associated with a given small molecule and its fragments, including specific versioned datasets to allow toolkits to be reproducible. When public users query for fragment data associated with a new molecule that does not already exist, calculations can be automatically queued on computing resources coupled to QCArchive via the QCFractal distributed workflow engine,[124] which will include local academic compute resources at our institutions, cloud computing resources, XSEDE, and other suitable compute grids, with results deposited in the public database and returned when complete.

2B: Compile high-quality experimental datasets of small molecule liquid mixtures relevant to biomolecules.

Nonbonded parameters have largely been derived from experimental data for a modest set of pure organic liquids.[37;45] However, a wealth of available experimental data has not yet been exploited. The properties of liquid mixtures—such as such as excess heats of mixing and excess densities[125–128]—provide information on diverse chemical interactions. Recently, we demonstrated how data can be automatically extracted from the NIST ThermoML Archive,[129] which aggregates high-quality thermophysical data deposited from papers published in multiple journals, for force field assessment.[130] We will extract significantly expanded liquid datasets for parameterization and assessment, initially focusings on densities, heats of mixing, partial molar volumes, and compressibilities, expanding to other properties in order of abundance and utility. These datasets have been curated and internally annotated by NIST scientists for self-consistency and physical validity, a vital step for high-quality force field development. The scope is significant; more than 600 unique molecules have measured densities of binary mixtures at varying conditions, and more than 35,000 measurements are available for enthalpies of mixing. We will also add new targeted experimental datasets to ThermoML using Bayesian experimental design (Aim 3).

2C: Use automation to create diverse experimental datasets of host-guest binding thermodynamics.

A key goal of computational chemistry is the accurate prediction of protein-ligand binding free energies to guide drug design. However, multiple issues preclude their direct inclusion in a force field training dataset: these calculations are very slow to converge[8;131] and pose significant challenges aside from force field considerations, such as the need to model missing loops, select tautomer/protonation states, refine low-resolution structures, and make modeling decisions regarding crystallographic waters.[7] Host-guest systems embody much of the same physical chemistry,[25–28;131] their affinities can rival those of protein-ligand complexes,[132–134] and high-quality binding measurements are possible via ITC and NMR.[135–137] Host-guest systems have therefore been widely adopted as tests of computational models,[26–28] and we have shown that host-guest systems provide insights into force field accuracy and that fitting to host-guest binding can result in improved force fields.[138–142] However, there is a need for increased diversity in the noncovalent interactions probed by host-guest systems in order to provide a dataset that provides information on the wide range of interactions relevant in protein-ligand binding. While it is straightforward to select and purchase new guest molecules, diversifying host molecules can be challenging.

Preliminary data. To address this need, we (Gilson lab) have developed new, simple, high-yield synthetic routes that provide facile access to multiple families of new β-cyclodextrin (CD) derivatives (Figure 8), mono-derivatized at the wider secondary opening of this host; our preliminary work indicates guest molecules can predominantly interact with host substituents[4;5] making this opening especially appealing for functionalization. In preliminary studies, we have also successfully mono-derivatized the more tractable primary face, and this affords further potential for generating diverse host-guest interactions. To date, we have synthesized $>$16 novel derivatives, some of which are shown in Figure 8, by routes that provide access to a far wider range of new derivatives. We have also used isothermal titration calorimetry (ITC) to measure the binding thermodynamics of a guest molecule with native β-CD and one of these derivatives, confirming that derivatization modulates affinity.[4;5]

Plan: We will generate a large, chemically diverse dataset of high-quality host-guest binding thermodynamics. Our facile, one-pot, one-step, CD derivatization routes[4;5] provide a strong foundation for the automated synthesis of a much larger library (low hundreds) of cyclodextrin derivatives. This work will be a close collaboration of the Gilson lab (synthetic innovation) with the Chodera lab (automated instrumentation). Starting materials will come from commercial suppliers like Enamine, and reactions will be carried out in open polypropylene plates overnight at room temperature using small volumes of dimethylformamide (DMF), a polypropylene-compatible solvent, according to our existing protocols.[4;5]

Stock solutions will be prepared via a Quantos gravimetric dosing system and combinatorial reactions conducted in 96-well plates using a low-cost OpenTrons OT-2 automated pipetting system. Following reaction, excess DMF will be evaporated using a Genevac HT-4x, and product precipitated, filtered, and washed with acetone, yielding essentially pure product for characterization via automated mass spectrometry. Guest molecules will be selected from catalogs, considering biomedical and pharmaceutical relevance of functional groups, aqueous solubility, and cost. Host-guest binding assays will be performed with an automated MicroCal PEAQ ITC; as CD derivatives often possess millimolar solubilities, their complexation thermodynamics can be characterized via high-heat ITC experiments[138] to yield thermodynamic parameters that can be directly computed for use in both force field validation[139] and parameterization.[140–142] From a single 96-well plate of diverse synthetic host derivatives, assayed binding interactions with a panel of small molecule β-CD ligands can produce a library of thousands of binding measurements for force field parameterization and assessment. Software will be developed to facilitate quality review; to move the data into publicly accessible repositories;[100] and to format datasets for use in automated parameterization.

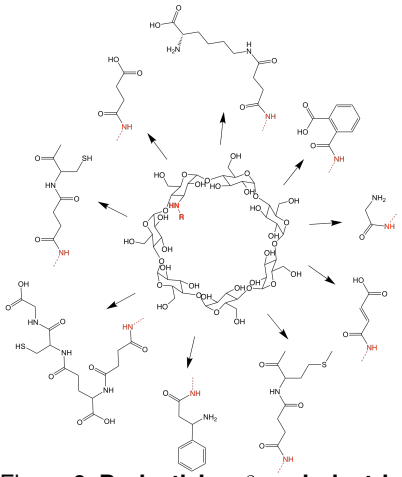

Figure 8. **Derivatizing β-cyclodextrin to generate a library of host compounds.** Automated synthesis will generate hosts for a large set of binding thermodynamics measurements against a library of guests with automated calorimetry. Existing syntheses use commercially available mono-3-amino-β-cyclodextrin[4;5] (center) to couple commercially available reagents to the secondary face, including peptides.[5]

2D: Curate experimental datasets suitable for force field assessment.

We will curate extensive experimental datasets that, while too expensive to use in parameterization, will be valuable in assessing biomolecular force fields produced by this and other efforts. In parallel, we will integrate new property computation plug-ins into our physical property computation pipeline from Aim 1 for each new property class, documenting best practices.

NMR data. We will curate published protein NMR data, including NOEs, through-bond ^{3}J couplings, Lipari-Szabo S^2 order parameters and residual dipolar couplings (RDCs) from the Biological Magnetic Resonance Data Bank (BMRB),[143] expanding on earlier work.[144] We will also collaborate with existing force field efforts to incorporate extensive protein and nucleic acid NMR datasets they have found useful in force field assessment (see letters),[24;66;145] and implement best-practices procedures for computing these NMR observables.

Miniprotein thermostablities. We are working with Gabriel Rocklin (see letter), who is conducting high-throughput stability studies of thousands of designed mini-proteins,[146] to utilize this unique and comprehensive dataset for protein force field assessment. We will also include melting curves for several small proteins and peptides, such as the villin headpiece, the helical $(AAQAA)_3$ peptide, Trp-cage, and the CLN025 hairpin, which have been experimentally obtained using standard techniques such as temperature-dependent circular dichroism and NMR chemical shift deviations.[147–149] We will integrate enhanced sampling methods (such as simulated tempering) into our physical property computation pipeline to compute miniprotein thermostabilities.

Binding thermodynamics. While protein-ligand binding free energies present a significant challenge for use in parameterization due to their expense, their inclusion in validation datasets is essential for benchmarking performance in a critical target application. We will curate systems from protein-ligand validation datasets available in BindingDB[100] to assess force field accuracy in modeling small molecule-protein interactions, along with more tractable existing host-guest datasets from the literature, and will integrate alchemical binding free energy calculations (using OpenMM[150] or GROMACS[151]) into our pipeline in order to take advantage of these data.

