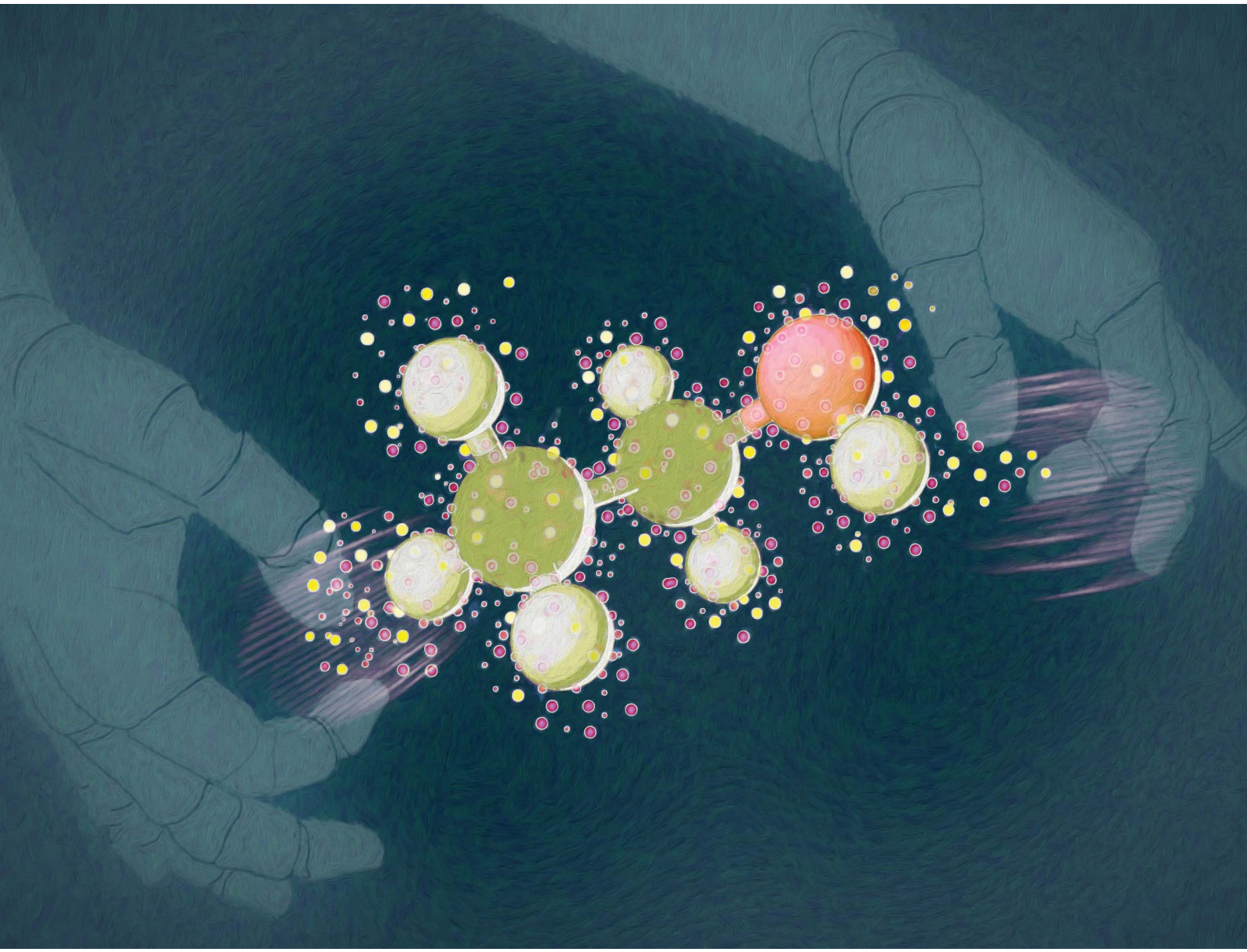
Volume 13
Number 41
7 November 2022
Pages 11953–12246

Chemical Science

rsc.li/chemical-science



ISSN 2041-6539

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE
Yuanqing Wang, John D. Chodera *et al.*
End-to-end differentiable construction of molecular mechanics force fields

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2022, **13**, 12016

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

End-to-end differentiable construction of molecular mechanics force fields†

Yuanqing Wang, *[acf] Josh Fass, ‡[ab] Benjamin Kaminow, [ab] John E. Herr, [a] Dominic Rufa, [ad] Ivy Zhang, [ab] Iván Pulido, [a] Mike Henry, [a] Hannah E. Bruce Macdonald, §[a] Kenichiro Takaba [ae] and John D. Chodera *[a]

Molecular mechanics (MM) potentials have long been a workhorse of computational chemistry. Leveraging accuracy and speed, these functional forms find use in a wide variety of applications in biomolecular modeling and drug discovery, from rapid virtual screening to detailed free energy calculations. Traditionally, MM potentials have relied on human-curated, inflexible, and poorly extensible discrete chemical perception rules (atom types) for applying parameters to small molecules or biopolymers, making it difficult to optimize both types and parameters to fit quantum chemical or physical property data. Here, we propose an alternative approach that uses graph neural networks to perceive chemical environments, producing continuous atom embeddings from which valence and nonbonded parameters can be predicted using invariance-preserving layers. Since all stages are built from smooth neural functions, the entire process—spanning chemical perception to parameter assignment—is modular and end-to-end differentiable with respect to model parameters, allowing new force fields to be easily constructed, extended, and applied to arbitrary molecules. We show that this approach is not only sufficiently expressive to reproduce legacy atom types, but that it can learn to accurately reproduce and extend existing molecular mechanics force fields. Trained with arbitrary loss functions, it can construct entirely new force fields self-consistently applicable to both biopolymers and small molecules directly from quantum chemical calculations, with superior fidelity than traditional atom or parameter typing schemes. When adapted to simultaneously fit partial charge models, espaloma delivers high-quality partial atomic charges orders of magnitude faster than current best-practices with low inaccuracy. When trained on the same quantum chemical small molecule dataset used to parameterize the Open Force Field ("Parsley") openff-1.2.0 small molecule force field augmented with a peptide dataset, the resulting espaloma model shows superior accuracy *vis-á-vis* experiments in computing relative alchemical free energy calculations for a popular benchmark. This approach is implemented in the free and open source package espaloma, available at https://github.com/choderalab/espaloma.

Received 16th May 2022
Accepted 5th September 2022

DOI: 10.1039/d2sc02739a

rsc.li/chemical-science

Molecular mechanics (MM) force fields—physical models that abstract molecular systems as atomic point masses that interact *via* nonbonded interactions and valence (bond, angle, torsion) terms—have powered *in silico* modeling to provide key insights and quantitative predictions in all aspects of chemistry, from drug discovery to materials science.[1–9] While recent work in quantum machine learning (QML) potentials has demonstrated how flexibility in functional forms and training strategies can lead to increased accuracy,[10–16] these QML potentials are orders of magnitude slower than popular molecular mechanics potentials even on expensive hardware accelerators, as they involve orders of magnitude more floating point operations per energy or force evaluation.

