Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2024, **15**, 12861

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 29th January 2024
Accepted 17th June 2024

DOI: 10.1039/d4sc00690a

rsc.li/chemical-science

Machine-learned molecular mechanics force fields from large-scale quantum chemical data†

Kenichiro Takaba, *[ab] Anika J. Friedman, [e] Chapin E. Cavender, [d] Pavan Kumar Behara, [c] Iván Pulido, [a] Michael M. Henry, [a] Hugo MacDermott-Opeskin, [f] Christopher R. Iacovella, [a] Arnav M. Nagle, [ag] Alexander Matthew Payne, [ai] Michael R. Shirts, [e] David L. Mobley, [h] John D. Chodera *[a] and Yuanqing Wang *[ja]

The development of reliable and extensible molecular mechanics (MM) force fields—fast, empirical models characterizing the potential energy surface of molecular systems—is indispensable for biomolecular simulation and computer-aided drug design. Here, we introduce a generalized and extensible machine-learned MM force field, espaloma-0.3, and an end-to-end differentiable framework using graph neural networks to overcome the limitations of traditional rule-based methods. Trained in a single GPU-day to fit a large and diverse quantum chemical dataset of over 1.1 M energy and force calculations, espaloma-0.3 reproduces quantum chemical energetic properties of chemical domains highly relevant to drug discovery, including small molecules, peptides, and nucleic acids. Moreover, this force field maintains the quantum chemical energy-minimized geometries of small molecules and preserves the condensed phase properties of peptides and folded proteins, self-consistently parametrizing proteins and ligands to produce stable simulations leading to highly accurate predictions of binding free energies. This methodology demonstrates significant promise as a path forward for systematically building more accurate force fields that are easily extensible to new chemical domains of interest.

Molecular mechanics (MM) force fields[1,2] are fast, empirical models that describe the potential energy surfaces of biomolecular systems by treating them as collections of atomic point masses. These point masses interact *via* non-bonded and valence (bond, angle, and torsion) terms, which are typically parametrized to reproduce quantum chemical conformational energetics and physical properties. Despite their simplified representation of the underlying physical model, MM force fields have proven to be indispensable for a multitude of tasks in biomolecular simulation and computer-aided drug design,[3,4] such as enumeration of putative bioactive conformations,[5] hit identification *via* virtual screening,[6] prediction of membrane permeability,[7] simulations of biomolecular dynamics,[8] and estimation of protein–ligand binding free energies *via* alchemical free energy calculations.[9]

1 Class I MM force fields have been a widely popular compromise between speed and accuracy

Class I MM force fields[1,2] are most widely used for proteins, lipids, nucleic acids, and other relevant biomolecules due to the computational efficiency afforded by the simple functional form:

[a]*Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. E-mail: john.chodera@choderalab.org; wangyq@wangyq.net*

[b]*Pharmaceuticals Research Center, Advanced Drug Discovery, Asahi Kasei Pharma Corporation, Shizuoka 410-2321, Japan. E-mail: takaba.kb@om.asahi-kasei.co.jp*

[c]*Center for Neurotherapeutics, Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA*

[d]*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA*

[e]*Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO, 80309, USA*

[f]*Open Molecular Software Foundation, Davis, CA 95618, USA*

[g]*Department of Bioengineering, University of California, Berkeley, Berkeley, CA, 94720, USA*

[h]*Department of Pharmaceutical Sciences, University of California, Irvine, California 92697, USA*

[i]*Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York 10065, USA*

[j]*Simons Center for Computational Physical Chemistry and Center for Data Science, New York University, New York, NY 10004, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4sc00690a

$$U_{\mathrm{MM}}(x;\Phi_{\mathrm{FF}}) = \sum_{\mathrm{bond}} \frac{K_r}{2}\left(r_{i,j} - r_0\right)^2 + \sum_{\mathrm{angle}} \frac{K_\theta}{2}\left(\theta_{i,j,k} - \theta_0\right)^2 + \sum_{\mathrm{torsion}} \sum_{n=1}^{n_{\max}} K_{\phi,n}\left[1 + \cos\left(n\phi_{i,j,k,l} - \phi_0\right)\right] + \sum_{\mathrm{Coulomb}} \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{i,j}} + \sum_{\mathrm{van\ der\ Waals}} 4\varepsilon\left[\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right], \quad (1)$$

where the total potential energy U_{MM} of a molecular system with coordinates x is defined by sets of force field parameters $\Phi_{\mathrm{FF}} = \{K_r, K_\theta, r_0, \theta_0, K_{\phi,n}, \phi_0, q, \sigma, \varepsilon\}_i$ specified for each atom i or valence term (bond, angle, torsion) of the system. An out-of-plane term (an improper torsion) can be also introduced with the torsion term to improve molecular planarity. The van der Waals interactions are usually described with Lennard-Jones 12–6 potentials using the Lorentz–Berthelot[10] combining rules to determine σ and ε between different atom types, but alternative combination rules are possible. In practice, such interactions usually require combining distinct force field parameters developed independently for specific chemical domains to complement the heterogeneity of biomolecular systems. Note that the functional forms of force fields can slightly differ among different Class I force fields, incorporating different scaling constants and additional functional terms, such as CMAP[2] and Urey-Bradley.[1] The minimalistic nature of Class I force fields has enabled them to achieve extraordinary speed on inexpensive hardware, with modern GPU-accelerated molecular simulation frameworks now able to generate more than 1 microsecond per day for many biomolecular drug targets[11–13] while still achieving useful accuracy in tasks such as predicting protein–ligand binding free energies for drug discovery.[14–16]

2 Traditional MM force field parametrization approaches struggle with complexity, limiting accuracy

Traditionally, the construction of MM force fields requires expert knowledge of physical organic chemistry to build atom-typing rules classifying atoms into discrete categories representing distinct chemical environments, enabling MM parameters to be subsequently assigned from a table of relevant atomic, bond, angle, and torsion parameters. This creates an intractable mixed discrete-continuous optimization problem, posing a labor-intensive challenge, heavily reliant on human effort. Force field accuracy is limited by the resolution of chemical perception, which in turn is limited by the number of distinct atom types. Attempting to improve accuracy by increasing the number of atom types results in a combinatorial explosion of bond, angle, and torsion parameters, which imposes strong practical limits.[17] As a result, modelers frequently turn to bespoke parameter generation tools—such as Paramfit,[18] FFBuilder[19] or OpenFF BespokeFit[20]—to assign individual parameters for molecules of interest for which high accuracy is needed, requiring expensive quantum chemical calculations to be performed *ad hoc* and diminishing the speed benefits of Class I force fields.

3 Traditional MM force field parametrization approaches often aim for divide-and-conquer, rather than self-consistency

To tame the explosion of atom type complexity, biomolecular force field efforts have frequently taken the approach of building separate but purportedly compatible models for proteins, small molecules, and other biomolecules independently. For example, the recent AmberTools 23 release[21] recommends combining independently developed force fields to simulate systems containing proteins,[22] DNA,[23,24] RNA,[25] water,[26–28] monovalent[29,30] and divalent[31–33] counterions, lipids,[34] carbohydrates,[35] glycoconjugates,[36,37] small molecules,[38,39] post-translational modifications,[40] and nucleic acid modifications[41]—which collectively might represent more than 100 person-years of effort. While this simplifies the subproblems of parametrizing each class of molecules, using these separate force fields together to treat complex, heterogeneous systems is neither simple nor optimal. There are often overlaps in the chemical space that each force field is designed to model, with no guarantee that the parameters in these regions are identical and remain entirely compatible. This introduces significant caveats when multiple classes of biomolecules interact, risking poor accuracy and greatly frustrating the cases where molecules of different classes must be covalently bonded. As such, extension or expansion to new classes of biomolecules or chemical spaces becomes a time-consuming ordeal, as combining force fields often results in a large combinatorial space of possible force field parameters where the quality of the resulting force field depends heavily on the choices made by the user.

There have been numerous efforts to systematize and automate the process of force field development.[17,19,42–45] For example, the Open Force Field Initiative has developed a number of modern, open-source tools,[20,46] datasets, and force fields[44,45] that employ a direct approach to chemical perception,[17] which use a standard SMARTS-based chemical substructure query to assign entire sets of valence parameters (atoms, bonds, angles, torsions) in a hierarchical manner, attempting to ameliorate the combinatorial explosion of parameters. There have also been extensive efforts to systematically optimize parameters using finite-difference methods[42,43] and machine learning approaches.[47,48] However, much of the work focuses on small molecules, and extending the force field to new chemical domains still requires human effort—jointly optimizing discrete chemical perception rules and continuous force field parameters remains intractable.

4 A graph neural network parametrization scheme can automate, simplify, and significantly improve the accuracy of MM force fields with no performance penalty

Recently, we proposed a novel approach—Espaloma[49] (extensible surrogate potential optimized by mes-sage passing)—which replaces the rule-based discrete atom-typing schemes with a continuous atomic representation generated by graph neural networks that operate on chemical graphs.[49–51] These atom representations are coupled with a set of symmetry-preserving pooling layers and feed-forward neural networks to enable fully end-to-end differentiable construction of MM force fields. The neural network parameters are optimized directly using standard machine learning frameworks to fit quantum chemical and/or experimental data. The expressiveness of Espaloma's continuous atomic representations eliminates the need to combine force fields developed for different chemical domains (it has been well known[52,53] that vanilla GNNs cannot realize some crucial local properties such as ring size, whereas in our implementation this is supplemented by chemo-informatics tools) Thus, Espaloma can self-consistently parametrize any system of molecules with elemental coverage in its training set.

Earlier work[49,50] demonstrated that this approach, in principle, parametrizes multiple classes of biomolecules—the open source Espaloma package was used to train a small Espaloma model for a Class I MM force field on a limited set of 45 000 quantum chemical calculations covering small molecules and amino acids.[49] While surprisingly robust in comparison to traditional small molecule and amino acid force fields, that model was far from providing comprehensive coverage of chemical space relevant to biomolecular modeling and drug discovery, and its potential usage for real-world applications remained unclear.

5 espaloma-0.3: a versatile, robust, and accurate machine-learned Class I MM force field retrainable in a single-GPU day

In this paper, we introduce a significantly enhanced Espaloma framework that incorporates energy and force matching with quantum chemical data, scalability to massive quantum chemical datasets, and stringent regularization for enhanced model stability. We demonstrate how this approach can easily fine-tune the valence terms of an existing Class I small molecule force field (see Section 8 for a discussion on the condensed-phase properties related to the non-bonded parameters) and extend to new chemical domains of interest without a performance penalty, resulting in a generalized and extensible machine-learned Class I MM force field, espaloma-0.3. Trained in a single GPU-day to fit a large and diverse curated quantum chemical dataset of over 1.1 M energy and force calculations for 17 000 unique molecular species, espaloma-0.3 reproduces quantum chemical energetic properties of chemical spaces of small molecules, peptides, and nucleic acids much more accurately than the well-established MM force fields widely used in the fields of biomolecular simulation and computer-aided drug design. Furthermore, it maintains the quantum chemical energy-minimized geometries of small molecules and preserves the condensed phase properties of peptides and folded proteins, thus self-consistently parametrizing proteins and ligands to produce stable simulations leading to highly accurate protein–ligand binding free energy predictions. To our knowledge, this study represents the first well-demonstrated example of a self-consistent MM force field capable of parametrizing a protein–ligand system that is applicable for real-world drug discovery purposes.

This enhanced Espaloma framework demonstrates significant promise as a path forward for systematically building more accurate and extensible force fields with additional quantum chemical data, similarly to how foundational large language models can be fine-tuned to perform better on domain tasks of interest.

6 Espaloma provides a flexible, end-to-end differentiable framework for assigning molecular mechanics (MM) parameters using graph neural networks (GNNs)

Espaloma[49] (Fig. 1) operates analogously to an atom-typing based force field, where chemical perceptions are predefined to generate MM force field parameters Φ_{FF}. However, instead of working with atom types, Espaloma operates on a chemical graph $\mathcal{G}$ using a graph neural network (GNN) parametrized by neural network model parameters Φ_{NN},

$$\Phi_{\mathrm{FF}} \leftarrow \mathrm{espaloma}(\mathcal{G}, \Phi_{\mathrm{NN}}). \quad (2)$$

The resulting parameters Φ_{FF} can then be subsequently used in a standard molecular mechanics package to compute the MM energy and forces for any conformation, as with a standard MM force field.

Espaloma parametrizes molecular systems in three sequential stages (Fig. 1).