Community-organized datasets. We will work with collaborators from biomolecular force field communities and pharma (see letters) to assemble shared datasets of high value across these communities for force field validation, coordinate with experimental collaborators on ways these datasets might be sourced.

Aim 3: Develop Bayesian inference techniques to provide statistical tools to address key questions in force field modeling and predict systematic error.

Traditional force field parameterization approaches seek a single model that best fits the training data. However, a given training set may not tightly constrain all parameters, particularly when parameters covary. For example, different combinations of ϵ and σ values can yield equally accurate hydration free energies.[87;152] Importantly, even if two models are equally consistent with a training set, they may lead to a very different predictions of observables not well-represented in the training set. Thus, uncertainties in the parameters can propagate to uncertainties in quantities of interest computed with the force field. A goal of this project is to develop a method that can identify and quantify such uncertainties, so that a force field's users can better judge how much confidence to place in the

results it yields. We are not aware of any other major force field development effort that addresses this issue. We also aim to bring a rigorous approach to regularization, i.e., to avoiding overfitting, by assigning an appropriate preference to simpler over more complex models, as models that avoid overfitting tend to be more generalizable.[153] We will therefore augment the optimization-based approach of Aim 1 with a new Bayesian approach to force field parameterization, where force field parameters and model choices are *sampled* from a posterior distribution. This Bayesian reframing provides many benefits: (1) sampling techniques that escape local optima and find broader, more robust fits to the data; (2) a disciplined approach to weighting simpler models more than complex ones; (3) the ability to use Bayesian model selection to select not only parameters but also structural aspects of the model, such as functional forms and chemical perception trees for type assignments; (4) a route to predicting the uncertainty in predictions arising from force field uncertainties; and (5) the use of Bayesian experimental design principles to identify *new* experimental data to collect that will maximize the expected information gain.

3A: Develop a Bayesian inference infrastructure for automated parameter sampling, typing determination, physical model selection, and uncertainty quantification

Building on earlier work,[3;80–82] we have reformulated the force field parameterization problem to shift from minimizing a target objective function to sampling force field models θ from the Bayesian posterior $p(\theta|\mathcal{D}_1, \ldots, \mathcal{D}_N) \propto p(\theta) \prod_{n=1}^{N} p(\mathcal{D}_n|\theta)$ where $\mathcal{D}_n$ denotes one datum in a set of N training data (physical measurements or quantum chemical properties), $p(\theta)$ denotes a prior over physically relevant parameters, and $p(\mathcal{D}_n|\theta)$ is the likelihood the experimental datum $\mathcal{D}_n$ was measured given force field model θ. This approach replaces human-crafted weighted objective functions with likelihood functions that utilize explicit models of experimental uncertainty. Many physical measurements are already associated with meaningful experimental uncertainties, such as those in the ThermoML Archive. For experiments lacking uncertainties, or for quantum chemical properties, we will infer the error using nuisance parameters with weak priors, where consistency with physical measurements drives inference of the error magnitude—a technique we have previously demonstrated for biophysical experiments.[154]

We will architect modular software for efficient Markov chain Monte Carlo (MCMC) schemes[79] within a Gibbs sampling framework[155] to sample parameters, θ, alongside reversible jump Monte Carlo (RJMC)[156] to sample hierarchical chemical perception trees and physically motivated functional forms. Bayesian methods naturally penalize complexity,[75–77] ensuring that the sampled typing trees lead to consistency with the training data without overfitting. This Bayesian approach will replace ad hoc human decision-making about the structure of the force field with a well-founded statistical method grounded in the data.

Preliminary results: We recently described SMIRKY,[2] an RJMC-based approach to automatically sample over chemical perception trees used to assign valence and nonbonded parameters, and reported its ability to posit and learn the diversity of human-crafted atom types from existing small molecule force fields.[2] In unpublished work, we have applied this approach to parameterizing a Generalized Born / surface area model based on the OBC2 model,[157] for the SMIRNOFF99Frosst.[1] This work used a Langevin integrator[158;159] to sample in parameter space and an RJMC method to sample over chemical perception trees for assigning GB parameters, demonstrating success in simultaneous sampling over both numerical parameters and typing (Figure 9).

Plan: We will develop a modular infrastructure for Bayesian inference of force field models from experimental and quantum chemical data. This software infrastructure will allow us to plug in various MCMC sampling schemes that can sample parameters with and without the use of parameter gradients, while also sampling over typing trees, physical model functional forms (e.g., Lennard-Jones, Halgren[160]), combining rules, electrostatics models, and other discrete physical modeling decisions. This infrastructure will take advantage of the efficient reweighting and surrogate modeling framework for computing physical properties and their parameter gradients from Aim 1B, to enable sampling over large numbers of parameters. We will integrate our SMIRKY[2] scheme to allow automated elaboration of bond, angle, torsion, and nonbonded types to achieve improved likelihoods on training datasets, while automatically penalizing complexity to prevent proliferation of types[75–77] (Figure 9).

We will develop experimentally-informed data likelihood functions $p(\mathcal{D}_n|\theta)$ for the experimental datasets developed in Aims 1 and 2, allowing us utilize the same datasets to sample families of force field models from the posterior $\theta \sim p(\theta|\mathcal{D})$. The spread of predictions from by these sampled models can be used to estimate the systematic error in predictions made by the force field,[3;80–82] error typically neglected when estimating prediction uncertainties. To allow these error estimates to be carried out efficiently for predictions of new physical properties made with our Bayesian-derived force fields, we will create tools to run simulations with a single central set of force field parameters and then apply post-simulation reweighting techniques[83] to estimate the spread in the predictions without any need for additional simulations. This will, for example, give an estimate of the uncertainty a predicted protein-ligand binding free energy due to the sampled uncertainty in the force field parameters.

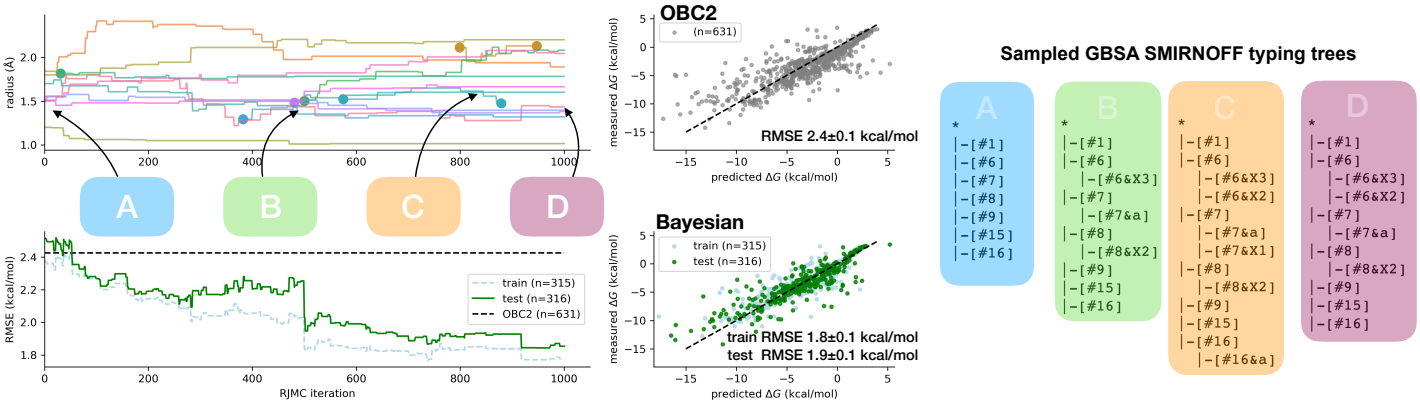

Figure 9. **We can learn chemical perception trees from data using Bayesian inference: learning a simple GBSA model.** *Left:* Using RJMC, Bayesian inference can sample over SMIRNOFF typing trees and their associated Born radii parameters (*top*) to fit experimental small molecule hydration free energies from FreeSolv[161] without overfitting; corresponding RMSE of sampled models shown below. *Middle:* A comparison of experimental and predicted hydration free energies for published OBC2 (`igb=5`) with `mbondi2` radii[157] (*top*) and final Bayesian (*bottom*) parameters. *Right:* Sampled SMIRNOFF typing trees illustrate how physically meaningful types are discovered.