On the other hand, the simpler physical energy functions of MM models are compatible with highly optimized implementations that can exploit a wide variety of hardware,[2,17–21] but

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

[b]*Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, Cornell University, New York 10065, NY, USA*

[c]*Physiology, Biophysics and System Biology PhD Program, Weill Cornell Medical College, Cornell University, New York 10065, NY, USA*

[d]*Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Medical College, Cornell University, New York 10065, NY, USA*

[e]*Pharmaceutical Research Center, Advanced Drug Discovery, Asahi Kasei Pharma Corporation, Shizuoka 410-2321, Japan*

[f]*MFA Program in Creative Writing, Division of Humanities and Arts, City College of New York, City University of New York, New York 10031, NY, USA*

† Electronic supplementary information (ESI) available. See https://doi.org/10.1039/d2sc02739a

‡ Present address: Relay Therapeutics, Cambridge, MA 02139, USA.

§ Present address: MSD R&D Innovation Centre,120 Moorgate, London, EC2M 6UR, UK.

rely on complex and inextensible legacy atom typing schemes for parameter assignment:[22]

- First, a set of rules is used to classify atoms into discrete atom types that must encode all information about an atom's chemical environment needed in subsequent parameter assignment steps.
- Next, a discrete set of bond, angle, and torsion types is determined by composing the atom types involved in the interaction.
- Finally, the parameters attached to atoms, bonds, angles, and torsions are assigned according to a look-up table of composed parameter types.

Consequently, atoms, bonds, angles, or torsions with distinct chemical environments that happen to fall into the same expert-derived discrete type class are forced to share the same set of MM parameters, potentially leading to low resolution and poor accuracy. Furthermore, the explosion of the number of discrete parameter classes describing equivalent chemical environments required by traditional MM atom typing schemes not only poses significant challenges to extending the space of atom types,[22] but optimizing these independently has the potential to compromise generalizability and lead to overfitting. Even with modern MM parameter optimization frameworks[23–25] and sufficient data, parameter optimization is only feasible in the continuous parameter space defined by these fixed atom types, while the mixed discrete-continuous optimization problem—jointly optimizing types and parameters—is intractable.

Here, we present a continuous alternative to traditional discrete atom typing schemes that permits full end-to-end differentiable optimization of both typing and parameter assignment stages, allowing an entire force field to be built, extended, and applied using standard automatic differentiation machine learning frameworks such as PyTorch,[32] TensorFlow,[33] or JAX[34] (Fig. 1). Graph neural networks have recently emerged as a powerful way to learn chemical properties of atoms, bonds, and molecules for biomolecular species (both small organic molecules and biopolymers), which can be considered isomorphic with their graph representations.[35–44] We hypothesize that graph neural networks operating on molecules have expressiveness that is at least equivalent to—and likely much greater than—expert-derived typing rules, with the advantage of being able to smoothly interpolate between representations of chemical environments (such as accounting for fractional bond orders[45]). We provide empirical evidence for this in Section 1.1.

Next, we demonstrate the utility of such a model (which we call the extensible surrogate potential optimized by message-passing, or espaloma) to construct end-to-end optimizable force fields with continuous atom types. Espaloma assigns molecular mechanics parameters from a molecular graph in three differentiable stages (Fig. 1):

- Stage 1: continuous atom embeddings are constructed using graph neural networks to perceive chemical environments (Section 1.1).
- Stage 2: continuous bond, angle, and torsion embeddings are constructed using pooling that preserves appropriate symmetries (Section 1.2).
- Stage 3: molecular mechanics force field parameters are computed from atom, bond, angle, and torsion embeddings using feed-forward networks (Section 1.3).

Additional molecular mechanics parameter classes (such as point polarizabilities, valence coupling terms, or even parameters for charge-transfer models[46]) can easily be added in a modular manner.

Similar to legacy molecular mechanics parameter assignment infrastructures, molecular mechanics parameters are assigned once for each system, and can be subsequently used to compute energies and forces or carry out molecular simulations with standard molecular mechanics packages. Unlike traditional legacy force fields, espaloma model parameters Φ_{NN}—which define the entire process by which molecular mechanics force field parameters Φ_{FF} are generated ad hoc for a given molecule—can easily be fit to data from scratch using standard, highly portable, high-performance machine learning frameworks that support automatic differentiation.

Here, we demonstrate that espaloma provides a sufficiently flexible representation to both learn to apply existing MM force fields and to generalize them to new molecules (Section 2). Espaloma's modular loss function enables force fields to be learned directly from quantum chemical energies (Section 3), partial charges (Section 4), or both. The resulting force fields can generate self-consistent parameters for small molecules, biopolymers (Section 5), and covalent adducts (Section 1). Finally, an espaloma model fit to the same quantum chemical dataset used to build the Open Force Field OpenFF ("Parsley") openff-1.2.0 small molecule force field, augmented with peptide quantum chemical data, can outperform it in an out-of-sample kinase : inhibitor alchemical free energy benchmark (Section A.4 in ESI†).

1 Espaloma: end-to-end differentiable MM parameter assignment

First, we describe how our proposed framework, espaloma (Fig. 1), operates analogously to legacy force field typing schemes to generate MM parameters Φ_{FF} from a molecular graph $\mathcal{G}$ and neural model parameters Φ_{NN},

$$\Phi_{\mathrm{FF}} \leftarrow \mathrm{espaloma}(\mathcal{G}, \Phi_{\mathrm{NN}}), \tag{1}$$

which can subsequently be used to compute the MM energy (as in eqn (14) in ESI†) for any conformation. A brief graph-theoretic overview of molecular mechanics force fields is provided in the Appendix (Section C in ESI†).

1.1 Stage 1: graph neural networks generate a continuous atom embedding, replacing legacy discrete atom typing

We propose to use graph neural networks[35–44] as a replacement for rule-based chemical environment perception.[22] These neural architectures learn useful representations of atomic chemical environments from simple input features by updating and pooling embedding vectors *via* message passing iterations to neighboring atoms.[44] As such, symmetries in chemical graphs