6.1 Stage 1

Graph neural networks generate a continuous vectorial atom embedding, replacing discrete atom-typing rules. First, using chemoinformatics toolkits such as RDKit,[57] the molecular system is abstracted as a graph, with nodes and edges denoted as atoms and covalent bonds, respectively. Espaloma uses GNNs[53,58–66] as a replacement for rule-based chemical environment perception[17] to operate on this graph. These neural architectures learn useful representations of atomic chemical environments from simple input features by updating and

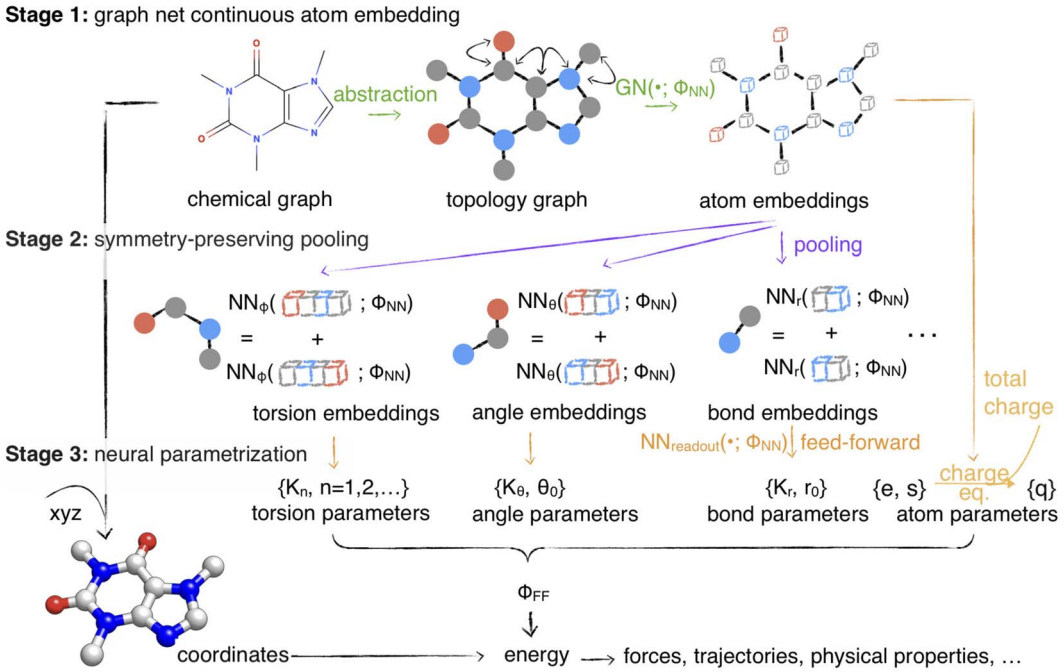

Fig. 1 Espaloma is an end-to-end differentiable molecular mechanics parameter assignment scheme for arbitrary organic molecules. Espaloma (extensible surrogate potential optimized by message-passing) is a modular approach for directly computing molecular mechanics force field parameters Φ_{FF} from a chemical graph $\mathcal{G}$ such as a small molecule or biopolymer *via* a process that is fully differentiable in the model parameters Φ_{NN}. In Stage 1, a graph neural network is used to generate continuous latent atom embeddings describing local chemical environments from the chemical graph. In Stage 2, these atom embeddings are transformed into feature vectors that preserve appropriate symmetries for atom, bond, angle, and proper/improper torsion inference *via* Janossy pooling.[54] In Stage 3, molecular mechanics parameters are directly predicted from these feature vectors using feed-forward neural networks. This parameter assignment process is performed once per molecular species, allowing the potential energy to be rapidly computed using standard molecular mechanics or molecular dynamics frameworks thereafter. The collection of parameters Φ_{NN} describing the espaloma model can be considered as the equivalent complete specification of a traditional molecular mechanics force field such as GAFF[38,39]/AM1-BCC[55,56] in that it encodes the equivalent of traditional typing rules, parameter assignment tables, and even partial charge models. Reproduced from ref. 49 with permission from the Royal Society of Chemistry.

pooling embedding vectors *via* message-passing iterations to neighboring atoms.[60] As such, symmetries in chemical graphs (chemical equivalencies) are inherently preserved, while a rich, continuous, and differentiable learnable representation of the atomic environment is derived.

6.2 Stage 2

Symmetry-preserving pooling generates continuous bond, angle, and torsion embeddings. The representations determined by GNNs in Stage 1 are used to come up with bond, angle, and torsion representations in a symmetry-preserving manner, where the relevant equivalent atom permutations are listed and summed up *via* Janossy pooling.[54]

6.3 Stage 3

Neural parametrization of atoms, bonds, angles, and torsions replaces tabulated parameter assignment. In the final stage, feed-forward neural networks learn the mapping from these symmetry-preserving invariant atom, bond, angle, and torsion embeddings to MM parameters Φ_{FF} that reflect the specific chemical environments appropriate for these terms. Each distinct parameter class (such as atom, bond, angle, and torsion parameters) is assigned by a separate neural network, making this stage fully modular. This stage is analogous to the final table lookup step in traditional force field construction, but it offers significant added flexibility due to the continuous embedding that captures the chemical environment specific to the assigned potential energy term.

The final output is a set of force field parameters Φ_{FF} uniquely determined by the neural network conditioned on its associated weights Φ_{NN}. This means that once the Φ_{NN} is optimized, biomolecular simulations can be performed as fast as those simulated with traditional MM force fields. Atomic partial charges can also be generated within the Espaloma framework, using a geometry-independent charge equilibration approach[67] to rapidly generate AM1-BCC[55,56] quality charges.[68,69]

Overall, the Espaloma framework is end-to-end differentiable—the error in energy (or the function thereof, such as forces) can be backpropagated to optimize the force field parameters Φ_{FF}, and thereby neural network parameters Φ_{NN} that govern how they are produced from the input molecule. Stage 3 is especially modular and flexible. New force field terms that act on atoms, bonds, angles, torsions, or combinations thereof can easily be added and the entire force field refit starting from either an existing Φ_{NN} or training from scratch. In this way, Espaloma provides a rapid and flexible approach to

Table 1 Espaloma-0.3 can directly fit quantum chemical potential energies and forces more accurately than baseline force fields. Espaloma was fit to quantum chemical (QC) potential energies and forces from various gas-phase QC datasets sourced from QCArchive,[70] covering a broad chemical space that includes small molecules, peptides, and RNA molecules (see ESI Section B). The entire dataset consists of 17 427 unique molecules and 1 188 317 conformations. These datasets were extracted from three different QCArchive workflows: BasicDataset, OptimizationDataset, and TorsionDriveDataset. The datasets were partitioned into train, validate, and test sets in an 80 : 10 : 10 ratio split by molecules, except for the RNA-Trinucleotide and RNA-Nucleoside datasets. Since RNA nucleosides and trinucleosides lack chemical diversity, the RNA-Nucleoside dataset was used for training, whereas the RNA-Trinucleotide dataset, which covers the same molecules as the RNA-Diverse dataset but with much more diverse conformers, was used as a test set. The number of molecules and total conformations for each dataset is annotated in the table. We report the root mean square error (RMSE) on the training and test sets, along with the performance of other force fields as baselines on the test set. The baseline force fields used were gaff-2.11,[71] openff-2.0.0,[72] and openff-2.1.0 (ref. 73) for small molecules, Amber ff14SB[22] for peptides, and Amber RNA.OL3 (ref. 25) for RNA molecules. All statistics are computed with predicted and reference energies centered to have a zero mean for each molecule similar to the previous work.[49] The 95% confidence intervals, as annotated in the results, were calculated by bootstrapping molecule replacement using 1000 replicates

					Espaloma-0.3		Baseline force field (test molecules)			
					Energy RMSE (kcal mol^{-1}) force RMSE (kcal mol^{-1} Å^{-1})		Energy RMSE (kcal mol^{-1}) force RMSE (kcal mol^{-1} Å^{-1})			
Dataset (QCArchive workflow)	Category	Mols	Confs	Split	Train (80%)	Test (10%)	gaff-2.11 (ref. 71)	openff-2.0.0 (ref. 72)	openff-2.1.0 (ref. 73)	ff14SB[22]/RNA.OL3 (ref. 25)
SPICE-Pubchem[74,75] (dataset)	Small molecule	14 110	608 436	80 : 10 : 10	$2.06^{2.07}_{2.04}$	$2.30^{2.36}_{2.25}$	$4.39^{4.48}_{4.30}$	$4.21^{4.30}_{4.13}$	$4.45^{4.53}_{4.37}$	—
					$6.22^{6.26}_{6.19}$	$6.81^{6.95}_{6.68}$	$14.02^{14.37}_{13.71}$	$13.95^{14.20}_{13.71}$	$15.45^{15.75}_{15.17}$	—
SPICE-DES-monomers[74,76] (dataset)	Small molecule	369	18 435	80 : 10 : 10	$1.39^{1.46}_{1.32}$	$1.36^{1.67}_{1.13}$	$1.88^{2.22}_{1.57}$	$2.34^{2.75}_{1.97}$	$2.43^{2.81}_{2.05}$	—
					$5.86^{6.02}_{5.69}$	$5.91^{6.42}_{5.49}$	$9.46^{10.91}_{8.09}$	$11.12^{12.47}_{9.86}$	$11.87^{13.15}_{10.57}$	—
Gen2-Opt (OptimizationDataset)	Small molecule	1024	244 989	80 : 10 : 10	$1.36^{1.48}_{1.26}$	$1.66^{2.29}_{1.21}$	$2.29^{2.82}_{1.88}$	$2.18^{2.77}_{1.73}$	$2.25^{2.85}_{1.78}$	—
					$3.94^{4.11}_{3.79}$	$4.47^{5.40}_{3.90}$	$10.51^{11.36}_{9.75}$	$10.53^{11.40}_{9.86}$	$11.67^{12.53}_{10.83}$	—
Gen2-torsion (TorsionDriveDataset)	Small molecule	729	25 832	80 : 10 : 10	$1.76^{1.91}_{1.61}$	$1.64^{2.01}_{1.32}$	$2.53^{3.21}_{1.95}$	$1.69^{2.06}_{1.38}$	$1.83^{2.24}_{1.46}$	—
					$4.31^{4.44}_{4.18}$	$4.71^{5.29}_{4.18}$	$10.50^{11.67}_{9.42}$	$11.11^{12.09}_{10.21}$	$11.92^{12.87}_{11.04}$	—
SPICE-dipeptide[74] (dataset)	Peptide	677	26 279	80 : 10 : 10	$3.21^{3.26}_{3.16}$	$3.09^{3.21}_{2.96}$	$4.24^{4.42}_{4.07}$	$4.11^{4.28}_{3.96}$	$4.28^{4.44}_{4.10}$	$4.36^{4.55}_{4.20}$
					$7.98^{8.07}_{7.88}$	$7.78^{8.02}_{7.55}$	$11.90^{12.32}_{11.50}$	$11.95^{12.32}_{11.62}$	$11.57^{11.88}_{11.26}$	$11.76^{12.09}_{11.40}$
Pepconf-Opt[77] (OptimizationDataset)	Peptide	557	166 291	80 : 10 : 10	$2.61^{2.83}_{2.43}$	$2.79^{3.13}_{2.45}$	$3.53^{3.82}_{3.03}$	$2.91^{3.39}_{2.56}$	$3.19^{3.66}_{2.73}$	$3.59^{4.17}_{3.00}$
					$3.83^{4.09}_{3.60}$	$4.01^{4.46}_{3.63}$	$8.07^{8.23}_{7.84}$	$8.74^{9.08}_{8.49}$	$8.79^{9.56}_{8.27}$	$9.13^{9.70}_{8.67}$
Protein-torsion (TorsionDriveDataset)	Peptide	62	48 999	80 : 10 : 10	$2.27^{2.50}_{2.06}$	$1.93^{2.14}_{1.73}$	$3.53^{3.82}_{3.03}$	$2.91^{3.39}_{2.56}$	$3.19^{3.66}_{2.73}$	$3.59^{4.17}_{3.00}$
					$3.94^{4.24}_{3.70}$	$3.49^{3.78}_{3.22}$	$8.07^{8.23}_{7.84}$	$8.74^{9.00}_{8.49}$	$8.79^{9.56}_{8.27}$	$9.13^{9.70}_{8.67}$
RNA-diverse (dataset)	RNA	64	3703	80 : 10 : 10	$4.12^{4.31}_{3.95}$	$4.17^{4.52}_{3.85}$	$5.65^{6.32}_{4.95}$	$5.79^{6.19}_{5.37}$	$6.26^{6.90}_{5.64}$	$6.06^{6.43}_{5.70}$
					$4.44^{4.47}_{4.40}$	$4.41^{4.51}_{4.29}$	$17.19^{17.71}_{16.71}$	$18.54^{19.10}_{17.85}$	$19.68^{20.15}_{19.19}$	$19.38^{19.83}_{18.77}$
RNA-trinucleotide (dataset)	RNA	64	35 811	0 : 0:100	—	$3.75^{3.94}_{3.59}$	$5.79^{5.98}_{5.61}$	$5.81^{5.96}_{5.67}$	$6.26^{6.42}_{6.10}$	$5.94^{6.12}_{5.77}$
					—	$4.28^{4.39}_{4.20}$	$17.15^{17.28}_{17.00}$	$18.88^{19.02}_{18.72}$	$19.97^{20.13}_{19.81}$	$19.82^{19.97}_{19.67}$
RNA-nucleoside (dataset)	RNA	4	9542	100 : 0:0	$1.32^{1.49}_{1.16}$	—	—	—	—	—
					$4.17^{4.47}_{3.86}$	—	—	—	—	—

experimenting with different potential functions (such as the addition of point polarizability or exploration of alternative functional forms) or retraining with augmented training datasets.