3B: Develop Bayesian inference approaches to identify targeted experimental measurements.

Plan: Targeted experimental data collection guided by Bayesian experimental design principles. We will develop Bayesian experimental design approaches [162–167] to identify sets of feasible experiments that will most rapidly reduce prediction uncertainties. As a practical test, we will guide the collection of temperature- and mole-fraction dependent densities and enthalpies of liquid mixtures predicted to minimize predictive uncertainty on host-guest binding thermodynamics. We will select inexpensive, high-purity (HPLC grade or better) miscible liquids for automated density measurements (Metter-Toledo DM40 with autosampler) and automated calorimetry measurements (MicroCal Automated PEAQ-ITC), using the automated gravimetric dispensation system (Mettler-Toledo Quantos) and liquid handling system in the Chodera lab.

Aim 4: Construct and validate open source, transferable, and self-consistent force fields for small molecule interactions with complex biomolecular systems.

The machinery and datasets constructed in Aims 1–3 provide a fully automated framework capable of reproducibly building and assessing consistent biomolecular force fields. We will apply this framework to develop iteratively refined, versioned biomolecular force fields built from versioned open datasets, with the aim of achieving both significantly improved accuracy and greater coverage of chemical space for modeling small molecule interactions with biopolymers and other biomolecules. Key to this approach will be the coordinated and systematic reparameterization of small molecule and biomolecular force fields *together* to achieve superior accuracy in modeling their interactions. In pursuit of this aim, we will address significant outstanding questions about which factors, such as choice of physical models and training datasets, impact accuracy and transferability.

Construction of a high-accuracy force field for both biopolymers and general organic molecules in an integrated manner. The SMIRNOFF chemical perception scheme[1] already allows our initial SMIRNOFF99Frosst small molecule force field[1]—which covers the chemical space of relevance to biomolecules—to apply complete valence parameters to biopolymers and other biomolecules. Using this as a starting point, we will use training datasets drawn from Aim 2 that inform *both* small organic molecule and biopolymer properties—as well as other specialized solvent and biomolecule datasets (such as lipid NMR data from NMRlipids[64])—to simultaneously optimize all parameters and refine the chemical perception trees to achieve high accuracy and broad coverage. Our aim is to produce a general force field that can cover standard biopolymers, covalently modified or nonnatural organic amino or nucleic acids, biomolecules like lipids and carbohydrates, and be consistent with pharmaceutically relevant small molecules and arbitrary solvents.

ForceBalance optimization. Initially, we will employ ForceBalance[94;168] to minimize an error function balancing fits to quantum chemical and experimental data, utilizing the infrastructure in Aim 1. Fitting of valence terms will largely be driven by quantum chemical 1D and 2D torsion drives, minimized geometries, vibrational modes, and valence force constant data generated in QCArchive (Aim 2A). Fitting of nonbonded terms will primarily be driven by experimental datasets curated in Aims 2B/2C: building on our earlier work,[130] we will use densities, static dielectric, heat capacities, partial molar volumes, and enthalpies of mixing for pure and binary mixtures (including water) near biologically-relevant temperatures and pressures extracted from the NIST ThermoML Archive.[169;170] These datasets subsume those we recently used to develop recent protein[98] and water[94;168] force fields. Host-guest thermodynamics can also be incorporated since parameter gradients of binding free energies and enthalpies are

available by simple post-processing of simulation trajectories for free and bound species. To accelerate parameter optimization, we will use both reweighting and hierarchical surrogate modeling (Aim 1B).

Automatically learning chemical perception trees. After developing the Bayesian framework in Aim 3, we will revisit the force field parameterization tasks above. We will use the likelihood functions discussed above to account for reported experimental error, while quantum chemical data will use inferred nuisance parameters to account for error in quantum chemical energetics. Using the RJMC framework from SMIRKY (Figure 9), we will introduce sampling over chemical perception trees to increase or decrease the number of parameters independently used for each valence and nonbonded term (bonds, angles, torsions, impropers, Lennard-Jones) in addition to the parameter sampling, in a matter similar to the GB parameter type data presented above. Because the Bayesian approach automatically penalizes complexity, this procedure will naturally balance the goals of achieving a good fit to the data while avoiding undue complexity. We will infer what the natural number of types is for each class of interactions, explore how this varies as training datasets of different compositions are used during fitting, and identify how generalizability (as judged by the validation dataset, described below) is impacted by complexity.

Learning new charge models. Initially, we will use a specific partial charge model adapted from the widely popular AM1-BCC[171;172] modified to apply to arbitrary amino acids (or variants) by capping them, charging the resulting molecule, zeroing charges on the caps, and scaling the resulting charges to ensure the residue retains integral charge. In subsequent experiments, under the guidance of collaborator Bayly (original developer of RESP and AM1-BCC, see letter), we will improve AM1-BCC by optimizing bond charge correction (BCC) parameters using training datasets. The Bayesian RJMC approach allows us to introduce additional discrete model choices beyond chemical perception trees: For example, we will use SMIRKY to determine *which* BCCs should be present, as well as which fast semiempirical method (e.g., AM1 vs PM3) is best supported by data. We will continue to explore new charge models such as those based on IPolQ approach (see letter from Senior Personnel Cerutti).

Prediction of systematic error. The Bayesian approach will allow estimates of properties to include force field uncertainty estimates which reflect the uncertainty in the underlying parameters given the limited training dataset[3;80–82] (Aim 3). We will systematically explore the fidelity of this estimate by determining how well the force field is able to predict its own systematic error on validation data or held-out training data.

Comparing physical functional forms. Bayesian inference allows comparison of statistical evidence for different families of force field models via Bayes factors,[173] with data-driven decisions for each modeling choice, even if model families possess different numbers of parameters (Figure 9). We will use this strategy to investigate (initially with subsets of the experimental data) the statistical support for alternative van der Waals functional forms, partial charge models, selective multipoles, selective off-center charges, site polarizability models, and deep learning models to replace valence terms. For example, this process will allow us to directly compare fixed-charge and polarizable (both point dipole and Drude) models fit to identical datasets, comparing computational cost, accuracy, and whether the data justifies the increase in number of parameters for any given dataset.

Assessment datasets. The scalable physical property estimation framework developed in Aim 1 will be used to assess our force fields, and existing force fields, where coverage permits, against physical property datasets from Aim 2. A fraction of each dataset used in training will be withheld for use in assessment, and physical datasets too costly to use within parameter optimization iterations will be reserved for assessment. Assessments will be performed for every versioned force field release and presented publicly for comparison with existing force fields.

Long-term outlook. This project will yield new technologies that will drive the development of new generations of force fields employing both standard and novel functional forms, via both traditional optimization and Bayesian approaches, spanning small molecules, proteins, carbohydrates, nucleic acids, lipids, and biologically-interacting molecules. This capability to automate force field development means that, as training datasets grow, so will force field accuracy. These tools will also provide needed insight into which changes in functional form will yield the best improvements in accuracy, allowing the field to manage the tradeoff between complexity and computational speed. By creating an open infrastructure and nucleating a community around open training and validation datasets, we will also enable other researchers to construct, extend, and refine transferable and self-consistent physical force fields for biomedically relevant compounds. The resulting ongoing improvements in accuracy will help researchers worldwide develop a deeper understanding biomolecular function and speed the discovery of new therapeutics.