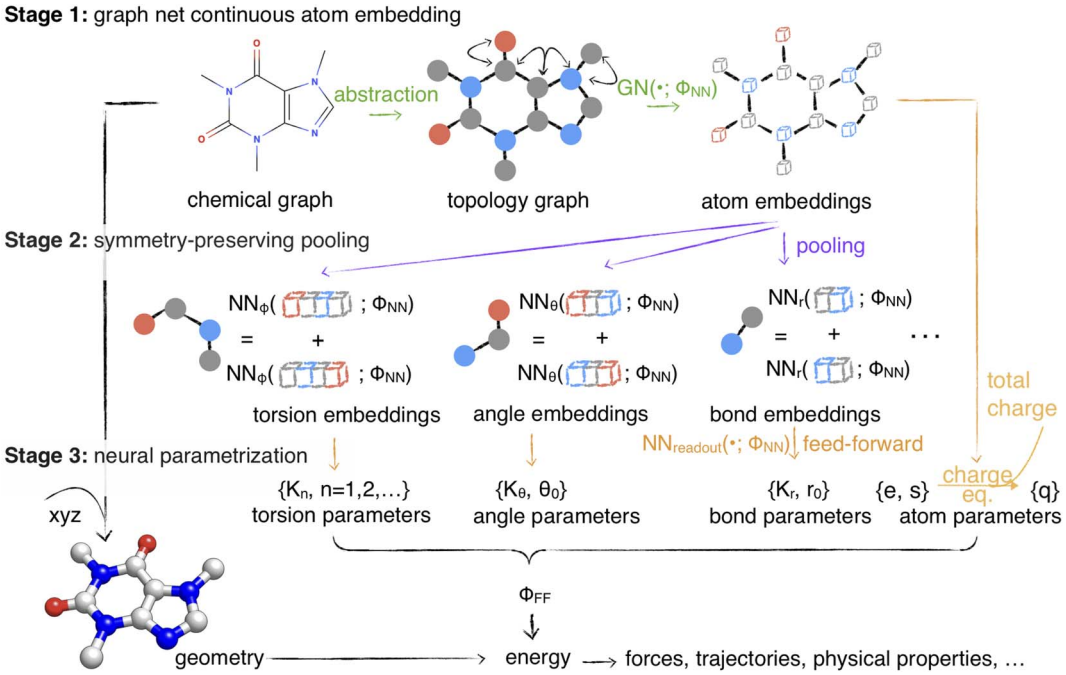


Fig. 1 Espaloma is an end-to-end differentiable molecular mechanics parameter assignment scheme for arbitrary organic molecules. Espaloma (extendable surrogate potential optimized by message-passing) is a modular approach for directly computing molecular mechanics force field parameters Φ_{FF} from a 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 (Section 1.1). 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 (Section 1.2). In Stage 3, molecular mechanics parameters are directly predicted from these feature vectors using feed-forward neural nets (Section 1.3). 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[26,27]/AM1-BCC[28,29] in that it encodes the equivalent of traditional typing rules, parameter assignment tables, and even partial charge models. This final stage is modular, and can be easily extended to incorporate additional molecular mechanics parameter classes, such as parameters for a charge-equilibration model (Section 4), point polarizabilities, or valence-coupling terms for class II molecular mechanics force fields.[30,31]

(chemical equivalencies) are inherently preserved, while a rich, continuous, and differentiably learnable representation of the atomic environment is derived. For a brief introduction to graph neural networks, see Appendix Section D in ESI †.

Traditional molecular mechanics force field parameter assignment schemes such as Antechamber/GAFF[26,27] or CGenFF[47,48] use attributes of atoms and their neighbors (such as atomic number, hybridization, and aromaticity) to assign discrete atom types. Subsequently, atom, bond, angle, and torsion parameters are assigned for specific combinations of these discrete types according to human chemical intuition.[22] On a closer look, this scheme resembles a two- or three-step Weisfeiler–Leman test,[49] which has been shown to be well approximated by some graph neural network architectures.[35] We hypothesize that graph neural network architectures can be at least as expressive as legacy atom typing rules.

To compute continuous atom embeddings, we start with a molecular graph $\mathcal{G}$ whose atoms (nodes) are labeled with simple chemical properties (here, we consider element, hybridization, aromaticity, formal charge, and membership in various ring sizes) easily computed in any cheminformatics toolkit. Sequential application of the graph neural network message-passing update rules (Appendix Section D in ESI†) then computes an updated set of atom (node) features in each graph neural network layer, and the final atom embeddings $h_v \in \mathbb{R}^{|\mathcal{G}| \times D}$ are extracted from the final layer. The loss on training data is then minimized by minimizing the cross-entropy loss between predicted and reference types.¶

We use a subset of ZINC[50] provided with parm@Frosst to validate atom typing implementations[51] (7529 small drug-like molecules, partitioned 80 : 10 : 10 into training : validation : test sets) for this experiment. Reference GAFF 1.81 [27] atom types are assigned using Antechamber[27] from AmberTools and are used for training and testing.

1.1.1 Graph neural networks can reproduce legacy atom types with high accuracy. The test set performance is reported in Fig. 2, where the overall accuracy between reference legacy types and learned types is very high—$98.31\%^{98.63\%}_{97.94\%}$, where sub- and superscripts represent a 95% confidence interval. In analyzing the infrequent failures, we find the model assigns atom types that correspond to the reference type more often when the atom type appears more frequently in the training data, whereas the discrepancies occur in assigning rare types and types whose definitions follow a more sophisticated (but

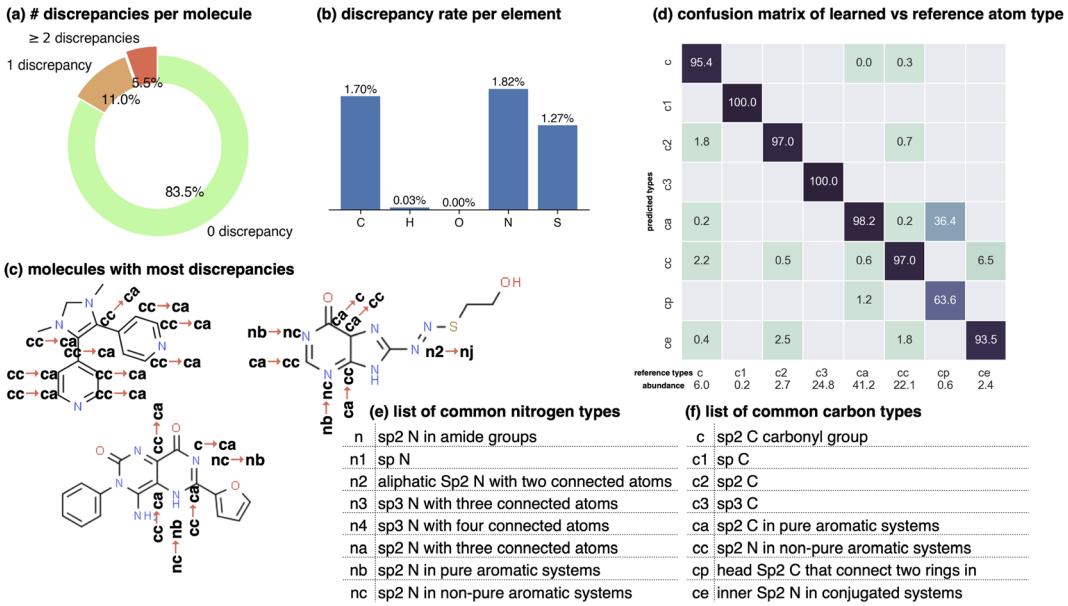


Fig. 2 Graph neural networks can reproduce legacy atom types with high accuracy. The Stage 1 graph neural network of espaloma (Section 1.1) chained to a discrete atom type readout was fit to GAFF 1.81 atom types[26,27] on a subset of ZINC[50] distributed with parm@Frosst[51] as a validation set. The 7529 molecules in this set were partitioned 80 : 10 : 10 into training : validation : test sets for this experiment. Within the test set, 99.07%$^{99.22\%}_{98.93\%}$ of atoms were correctly typed, with 1000 bootstrap replicates used to estimate the confidence intervals arising from finite test set size effects. (a) The distribution of the number of atom type discrepancies per molecule on the test set demonstrates that only a minority of atoms are incorrectly typed. (b) The error rate per element is primarily concentrated within carbon, nitrogen, and sulfur types. (c) Examining atom type failures in detail on molecules with the largest numbers of discrepancies shows that the atom types are easily confusing even to a human expert, since they represent qualities that are difficult to precisely define. (d) The distribution of predicted atom types for each reference atom type for carbon types are shown; on-diagonal values indicate agreement. The percentages annotated under *x*-axis denote the relative abundance within the test set. Only the common carbon types are included in the confusion matrix here; for full confusion matrix across all atom types, see ESI Fig. 14†. (e) A list of common nitrogen types in GAFF-1.81.[52] (f) A list of common carbon types in GAFF-1.81.[52]

chemically arbitrary) logic. For instance, one of the most frequent confusions is the misassignment of ca (sp^2 carbon in pure aromatic systems) to cp (head sp^2 carbon that connects two rings in biphenyl systems, occurring in only 0.6% of the dataset). The relative ambiguity of the various types that are most frequently confused is suggestive that the graph net makes human-like errors in perceiving subtle differences between distinct chemical environments.