7 Extensive open quantum chemical dataset curated to provide coverage of biomolecules: small molecules, proteins, and nucleic acids

To develop a self-consistent MM force field broadly applicable to biomolecular modeling, we first curate a high-quality gas-phase quantum chemical dataset deposited in QCArchive[70] (Table 1). The curated quantum chemical dataset is built from several components that provide complementary coverage of relevant biomolecular chemistries: from the foundational SPICE dataset,[74] we extracted a large set of drug-like small molecules selected from PubChem,[78] dipeptides (capped 2-mers) and their common protonation and tautomeric variants, and diverse molecular fragments providing broad coverage of biomolecules from the DES370K dataset;[76] from the OpenFF 1.x ("Parsley")[44] and 2.x ("Sage")[45] datasets, we extracted optimization and torsion-drive datasets for diverse small molecules; a diverse set of dipeptide (capped 2-mers), tripeptides (capped 3-mers), disulfide-bridged, bioactive, and cyclic peptides from the PepConf dataset;[77] a peptide torsion scan set generated by the Open Force Field Consortium for the OpenFF 3.x ("Rosemary") force field;[79] and a new set of RNA nucleosides, trinucleotides, and diverse experimental RNA fragments sourced from the Nucleic Acid Database[80] and RNA Structure Atlas[81] to extend coverage to this important and growing class of drug targets.

To capture the rugged conformational energy surface of biomolecules, the quantum chemical datasets were extracted from three different QCArchive workflows: Dataset, OptimizationDataset, and TorsionDriveDataset. A Dataset contains single-point energy calculations of structures that are not necessarily at their local quantum energy minima, generated using MD simulations or conformer generators. An OptimizationDataset is a collection of QM optimization trajectories for a given structure. A TorsionDriveDataset involves torsion scans performed on a set of rotatable torsions, followed by QM optimization.

The curated dataset consists of 1 188 317 conformations of 17 427 unique molecules in total. We also computed the AM1-BCC ELF10 partial charges using the OpenEye Toolkits to train and generate AM1-BCC[55,56] quality partial charges with Espaloma. Complete details of the dataset construction and composition are given in ESI Section B.† All quantum chemical energies are computed with the Open Force Field (OpenFF) standard level of quantum chemical theory (B3LYP-D3BJ/DZVP),[44,45] which balances the computational efficiency and accuracy to reproduce the conformations generated by higher levels of theories.[82] These quantum chemical datasets were generated with the open source psi4 quantum chemistry packag[83] using the QCArchive[70] QCFractal infrastructure *via* OpenFF QCSubmit[84] workflows.

8 Espaloma force field reproduces quantum chemical energies and forces

Leveraging the curated gas-phase quantum chemical datasets discussed in Section 2, we fine-tune and extend the OpenFF 2.0 ("Sage") force field, openff-2.0.0—a Class I MM force field originally developed for small molecules—into new chemical domains of interest, resulting in a novel Class I MM force field termed espaloma-0.3. Similar to the original implementation[49] and historic practice in MM force field parametrization,[22,24,25,43,44,85] we optimized the valence parameters (bonds, angles, and proper/improper torsions) and use the Lennard-Jones parameters from openff-2.0.0.[45] While it is possible to optimize Lennard-Jones parameters as well, it is critical to include more computationally expensive condensed-phase simulations when doing so.[86,87] For partial charges, following the protocol of Wang *et al.*[68] we predict the electronegativity and hardness of atoms used in a charge equilibration[67] to predict atomic partial charges while preserving the total charge of a given molecule. We utilize the AM1-BCC ELF10 partial charges computed with the OpenEye Toolkits as our target partial charges.

We enhance the original Espaloma framework to improve the model stability and data efficiency (see ESI Section C† for further details):

- quantum chemical forces are incorporated into training to provide more information about the quantum chemical potential surface;
- L2 regularization is applied to proper and improper torsion force constants to suppress spurious features in torsion profiles;
- improper torsion terms expressed using $n = 1, 2$ periodicities to reduce the complexity of the model and to align with other conventional force fields which usually employs $n = 1, 2$ periodicities;
- node features that were sensitive to resonance form have been eliminated to ensure chemically equivalent representations of the same molecule receive identical parameters.

To train espaloma-0.3, we randomly shuffle the datasets and split each dataset by molecules into train, test, and validation sets (80%, 10%, and 10%, respectively) based on unique isomeric SMILES strings. Since the MM force field is incapable of reproducing quantum chemical heats of formation, which are reflected as an additive offset in quantum chemical energy targets for each molecule, we shift the reference quantum chemical energy of each molecule to have zero mean; note that when deployed, the absolute value of MM energy is not physically meaningful and traditional MM force fields are never used to simulate bond-breaking events. The loss function used in training included deviations from quantum chemical snapshot energies and forces, as well as deviations from target partial charges for each molecule in the training set (see ESI Section C† for complete details).

As shown in Table 1, espaloma-0.3 significantly outperforms all baseline force fields (gaff-2.11,[71] openff-2.0.0,[72] openff-2.1.0,[73] Amber ff14SB,[22] Amber RNA. OL3 (ref. 25)) in reproducing quantum chemical energies and forces, demonstrating the ability of espaloma-0.3 to recapitulate the quantum chemical energy surface more accurately than current-generation Class I MM potentials for biomolecules and organic chemistry despite using the same functional form. In contrast, the baseline force fields widely popular in the field of biomolecular simulations yield considerable energy errors and huge force errors (on average twice to thrice that of espaloma-0.3) with respect to quantum chemical calculations. The performance superiority holds true across diverse chemical categories, suggesting the general utility of espaloma-0.3 in a wide array of chemical and biochemical modeling tasks, as evidenced in Sections 11 and 12. These observations hold true when Espaloma is trained with different data splitting strategies (ESI Table 1†).

Notably, the backbone torsion parameters for ff14SB are empirically adjusted to improve agreement with condensed-phase NMR data. Therefore, it might be expected to perform less effectively when benchmarked against quantum chemical energetic properties. For a more rigorous comparison, we conducted the same benchmark experiment using ff14SBonlysc,[88] which is the same model as ff14SB but without the empirical backbone corrections. The resulting energy RMSE on test datasets for SPICE-Dipeptide, Pepconf-Opt, and Protein-Torsion were 4.36 [95% CI: 4.52, 4.19], 3.93 [95% CI: 3.58, 4.23], and 3.59 [95% CI: 3.00, 4.18] kcal mol^{-1} respectively, with corresponding force RMSE values of 11.76 [95% CI: 11.41, 12.12], 10.22 [95% CI: 9.82, 10.68], 9.13 [95% CI: 8.67, 9.70] kcal mol^{-1} Å^{-1}; espaloma-0.3 performed superiorly better for all three datasets.

9 Espaloma force field preserves quantum chemical energy minima

We next examined whether the ability of espaloma-0.3 to quantitatively reproduce the quantum chemical equilibrium conformational energetics extends to an ability to qualitatively preserve the conformations of quantum chemical local energy minima—important for accurately representing geometries for phenomena like ligand binding docking studies, simulations, or free energy calculations. To assess this, we used a standardized industry benchmark of gas-phase QM-optimized geometries (the OpenFF Industry Benchmark Season 1 v1.1 (ref. 89)‡ obtained from QCArchive) to compare the structures and energetics of conformers optimized using espaloma-0.3 and baseline force fields (openff-2.0.0, openff-2.1.0, and gaff-2.11) with respect to their QM-optimized geometries at the B3LYP-D3BJ/DZVP level of theory. The dataset is a collection of drug-like molecules selected by industry partners of the Open Force Field Consortium and is representative of their current interests in chemical spaces, serving as an out-of-distribution test dataset. It contains 9728 unique molecules and 73 301 conformers after filtering out any quantum chemical calculation failures due to convergence issues and connectivity changes during geometry optimization.

As shown in Fig. 2(a and b), the geometries and relative conformer energies with respect to their quantum chemical reference values showed better agreement with espaloma-0.3 than with the baseline force fields—openff-2.0.0, openff-2.1.0, and gaff-2.11. Additionally, the bonds, angles, and torsions in MM-optimized geometries obtained using espaloma-0.3 show close agreement with quantum chemical values (Fig. 2(c)), resulting in an overall performance compatible or slightly better than the baseline force fields. The bond outliers (>0.1 Å) with espaloma-0.3 arise from three sulfonamides connected to aliphatic carbons, comprising a total of 30 conformers—0.04% of the conformers in the entire benchmark dataset—exhibiting ~0.4 Å elongated S–N bond distances in the sulfonamide groups compared to the QM-optimized geometries (ESI Fig. 4(a)†). 12 other molecules containing sulfonamide groups, excluding the bond RMSD outliers were found within the benchmark dataset with each molecular conformer featuring reasonable bond distances within the sulfonamide group (ESI Fig. 4(b)†). However, the nitrogen geometry of pyrazoles and imidazoles substituted with sulfonamides became trigonal pyramidal when minimized with espaloma-0.3, rather than preserving a flat ring geometry and losing their sp2 hybridized features, as observed with QM-optimized geometries (ESI Fig. 4(c)†). The angle outlier is also related to a sulfonamide but was a singleton of a non-druglike molecule containing a single conformer, with ~40° deviation from its original QM-optimized geometry (ESI Fig. 4(a)†).

Nonetheless, the degree of improvement of espaloma-0.3 relative to openff-2.0.0 is surprising and intriguing, considering that the Lennard-Jones parameters are transferred from openff-2.0.0 and the overlap in the underlying Optimization and TorsionDrive datasets used for optimizing both force fields. This is notable, despite espaloma-0.3 was trained on quantum chemical dataset comprising larger and broader chemical species.

10 Espaloma force field reproduces experimental NMR observables for peptides and folded proteins

10.1 Peptides

To quantitatively assess the ability of espaloma-0.3 to model the intrinsic backbone preferences of amino acids, we performed MD simulations of thirteen short, unstructured peptides for which NMR observables have been experimentally measured.[91,92] The peptides are composed of 3 to 5 residues, uncapped, and have protonated C Termini due to the low pH of the NMR experiments. Measured vicinal scalar couplings inform on the backbone dihedral preferences of these peptides. Scalar couplings were computed from 500 ns trajectories using a Karplus model,[98,102,104–107] and agreement with experimental observables was quantified using a χ^2 value.

Overall, espaloma-0.3 produces closer agreement with experiment than ff14SB, as evidenced by the low χ^2 value (Fig. 3(a)). With note, ff14SB tends to exhibit closer agreement with experiments on amino acids with short side chains such as