References

[1] **Mobley DL**, Bannan CC, Rizzi A, Bayly CI, Chodera JD, Lim VT, Lim NM, Beauchamp KA, Slochower DR, Shirts MR, et al. Escaping Atom Types in Force Fields Using Direct Chemical Perception. J Chem Theory Comput. 2018; 14(11):6076–6092.

[2] **Zanette C**, Bannan CC, Bayly CI, Fass J, Gilson MK, Shirts MR, Chodera JD, Mobley DL. Toward learned chemical perception of force field typing rules. J Chem Theory Comput. 2018; 15(1):402–423.

[3] **Messerly RA**, Shirts MR, Kazakov AF. Uncertainty quantification confirms unreliable extrapolation toward high pressures for united-atom Mie λ-6 force field. J Chem Phys. 2018; 149(11):114109.

[4] **Kellett K**, Kantonen SA, Duggan BM, Gilson MK. Toward Expanded Diversity of Host–Guest Interactions via Synthesis and Characterization of Cyclodextrin Derivatives. Journal of Solution Chemistry. 2018; 47(10):1597–1608.

[5] **Kellett K**, Duggan BM, Gilson M. Facile synthesis of a diverse library of mono-3-substituted β-cyclodextrin analogues. Supramolecular Chemistry. 2019; p. 1–9.

[6] **Smith JS**, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017; 8(4):3192–3203.

[7] **Sherborne B**, Shanmugasundaram V, Cheng AC, Christ CD, DesJarlais RL, Duca JS, Lewis RA, Loughney DA, Manas ES, McGaughey GB, et al. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. J Comput Aided Mol Des. 2016; 30(12):1139–1141.

[8] **Cournia Z**, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. J Chem Inf Model. 2017; 57(12):2911–2937. https://doi.org/10.1021/acs.jcim.7b00564.

[9] **Abel R**, Wang L, Harder ED, Berne B, Friesner RA. Advancing drug discovery through enhanced free energy calculations. Accounts of chemical research. 2017; 50(7):1625–1632.

[10] **Geballe MT**, Guthrie JP. The SAMPL3 Blind Prediction Challenge: Transfer Energy Overview. J Comput Aided Mol Des. 2012; 26(5):489–496. https://doi.org/10.1007/s10822-012-9568-8.

[11] **Mobley DL**, Liu S, Cerutti DS, Swope WC, Rice JE. Alchemical Prediction of Hydration Free Energies for SAMPL. J Comput Aided Mol Des. 2012; 26(5):551–562. https://doi.org/10.1007/s10822-011-9528-8.

[12] **Bannan CC**, Burley KH, Chiu M, Shirts MR, Gilson MK, Mobley DL. Blind Prediction of Cyclohexane–water Distribution Coefficients from the SAMPL5 Challenge. J Comput Aided Mol Des. 2016; 30(11):1–18. https://doi.org/10.1007/s10822-016-9954-8.

[13] **Park J**, Nessler I, McClain B, Macikenas D, Baltrusaitis J, Schnieders MJ. Absolute Organic Crystal Thermodynamics: Growth of the Asymmetric Unit into a Crystal via Alchemy. J Chem Theory Comput. 2014; 10(7):2781–2791. https://doi.org/10.1021/ct500180m.

[14] **Schnieders MJ**, Baltrusaitis J, Shi Y, Chattree G, Zheng L, Yang W, Ren P. The Structure, Thermodynamics, and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. J Chem Theory Comput. 2012; 8(5):1721–1736. https://doi.org/10.1021/ct300035u.

[15] **Matos GDR**, Mobley DL. Challenges in the use of atomistic simulations to predict solubilities of drug-like molecules. F1000Research. 2018; 7.

[16] **Aldeghi M**, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. J Am Chem Soc. 2017; 139(2):946–957. https://doi.org/10.1021/jacs.6b11467.

[17] **Comer J**, Schulten K, Chipot C. Calculation of Lipid-Bilayer Permeabilities Using an Average Force. J Chem Theory Comput. 2014; 10(2):554–564. https://doi.org/10.1021/ct400925s.

[18] **Lee CT**, Comer J, Herndon C, Leung N, Pavlova A, Swift RV, Tung C, Rowley CN, Amaro RE, Chipot C, Wang Y, Gumbart JC. Simulation-Based Approaches for Determining Membrane Permeability of Small Compounds. J Chem Inf Model. 2016; 56(4):721–733. https://doi.org/10.1021/acs.jcim.6b00022.

[19] **Salomon-Ferrer R**, Gotz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comp. 2013; 9(9):3878–3888.

[20] **Wang L**, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. J Am Chem Soc. 2015; 137(7):2695–2703. https://doi.org/10.1021/ja512751q.

[21] **Nerenberg PS**, Head-Gordon T. New Developments in Force Fields for Biomolecular Simulations. Curr

Opin Struct Biol. 2018; 49:129–138. https://doi.org/10.1016/j.sbi.2018.02.002.
[22] **Shaw DE**, Grossman J, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In: *Proceedings of the international conference for high performance computing, networking, storage and analysis* IEEE Press; 2014. p. 41–53.
[23] **Lindorff-Larsen K**, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010; 78(8):1950–1958.
[24] **Robustelli P**, Piana S, Shaw DE. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. Proc Natl Acad Sci USA. 2018; 115:E4758–E4766. https://doi.org/10.1073/pnas.1800690115.
[25] **Muddana HS**, Varnado CD, Bielawski CW, Urbach AR, Isaacs L, Geballe MT, Gilson MK. Blind Prediction of Host–guest Binding Affinities: A New SAMPL3 Challenge. J Comput Aided Mol Des. 2012; 26(5):475–487. https://doi.org/10.1007/s10822-012-9554-1.
[26] **Skillman AG**. SAMPL3: Blinded Prediction of Host–guest Binding Affinities, Hydration Free Energies, and Trypsin Inhibitors. J Comput Aided Mol Des. 2012; 26(5):473–474. https://doi.org/10.1007/s10822-012-9580-z.
[27] **Muddana HS**, Fenley AT, Mobley DL, Gilson MK. The SAMPL4 Host–guest Blind Prediction Challenge: An Overview. J Comput Aided Mol Des. 2014; 28(4):305–317. https://doi.org/10.1007/s10822-014-9735-1.
[28] **Yin J**, Henriksen NM, Slochower DR, Shirts MR, Chiu MW, Mobley DL, Gilson MK. Overview of the SAMPL5 Host–guest Challenge: Are We Doing Better? J Comput Aided Mol Des. 2017; 31(1):1–19. https://doi.org/10.1007/s10822-016-9974-4.
[29] **Rizzi A**, Murkli S, McNeill JN, Yao W, Sullivan M, Gilson MK, Chiu MW, Isaacs L, Gibb BC, Mobley DL, et al. Overview of the SAMPL6 host–guest binding affinity prediction challenge. J Comput Aided Mol Des. 2018; 32(10):937–963.
[30] **Matos GDR**, Calabr G, Mobley D. Infinite Dilution Activity Coefficients as Constraints for Force Field Parameterization and Method Development. chemRxiv. ; .
[31] **Liu S**, Cao S, Hoang K, Young KL, Paluch AS, Mobley DL. Using MD simulations to calculate how solvents modulate solubility. J Chem Theory Comput. 2016; 12(4):1930–1941.
[32] **Wang J**, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General Amber Force Field. J Comput Chem. 2004; 25(9):1157–1174. https://doi.org/10.1002/jcc.20035.
[33] **Wang J**, Wang W, Kollman PA, Case DA. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. J Mol Graph Model. 2006; 25(2):247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.
[34] **Wang J**, A Snapshot of GAFF2 Development; 2017.
[35] **Vanommeslaeghe K**, MacKerell AD. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. J Chem Inf Model. 2012; 52(12):3144–3154. https://doi.org/10.1021/ci300363c.
[36] **Vanommeslaeghe K**, Raman EP, MacKerell AD. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. J Chem Inf Model. 2012; 52(12):3155–3168. https://doi.org/10.1021/ci3003649.
[37] **Boulanger E**, Huang L, Rupakheti C, MacKerell Jr AD, Roux B. Optimized Lennard-Jones parameters for druglike small molecules. J Chem Theory Comput. 2018; 14(6):3121–3131.
[38] **Malde AK**, Zuo L, Breeze M, Stroet M, Poger D, Nair PC, Oostenbrink C, Mark AE. An automated force field topology builder (ATB) and repository: version 1.0. J Chem Theory Comput. 2011; 7(12):4026–4037.
[39] **Koziara KB**, Stroet M, Malde AK, Mark AE. Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies. J Comput Aided Mol Des. 2014; 28(3):221–233.
[40] **Harder E**, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput. 2015; 12(1):281–296.
[41] **Roos K**, Wu C, Damm W, Reboul M, Stevenson JM, Lu C, Dahlgren MK, Mondal S, Chen W, Wang L, et al. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J Chem Theory Comput. 2019; p. ASAP.
[42] **Steinbrecher TB**, Dahlgren M, Cappel D, Lin T, Wang L, Krilov G, Abel R, Friesner R, Sherman W. Accurate binding free energy predictions in fragment optimization. J Chem Inf Model. 2015; 55(11):2411–2420.
[43] **Chen W**, Deng Y, Russell E, Wu Y, Abel R, Wang L. Accurate calculation of relative binding free energies between ligands with different net charges. J Chem Theory Comput. 2018; 14(12):6346–6358.