The benefits of neural embedding compared to legacy discrete typing are many-fold:

• Legacy typing schemes are generally described in text form in published work (for example ref. 26 and 27), creating the potential for discrepancies between implementations when different cheminformatics toolkits are used. By contrast, with the knowledge to distinguish chemical environments stored in latent vectors and not dependent on any manual coding, our approach is deterministic once trained, and is portable across platforms thanks to modern machine learning frameworks.

• Both the chemical perception process and the application of force field parameters Φ_{FF} can be optimized simultaneously *via* gradient-based optimization of Φ_{NN} using standard machine learning frameworks that support automatic differentiation.

• While extending a legacy force field by adding new atom types can lead to an explosion in the number of parameter types, continuous neural embeddings do not suffer from this limitation; expansion of the typing process occurs automatically as more diverse training examples are introduced.

1.2 Stage 2: symmetry-preserving pooling generates continuous bond, angle, and torsion embeddings, replacing discrete types

Terms in a molecular mechanics potential are symmetric with respect to certain permutations of the atoms involved in the interaction. For example, harmonic bond terms are symmetric with respect to the exchange of atoms involved in the bond. More elaborate symmetries are frequently present, such as in the three-fold terms representing improper torsions for the Open Force Field "Parsley" generation of force fields (*k*–*i*–*j*–*l*, *k*–*j*–*l*–*i*, and *k*–*l*–*i*–*j*, where *k* is the central atom).[53] Traditional force fields, for bond, angle, and proper torsion terms, enforce this by ensuring equivalent orderings of atom types receive the same parameter value.‖

For neural embeddings, the invariances of valence terms with respect to these atom ordering symmetries must be considered while searching for expressive latent representations to feed into a subsequent parameter prediction network stage. Inspired by Janossy pooling,[55] we enumerate the relevant equivalent atom permutations to derive bond, angle, and torsion embeddings h_r, h_θ, h_ϕ that respect these symmetries from atom embeddings h_v (see Fig. 1 Stage 2),

$$h_{r_{ij}} = \mathrm{NN}_r([h_{v_i} : h_{v_j}]) + \mathrm{NN}_r([h_{v_j} : h_{v_i}]); \quad (2)$$

$$h_{\theta_{ijk}} = \mathrm{NN}_\theta([h_{v_i} : h_{v_j} : h_{v_k}]) + \mathrm{NN}_\theta([h_{v_k} : h_{v_j} : h_{v_i}]); \quad (3)$$

$$h_{\phi_{ijkl_{\mathrm{proper}}}} = \mathrm{NN}_{\phi_{\mathrm{proper}}}\left(\left[h_{v_i} : h_{v_j} : h_{v_k} : h_{v_l}\right]\right) + \mathrm{NN}_\phi\left(\left[h_{v_l} : h_{v_k} : h_{v_j} : h_{v_i}\right]\right); \quad (4)$$

$$h_{\phi_{ijkl_{\mathrm{improper}}}} = \mathrm{NN}_{\phi_{\mathrm{improper}}}\left(\left[h_{v_k} : h_{v_i} : h_{v_j} : h_{v_l}\right]\right) + \mathrm{NN}_\phi\left(\left[h_{v_k} : h_{v_j} : h_{v_i} : h_{v_l}\right]\right) + \mathrm{NN}_{\phi_{\mathrm{improper}}}\left(\left[h_{v_k} : h_{v_l} : h_{v_i} : h_{v_j}\right]\right), \quad (5)$$

where columns (:) denote concatenation**. As such, the order invariance is evident, *i.e.*, $h_{r_{ij}} = h_{r_{ji}}$,$h_{\theta_{ijk}} = h_{\theta_{kji}}$, and $h_{\phi_{ijkl}} = h_{\phi_{lkji}}$††.

1.3 Stage 3: neural parametrization of atoms, bonds, angles, and torsions replaces tabulated parameter assignment

In the final stage, each feed-forward neural network modularly learns the mapping from these symmetry-preserving atom, bond, angle, and torsion encodings to MM parameters Φ_{FF} that reflect the specific chemical environments appropriate for these terms:

$$\{\varepsilon_v,\sigma_v\} = \mathrm{NN}_{v_{\mathrm{readout}}}(h_v) \text{ atom parameters} \quad (6)$$

$$\{k_r,r_0\} = \mathrm{NN}_{r_{\mathrm{readout}}}(h_r) \text{ bond parameters} \quad (7)$$

$$\{k_\theta,\theta_0\} = \mathrm{NN}_{\theta_{\mathrm{readout}}}(h_\theta) \text{ angle parameters} \quad (8)$$

$$\{k_\phi\} = \mathrm{NN}_{\phi_{\mathrm{readout}}}(h_\phi) \text{ torsion parameters} \quad (9)$$

where k_r and k_θ denote force constants for bonds and angles, r_0 and θ_0 denote equilibrium bond lengths and angles, $k_{\phi,n}$ denotes a torsion energy factor (which can be positive or negative) for periodicity n, σ_v is the effective radius and ε_v is the effective well depth for Lennard-Jones interactions. (For a brief review of the molecular mechanics force fields functional term, see Appendix Section C in ESI.†) This stage is analogous to the final table lookup step in traditional force field construction, but with significant added flexibility arising from the continuous embedding that captures the chemical environment specific to the potential energy term being assigned.

Here, we use Lennard-Jones parameters from legacy force fields (here, the Open Force Field 1.2.0 "Parsley" small molecule force field[56]) to avoid having to include condensed-phase physical properties in the fitting procedure. While including condensed-phase physical properties in the loss function is possible, it is very expensive to do so, and as our experiments demonstrate, may not be necessary for achieving increased accuracy over legacy force fields.

We also found producing bond and angle parameters directly in Stage 3 to frustrate optimization, so we employ a mixture of linear bases to represent harmonic energies that can be translated back to the original functional form (see Appendix Section D.1.1 in ESI†). Similarly, we do not fit phases and periodicities of torsions as they are discrete. We instead fix phases at $\phi_0 = 0$ and fit all periodicities $n = 1, \ldots, 6$. This allows the corresponding torsion barriers K_n to assume the entire continuum of positive or negative values; as a result, $K_n < 0$ mimics the effect of $\phi_0 = \pi$.

As a result of using the continuous atom embedding vectors to represent chemical environments for each atom, it is possible to intelligently interpolate between relevant chemical environments seen during training. This interpolation produces more nuanced varieties of parameters than either traditional atom typing or direct chemical perception, and is capable of capturing subtle effects arising from fractional bond order perturbation.[45] Due to the modularity of this stage, it is easy to add new modules or swap out existing ones to explore other force field functional forms, such as alternative vdW interactions;[57] pair-specific Lennard-Jones interaction parameters;[58,59] point polarizabilities for instantaneous dipole,[60] Drude oscillator,[61] or Gaussian charge[62] polarizability models; class II valence couplings;[63–65] charge transfer;[46] or other potential energy terms of interest.