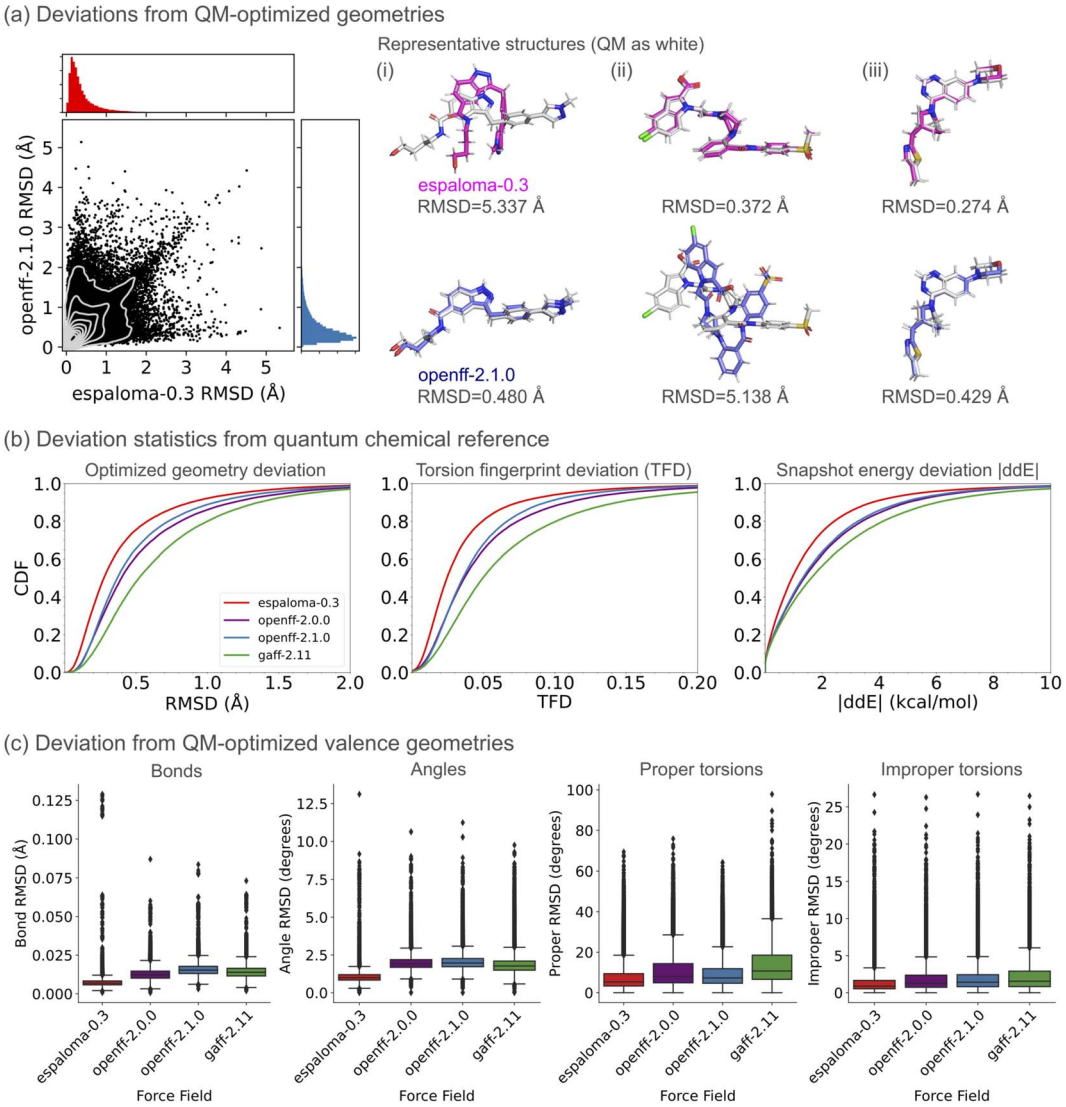

Fig. 2 Espaloma-0.3 preserves the location of quantum chemical energy minima. An industry standard benchmark of gas-phase QM-optimized geometries (the OpenFF Industry Benchmark Season 1 v1.1 (ref. 89) from QCArchive), comprising 9728 unique molecules and 73 301 conformers, was used to compare the structures and energetics of conformers optimized using espaloma-0.3, openff-2.0.0,[72] openff-2.1.0,[73] and gaff-2.11 (ref. 71) with respect to their QM-optimized geometries at the B3LYP-D3BJ/DZVP level of theory. (a) Representative scatter plot of root-mean-square deviation (RMSD) of atomic positions between espaloma-0.3 and openff-2.1.0. The superposed structures between the QM-optimized (white) and MM-optimized geometries with the maximum RMSD obtained by (i) espaloma-0.3, (ii) openff-2.1.0, and (iii) the median RMSD of espaloma-0.3 are shown. (b) The cumulative distribution functions of root-mean-square deviation (RMSD) of atomic positions, torsion fingerprint deviation (TFD) score, and relative energy differences (ddE) as described in a previous work[90] are reported. (c) Distributions of bond, angle, proper torsion, and improper torsion RMSD within each conformer with respect to its QM-optimized geometries are shown as quartile box plots. Lower values for all metrics indicate that the MM-optimized geometry is close to the quantum chemical reference structure.

glycine and alanine (Fig. 3(b)). This is unsurprising as the backbone torsion parameters for ff14SB were tuned to reproduce the NMR scalar couplings for the alanine 5-mer peptide included in this dataset.[22] However, espaloma-0.3 tends to have closer agreement with experiments on more challenging amino acids with charged (*e.g.* lysine), bulky (*e.g.* methionine), or β-branched (*e.g.* valine) side chains.

10.2 Folded proteins

The intrinsic dynamics of both the backbone and χ_1 side chains in folded proteins were assessed by conducting MD simulations of four globular proteins: the third IgG-binding domain of protein G (GB3), bovine pancreatic trypsin inhibitor (BPTI), lysozyme, and ubiquitin (Fig. 4). These proteins, for which scalar couplings have been measured by NMR experiments, have been extensively studied for the development of protein force fields.[22] Scalar couplings were computed from 10 μs trajectories using the same Karplus model[96,98,102,104–109] that was employed in the peptide analysis. Additionally, inter-residue scalar couplings between backbone–backbone hydrogen bonds[109] were computed for GB3 and ubiquitin. The agreement with experimental observables was quantified using an average

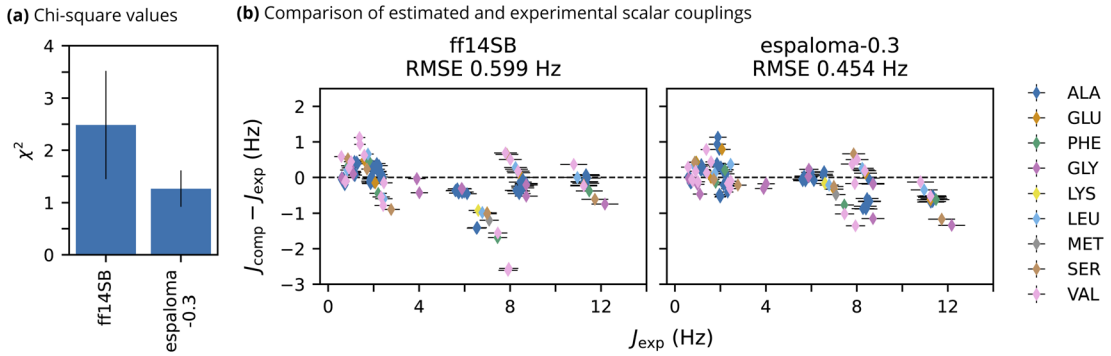

Fig. 3 espaloma-0.3 reproduces experimental NMR scalar couplings of unstructured peptides better than well-established biomolecular force field, ff14sb. (a) χ^2 values (lower is better) quantifying deviations of simulated NMR scalar couplings computed from 500 ns trajectories from experimental NMR measurements.[91,92] Error bars represent a 95% confidence interval constructed from the critical values of a Student's *t* distribution and the standard error of the mean across the NMR observables. (b) Comparison of the error in computed estimates of NMR scalar couplings *versus* experiment. Colors represent the identity of the amino acid associated with each scalar coupling. Horizontal error bars represent the estimate of the systematic error in the experimental scalar coupling, and vertical error bars represent the uncertainty due to the computed estimate (standard error of the mean across 3 replicates) and the uncertainty due to the experimental value (systematic error) added in quadrature.

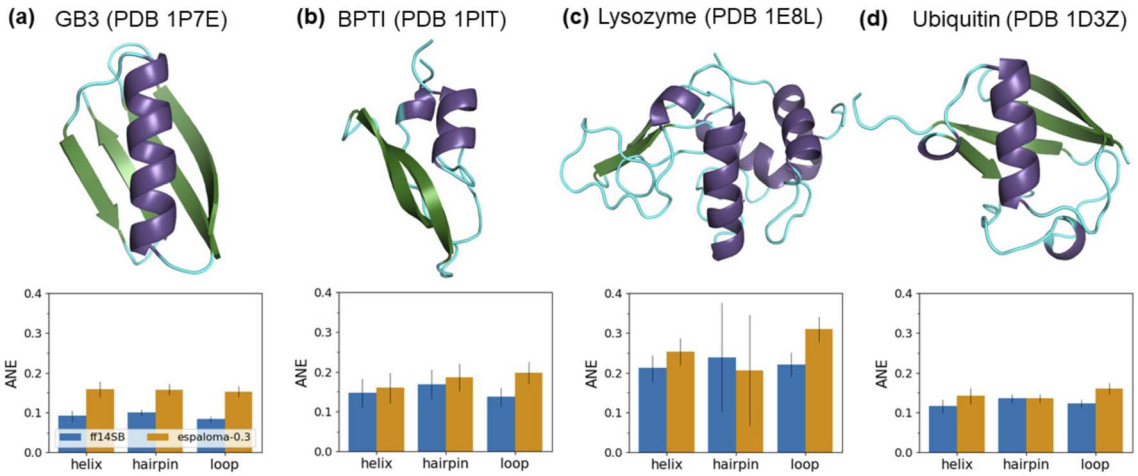

Fig. 4 Espaloma-0.3 reproduces experimental NMR scalar couplings of folded globular proteins with a slightly higher error compared to the well-established biomolecular force field, ff14sb. The absolute normalized error (ANE) values (lower the better),[22] quantifying the deviations of simulated NMR scalar couplings from 10 μs trajectories compared to experimental measurements,[93–103] are compared for (a) GB3, (b) BPTI, (c) lysozyme, and (d) ubiquitin. The regions of helix, hairpin, and loop are depicted in purple, green, and cyan, respectively, as defined by Ramachandran angles from the crystal structures. Error bars represent a 95% confidence interval, constructed from the critical values of a Student's *t*-distribution and the standard error of the mean across the NMR observables, based on three replicates of the 10 μs simulation. Note that ANEs were applied instead of χ^2 values to address the potential underestimation of experimental and Karplus model inaccuracies, as well as the significant variance in the scalar coupling value range across different coupling types,[22] resulting in a more intuitive metric (see ESI Section E.2† for more details). A comparison with χ^2 values can be found in ESI Table 2.†

normalized error (ANE) metric, motivated by the work of Maier *et al.*[22] The ANE metric was introduced to address the potential underestimation of experimental errors and Karplus model inaccuracies, as well as the significant variance in the scalar coupling value range across different coupling types, resulting in a more intuitive metric than the χ^2 value (see ESI Section E.2† for more details). Here, 0 indicates the best possible agreement, while 1 indicates maximum deviation.

Overall, espaloma-0.3 accurately replicates experimental NMR scalar couplings for all four folded proteins, but with slightly higher ANE values compared to ff14sb (ESI Table 2, Fig. 5† and 4). The larger deviations tend to arise from the side chain scalar couplings (ESI Fig. 6†). Simulations with espaloma-0.3 also indicate a greater decrease in the occupancy of defined folded regions, such as the alpha (α) and beta (β) backbone structures (ESI Fig. 7†), leading to slightly increased backbone flexibility, as suggested by the Cα RMSD plots (ESI Fig. 8†).

Although folded proteins tend to be more flexible and less reproducible regarding the experimental NMR scalar couplings when simulated with espaloma-0.3 than with ff14sb, the above

results, along with the peptide benchmark results, reflect the transferability of Espaloma's neural network parameters—which were trained on gas-phase quantum chemistry data—to the condensed phase.

11 Espaloma force field accurately describes protein–ligand binding free energies

To evaluate espaloma-0.3 for real-world drug discovery applications, we performed relative alchemical free energy calculations on a curated protein–ligand binding benchmark dataset, which was adopted from the Open Force Field protein–ligand benchmark dataset (see ESI Section F†).§ We selected target systems from available datasets based on several criteria: firstly, we prioritized systems with ligands that can be effectively modeled to alleviate the potential sampling issues arising from poor initial ligand poses; secondly, we excluded systems with cofactors and ions near the ligand binding site to simplify the evaluation; thirdly, we considered systems with diverse structure–activity relationships, including ligand net charges, multiple R-group enumeration, and scaffold hopping. As a result, we selected four well-studied protein–ligand binding benchmark systems. The protein structures, ligand poses, and ligand transformation networks were manually curated to ensure the free energy benchmark was an accurate and reproducible assessment of force field accuracy.

• Tyk2 (PDB: 4GIH),[112] a non-receptor tyrosine-protine kinase, has therapeutic significance in inflammatory bowel diseases (IBD). This particularly popular system has good convergence and served as a control experiment.

• Cdk2 (PDB: 1H1Q),[113] a cyclin-dependent kinase, is involved in molecular pathology of cancer and is, therefore, a popular target for structure-based drug design. We use this system, which is complexed with cyclin A, to test the capability of parametrizing multiple protein subunits.

• P38 (PDB: 3FLY)[114] is a mitogen-activated protein (MAP) kinase which is a central component in signaling networks in mammalian cell types. This target is another well-studied system, but is expected to be more challenging compared to Tyk2 and Cdk2 because of the larger ligand transformations and exploration of structure–activity relationships with multiple R-groups from different scaffold positions.

• Mcl1 (PDB: 4HW3)[115] (myeloid cell leukemia 1) is a member of the Bcl-2 family of proteins, which is overexpressed in various cancers and promotes aberrant survival of tumor cells. This target entails all ligands with a net charge of −1 and includes scaffold hopping; thus, chosen to test the capability to simulate free energy calculations for charged ligands and scaffold hopping.