[44] **Jorgensen WL**, Tirado-Rives J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. J Am Chem Soc. 1988; 110(6):1657–1666. https://doi.org/10.1021/ja00214a001.

[45] **Jorgensen WL**, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc. 1996; 118(45):11225–11236.

[46] **Tan D**, Piana S, Dirks RM, Shaw DE. RNA Force Field with Accuracy Comparable to State-of-the-Art Protein Force Fields. Proc Natl Acad Sci USA. 2018; 115:E1346–E1355. https://doi.org/10.1073/pnas.1713027115.

[47] **Hornak V**, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. Proteins. 2006; 65(3):712–725. https://doi.org/10.1002/prot.21123.

[48] **Piana S**, Klepeis JL, Shaw DE. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. Curr Opin Struct Bio. 2014; 24:98–105. https://doi.org/10.1016/j.sbi.2013.12.006.

[49] **Maier JA**, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11(8):3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

[50] **Dauber-Osguthorpe P**, Hagler A. 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. 2018; p. 1–71.

[51] **Sun H**, Jin Z, Yang C, Akkermans RLC, Robertson SH, Spenley NA, Miller S, Todd SM. COMPASS II: Extended Coverage for Polymer and Drug-like Molecule Databases. J Mol Model. 2016; 22(2):47. https://doi.org/10.1007/s00894-016-2909-0.

[52] **Ponder JW**, Wu C, Ren P, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T. Current Status of the AMOEBA Polarizable Force Field. J Phys Chem B. 2010; 114(8):2549–2564. https://doi.org/10.1021/jp910674d.

[53] **Zhang C**, Lu C, Jing Z, Wu C, Piquemal JP, Ponder JW, Ren P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. J Chem Theory Comput. 2018; 14(4):2084–2108. https://doi.org/10.1021/acs.jctc.7b01169.

[54] **Lin FY**, Lopes PEM, Harder E, Roux B, MacKerell AD. Polarizable Force Field for Molecular Ions Based on the Classical Drude Oscillator. J Chem Inf Model. 2018; 58:993–1004. https://doi.org/10.1021/acs.jcim.8b00132.

[55] **Mobley DL**, Bayly CI, Cooper MD, Shirts MR, Dill KA. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J Chem Theory Comput. 2009; 5(2):350–358. https://doi.org/10.1021/ct800409d.

[56] **Fennell CJ**, Wymer KL, Mobley DL. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. J Phys Chem B. 2014; 118(24):6438–6446. https://doi.org/10.1021/jp411529h.

[57] **Hagler A**. 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. 2018; p. 1–60.

[58] **Galindo-Murillo R**, Robertson JC, Zgarbová M, Šponer J, Otyepka M, Jurečka P, Cheatham TE. Assessing the Current State of Amber Force Field Modifications for DNA. J Chem Theory Comput. 2016; 12(8):4114–4127. https://doi.org/10.1021/acs.jctc.6b00186.

[59] **MacKerell Jr AD**, Nilsson L. Molecular dynamics simulations of nucleic acid–protein complexes. Curr Opin Struc Biol. 2008; 18(2):194–199.

[60] **Palermo G**, Miao Y, Walker RC, Jinek M, McCammon JA. CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. Proc Nat Acad Sci USA. 2017; 114(28):7260–7265.

[61] **Palermo G**, Ricci CG, Chen JS, Miao Y, Jinek M, Doudna JA, McCammon JA. Molecular Mechanism of Off-Target Effects in CRISPR-Cas9. Biophys J. 2019; 116(3):319a.

[62] **Lyubartsev AP**, Rabinovich AL. Force Field Development for Lipid Membrane Simulations. BBA-Biomembranes. 2016; 1858(10):2483–2497. https://doi.org/10.1016/j.bbamem.2015.12.033.

[63] **Poger D**, Caron B, Mark AE. Validating Lipid Force Fields against Experimental Data: Progress, Challenges and Perspectives. BBA-Biomembranes. 2016; 1858(7, Part B):1556–1565. https://doi.org/10.1016/j.bbamem.2016.01.029.

[64] **Ollila OS**, Pabst G. Atomistic resolution structure and dynamics of lipid bilayers in simulations and experiments. BBA-Biomembranes. 2016; 1858(10):2512–2528.

[65] **Kirschner KN**, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ. GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. J Comput Chem. 2008; 29(4):622–655.

https://doi.org/10.1002/jcc.20820.
[66] **Huang J**, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, Jr ADM. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. Nat Methods. 2017; 14(1):71–73. https://doi.org/10.1038/nmeth.4067.
[67] **Rauscher S**, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmüller H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. J Chem Theory Comput. 2015; 11(11):5513–5524. https://doi.org/10.1021/acs.jctc.5b00736.
[68] QCArchive: An open community database for quantum chemistry;. https://qcarchive.molssi.org.
[69] SMARTS - A Language for Describing Molecular Patterns;. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
[70] SMIRKS - A Reaction Transform Language;. http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html.
[71] **Botu V**, Batra R, Chapman J, Ramprasad R. Machine learning force fields: construction, validation, and outlook. J Phys Chem C. 2016; 121(1):511–522.
[72] **Chmiela S**, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. Machine learning of accurate energy-conserving molecular force fields. Science Adv. 2017; 3(5):e1603015.
[73] **Podryabinkin EV**, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mat Sci. 2017; 140:171–180.
[74] **Pathak H**, Palmer JC, Schlesinger D, Wikfeldt KT, Sellberg JA, Pettersson LGM, Nilsson A. The structural validity of various thermodynamical models of supercooled water. J Chem Phys. 2016; 145(13). https://doi.org/–10.1063/1.4963913˝.
[75] **MacKay DJ**. Bayesian methods for adaptive models. PhD thesis, California Institute of Technology; 1992.
[76] **Ghahramani IMZ**. A note on the evidence and Bayesian Occams razor. Gatsby Unit Technical Report GCNU-TR 2005–003; 2005.
[77] **Zhang Z**, Chan KL, Wu Y, Chen C. Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Statistics and Computing. 2004; 14(4):343–355.
[78] **Box GEP**, Tiao GC. Bayesian Inference in Statistical Analysis. Wiley Classics Library, Wiley; 2011. http://books.google.com/books?id=T8Askeyk1k4C.
[79] **Liu JS**. Monte Carlo strategies in scientific computing. 2nd ed. ed. New York: Springer-Verlag; 2002.
[80] **Rizzi F**, Najm HN, Debusschere BJ, Sargsyan K, Salloum M, Adalsteinsson H, Knio OM. Uncertainty quantification in MD simulations. Part II: Bayesian inference of force-field parameters. Multiscale Model Sim. 2012; 10(4):1460–1492.
[81] **Angelikopoulos P**, Papadimitriou C, Koumoutsakos P. Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework. J Chem Phys. 2012; 137(14):144103.
[82] **Wu S**, Angelikopoulos P, Papadimitriou C, Moser R, Koumoutsakos P. A hierarchical Bayesian framework for force field selection in molecular dynamics simulations. Philos T R Soc A. 2016; 374(2060):20150032.
[83] **Shirts MR**, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys. 2008; 129(12):124105.
[84] **Paliwal H**. Efficient multistate reweighting and configurational mapping algorithms for very large scale thermodynamic property prediction from molecular simulations. PhD thesis, University of Virgina School of Engineering and Applied Science; 2014.
[85] **Paliwal H**, Shirts MR. Using Multistate Reweighting to Rapidly and Efficiently Explore Molecular Simulation Parameters Space for Nonbonded Interactions. J Chem Theory Comput. 2013; 9(11):4700–4717. https://doi.org/10.1021/ct4005068.
[86] **Paliwal H**, Shirts MR. Multistate Reweighting and Configuration Mapping Together Accelerate the Efficiency of Thermodynamic Calculations as a Function of Molecular Geometry by Orders of Magnitude. J Chem Phys. 2013; 138(15):154108. https://doi.org/10.1063/1.4801332.
[87] **Naden LN**, Shirts MR. Rapid Computation of Thermodynamic Properties over Multidimensional Nonbonded Parameter Spaces Using Adaptive Multistate Reweighting. J Chem Theory Comput. 2016; 12(4):1806–1823. https://doi.org/10.1021/acs.jctc.5b00869.
[88] **Naden LN**, Shirts MR. Linear Basis Function Approach to Efficient Alchemical Free Energy Calculations. 2. Inserting and Deleting Particles with Coulombic Interactions. J Chem Theory Comput. 2015; 11(6):2536–2549. https://doi.org/10.1021/ct501047e.
[89] **Messerly RA**, Razavi SM, Shirts MR. Configuration-sampling-based surrogate models for rapid parameterization of non-bonded interactions. J Chem Theory Comput. 2018; 14(6):3144–3162.