2 Espaloma can learn to mimic existing molecular mechanics force fields from snapshots and associated potential energies

Having established that graph neural networks have the capacity to learn to reproduce legacy atom types describing distinct chemical environments, we ask whether espaloma is capable of learning to reproduce traditional molecular mechanics (MM) force fields assigned *via* standard atom typing schemes. In addition to quantifying how well a force field can be learned when the exact parameters of the model being learned are known, being able to accurately learn existing MM force fields would have numerous applications, including replacing legacy non-portable parameter assignment codes with modern portable machine learning frameworks, learning to generalize to new molecules that contain familiar chemical environments, and permitting simplified parameter assignment for complex, heterogeneous systems involving post-translational modifications, covalent ligands, or heterogeneous combinations of biopolymers and small molecules.

To assess how well espaloma can learn to reproduce a molecular mechanics force field from a limited amount of data, we selected a dataset with limited chemical complexity—PhAlkEthOH[66,67]—which consists of 7408 linear and cyclic molecules containing phenyl rings, small alkanes, ethers, and alcohols composed of only the elements carbon, oxygen, and hydrogen. Three- and four-membered rings are excluded in the dataset since they would cause instability in the prediction of energies (see Section H in ESI†). We generated a set of 100 conformational snapshots for each molecule using short molecular dynamics simulations at 300 K initiated from multiple conformations to ensure adequate sampling of conformers. The PhAlkEthOH dataset was randomly partitioned (by molecules) into 80% training, 10% validation, and 10% test molecules, and an espaloma model was trained with early stopping *via* monitoring for a decrease in accuracy in the

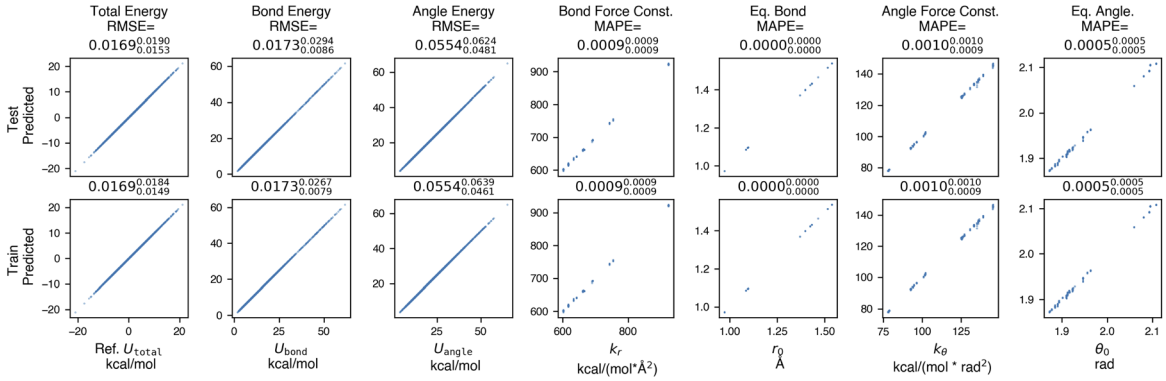


Fig. 3 Espaloma accurately learns molecular mechanics parameters when fit to snapshot energies from a molecular mechanics force field. In this experiment, espaloma was used to fit molecular mechanics (GAFF 1.81) potential energies of snapshots generated from short molecular dynamics (MD) simulations initiated from multiple conformers of molecules from the PhAlkEthOH dataset, which uses only the elements carbon, oxygen, and hydrogen.[66] The dataset contains 7408 molecules with 100 snapshots each, and was partitioned by molecules 80 : 10 : 10 into train : validate : test sets. We excluded three- and four-membered rings in the dataset; for a detailed study on the roles of these two factors, see Section H in ESI†. Statistics quoted above the plots provide the root mean squared error (RMSE) between reference and predicted MM energies, and mean absolute percentage error (MAPE) (in fractional form) between reference and predicted force field parameters. The sub- and superscripts report the 95% confidence interval of each statistics estimated from 1000 bootstrapped replicates over molecules in the test set. Energy terms have kcal mol^{-1} units whereas the units of the force field parameters do not affect the statistics reported here. Force field parameters (Φ_{FF}): k_r: bond force constant; r_0: equilibrium bond length; k_θ: angle force constant; θ_0: equilibrium angle value. Torsion parameters are not shown because of the potential for degeneracy of fit given that all periodicities $n = 1, 2, \ldots, 6$ are learned by espaloma. See Section 2 for details.

validation set. The performance of the resulting model is shown in Fig. 3.

2.1 Espaloma can learn existing force fields and generalize to new molecules with low error

Espaloma is able to achieve very low total energy and parameter error on the training set, suggesting that espaloma can learn the parameters of typed molecules from energies alone. In addition, error on the out-of-sample test set of molecules is comparable—less than 0.02 kcal mol^{-1}—suggesting that espaloma can effectively generalize to new molecules within the same chemical space. Surprisingly, the total energy RMSE is lower than the angle energy RMSE, suggesting that there is some degeneracy in how energy contributions are distributed among valence energy terms. It is worth noting that espaloma requires very few conformations per molecule to achieve high accuracy; we closely study its data efficiency in Appendix Section 8 in ESI†.

3 Espaloma can fit quantum chemical energies directly to build new molecular mechanics force fields

Since espaloma can derive a force field solely by fitting to energies (and optionally gradients), we repeat the end-to-end fitting experiment (Section 2) directly using quantum chemical (QM) datasets used to build and evaluate MM force fields. We assessed the ability of Espaloma to learn several distinct quantum chemical datasets generated by the Open Force Field Initiative[70] and deposited in the MolSSI QCArchive[71] with B3LYP-D3BJ/DZVP level of theory:

- PhAlkEthOH[66] is a collection of compounds containing only the elements carbon, hydrogen, and oxygen in compounds containing phenyl rings, alkanes, ketones, and alcohols. Limited in elemental and chemical diversity, this dataset is chosen as a proof-of-concept to demonstrate the capability of espaloma to fit and generalize quantum chemical energies when training data is sufficient to exhaustively cover the breadth of chemical environments.
- OpenFF Gen2 Optimization[72] consists of druglike molecules used in the parametrization of the Open Force Field 1.2.0 ("Parsley") small molecule force field.[73] This set was constructed by the Open Force Field Consortium from challenging molecule structures provided by Pfizer, Bayer, and Roche, along with diverse molecules selected from eMolecules to achieve useful coverage of chemical space.
- VEHICLe[74] or virtual exploratory heterocyclic library, is a set of heteroaromatic ring systems of interest to drug discovery enumerated by Pitt *et al.*[75] The atoms in the molecules in this dataset have interesting chemical environments in heteroarmatic rings that present a challenge to traditional atom typing schemes, which cannot easily accommodate the nuanced distinctions in chemical environments that lead to perturbations in heterocycle structure. We use this dataset to illustrate that espaloma performs well in situations challenging to traditional force fields.
- PepConf[76] from Prasad *et al.*[77] contains a variety of short peptides, including capped, cyclic, and disulfide-bonded peptides. This dataset—regenerated as an OptimizationDataset (quantum chemical optimization trajectories initiated from multiple conformers) using the Open Force Field QCSubmit tool[78]—explores the applicability of espaloma to biopolymers, such as proteins.