Within each system, we benchmarked three approaches of parametrization to evaluate the accuracy of espaloma-0.3 in modeling either the ligand alone or the entire protein–ligand complex:

• Protein: ff14SB/ligand: openff-2.1.0 (ff14SB + openff-2.1.0): as a baseline, we parametrize the ligand region using a well-established small molecule force field openff-2.1.0 (ref. 73) and use the Amber ff14SB[22] to parametrize the protein.

• Protein: ff14SB/ligand: espaloma-0.3 (ff14SB + espaloma-0.3): we parametrize the ligand region using espaloma-0.3 and use the Amber ff14SB[22] to parametrize the protein. We only parametrize the ligand region with espaloma-0.3 to provide a head-to-head comparison with openff-2.1.0.

• Protein: espaloma-0.3/ligand: espaloma-0.3 (espaloma-0.3): we apply espaloma-0.3 to both the ligand and protein regions of the system. This is to test the capability of espaloma-0.3 to entirely replace the force field parametrization pipeline. Instead of using two separate force fields for small molecules and proteins, each developed independently, we aim to apply a self-consistently developed force field that covers different chemical domains.

As our training dataset does not yet include water and ions, all systems were solvated with TIP3P water[26] and neutralized with the Joung and Cheatham monovalent counterions.[29] The perses 0.10.1 infrastructure[110] was used to perform the alchemical protein–ligand binding free energy calculations (see ESI Section G†).

In Fig. 5 and Table 2, we illustrate that espaloma-0.3, which parametrizes both the protein and ligand self-consistently, has comparable protein–ligand binding free energy performance to ff14SB + openff-2.1.0. espaloma-0.3 achieves absolute (ΔG) and relative ($\Delta\Delta G$) free energy RMSE of 1.02 [95% CI: 0.74, 1.37] kcal mol^{-1} and 1.12 [95% CI: 0.88, 1.41] kcal mol^{-1}, respectively. Correspondingly, the ΔG and $\Delta\Delta G$ RMSE for ff14SB + openff-2.1.0 were 1.01 [95% CI: 0.73, 1.33] kcal mol^{-1} and 1.21 [95% CI: 0.93, 1.54] kcal mol^{-1}, respectively. Although, the reported error and correlation statistics have overlapping confidence intervals, these results are encouraging as espaloma-0.3 demonstrates its capability to cover different chemical domains, which traditional force fields have struggled for decades and have not accomplished.

Notably, a large outlier for the Mcl1 system for all three cases was observed as shown in Fig. 5. The problematic ligand transformation and the initial ligand pose is illustrated in ESI Fig. 11.† The relative binding affinity $\Delta\Delta G$ computed with ff14SB + espaloma-0.3 was 4.05 kcal mol^{-1} (Fig. 5(b)). However, we found that the error can be reduced to 2.60 kcal mol^{-1} when the alchemical binding free energy calculation was performed from a flipped binding pose, which is in better agreement with the experimental difference ($\Delta\Delta G = -0.54$ kcal mol^{-1}).

We also conducted another set of free energy calculations for the four target systems, each with three parametrization approaches (ESI Fig. 9†). In most cases, the absolute (ΔG) and relative ($\Delta\Delta G$) binding free energies from the two independent trials were within 1.0 kcal mol^{-1}, demonstrating reasonable reproducibility; except for P38, which tends to be a more challenging target for the free energy calculations to reproduce.

It is worth noting that the ligands from the protein–ligand binding benchmark dataset are highly dissimilar to the molecules used in developing espaloma-0.3, with a maximum Tanimoto similarity of 0.5 between the two sources, suggesting the high generalizability of Espaloma (ESI Fig. 10†).

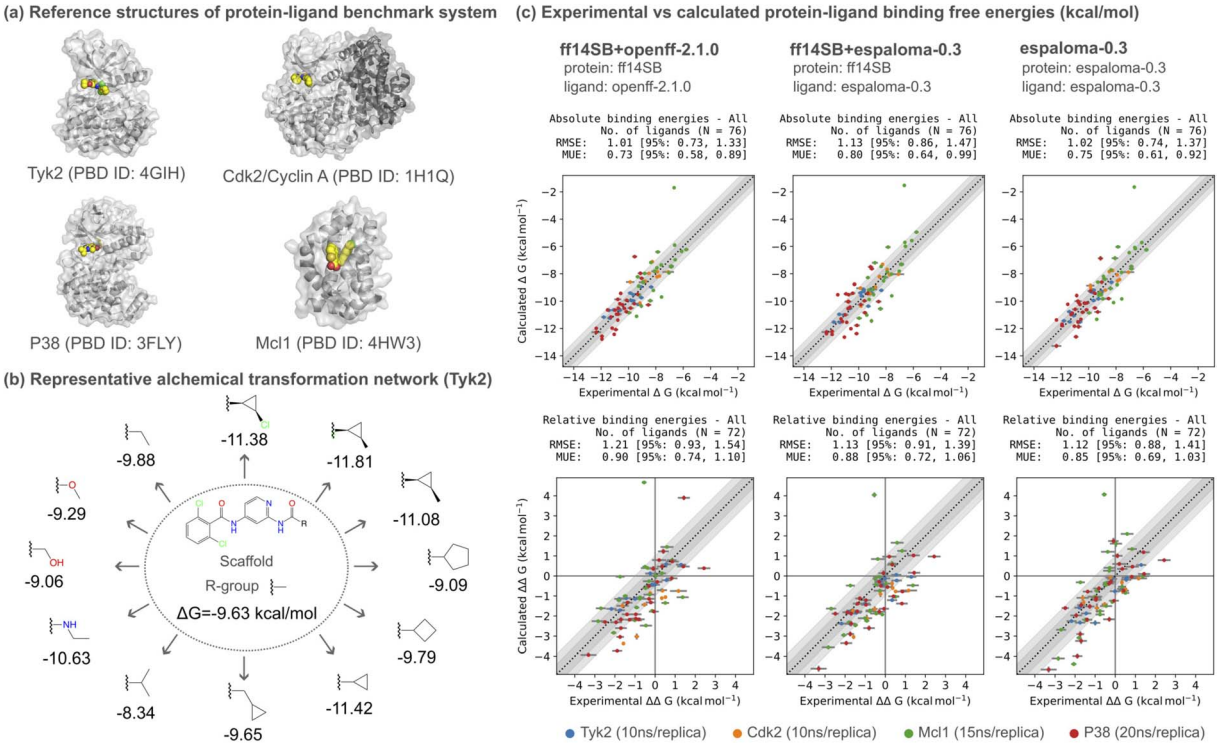

Fig. 5 espaloma-0.3 can be used for accurate protein–ligand alchemical free energy calculations. (a) Protein–ligand (PL) alchemical free energy calculations were calculated for Tyk2 (10 ns/replica), Cdk2 (10 ns/replica), Mcl1 (15 ns/replica), P38 (20 ns/replica) using a curated PL-benchmark dataset (see ESI Section F†) which comprises 76 ligands in total. The PL structures used to setup the alchemical free energy calculations for each target system is shown. Here, we used Perses 0.10.1 relative free energy calculation infrastructure,[110] based on OpenMM 8.0.0,[111] to assess the accuracy of espaloma-0.3 and openff-2.1.0 (ref. 73) combined with Amber ff14SB force field[22] for comparison. (b) Schematic illustration of the alchemical ligand transformation network for Tyk2. The methyl R-group in the center is alchemically transformed into various R-groups. The binding free energy for each R-group is annotated alongside the respective R-groups. (c) The openff-2.1.0 (ref. 73) with protein parametrized with Amber ff14SB force field (ff14SB + openff-2.1.0) achieves an absolute free energy (ΔG) RMSE of 1.01 [95% CI: 0.73, 1.33] kcal mol^{-1}. The espaloma-0.3 for predicting valence parameters and partial charges of small molecules combined with Amber ff14SB force field for proteins (ff14SB + espaloma-0.3) achieves an absolute free energy (ΔG) RMSE of 1.13 [95% CI: 0.86, 1.47] kcal mol^{-1}. Parametrizing small molecule and protein self-consistently with espaloma-0.3 (espaloma-0.3) achieves absolute free energy (ΔG) RMSE of 1.02 [95% CI: 0.74, 1.37] kcal mol^{-1} which is comparable to those obtained by (ff14SB + openff-2.1.0) and (ff14SB + espaloma-0.3). All systems were solvated with TIP3P water[26] and neutralized with 300 mM NaCl salt using Joung and Cheatham monovalent counterions.[29] The light and dark gray regions depict the confidence bounds of 0.5 kcal mol^{-1} and 1.0 kcal mol^{-1}, respectively.

11.1 Regularization and larger training dataset significantly improve performance

To assess the impact of dataset scale and the regularization procedures introduced here for training espaloma-0.3, we compared the protein–ligand binding free energy calculations using the first-generation Espaloma force field (0.2.2),[49] which was trained on a limited quantum chemical dataset and without regularization compared to 0.3. The free energy calculations were conducted for all four target systems and were prepared similarly to those described above. In ESI Fig. 12,† espaloma-0.2.2 significantly underperforms compared to espaloma-0.3 for the Cdk2 system due to a large outlier. espaloma-0.2.2 also demonstrates lesser performance on the Tyk2 system, as illustrated in ESI Fig. 13.† Importantly, the protein–ligand binding free energy calculations were unstable for Mcl1 and P38, with many of the ligand transformations being suspended during the simulation. These results indicate that espaloma-0.3, trained on an extensive quantum chemical dataset and with an improved training strategy, has resulted in the development of a robust and stable Espaloma force field.

12 Espaloma force field produces stable long-time molecular dynamics of protein–ligand complex system

Recent benchmarks of machine learned force fields demonstrated that many of these potentials are accurate but cannot produce stable molecular dynamics simulations.[116] To assess whether espaloma-0.3 was sufficiently stable and robust for general use in molecular dynamics simulations, we performed multiple replicates of a 3 microsecond MD simulation of a solvated protein–ligand complex (Tyk2 complexed with ligand #1, ESI Fig. 13†) and monitored the root-mean square deviation (RMSD) of the ligand and Cα protein atoms, as well as the root-mean square fluctuation (RMSF) profiles of the Cα protein atoms, as shown in ESI Fig. 14.†

Table 2 Protein–ligand alchemical free energy calculation benchmarks show espaloma-0.3 achieves high accuracy that is competitive to well-established force fields. Here, we report several different metrics to assess the performance of the protein–ligand binding benchmark results including root mean square error (RMSE), mean unsigned error (MUE), the square of the correlation coefficient (R^2), and the Spearman's rank correlation coefficient (ρ) along with 95% CI for each metric. The initial PDB ID, number of compounds, number of edges (ligand transformations), the binding affinity range, and the simulation time per replica are reported in the table