[90] **Lyu J**, Wang S, Balius TE, Singh I, Levit A, Moroz YS, OMeara MJ, Che T, Algaa E, Tolmachova K, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019; 566:224–229.
[91] **Irwin JJ**, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model. 2012; 52(7):1757–1768. https://doi.org/10.1021/ci3001277.
[92] **Betz RM**, Walker RC. Paramfit: automated optimization of force field parameters for molecular dynamics simulations. J Comput Chem. 2015; 36(2):79–87.
[93] **Wu JC**, Chattree G, Ren P. Automation of AMOEBA polarizable force field parameterization for small molecules. Theor Chem Acct. 2012; 131(3):1138.
[94] **Wang LP**, Head-Gordon TL, Ponder JW, Ren P, Chodera JD, Eastman PK, Martinez TJ, Pande VS. Systematic improvement of a classical molecular model of water. J Phys Chem B. 2013; 117:9956.
[95] **Wang LP**, Martinez TJ, Pande VS. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. J Phys Chem Lett. 2014; 5(11):1885–1891. https://doi.org/10.1021/jz500737m.
[96] **Laury ML**, Wang LP, Pande VS, Head-Gordon T, Ponder JW. Revised parameters for the AMOEBA polarizable atomic multipole water model. The Journal of Physical Chemistry B. 2015; 119(29):9423–9437.
[97] **McKiernan KA**, Wang LP, Pande VS. Training and Validation of a Liquid-Crystalline Phospholipid Bilayer Force Field. J Chem Theory Comput. 2016; 12(12):5960–5967. https://doi.org/–10.1021/acs.jctc.6b00801˝.
[98] **Wang LP**, McKiernan KA, Gomes J, Beauchamp KA, Head-Gordon T, Rice JE, Swope WC, Martínez TJ, Pande VS. Building a more predictive protein force field: a systematic and reproducible route to AMBER-FB15. J Phys Chem B. 2017; 121(16):4023–4039.
[99] **Moine E**, Privat R, Sirjean B, Jaubert JN. Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes. J Phys Chem Ref Data. 2017; 46(3):033102.
[100] **Liu T**, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res. 2006; 35(suppl_1):D198–D201.
[101] **Wang J**, Kollman PA. Automatic parmaeterization of force field by systematic search and genetic algorithms. J Comput Chem. 2001; 22:1219–1228.
[102] **Duarte Ramos Matos G**, Kyu DY, Loeffler HH, Chodera JD, Shirts MR, Mobley DL. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. J Chem Eng Data. 2017; 62(5):1559–1569. https://doi.org/10.1021/acs.jced.7b00104.
[103] **Rekharsky MV**, Goldberg RN, Schwarz FP, Tewari YB, Ross PD, Yamashoji Y, Inoue Y. Thermodynamic and nuclear magnetic resonance study of the interactions of. alpha.-and. beta.-cyclodextrin with model substances: phenethylamine, ephedrines, and related substances. J Am Chem Soc. 1995; 117(34):8830–8840.
[104] An Informal AMBER Small Molecule Force Field: parm@Frosst;. http://www.ccl.net/cca/data/parm_at_Frosst/.
[105] **Faller R**, Schmitz H, Biermann O, Muller-Plathe F. Automatic parameterization of force fields for liquids by simplex optimization. J Comput Chem. 1999; 20(10):1009–1017. https://doi.org/10.1002/(sici)1096-987x(19990730)20:10¡1009::aid-jcc3¿3.0.co;2-c.
[106] **Brommer P**, Gahler F. Potfit: effective potentials from ab initio data. Model Sim Mat Sci Eng. 2007; 15(3):295–304. https://doi.org/10.1088/0965-0393/15/3/008.
[107] **Hulsmann M**, Kirschner KN, Kramer A, Heinrich DD, Kramer-Fuhrmann O, Reith D. In: Snurr RQ, Adjiman CS, Kofke DA, editors. Optimizing Molecular Models Through Force-Field Parameterization via the Efficient Combination of Modular Program Packages Molecular Modeling and Simulation-Applications and Perspectives, Singapore: Springer-Verlag Singapore Pte Ltd; 2016. p. 53–77. https://doi.org/10.1007/978-981-10-1128-3˙4.
[108] **Qiu Y**, Schwegler BR, Wang LP. Polarizable Molecular Simulations Reveal How Silicon-Containing Functional Groups Govern the Desalination Mechanism in Nanoporous Graphene. J Chem Theory Comput. 2018; 14(8):4279–4290. https://doi.org/–10.1021/acs.jctc.8b00226˝.
[109] **Angelikopoulos P**, Papadimitriou C, Koumoutsakos P. X-TMCMC: Adaptive kriging for Bayesian inverse modeling. Comput Method Appl Mechan Eng. 2015; 289:409–428. https://doi.org/10.1016/j.cma.2015.01.015.
[110] **Deng H**, Shao W, Ma Y, Wei Z. Bayesian Metamodeling for Computer Experiments Using the Gaussian Kriging Models. Qual Reliab Engng Int. 2012; 28(4):455–466. https://doi.org/10.1002/qre.1259.
[111] **Choi SK**, Grandhi RV, Canfield RA, Pettit CL. Polynomial Chaos Expansion with Latin Hypercube Sampling for Estimating Response Variability. AIAA J. 2004; 42(6):1191–1198. https://doi.org/10.2514/1.2220.
[112] **Sudret B**. Global sensitivity analysis using polynomial chaos expansions. Reliabil Eng & Sys Safety. 2008;