Since nonbonded terms are generally optimized to fit other condensed-phase properties, we focused here on optimizing only the valence parameters (bond, angle, and proper and improper torsion) to fit these gas-phase quantum chemical datasets, fixing the non-bonded energies using a legacy force field.[70] In this experiment, all the non-bonded energies (Lennard-Jones and electrostatics) were computed using Open Force Field 1.2 Parsley,[79] with AM1-BCC charges generated by the OpenEye Toolkit back-end for the Open Force Field toolkit 0.10.0.[69] Because we are learning an MM force field that is incapable of reproducing quantum chemical heats of formation, which are reflected as an additive offset in the quantum chemical energy targets, snapshot energies for each molecule in both the training and test sets are shifted to have zero mean. All datasets are randomly shuffled and split (by molecules) into training (80%), validation (10%), and test (10%) sets.

3.1 Espaloma generalizes to new molecules better than widely-used traditional force fields

To assess how well espaloma is able to generalize to new molecules, the performance for espaloma on test (and training) sets was compared to a legacy atom typing based force field (GAFF 1.81 and 2.11,[26,27] which collectively have been cited over 13 066 times) and a modern force field based on direct chemical perception[22] (the Open Force Field 1.2.0 ("Parsley") small molecule force field,[56] downloaded over 150 000 times).

The results of this experiment are reported in Fig. 4. As can be readily seen by the reported test set root mean squared error (RMSE), espaloma can produce MM force fields with generalization performance consistently better than legacy force fields based on discrete atom typing (GAFF[26,27]). In chemically well-represented datasets like PhAlkEthOH—which contains only simple molecules constructed from elements C, H, and O—espaloma is able to significantly improve on the accuracy of traditional force fields such as OpenFF 1.2.0, GAFF-1.81, and GAFF-2.11 on the test set.

Surprisingly, even though OpenFF 1.2.0 included the "Open FF Gen 2" dataset in training, espaloma is able to achieve superior test set performance on this dataset, suggesting that both the flexibility and generalizability of continuous atom typing have significant advantages over even direct chemical perception.[22]

Even compared to highly optimized late-generation protein force fields such as Amber ff14SB[80]—which was highly optimized to reproduce quantum chemical torsion drive data—espaloma achieves significantly higher accuracy, improving on Amber ff14SB error of $3.1502^{3.1859,*}_{3.1117}$ kcal mol^{-1} to achieve $1.8727^{1.9749}_{1.7309}$ kcal mol^{-1} on the PepConf peptide dataset.[76,81] This suggests that espaloma is capable of effectively parameterizing both small molecule and biopolymer force fields. Indeed, when we train an espaloma model using both the OpenFF Gen2 Optimization and PepConf datasets (joint in Fig. 4(a)), we see

(a) dataset		# mols	# trajs	# snapshots	Espaloma RMSE Train	Espaloma RMSE Test	Legacy FF RMSE (kcal/mol) (Test molecules) OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)		7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)		792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)		24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	
PepConf (peptides)		736	7560	22154	$1.2714^{1.3616}_{1.1899}$	$1.8727^{1.9749}_{1.7309}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$
joint	OpenFF Gen2 Optimization PepConf	1528	11537	45902	$0.8264^{0.9007}_{0.7682}$	$1.8764^{1.9947}_{1.7827}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
					$1.2038^{1.3056}_{1.1178}$	$1.7307^{1.8439}_{1.6053}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$

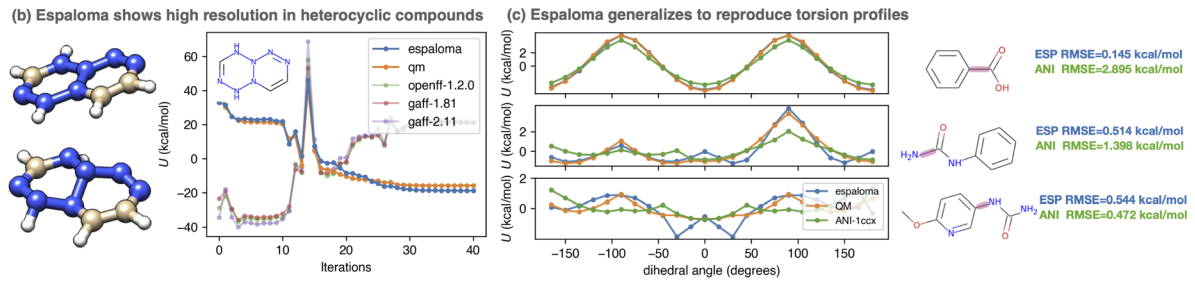


Fig. 4 Espaloma can directly fit quantum chemical energies to produce new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories initiated from distinct conformers in various datasets from QCArchive.[27] All datasets were partitioned by molecules 80 : 10 : 10 into train : validate : test sets. The number of molecules, optimization trajectories, and total snapshots are annotated in the table. (a) We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule, in order to focus on errors in relative conformational energetics rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by bootstrapping molecules with replacement using 1000 replicates. (b) Optimization trajectory of a representative (with highest OpenFF 1.2.0 RMSE) heterocyclic compound in VEHICLe dataset with SMILES string [H]C1=C(N2N(N=C(N(N2N=N1)[H])[H])[H])[H]. Legacy force fields, because of their limited chemical typing rules, were not able to perceive the chemical environment of the nitrogen atoms, which were not aromatic. (c) Espaloma is able to predict energies for quantum chemical torsion scans for an out-of-sample torsion scan dataset (the OpenFF Phenyl Torsion Drive Dataset,[45,68] dihedral angle profiled marked in rouge) to high accuracy even though it was not trained on torsion scans (only optimization trajectories) or any of the molecules in the torsion scan set. We also include the torsion energy profile computed by a popular machine learning force field, ANI-1ccx.[12] *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine[69] are not included.

that a single espaloma model is capable of achieving superior accuracy to traditional small molecule and protein force fields simultaneously.

3.3 Espaloma can reliably learn torsion profiles from optimization trajectories

We wondered whether espaloma could faithfully recover torsion energy profiles—which are traditionally expensive to generate using methods like wavefront propagation[82]—from the inexpensive optimization trajectories used to train espaloma models. We therefore examined some representative dihedral energy profiles for molecules outside of the dataset used to train espaloma. In Fig. 4(c), we use the espaloma model trained on OpenFF Gen2 Optimization and PepConf to predict the energy profiles of several torsion drive experiments in the OpenFF Phenyl Torsion Drive Dataset[68]—which does not contain any of the molecules in the training set—and observed that the locations and heights of torsion energy barriers are recapitulated with reasonable accuracy. This suggests that optimization trajectories are sufficient to capture the locations and relative heights of torsion barriers—a highly useful finding given the relative expense of generating accurate torsion profiles compared to simple optimization trajectories.[82]

3.2 Espaloma can automatically learn distinct atom environments overlooked by traditional force fields

It is worth noting that the traditional, widely used force fields considered here uniformly perform poorly on the VEHICLe dataset[75] ("Heteroaromatic Rings of the Future", containing heterocyclic scaffolds of interest to future drug discovery programs). In Fig. 4(b), we show the most common mode of failure of legacy force fields by examining their predicted energy over the QM optimization trajectory of the compound with largest RMSE (with SMILES string [H]C1=C(N2N(N=C(N(N2N=N1)[H])[H])[H])[H]). The initial conformation of the molecule, generated by OpenEye Toolkit, was planar. As the conformation was optimized by quantum chemical methods, the tertiary nitrogens in the system become pyramidal. In a closer examination, GAFF-2.11, for instance, assigned all carbons to be of type cc and all nitrogens na, indicating that they were perceived as aromatic, whereas there is no conjugated system present in the molecule. This also reflects the limitation in resolution of legacy force fields. Espaloma, on the other hand, provides a high-resolution atom embedding that can flexibly characterize the chemical environments, provided that similar environments existed in the training data.