						Protein: ff14SB/ligand: openff-2.1.0					
						Relative ($\Delta\Delta G$)		Absolute (ΔG)			
System	PDB ID	Compds	Edges	Range (kcal mol^{-1})	ns/replica	RMSE	MUE	RMSE	MUE	R^2	Spearman ρ
Tyk2	4GIH	13	12	3.47	10	$0.54^{0.71}_{0.36}$	$0.45^{0.62}_{0.28}$	$0.50^{0.64}_{0.36}$	$0.42^{0.57}_{0.27}$	$0.80^{0.93}_{0.53}$	$0.89^{0.96}_{0.75}$
Cdk2	1H1Q	10	9	2.78	10	$1.43^{1.75}_{1.04}$	$1.29^{1.67}_{0.80}$	$0.74^{0.93}_{0.50}$	$0.63^{0.86}_{0.41}$	$0.48^{0.85}_{0.13}$	$0.69^{0.92}_{0.30}$
Mcl1	4HW3	25	24	4.19	15	$1.50^{2.12}_{0.83}$	$1.02^{1.55}_{0.63}$	$1.36^{2.01}_{0.77}$	$0.97^{1.41}_{0.66}$	$0.50^{0.73}_{0.35}$	$0.71^{0.86}_{0.57}$
P38	3FLY	28	27	3.81	20	$1.06^{1.30}_{0.81}$	$0.87^{1.09}_{0.65}$	$0.90^{1.19}_{0.60}$	$0.69^{0.92}_{0.50}$	$0.57^{0.78}_{0.38}$	$0.76^{0.89}_{0.63}$
						Protein: ff14SB/ligand: espaloma-0.3					
						Relative ($\Delta\Delta G$)		Absolute (ΔG)			
System	PDB ID	Compds	Edges	Range (kcal mol^{-1})	ns/replica	RMSE	MUE	RMSE	MUE	R^2	Spearman ρ
Tyk2	4GIH	13	12	3.47	10	$0.70^{0.98}_{0.34}$	$0.52^{0.80}_{0.28}$	$0.48^{0.65}_{0.29}$	$0.37^{0.55}_{0.23}$	$0.79^{0.95}_{0.49}$	$0.89^{0.97}_{0.71}$
Cdk2	1H1Q	10	9	2.78	10	$1.15^{1.44}_{0.85}$	$1.05^{1.36}_{0.73}$	$0.56^{0.74}_{0.32}$	$0.46^{0.66}_{0.27}$	$0.63^{0.92}_{0.27}$	$0.80^{0.96}_{0.53}$
Mcl1	4HW3	25	24	4.19	15	$1.38^{1.96}_{0.90}$	$1.06^{1.44}_{0.76}$	$1.51^{2.15}_{0.90}$	$1.08^{1.56}_{0.74}$	$0.60^{0.80}_{0.42}$	$0.77^{0.90}_{0.63}$
P38	3FLY	28	27	3.81	20	$1.03^{1.26}_{0.81}$	$0.82^{1.05}_{0.59}$	$1.10^{1.32}_{0.86}$	$0.88^{1.13}_{0.63}$	$0.38^{0.64}_{0.11}$	$0.62^{0.80}_{0.34}$
						Protein: espaloma-0.3/ligand: espaloma-0.3					
						Relative ($\Delta\Delta G$)		Absolute (ΔG)			
System	PDB ID	Compds	Edges	Range (kcal mol^{-1})	ns/replica	RMSE	MUE	RMSE	MUE	R^2	Spearman ρ
Tyk2	4GIH	13	12	3.47	10	$0.67^{0.87}_{0.45}$	$0.56^{0.76}_{0.35}$	$0.46^{0.58}_{0.33}$	$0.40^{0.53}_{0.28}$	$0.81^{0.94}_{0.64}$	$0.90^{0.97}_{0.79}$
Cdk2	1H1Q	10	9	2.78	10	$0.84^{1.05}_{0.58}$	$0.75^{0.99}_{0.51}$	$0.63^{0.76}_{0.48}$	$0.58^{0.74}_{0.41}$	$0.47^{0.82}_{0.14}$	$0.68^{0.90}_{0.41}$
Mcl1	4HW3	25	24	4.19	15	$1.44^{1.99}_{0.96}$	$1.10^{1.50}_{0.76}$	$1.40^{2.09}_{0.78}$	$1.00^{1.43}_{0.67}$	$0.56^{0.78}_{0.40}$	$0.75^{0.88}_{0.63}$
P38	3FLY	28	27	3.81	20	$1.02^{1.24}_{0.77}$	$0.79^{1.04}_{0.56}$	$0.91^{1.13}_{0.68}$	$0.75^{0.95}_{0.57}$	$0.47^{0.68}_{0.24}$	$0.68^{0.82}_{0.49}$

The simulations parametrized with espaloma-0.3 remained comparably stable to those generated with ff14SB + openff-2.1.0, with both protein and ligand RMSD generally remaining below 2.0 Å. The averaged RMSF profiles, simulated using espaloma-0.3 and ff14SB + openff-2.1.0, showed a similar trend, with a Pearson correlation coefficient of 0.76. However, espaloma-0.3 exhibited higher peaks, indicating greater protein flexibility with this force field—a finding that aligns with those described in Section 5.

13 Discussion

In this study, we introduced an enhanced graph neural network approach to rapidly construct a new generation of accurate, robust, and generalizable machine-learned MM force field, espaloma-0.3, capable of fine-tuning and extending to new chemical domains of interest. The newly developed force field captures both quantitative and qualitative behavior of quantum chemical conformational energetics for a wide range of chemical species. As a result, it not only recapitulates quantum chemical conformational energetics and geometries, but it also reproduces experimental NMR observables for peptides and folded proteins, leading to accurate predictions of protein–ligand binding free energies when both the protein and ligand are self-consistently parametrized with espaloma-0.3. We hope this work will lay the foundations to inspire the design of new generations of machine learning-empowered molecular mechanics force fields that can self-consistently describe the wide chemical domains relevant to biomolecular modeling and drug discovery.

13.1 An open chemically and conformationally diverse quantum chemical dataset was curated to construct espaloma-0.3

In this paper, we have curated a high-quality open dataset covering chemical spaces and conformational regions of interest to biomolecular modeling, including small molecules, peptides, and RNA. We demonstrated how our enhanced Espaloma framework can scale to foundational quantum chemical datasets, enabling the achievement of a stable machine-learned MM force field. We released this dataset along with our implementation in the hope that this will enable the community to further optimize MM force fields by building on this dataset, or fine-tuning the espaloma-0.3 model with additional data much the way foundational large language models (LLMs) can be fine-tuned to perform better on domain tasks of interest.

13.2 Espaloma-0.3 quantitatively and qualitatively recapitulates quantum chemical conformational energy landscapes

We demonstrated that current force fields typically exhibit considerable disagreement with quantum chemical calculations in terms of reproducing conformational energies and forces (Table 1). With carefully crafted training and regularization strategies, we show that espaloma-0.3 not only quantitatively agrees more closely with quantum chemical conformational energetics for a wide variety of chemical species, but also behaves qualitatively similarly with quantum chemistry, even in low data regimes (ESI Fig. 1†). Although espaloma-0.3 poses a challenge in preserving the quantum chemical energy minima for some sulfonamide groups (ESI Fig. 4†), more rigorous hyperparameter tuning of the Espaloma framework may help resolve this problem, especially adjusting the weights for each loss component, as we find this to be sensitive to the overall performance.

13.3 Chemical diversity and high-energy conformers are important for accurately capturing quantum chemical energies and forces with Espaloma

The cross-validation experiment (ESI Fig. 2†), in which Espaloma is trained without certain categories of chemical species (small molecules, peptides, or RNA), suggests that quantum chemical datasets with broad chemical coverage—specifically, the SPICE-Pubchem (small molecules) dataset—can perceive and extrapolate the chemical environments for out-of-distributed chemical domains. A lack of chemical diversity leads to large quantum chemical force errors, whereas reproducing energies is easier (ESI Fig. 2(a)†). Similarly, cross-validating certain dataset classes (single-point energies generated by MD [Dataset], optimization trajectories of enumerated conformers [OptimizationDataset], or one-dimensional torsion drives [TorsionDriveDataset]) suggests that high-energy conformers may be important to accurately capture the quantum chemical energies and forces with Espaloma and other machine learning-based methods (ESI Fig. 2(b)†).[117] The quantum chemical forces of peptide datasets, including local energy minima conformers (Pepconf-Opt dataset from [OptimizationDataset]), were poorly reproduced when trained without datasets storing relatively high energy conformers (SPICE-Dipeptide dataset from [Dataset]).

13.4 Espaloma-0.3 can be easily extended to other chemical spaces of interest

The chemical space covered by an Espaloma force field can easily be extended to spaces highly relevant in other areas of biomolecular modeling, such as lipids, DNA, and glycans, by simply augmenting the quantum chemical dataset used in training. In constructing espaloma-0.3, we demonstrated that this approach easily scales to 1.1 million energies and forces, representing nearly 17 000 chemical species, in less than a single GPU-day. Because loss function is easily parallelizable, this approach should scale gracefully to much larger datasets by simply distributing gradient computation across multiple GPUs, enabling rapid parametrization on much larger datasets or extension to new chemical domains of interest.

13.5 Espaloma offers a modular and extensible approach to building MM force fields

Since the Espaloma architecture and loss function are modular[49] and, as demonstrated here, new force fields can be trained in a single GPU-day, Espaloma offers the opportunity to rapidly explore different MM functional forms. For example, many molecular mechanics simulation packages support atom-pair specific 1–4 Lennard-Jones and electrostatic parameters, alternative Lennard-Jones mixing rules, or alternative functional forms for van der Waals treatment. Of particular interest are Class II force fields,[1,2] where higher-order couplings between valence terms are introduced to reproduce the bond and angle vibrations more accurately—while the combinatorial explosion of these terms presents a problem for atom type based force fields, Espaloma does not suffer from the same issue and may provide a robust way to parametrize these force fields.

13.6 Espaloma fit to condensed-phase properties can further improve accuracy

While we have demonstrated the ability to create a force field capable of reproducing NMR observables for peptides and folded proteins, as well as predicting accurate protein–ligand binding free energies solely from fitting to quantum chemical data, further assessment is needed to confirm its ability to accurately reproduce condensed-phase properties. Since non-bonded interactions are generally optimized to fit condensed-phase properties, training against these properties may be necessary to further improve the predictive accuracy of such properties. An earlier study has shown that optimizing against condensed-phase mixture properties, rather than properties of pure systems, is better suited to improve force field accuracy for biomolecular systems.[87] The end-to-end differentiable nature of Espaloma makes it possible to employ reweighting approaches to directly fit to experimental free energies or thermodynamics[118–120] or other thermophysical properties.[87] This could either be done directly during fitting or during a second-stage fine-tuning procedure that adapts an Espaloma force field to specific applications of interest by jointly fitting the valence and non-bonded terms. The challenge of this endeavor lies in the difficulty of analytically taking derivatives of condensed phase properties w.r.t. the force field, and thereby the neural network parameters, in order to constrain them within the physically feasible range.

13.7 Quantifying force field uncertainty could help generate more robust force fields

One of the challenges in force field development is quantifying the contribution of errors in the force field to predicted quantities. While statistical error is generally reported, this systematic force field error is frequently the major source of error in biomolecular simulations. In recent years, several approaches have emerged to quantify uncertainty in deep learning

methods, including mean-variance estimation, Bayesian methods, and ensemble methods.[121,122] Employing these methods to propagate force field uncertainty into predicted free energies and physical properties could enable Espaloma to provide a quantitative assessment of force field uncertainty. With a better understanding of how this uncertainty propagates to task predictions, we envision that uncertainty-based active learning[123] or adversarial attacks[124] could be employed to identify the most valuable new data to be generated in future efforts to train more robust Espaloma force fields.¶

Code availability

The Python code to download the quantum chemical data from QCArchive is available from **https://github.com/choderalab/download-qca-datasets**. The scripts used to train and evaluate espaloma-0.3 is available from **https://github.com/choderalab/refit-espaloma**. The scripts used to perform the small molecule geometry benchmark is available from **https://github.com/choderalab/geometry-benchmark-espaloma**. The curated protein–ligand benchmark dataset can be found from **https://github.com/kntkb/protein-ligand-benchmark-custom**, and the scripts to perform and analyze the alchemical protein–ligand binding affinity calculation with Perses is available from **https://github.com/choderalab/pl-benchmark-espaloma-experiment**. The scripts used to perform the MD simulation of Tyk2 protein–ligand system is available from **https://github.com/choderalab/vanilla-espaloma-experiment**. These python codes are also summarized in **https://github.com/choderalab/espaloma-0.3.0-manuscript**. The code used for the peptide benchmark study is available from **https://github.com/openforcefield/proteinbenchmark**.

Data availability

The raw quantum chemical datasets downloaded from QCArchive is deposited in Zenodo (**https://zenodo.org/record/8148817**). The pre-processed input data used to train espaloma-0.3 is deposited in Zenodo (**https://zenodo.org/record/8150601**). The QM- and MM-minimized structures used for the small molecule geometry benchmark study is deposited in Zenodo (**https://doi.org/10.5281/zenodo.8378216**).

Author contributions

Conceptualization: KT, YW, JDC; methodology: KT, YW; investigation of espaloma RMSE metric: KT, YW; investigation of small molecule geometry benchmark: KT, PKB; investigation of peptide benchmark: CEC; investigation of globular protein benchmark: AJF, CEC, KT; investigation of protein–ligand binding free energy calculation: KT; investigation of protein–ligand standard MD simulation: KT; software: KT, YW, IP, MMH, HM, CRI; writing – original draft: KT, YW; writing – review & editing: KT, IP, PKB, CEC, AJF, MMH, HM, CRI, AMN, AMP, MRS, DLM, JDC, YW; funding acquisition: JDC; resources: JDC; supervision: JDC, YW.

Conflicts of interest

J. D. C. is a current member of the Scientific Advisory Board of OpenEye Scientific Software, Redesign Science, Ventus Therapeutics, and Interline Therapeutics, and has equity interests in Redesign Science and Interline Therapeutics. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, Bayer, XtalPi, Interline Therapeutics, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute. A complete funding history for the Chodera lab can be found at **http://choderalab.org/funding**. Y. W. has limited financial interests in Flagship Pioneering, Inc. and its subsidiaries. M. R. S. is an Open Science Fellow with Psivant Sciences and consults for Relay Therapeutics. D. L. M. serves on the scientific advisory boards of Anagenex and OpenEye Scientific Software, Cadence Molecular Sciences, and is an Open Science Fellow with Psivant.