93(7):964–979. https://doi.org/10.1016/j.ress.2007.04.002.
[113] **Marzouk YM**, Najm HN. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. J Comput Phys. 2009; 228(6):1862–1902. https://doi.org/10.1016/j.jcp.2008.11.024.
[114] **Kersaudy P**, Sudret B, Varsier N, Picon O, Wiart J. A new surrogate modeling technique combining Kriging and polynomial chaos expansions—Application to uncertainty analysis in computational dosimetry. J Comput Phys. 2015; 286:103–117. https://doi.org/10.1016/j.jcp.2015.01.034.
[115] **Elsheikh AH**, Hoteit I, Wheeler MF. Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates. Comput Method Appl Mechan Eng. 2014; 269:515–537. https://doi.org/10.1016/j.cma.2013.11.001.
[116] A JSON Schema for Quantum Chemistry;. https://github.com/MolSSI/QC_JSON_Schema.
[117] A client interface to the QCArchive Project;. https://github.com/MolSSI/QCPortal.
[118] Dihedral scanner with wavefront propagation;. https://github.com/lpwgroup/torsiondrive.
[119] **Wang LP**, Song C. Geometry optimization made simple with translation and rotation coordinates. J Chem Phys. 2016; 144(21):214108.
[120] Geometry optimization code that includes the TRIC coordinate system;. https://github.com/leeping/geomeTRIC.
[121] implementation of open-source version of RESP method;. https://github.com/lpwgroup/respyte.
[122] **Bayly CI**, Cieplak P, Cornell WD, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. J Phys Chem. 1993; 97:10269–10280.
[123] Fragmenter: Chemically-aware fragmentation of molecules;. https://github.com/openforcefield/fragmenter.
[124] QCFractal: A distributed compute and database platform for quantum chemistry;. https://github.com/MolSSI/QCFractal.
[125] **González-Salgado D**, Nezbeda I. Excess properties of aqueous mixtures of methanol: Simulation versus experiment. Fluid Phase Equil. 2006; 240(2):161–166.
[126] **Wensink EJW**, Hoffmann AC, van Maaren PJ, van der Spoel D. Dynamic properties of water/alcohol mixtures studied by computer simulation. J Chem Phys. 2003; 119:7308.
[127] **Chen B**, Potoff JJ, Siepmann JI. Monte Carlo calculations for alcohols and their mixtures with alkanes. Transferable potentials for phase equilibria. 5. United-atom description of primary, secondary, and tertiary alcohols. J Phys Chem B. 2001; 105(15):3093–3104.
[128] **Stubbs JM**, Potoff JJ, Siepmann JI. Transferable Potentials for Phase Equilibria. 6. United-Atom Description for Ethers, Glycols, Ketones, and Aldehydes. J Phys Chem B. 2004; 108(45):17596–17605.
[129] The ThermoML Archive;. http://trc.nist.gov/ThermoML.html.
[130] **Beauchamp KA**, Behr JM, Rustenburg AS, Bayly CI, Kroenlein K, Chodera JD. Toward Automated Benchmarking of Atomistic Force Fields: Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive. J Phys Chem B. 2015; 119(40):12912–12920. https://doi.org/10.1021/acs.jpcb.5b06703.
[131] **Mobley DL**, Gilson MK. Predicting Binding Free Energies: Frontiers and Benchmarks. Annu Rev Biophys. 2017; 46(1):531–558. https://doi.org/10.1146/annurev-biophys-070816-033654.
[132] **Rekharsky MV**, Mori T, Yang C, Ko YH, Selvapalam N, Kim H, Sobransingh D, Kaifer AE, Liu S, Isaacs L, Chen W, Moghaddam S, Gilson MK, Kim K, Inoue Y. A Synthetic Host-Guest System Achieves Avidin-Biotin Affinity by Overcoming Enthalpy–entropy Compensation. Pro Natl Acad Sci USA. 2007; 104(52):20737–20742. https://doi.org/10.1073/pnas.0706407105.
[133] **Moghaddam S**, Inoue Y, Gilson MK. Host-Guest Complexes with Protein-Ligand-like Affinities: Computational Analysis and Design. J Am Chem Soc. 2009; 131(11):4012–4021. https://doi.org/10.1021/ja808175m.
[134] **Moghaddam S**, Yang C, Rekharsky M, Ko YH, Kim K, Inoue Y, Gilson MK. New Ultrahigh Affinity Host-Guest Complexes of Cucurbit[7]Uril with Bicyclo[2.2.2]Octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations. J Am Chem Soc. 2011; 133(10):3570–3581. https://doi.org/10.1021/ja109904u.
[135] **Gibb CLD**, Gibb BC. Binding of Cyclic Carboxylates to Octa-Acid Deep-Cavity Cavitand. J Comput Aided Mol Des. 2013; 28(4):319–325. https://doi.org/10.1007/s10822-013-9690-2.
[136] **Cao L**, Isaacs L. Absolute and Relative Binding Affinity of Cucurbit[7]Uril towards a Series of Cationic Guests. Supramol Chem. 2014; 26(3-4):251–258. https://doi.org/10.1080/10610278.2013.852674.
[137] **Sullivan MR**, Sokkalingam P, Nguyen T, Donahue JP, Gibb BC. Binding of Carboxylate and Trimethylammonium Salts to Octa-Acid and TEMOA Deep-Cavity Cavitands. J Comput Aided Mol Des. 2017; 31(1):1–8. https://doi.org/10.1007/s10822-016-9925-0.
[138] **Henriksen NM**, Fenley AT, Gilson MK. Computational Calorimetry: High-Precision Calculation

of Host–Guest Binding Thermodynamics. J Chem Theory Comput. 2015; 11(9):4377–4394. https://doi.org/10.1021/acs.jctc.5b00405.

[139] **Henriksen NM**, Gilson MK. Evaluating Force Field Performance in Thermodynamic Calculations of Cyclodextrin Host–Guest Binding: Water Models, Partial Charges, and Host Force Field Parameters. J Chem Theory Comput. 2017; 13(9):4253–4269. https://doi.org/10.1021/acs.jctc.7b00359.

[140] **Yin J**, Fenley AT, Henriksen NM, Gilson MK. Toward Improved Force-Field Accuracy through Sensitivity Analysis of Host-Guest Binding Thermodynamics. J Phys Chem B. 2015; 119(32):10145–10155. https://doi.org/10.1021/acs.jpcb.5b04262.

[141] **Wickstrom L**, Deng N, He P, Mentes A, Nguyen C, Gilson MK, Kurtzman T, Gallicchio E, Levy RM. Parameterization of an Effective Potential for Protein–ligand Binding from Host–guest Affinity Data. J Mol Recognit. 2016; 29(1):10–21. https://doi.org/10.1002/jmr.2489.

[142] **Yin J**, Henriksen NM, Muddana HS, Gilson MK. Bind3p: optimization of a water model based on host–guest binding data. J Chem Theory Comput. 2018; 14(7):3621–3632.

[143] Biological Magnetic Resonance Bank;. `http://brmb.wisc.edu`.

[144] **Beauchamp KA**, Lin YS, Das R, Pande VS. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J Chem Theory Comp. 2012; 8(4):1409–1414.

[145] **Best RB**, Zheng W, Mittal J. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. J Chem Theory Comput. 2014; 10(11):5113–5124. https://doi.org/10.1021/ct500569b.

[146] **Rocklin GJ**, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017; 357(6347):168–175.

[147] **Shalongo W**, Dugad L, Stellwagen E. Distribution of Helicity within the Model Peptide Acetyl(AAQAA)(3)amide. J Am Chem Soc. 1994; 116(18):8288–8293. https://doi.org/–10.1021/ja00097a039˝.

[148] **Streicher WW**, Makhatadze GI. Unfolding thermodynamics of Trp-cage, a 20 residue miniprotein, studied by differential scanning calorimetry and circular dichroism spectroscopy. Biochemistry. 2007; 46(10):2876–2880. https://doi.org/–10.1021/bi602424x˝.

[149] **Davis CM**, Xiao S, Raeigh DP, Dyer RB. Raising the Speed Limit for beta-Hairpin Formation. J Am Chem Soc. 2012; 134(35):14476–14482. https://doi.org/–10.1021/ja3046734˝.