4 Espaloma can learn self-consistent charge models in an end-to-end differentiable manner

Historically, biopolymer force fields derive partial atomic charges *via* fits to high-level multiconformer quantum chemical electrostatic potentials on capped model compounds, adjusted to ensure the repeating biopolymer units have integral charge (often incorporating constraints to share identical backbone partial charges).[83–85] Some approaches to the derivation of partial atomic charges are enormously expensive, requiring iterative QM/MM simulations in explicit solvent to derive partial charges for new molecules.[86,87] For small molecules, state-of-the-art methods range from fast bond charge corrections applied to charges derived from semiempirical quantum chemical methods (such as AM1-BCC[28,29] or CGenFF charge increments[47]) to expensive multiconformer restrained electrostatic potential (RESP) fits to high-level quantum chemistry.[88,89] Surprisingly little attention has been paid to the divergence of methods used for assigning partial charges to small molecules and biopolymers, and the potential impact this inconsistency has on accuracy or ease of use—indeed, developing charges for post-translational modifications to biopolymer residues[90,91] or covalent ligands can prove to be a significant technical challenge in attempting to bridge these two worlds.

While machine learning approaches have begun to find application in determining small molecule partial charges,[92–94] methods such as random forests are not fully continuously differentiable, rendering them unsuitable for a fully end-to-end differentiable parameter assignment framework. Recently, a fast (500× speed up for small molecules) approach has been proposed that uses graph neural networks as part of a charge-equilibration[95,96] scheme (inspired by the earlier VCharge model[97]) to self-consistently assign partial charges to small molecules, biopolymers, and arbitrarily complex hybrid molecules in a conformation-independent manner that only makes use of molecular topology.[36] Perhaps unsurprisingly, due to the requirement that molecules retain their integral net charge, directly predicting partial atomic charges from latent atom embeddings and subsequently renormalizing charges leads to poor performance (0.28e (ref. 36)).

Instead, predicting the parameters of a simple physical topological charge-equilibration model[95,97] can produce geometry-independent partial charges capable of reproducing charges derived from quantum chemical electrostatic potential fits.[36] Note that, unlike Wang *et al.*,[36] here we fit AM1-BCC charges rather than higher level of quantum mechanics theory due to their high cost. Specifically, we use our atom latent representation to instead predict the first- and second-order derivatives of a pseudopotential energy E with respect to the partial atomic charge q_i on atom i:

$$e_i \equiv \frac{\partial E}{\partial q_i}, \quad s_i \equiv \frac{\partial E^2}{\partial^2 q_i} \tag{10}$$

here, the electronegativity e_i quantifies the desire for an atom to take up negative charge, while the hardness s_i quantifies the resistance to gaining or losing too much charge. A module is added to Stage III of espaloma to predict the chemical environment adapted (e_i, s_i) parameters for each atom from the latent atom embeddings.

The partial charges for all atoms can then be obtained by minimizing the second-order Taylor expansion of the potential pseudoenergy contributed by atomic charges:

$$\{\hat{q}_i\} = \underset{q_i}{\operatorname{argmin}} \sum_i \hat{e}_i q_i + \frac{1}{2}\hat{s}_i {q_i}^2, \tag{11}$$

subject to

$$\sum_i \hat{q}_i = \sum_i q_i = Q, \tag{12}$$

where Q is the total (net) charge of the molecule.

Using Lagrange multipliers, the solution to 11 can be given analytically by:

$$\hat{q}_i = -e_i {s_i}^{-1} + {s_i}^{-1} \frac{Q + \sum_i e_i {s_i}^{-1}}{\sum_j {s_j}^{-1}}, \tag{13}$$

whose Jacobian and Hessian are trivially easy to calculate. As a result, the prediction of $\{\hat{e}_i, \hat{s}_i\}$ could be optimized end-to-end using backpropagation.

When predicting the partial charges independently, we observe that the RMSE error on the test set ($0.0072^{0.0073}_{0.0071}$ e), is smaller than the difference between the discrepancy between AM1-BCC charges assigned by two popular cheminformatics toolkits, Ambertools 21[80] and OpenEye Toolkit ($0.0126^{0.0129}_{0.0124}$ e). As shown in Appendix Table 1 in ESI †, there is only a slight decrease in energy performance (within confidence interval) when we switch from AM1-BCC charges to this neural charge equilibration model. Additionally, we can integrate the charge equilibration model into the valence parameter prediction pipeline outlined in Section 1 to have the atom embeddings shared across two tasks and curate a single, combined model. With this approach, we observed a test set total RMSE of $1.2216^{1.2766}_{1.1529}$ kcal mol^{-1} and charge RMSE of $0.0072^{0.0073}_{0.0070}$ e, which is within the confidence interval of the performance when predicting separately. We provide a detailed study of the joint learning in Appendix Section B in ESI †.

5 Espaloma can parameterize biopolymers

We have so far established that espaloma, as a method to construct MM force fields, shows great versatility and flexibility. In the following sections, we showcase its utility with a model predicting both valence parameters and partial charges trained on OpenFF Gen2 Optimization Dataset as well as PepConf dataset, which we released as 'espaloma-0.2.2' with the package.

The speed and flexibility of graph convolutional networks allows espaloma to parameterize even very large biopolymers, treating them as (large) small molecules in a graph-theoretical manner. While graph neural networks perceive nonlocal aspects of the chemical environment around each atom, the limited number of rounds of message passing ensures stability of the resulting parameters when parameterizing systems that consist of repeating residues, like proteins and nucleic acids.

To demonstrate this, we considered the simple polypeptide system ACE-ALA$_n$-NME, consisting of n alanine residues terminally capped by acetyl- and N-methyl amide capping groups. Using the joint charge and valence term espaloma model, we assigned parameters to ACE-ALA$_n$-NME systems with n = 1, 2, …, 500, showing illustrative parameters in Fig. 5. Espaloma stably assigns parameters to the interior residues of peptides even as they increase in length, with parameters of the central residue unchanged after n > 3. This pleasantly resembles the behavior of traditional residue template based protein force fields, even though no templates are used within espaloma's parameter assignment process.

5.1 Espaloma can generate self-consistent valence parameters and partial charges for large biopolymers in less than a second

Despite its use of a sophisticated graph net machine learning model, the wall time required to parameterize large proteins scales linearly with respect to the number of residues (and hence system size) on a CPU (Fig. 5, lower right). On a GPU, the wall clock time needed to parameterize systems of this size stays roughly constant (due to overhead in executing models on the GPU) at less than 100 microseconds. Since espaloma applies standard molecular mechanics force fields, the energy evaluation times for an Espaloma-generated force field are identical to traditional force fields.

6 Espaloma can produce self-consistent biopolymer and small molecule force fields that result in stable simulations

Traditionally, in a protein–ligand system, separate (but hopefully compatible) force fields and charge models have been assigned to small molecules (which are treated as independent entities parameterized holistically) and proteins (which are treated as collections of templated residues parameterized piecemeal).[80] This practice both has the potential to allow significant inconsistencies while also introducing significant complexity in parameterizing heterogeneous systems.