Acknowledgements

Y. W. acknowledges support from the Schmidt Science Fellowship, in partnership with the Rhodes Trust, and the Simons Center for Computational Physical Chemistry at New York University. J. D. C. acknowledges support from NIH grant P30 CA008748, NIH grant R01 GM132386, NIH grant R01 GM121505, and the Sloan Kettering Institute. M. R. S. acknowledges support from National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. D. L. M. appreciates support from the NIH grant R35GM148236 and R01GM132386. P. K. B. appreciates financial support from the NIH NIGMS R01GM132386. The authors thank OpenEye Scientific Software for providing a free academic license to the OpenEye Toolkits. The authors are also grateful for the OpenFF R01 grant and to all those that provided feedback on versions of the manuscript, including (but not limited to) Michael K. Gilson, Demetri Moustakas, Yutong Zhao, Tristan Croll, Domenico Bonanni, Timothy Bernat, and Mark E. Tuckerman. This research was carried out on high performance computing resources at Memorial Sloan Kettering Cancer Center and the Washington Square and Abu Dhabi campuses of New York University. This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation Accelerate ACCESS BIO230106 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

Notes and references

‡ **https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2021-06-04-OpenFF-Industry-Benchmark-Season-1-v1.1.**
§ **https://github.com/openforcefield/protein-ligand-benchmark/tree/d3387602bbeb0167abf00dfb81753d8936775dd2.**

¶ Implementation, experiment, and dataset details, as well as additional results, are deferred to the ESI, where ref. 125–151 are cited.

1 P. Dauber-Osguthorpe and A. T. Hagler, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 133–203.
2 A. T. Hagler, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 205–264.
3 A. R. Leach, *Molecular modelling: principles and applications*, Pearson education, 2001.
4 T. Schlick, *Molecular modeling and simulation: an interdisciplinary guide*, Springer, 2010, vol. 2.
5 E. A. Coutsias, K. W. Lexa, M. J. Wester, S. N. Pollock and M. P. Jacobson, *J. Chem. Theory Comput.*, 2016, **12**, 4674–4687.
6 B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin, *et al.*, *Nat. Protoc.*, 2021, **16**, 4799–4832.
7 C. H. Tse, J. Comer, S. K. Sang Chu, Y. Wang and C. Chipot, *J. Chem. Theory Comput.*, 2019, **15**, 2913–2924.
8 S. Prasad, D. Mobley, E. Braun, H. Mayes, J. Monroe, D. Zuckerman, *et al.*, *Living Journal of Computational Molecular Science*, 2018, **1**, 1–28.
9 A. S. Mey, B. K. Allen, H. E. B. Macdonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, *et al.*, *Living Journal of Computational Molecular Science*, 2020, **2**, year.
10 J. Delhommelle and P. Millié, *Mol. Phys.*, 2001, **99**, 619–625.
11 M. J. Harvey, G. Giupponi and G. D. Fabritiis, *J. Chem. Theory Comput.*, 2009, **5**, 1632–1639.
12 R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2013, **9**, 3878–3888.
13 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, *et al.*, *PLoS Comput. Biol.*, 2017, **13**, e1005659.
14 L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, *et al.*, *J. Am. Chem. Soc.*, 2015, **137**, 2695–2703.
15 C. E. Schindler, H. Baumann, A. Blum, D. Bose, H.-P. Buchstaller, L. Burgdorf, D. Cappel, E. Chekler, P. Czodrowski, D. Dorsch, *et al.*, *J. Chem. Inf. Model.*, 2020, **60**, 5457–5474.
16 V. Gapsys, D. F. Hahn, G. Tresadern, D. L. Mobley, M. Rampp and B. L. de Groot, *J. Chem. Inf. Model.*, 2022, **62**, 1172–1177.
17 D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, *et al.*, *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.
18 R. M. Betz and R. C. Walker, *J. Comput. Chem.*, 2015, **36**, 79–87.
19 E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, *et al.*, *J. Chem. Theory Comput.*, 2016, **12**, 281–296.
20 J. T. Horton, S. Boothroyd, J. Wagner, J. A. Mitchell, T. Gokey, D. L. Dotson, P. K. Behara, V. K. Ramaswamy, M. Mackey, J. D. Chodera, J. Anwar, D. L. Mobley and D. J. Cole, *J. Chem. Inf. Model.*, 2022, **62**, 5622–5633.
21 D. Case, H. Aktulga, K. Belfon, I. Ben-Shalom, J. Berryman, S. Brozell, D. Cerutti, T. Cheatham III, G. Cisneros, V. Cruzeiro, T. Darden, N. Forouzesh, G. Giambasu, T. Giese, M. Gilson, H. Gohlke, A. Goetz, J. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, J. Wang, H. Wei, X. Wu, Y. Wu, Y. Xiong, Y. Xue, D. York, S. Zhao, Q. Zhu and P. A. Kollman, *Amber 2023*, 2023.
22 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
23 M. Zgarbová, J. Sponer, M. Otyepka, T. E. Cheatham III, R. Galindo-Murillo and P. Jurecka, *J. Chem. Theory Comput.*, 2015, **11**, 5723–5736.
24 R. Galindo-Murillo, J. C. Robertson, M. Zgarbová, J. Šponer, M. Otyepka, P. Jurečka and T. E. Cheatham III, *J. Chem. Theory Comput.*, 2016, **12**, 4114–4127.
25 M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. Cheatham III and P. Jurečka, *J. Chem. Theory Comput.*, 2011, **7**, 2886–2902.
26 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
27 H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, 2004, **120**, 9665–9678.
28 S. Izadi, R. Anandakrishnan and A. V. Onufriev, *J. Phys. Chem. Lett.*, 2014, **5**, 3863–3871.
29 I. S. Joung and T. E. Cheatham III, *J. Phys. Chem. B*, 2008, **112**, 9020–9041.
30 I. S. Joung and T. E. Cheatham III, *J. Phys. Chem. B*, 2009, **113**, 13279–13290.
31 P. Li, B. P. Roberts, D. K. Chakravorty and K. M. Merz Jr, *J. Chem. Theory Comput.*, 2013, **9**, 2733–2748.
32 P. Li and K. M. Merz Jr, *J. Chem. Theory Comput.*, 2014, **10**, 289–297.
33 P. Li, L. F. Song and K. M. Merz Jr, *J. Phys. Chem. B*, 2015, **119**, 883–895.
34 C. J. Dickson, R. C. Walker and I. R. Gould, *J. Chem. Theory Comput.*, 2022, **18**(3), 1726–1736.
35 K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley and R. J. Woods, *J. Comput. Chem.*, 2008, **29**, 622–655.
36 M. L. DeMarco and R. J. Woods, *Glycobiology*, 2009, **19**, 344–355.
37 M. L. DeMarco, R. J. Woods, J. H. Prestegard and F. Tian, *J. Am. Chem. Soc.*, 2010, **132**, 1334–1338.
38 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

39 J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J. Mol. Graphics Modell.*, 2006, **25**, 247–260.
40 G. A. Khoury, J. P. Thompson, J. Smadbeck, C. A. Kieslich and C. A. Floudas, *J. Chem. Theory Comput.*, 2013, **9**, 5653–5674.
41 R. Aduri, B. T. Psciuk, P. Saro, H. Taniga, H. B. Schlegel and J. SantaLucia, *J. Chem. Theory Comput.*, 2007, **3**, 1464–1475.
42 L.-P. Wang, T. J. Martinez and V. S. Pande, *J. Phys. Chem. Lett.*, 2014, **5**, 1885–1891.
43 L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez and V. S. Pande, *J. Phys. Chem. B*, 2017, **121**, 4023–4039.
44 Y. Qiu, D. G. Smith, S. Boothroyd, H. Jang, D. F. Hahn, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. D. Stern, *et al.*, *J. Chem. Theory Comput.*, 2021, **17**, 6262–6280.
45 S. Boothroyd, P. K. Behara, O. C. Madin, D. F. Hahn, H. Jang, V. Gapsys, J. R. Wagner, J. T. Horton, D. L. Dotson, M. W. Thompson, *et al.*, *J. Chem. Theory Comput.*, 2023, **19**, 3251–3275.
46 S. Boothroyd, L.-P. Wang, D. L. Mobley, J. D. Chodera and M. R. Shirts, *J. Chem. Theory Comput.*, 2022, **18**, 3566–3576.
47 B. J. Befort, R. S. DeFever, G. M. Tow, A. W. Dowling and E. J. Maginn, *J. Chem. Inf. Model.*, 2021, **61**, 4400–4414.
48 X. Wang, J. Li, L. Yang, F. Chen, Y. Wang, J. Chang, J. Chen, W. Feng, L. Zhang and K. Yu, *J. Chem. Theory Comput.*, 2023, **19**, 5897–5909.
49 Y. Wang, J. Fass, B. Kaminow, J. E. Herr, D. Rufa, I. Zhang, I. Pulido, M. Henry, H. E. Bruce Macdonald, K. Takaba and J. D. Chodera, *Chem. Sci.*, 2022, **13**, 12016–12033.
50 M. Thurlemann, L. Boselt and S. Riniker, *J. Chem. Theory Comput.*, 2023, **19**, 562–579.
51 Y. Wang, PhD thesis, Cornell University, 2023.
52 Y. Wang and T. Karaletsos, *Stochastic Aggregation in Graph Neural Networks*, 2021.
53 K. Xu, W. Hu, J. Leskovec and S. Jegelka, *arXiv*, 2018, preprint, arXiv:181000826, DOI: **10.48550/arXiv.1810.00826**.
54 R. Murphy, B. Srinivasan, V. Rao, B. Ribeiro, *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 2019, vol. 97, pp. 4663–4673.
55 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2000, **21**, 132–146.
56 A. Jakalian, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.
57 *RDKit: open-source cheminformatics*, 2013, **http://www.rdkit.org**, accessed 11-April-2013.
58 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, *Advances in neural information processing systems*, 2015, pp. 2224–2232.
59 T. N. Kipf, M. Welling, *arXiv*, 2016, preprint, arXiv:1609.02907, DOI: **10.48550/arXiv.1609.02907**.
60 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *International conference on machine learning*, 2017, pp. 1263–1272.
61 P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, *arXiv*, 2018, preprint, arXiv:180601261, DOI: **10.48550/arXiv.1806.01261**.
62 J. Du, S. Zhang, G. Wu, J. M. F. Moura and S. Kar, *arXiv*, 2018, preprint, arXiv:171010370 [cs, stat], DOI: **10.48550/arXiv.1710.10370**.
63 F. Wu, T. Zhang, A. H. d. Souza Jr, C. Fifty, T. Yu and K. Q. Weinberger, *arXiv*, 2019, preprint, arXiv:1902.07153, DOI: **10.48550/arXiv.1902.07153**.
64 M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, *et al.*, *arXiv*, 2019, preprint, arXiv:190901315, DOI: **10.48550/arXiv.1909.01315**.
65 Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein and J. M. Solomon, *ACM Trans. Graph.*, 2019, **38**, 1–12.
66 C. K. Joshi, C. Bodnar, S. V. Mathis, T. Cohen and P. Lió, *arXiv*, 2023, preprint, arXiv:230109308, DOI: **10.48550/arXiv.2301.09308**.
67 M. K. Gilson, H. S. Gilson and M. J. Potter, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1982–1997.
68 Y. Wang, J. Fass, C. D. Stern, K. Luo, and J. Chodera, *arXiv*, 2019, preprint, arXiv:190907903, DOI: **10.48550/arXiv.1909.07903**.
69 Y. Wang, I. Pulido, K. Takaba, B. Kaminow, J. Scheen, L. Wang and J. D. Chodera, *arXiv*, 2023, preprint, arXiv:230206758, DOI: **10.48550/arXiv.2302.06758**.
70 D. G. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard and T. D. Crawford, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1491.
71 X. He, V. H. Man, W. Yang, T.-S. Lee and J. Wang, *J. Chem. Phys.*, 2020, **153**, 114502.
72 J. Wagner, M. Thompson, D. Dotson, S. B. Hyejang and J. Rodríguez-Guerra, *openforcefield/openff-forcefields: Version 2.0.0 "Sage" (2.0.0)*, Zenodo, 2021, DOI: **10.5281/zenodo.5214478**.
73 P. K. Behara, T. Gokey, C. Cavender, J. Horton, L. Wang, H. Jang, J. Wagner, D. Cole, C. Bayly and D. Mobley, *openforcefield/openff-forcefields (2023.05.1)*, Zenodo, 2023, DOI: **10.5281/zenodo.7889050**.
74 P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, *et al.*, *Sci. Data*, 2023, **10**, 11.
75 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, *Nucleic Acids Res.*, 2023, **51**, D1373–D1380.
76 A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, *et al.*, *Sci. Data*, 2021, **8**, 55.
77 V. K. Prasad, A. Otero-de La-Roza and G. A. DiLabio, *Sci. Data*, 2019, **6**, 1–9.
78 Q. Li, T. Cheng, Y. Wang and S. H. Bryant, *Drug Discov. Today*, 2010, **15**, 1052–1057.
79 C. E. Cavender, P. K. Behara, S. Boothroyd, D. L. Dotson, J. T. Horton, J. A. Mitchell, I. J. Pulido, M. W. Thompson, J. Wagner, L. Wang, J. D. Chodera, D. J. Cole, D. L. Mobley, M. R. Shirts and M. K. Gilson, *Development and benchmarking of an open, self-consistent force field for*