[150] **Eastman P**, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS computational biology. 2017; 13(7):e1005659.

[151] **Van Der Spoel D**, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem. 2005; 26(16):1701–1718.

[152] **Joung IS**, Cheatham TE. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. J Phys Chem B. 2008; 112(30):9020–9041. https://doi.org/10.1021/jp8001614.

[153] **Myung IJ**. The importance of complexity in model selection. J Math Psych. 2000; 44:190–204.

[154] **Nguyen TH**, Rustenburg AS, Krimmer SG, Zhang H, Clark JD, Novick PA, Branson K, Pande VS, Chodera JD, Minh DD. Bayesian analysis of isothermal titration calorimetry for binding thermodynamics. PloS one. 2018; 13(9):e0203224.

[155] **Chodera JD**, Shirts MR. Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing. J Chem Phys. 2011; 135(19):194110.

[156] **Green PJ**. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82:711–732.

[157] **Onufriev A**, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Structure, Function, and Bioinformatics. 2004; 55(2):383–394.

[158] **Leimkuhler B**, Matthews C. Robust and efficient configurational molecular sampling via Langevin dynamics. J Chem Phys. 2013; 138(17):174102.

[159] **Fass J**, Sivak D, Crooks G, Beauchamp K, Leimkuhler B, Chodera J. Quantifying configuration-sampling error in Langevin simulations of complex molecular systems. Entropy. 2018; 20(5):318.

[160] **Halgren TA**. Representation of van der waals (vdW) interactions in molecular mechanics force fields: Potential form, combination rules, and vdW parameters. J Am Chem Soc. 1992; 114:7827–7843.

[161] **Mobley DL**, Guthrie JP. FreeSolv: A database of experimental and calculated hydration free energies, with

input files. J Comput Aided Mol Des. 2014; 28:711–720.
[162] **Chaloner K**, Verdinelli I. Bayesian experimental design: A review. Statistical Science. 1995; p. 273–304.
[163] **Toman B**. Bayesian experimental design. Encyclopedia of Statistical Sciences. 1999; .
[164] **Sebastiani P**, Wynn HP. Maximum entropy sampling and optimal Bayesian experimental design. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62(1):145–157.
[165] **Seeger MW**, Nickisch H. Compressed sensing and Bayesian experimental design. In: *Proceedings of the 25th international conference on Machine learning* ACM; 2008. p. 912–919.
[166] **Huan X**, Marzouk YM. Simulation-based optimal Bayesian experimental design for nonlinear systems. J Comput Phys. 2013; 232(1):288–317.
[167] **Ryan EG**, Drovandi CC, McGree JM, Pettitt AN. A Review of Modern Computational Algorithms for Bayesian Optimal Design: A Review of Modern Algorithms for Bayesian Design. Int Stat Rev. 2016; 84(1):128–154. https://doi.org/10.1111/insr.12107.
[168] **Qi R**, Wang LP, Wang Q, Pande VS, Ren P. United Polarizable Multipole Water Model for Molecular Mechanics Simulation. J Chem Phys. 2015; 143(1):014504. https://doi.org/10.1063/1.4923338.
[169] **Frenkel M**, Chirico RD, Diky VV, Dong Q, Frenkel S, Franchois PR, Embry DL, Teague TL, Marsh KN, Wilholt RC. ThermoML—An XML-based approach for storage and exchange of experimental and critically valuated thermophysical and thermochemical property data. 1. Experimental data. J Chem Eng Data. 2003; 48:2–13.
[170] **Chirico RD**, Frenkel M, Diky VV. ThermoML—An XML-based approach for storage and exchange of experimental and critically valuated thermophysical and thermochemical property data. 2, Uncertainties. J Chem Eng Data. 2003; 48:1344–1359.
[171] **Jakalian A**, Bush BL, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. J Comput Chem. 2000; 21(2):132–146. https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2¡132::AID-JCC5¿3.0.CO;2-P.
[172] **Jakalian A**, Jack DB, Bayly CI. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. J Comput Chem. 2002; 23(16):1623–1641. https://doi.org/10.1002/jcc.10128.
[173] **Wasserman L**. Bayesian model selection and model averaging. Journal of mathematical psychology. 2000; 44(1):92–107.
[174] **Brand A**, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. Learned Publishing. 2015; 28(2):151–155.
[175] Open Researcher and Contributor ID;. `https://orcid.org/`.

RESOURCE SHARING PLAN

Numerous useful and shareable resources will be generated during the course of these project, all of which will be made freely available to the research community at the earliest opportunity. The Open Force Field Initiative website [http://openforcefield.org] will act as an index to the various resources we freely provide.

Software. All computer software developed for this project will be free and open source, licensed under a permissive Open Source Initiative (OSI) approved license, such as the MIT license. Software will be made available on online collaborative public code repositories such as GitHub [http://github.com], where codes produced by our collaboration are currently hosted [http://github.com/openforcefield/].

We aim to develop *modern* and *open* software as part of this project. Specifically, by *modern*, we mean

- Written in Python, a language current students can easily understand, modify, and quickly become proficient with; not written in Fortran, C, or other languages with larger barriers to productivity
- Modular and extensible, where adding new capabilities requires minimal effort
- Making use of well maintained conda-installable libraries when available to minimize maintenance burden throughout the software life cycle
- Hewing to open standards whenever possible to minimize specialized frameworks, APIs, or formats developers must learn
- With a focus on creating simple, easy to use APIs
- Portable to various systems without recompilation
- Accompanied by extensive documentation describing the theory underlying the software and the APIs
- Clearly described best practices that we will simultaneously engage the community to describe in journals such as the Living Journal of Computational Molecular Sciences (LiveCoMS)
- Accompanied by tutorials that clearly describe common use cases conforming to these best practices

and by *open*, we mean

- Free (libre) open source, as described above
- Developed on a public community oriented platform like GitHub, with a community organized mindset
- Welcoming of new users and developers, and open to contributions from others
- Easily extensible by others for their own applications

Force fields. All force fields generated by the project will be made available through revision-controlled public repositories such as GitHub [http://github.com] under the Creative Commons By Attribution 4.0 (CC-BY) license. As an example, smirnoff99Frosst can be found at https://github.com/openforcefield/smirnoff99Frosst

Preprints. We aim to report our progress rapidly via preprints on and http://chemrxiv.org, which we aim to publish as open-access publications in journals when feasible (with emphasis on complete inclusion of all primary data). We will make every attempt to also release materials and data as they are generated, ensuring all reported data, software, and parameter sets are available prior to publication.

Quantum chemical datasets. The public quantum chemical database (QCArchve[68]) will be made publicly accessible for retrieval access in collaboration with the Molecular Sciences Software Institute (MolSSI). Public users will be able to perform compute-intensive requests (such as the generation of new quantum chemical torsion profile datasets from submitted molecules) as resource availability permits.

Experimental datasets and materials. All experimental datasets—including primary data, where applicable—will be made available through online, version-controlled repositories such as GitHub [http://github.com], the Open Science Framework [http://osf.io], and FigShare [http://figshare.com]. Affinity measurements will be submitted (along with primary data, when possible) to BindingDB [http://www.bindingdb.org]. All datasets generated by this collaboration will be made freely available under the Creative Commons By Attribution 4.0 (CC-BY) license. In addition, we will make all reasonable effort to meet requests for samples of our synthetic hosts and any other unique materials generated during this project.

Simulation datasets. Simulation and model datasets will be shared, when practical, through the online repositories such as GitHub [http://github.com], the Open Science Framework [http://osf.io], and FigShare [http://figshare.com].

3D printable laboratory parts. Numerous useful 3D printed parts are fabricated in the Chodera laboratory to aid in our automation research projects. Electronic printable versions of these parts are made available on both [http://www.choderalab.org/3dparts/] and the NIH 3D Print Exchange [http://3dprint.nih.gov].