proteins and small molecules from the Open Force Field Initiative, Zenodo, 2023, DOI: 10.5281/zenodo.7696579.
80 B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A. I. Petrov, B. Sweeney, C. L. Zirbel, N. B. Leontis and H. M. Berman, *Nucleic Acids Res.*, 2014, **42**, D114–D122.
81 L. G. Parlea, B. A. Sweeney, M. Hosseini-Asanjan, C. L. Zirbel and N. B. Leontis, *Methods*, 2016, **103**, 99–119.
82 P. K. Behara, H. Jang, J. Horton, D. Dotson, S. Boothroyd, C. Cavender, V. Gapsys, T. Gokey, D. Hahn, J. Maat, O. Madin, I. Pulido, M. Thompson, J. Wagner, L. Wang, J. Chodera, D. Cole, M. Gilson, M. Shirts, C. Bayly, L.-P. Wang and D. Mobley, *Benchmarking QM theory for drug-like molecules to train force fields, OpenEye CUP XII*, Zenodo, Santa Fe, NM, 2022, DOI: 10.5281/zenodo.7548777.
83 D. G. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, *et al.*, *J. Chem. Phys.*, 2020, **152**(18), 184108.
84 J. Horton, *openforcefield/openff-qcsubmit: 0.3.1 (0.3.1)*, Zenodo, 2022, DOI: 10.5281/zenodo.6338096.
85 C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu and C. Simmerling, *J. Chem. Theory Comput.*, 2020, **16**, 528–552.
86 E. Boulanger, L. Huang, C. Rupakheti, A. D. MacKerell Jr and B. Roux, *J. Chem. Theory Comput.*, 2018, **14**, 3121–3131.
87 S. Boothroyd, O. C. Madin, D. L. Mobley, L.-P. Wang, J. D. Chodera and M. R. Shirts, *J. Chem. Inf. Model.*, 2022, **18**, 3577–3592.
88 H. Nguyen, J. Maier, H. Huang, V. Perrone and C. Simmerling, *J. Am. Chem. Soc.*, 2014, **136**, 13959–13962.
89 L. D'Amore, D. F. Hahn, D. L. Dotson, J. T. Horton, J. Anwar, I. Craig, T. Fox, A. Gobbi, S. K. Lakkaraju, X. Lucas, K. Meier, D. L. Mobley, A. Narayanan, C. E. M. Shindler, W. C. Swope, P. J. in 't Veld, J. Wagner, B. Xue and G. Tresadern, *J. Chem. Inf. Model.*, 2022, **62**, 6094–6104.
90 V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly and D. L. Mobley, *F1000Research*, 2020, **9**, 1390.
91 J. Graf, P. H. Nguyen, G. Stock and H. Schwalbe, *J. Am. Chem. Soc.*, 2007, **129**, 1179–1189.
92 A. Hagarman, T. J. Measey, D. Mathieu, H. Schwalbe and R. Schweitzer-Stenner, *J. Am. Chem. Soc.*, 2010, **132**, 540–551.
93 A. Pardi, M. Billeter and K. Wüthrich, *J. Mol. Biol.*, 1984, **180**, 741–751.
94 K. D. Berndt, P. Güntert, L. P. Orbons and K. Wüthrich, *J. Mol. Biol.*, 1992, **227**, 757–775.
95 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950–1958.
96 J. J. Chou, D. A. Case and A. Bax, *J. Am. Chem. Soc.*, 2003, **125**, 8959–8966.
97 E. Miclet, J. Boisbouvier and A. Bax, *J. Biomol. NMR*, 2005, **31**, 201–216.
98 B. Vögeli, J. Ying, A. Grishaev and A. Bax, *J. Am. Chem. Soc.*, 2007, **129**, 9377–9385.
99 G. Cornilescu, B. E. Ramirez, M. K. Frank, G. M. Clore, A. M. Gronenborn and A. Bax, *J. Am. Chem. Soc.*, 1999, **121**, 6275–6279.
100 H. Schwalbe, S. B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C. M. Dobson, C. Redfield and L. J. Smith, *Protein Sci.*, 2001, **10**, 677–688.
101 A. C. Wang and A. Bax, *J. Am. Chem. Soc.*, 1996, **118**, 2483–2494.
102 J.-S. Hu and A. Bax, *J. Am. Chem. Soc.*, 1997, **119**, 6360–6368.
103 F. Cordier and S. Grzesiek, *J. Am. Chem. Soc.*, 1999, **121**, 1601–1602.
104 M. Karplus, *J. Am. Chem. Soc.*, 1963, **85**, 2870–2871.
105 M. Hennig, W. Bermel, H. Schwalbe and C. Griesinger, *J. Am. Chem. Soc.*, 2000, **122**, 6268–6277.
106 J. Wirmer and H. Schwalbe, *J. Biomol. NMR*, 2002, **23**, 47–55.
107 K. Ding and A. M. Gronenborn, *J. Am. Chem. Soc.*, 2004, **126**, 6232–6233.
108 C. Pérez, F. Löhr, H. Rüterjans and J. M. Schmidt, *J. Am. Chem. Soc.*, 2001, **123**, 7081–7093.
109 M. Barfield, *J. Am. Chem. Soc.*, 2002, **124**, 4158–4168.
110 D. A. Rufa, I. Zhang, H. E. Bruce Macdonald, P. B. Grinaway, I. Pulido, M. M. Henry, J. Rodríguez-Guerra, M. Wittmann, S. K. Albanese, W. G. Glass, A. Silveira, D. Schaller, L. N. Naden and J. D. Chodera, *Perses (0.10.1)*, Zenodo, 2022, DOI: 10.5281/zenodo.6757402.
111 P. Eastman, R. Galvelis, R. P. Peláez, C. R. Abreu, S. E. Farr, E. Gallicchio, A. Gorenko, M. M. Henry, F. Hu, J. Huang, *et al.*, *J. Phys. Chem. B*, 2024, **128**, 109–116.
112 J. Liang, V. Tsui, A. Van Abbema, L. Bao, K. Barrett, M. Beresini, L. Berezhkovskiy, W. S. Blair, C. Chang, J. Driscoll, *et al.*, *Eur. J. Med. Chem.*, 2013, **67**, 175–187.
113 T. G. Davies, J. Bentley, C. E. Arris, F. T. Boyle, N. J. Curtin, J. A. Endicott, A. E. Gibson, B. T. Golding, R. J. Griffin, I. R. Hardcastle, *et al.*, *Nat. Struct. Biol.*, 2002, **9**, 745–749.
114 P. Labute and M. Ebert, *Free Energy Methods in Drug Discovery: Current State and Future Directions*, ACS Publications, 2021, pp. 227–245.
115 A. Friberg, D. Vigil, B. Zhao, R. N. Daniels, J. P. Burke, P. M. Garcia-Barrantes, D. Camper, B. A. Chauder, T. Lee, E. T. Olejniczak, *et al.*, *J. Med. Chem.*, 2013, **56**, 15–30.
116 X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli and T. Jaakkola, 2022, preprint, arXiv:221007237, DOI: 10.48550/arXiv.2210.07237.
117 Y. Wang, C. Xu, Z. Li and A. B. Farimani, *J. Chem. Theory Comput.*, 2023, **19**, 5077–5087.
118 M. Wieder, J. Fass and J. D. Chodera, *Chem. Sci.*, 2021, **12**, 11364–11381.
119 M. Wieder, J. Fass and J. D. Chodera, *bioRxiv*, 2021, preprint, DOI: 10.1101/2021.08.24.457513.
120 J. Setiadi, S. Boothroyd, D. Slochower, D. Dotson, M. Thompson, J. Wagner, L.-P. Wang and M. K. Gilson, *J. Chem. Theory Comput.*, 2024, **20**(1), 239–252.
121 L. Hirschfeld, K. Swanson, K. Yang, B. Regina and C. W. Coley, *J. Chem. Inf. Model.*, 2020, **60**, 3770–3780.
122 J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler and X. X. Zhu, *arXiv*,

2022, preprint, arXiv:210703342, DOI: 10.48550/arXiv.2107.03342.
123 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, *J. Chem. Phys.*, 2018, **148**, 241733.
124 D. Schwalbe-Koda, A. R. Tan and R. Gómez-Bombarelli, *Nat. Commun.*, 2021, **12**, 5104.
125 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.
126 J. Wagner, M. Thompson, D. L. Mobley, J. Chodera, C. Bannan, A. Rizzi, D. Dotson, J. Rodríguez-Guerra, J. A. Mitchell, *et al.*, *openforcefield/openff-toolkit: 0.10.6 Bugfix release (0.10.6)*, Zenodo, 2022, DOI: 10.5281/zenodo.6483648.
127 J. Chodera, R. Wiewiora, C. Stern and P. Eastman, *openmm/openmm-forcefields: Fix GAFF AM1-BCC charging bug for some molecules (0.7.1), Zenodo*, 2020, 10.5281.
128 H. B. Macdonald, D. F. Hahn, M. Henry, J. Chodera, D. Dotson, W. Glass, I. Pulido, *openforcefield/openff-arsenic: v0.2.1 (0.2.1)*, Zenodo, 2022, DOI: 10.5281/zenodo.6210305.
129 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.
130 G. Landrum, P. Tosco, B. Kelleyand, D. Ric, R. sriniker, N. Schneider, E. Kawashima, D. N, G. Jones, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, K. Ujihara, V. F. Scalfani, G. Godin, J. Lehtivarjo, A. Pahl, R. Walker and F. Berenger, jasondbiggs and strets123, rdkit/rdkit: 2023_03_2 (Q1 2023) Release (Release_2023_03_2), *Zenodo*, 2023, DOI: 10.5281/zenodo.8053810.
131 W. Hamilton, Z. Ying and J. Leskovec, *Adv. Neural Inf. Process. Syst.*, 2017, 1024–1034.
132 D. P. Kingma and J. Ba, *arXiv*, 2014, preprint, arXiv:14126980, DOI: 10.48550/arXiv.1412.6980.
133 T. Schulz-Gasch, C. Schärfer, W. Guba and M. Rarey, *J. Chem. Inf. Model.*, 2012, **52**, 1499–1512.
134 V. Gapsys, S. Michielssens, D. Seeliger and B. L. De Groot, *J. Comput. Chem.*, 2015, **19**, 348–354.
135 C. R. Søndergaard, M. H. Olsson, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 2284–2295.
136 E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, C. Minju, P. Li, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M. J. Holst, J. A. McCammon and N. A. Baker, *Protein Sci.*, 2018, **27**, 112–128.
137 P. S. Nerenberg and T. Head-Gordon, *J. Chem. Theory Comput.*, 2011, **7**, 1220–1230.
138 Z. Zhang, X. Liu, K. Yan, M. E. Tuckerman and J. Liu, *J. Phys. Chem. A*, 2019, **123**, 6056–6079.
139 M. Bernetti and G. Bussi, *J. Chem. Phys.*, 2020, **153**, 114107.
140 C. W. Hopkins, S. L. Grand, R. C. Walker and A. E. Roitberg, *J. Chem. Theory Comput.*, 2015, **11**, 1864–1874.
141 T. D. Romo, N. Leioatts and A. Grossfield, *J. Comput. Chem.*, 2014, **35**, 2305–2318.
142 P. Bauer, B. Hess and E. Lindahl, *GROMACS 2022.5 Manual (2022.5)*, Zenodo, 2023, DOI: 10.5281/zenodo.7586765.
143 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
144 J. Chodera, A. Rizzi, L. Naden, K. Beauchamp, P. Grinaway, J. Fass, A. Wade, I. Pulido, M. Henry, G. A. Ross, A. Krämer, H. B. Macdonald, J. Rodríguez-Guerra Pedregal, *et al.*, *choderalab/openmmtools: 0.22.1 (0.22.1)*, Zenodo, 2023, DOI: 10.5281/zenodo.7843902.
145 J. D. Chodera and M. R. Shirts, *J. Chem. Phys.*, 2011, **135**, 194110.
146 B. Leimkuhler and C. Matthews, *Proc. R. Soc. A*, 2016, **472**, 20160138.
147 M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.
148 H. Xu, *J. Chem. Inf. Model.*, 2019, **59**, 4720–4728.
149 P. A. Janowski, C. Liu, J. Deckman and D. A. Case, *Protein Sci.*, 2015, **25**, 87–102.
150 Z. Sun, Q. Liu, G. Qu, Y. Feng and M. T. Reetz, *Chem. Rev.*, 2019, **119**, 1626–1665.
151 L. Wickstrom, A. Okur and C. Simmerling, *Biophys. J.*, 2009, **97**, 853–856.