Using the joint espaloma model trained on both the "OpenFF Gen 2 Optimization" small molecule and "PepConf" peptide quantum chemical datasets (Section 3)[81,104]), we can apply a consistent set of parameters to both protein and small molecule components of a kinase : inhibitor system. Fig. 5f shows the ligand heavy-atom RMSD after aligning on protein heavy atoms for 0.5 ns trajectories of the Tyk2 : inhibitor system from the Alchemical Best Practices Benchmark Set 1.0.[105] It is readily apparent that the espaloma-derived parameters lead to trajectories that are comparably stable to simulations that utilize the Amber ff14SB protein force field[106] with GAFF 1.81, GAFF 2.11,[26,27] or OpenFF 1.2.0 (ref. 56) small molecule force fields. All systems are explicitly solvated with a 9 Å buffer around the protein with TIP3P water[107] and use the Joung and Cheatham monovalent counterion parameters[108] to model a neutral system with 300 mM NaCl salt.

Additionally, the espaloma model also provides sufficient coverage to model more complex and heterogeneous protein-

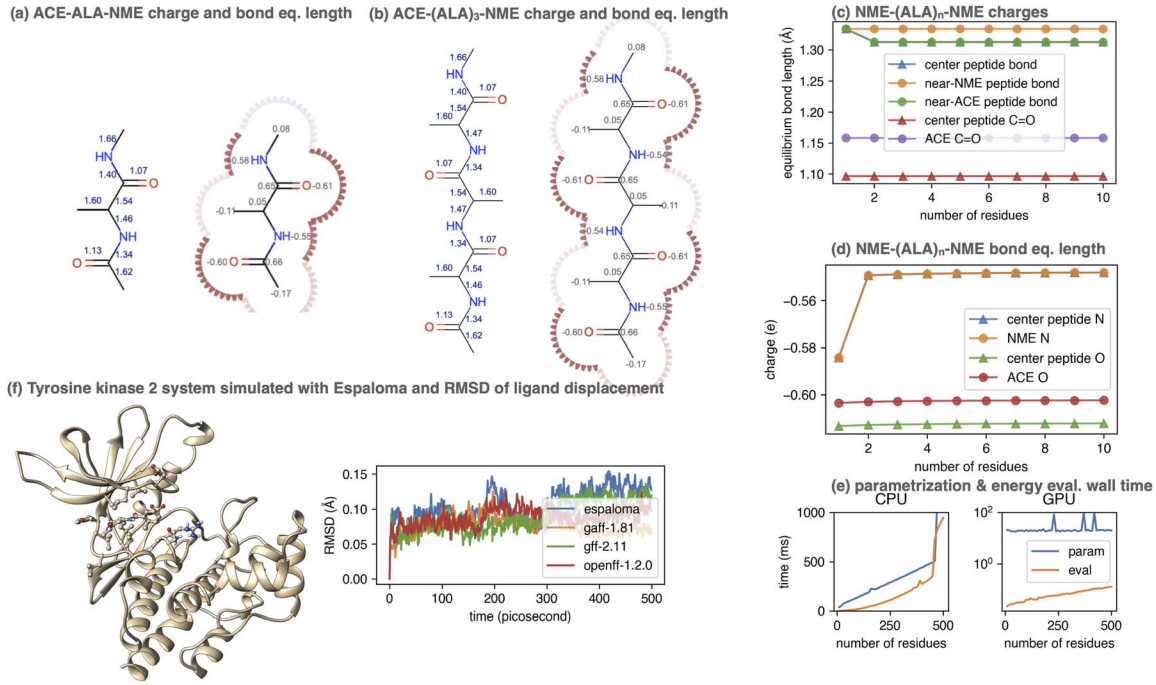


Fig. 5 Espaloma can be used to generate self-consistent force fields for biopolymers and small molecules. (a) Espaloma-assigned equilibrium heavy atom bond lengths (left, in angstrom) and heavy atom partial charges (right, in elemental charge unit) for ACE-$(ALA)_1$-NME. (b) Espaloma-assigned equilibrium heavy atom bond lengths (left) and heavy atom partial charges (right) for ACE-$(ALA)_3$-NME, which shows minimal deviation from ACE-$(ALA)_1$-NME and consistent parameters for amino acid residues. (c) Partial atomic charges of selected atoms shown for ACE-$(ALA)_n$-NME for $n = 1, \ldots, 10$. (d) Equilibrium bond lengths of selected bonds for ACE-$(ALA)_n$-NME for $n = 1, \ldots, 10$. (e) CPU (left) and GPU (right) parameter assignment (blue curve) and energy evaluation on OpenMM 7.5.1 (ref. 98) (orange curve) wall times. (f) Espaloma simultaneously parametrizes macro- and small molecules in a protein-ligand system. Left: Tyrosine kinase 2 system parametrized by espaloma and minimized and equilibrated with TIP3P water model[99] and counterions. Right: Root mean-squared displacement (RMSD) of ligand w.r.t. the initial position, in systems parametrized by espaloma and traditional force fields.

ligand covalent conjugates, which was highly non trivial in traditional force fields where protein and ligand are parametrized separately. We provide a detailed study of this capability in Appendix Section I in ESI†.

7 Espaloma small molecule parameters and charges provide accuracy improvements in alchemical free energy calculations

To assess whether the small molecule parameters and charges generated by espaloma achieve competitive performance to traditional force fields, we used the perses 0.9.5 relative alchemical free energy calculation infrastructure[103] (based on OpenMM 7.7 (ref. 17) and openmmtools 0.21.2 (ref. 109) to compare performance on the Tyk2 kinase : inhibitor benchmark set from the Schrodinger JACS benchmark set[100] as curated by the OpenFF protein-ligand benchmark 0.2.0.[110] In order to assess the impact of espaloma small molecule parameters and charges in isolation, we used the Amber ff14SB protein force field,[106] and performed simulations with either OpenFF 1.2.0 (openff-1.2.0) or the espaloma joint model trained on OpenFF Gen2 Optimization and PepConf datasets (espaloma-0.2.2) available through the openmmforcefields 0.11.0 package.[111] Notably, none of the ligands appearing in this set appear in the training set for either force field. All systems were explicitly solvated with a 9 Å buffer around the protein with TIP3P water[107] and use the Joung and Cheatham monovalent counterion parameters[108] to model a neutral system with 300 mM NaCl salt. The same transformation network provided in the OpenFF protein-ligand benchmark set was used to compute alchemical transformations, and absolute free energies up to an additive constant were estimated from a least-squares estimation strategy[112] as implemented in the OpenFF arsenic package.[113] Both experimental and calculated absolute free energies were shifted to their respective means before computing statistics, as in ref. 100.

Fig. 6 shows a comparison of both relative ($\Delta\Delta G$) and absolute (ΔG) free energy error statistics. While the OpenFF 1.2.0 force field achieves an impressive RMSE of $0.91^{1.17}_{0.66}$ kcal mol^{-1}, using espaloma valence and charge parameters improves the accuracy to $0.47^{0.70}_{0.30}$ kcal mol^{-1}. Additionally, the Spearman ρ correlation coefficient improves from $0.69^{0.89}_{0.28}$ (openff-1.2.0) to $0.93^{0.98}_{0.80}$ (espaloma-0.2.2). While more extensive benchmarking is necessary to establish the generality of these improvements, this represents a first demonstration that performance can be on par with, if not superior to, traditionally constructed force fields.

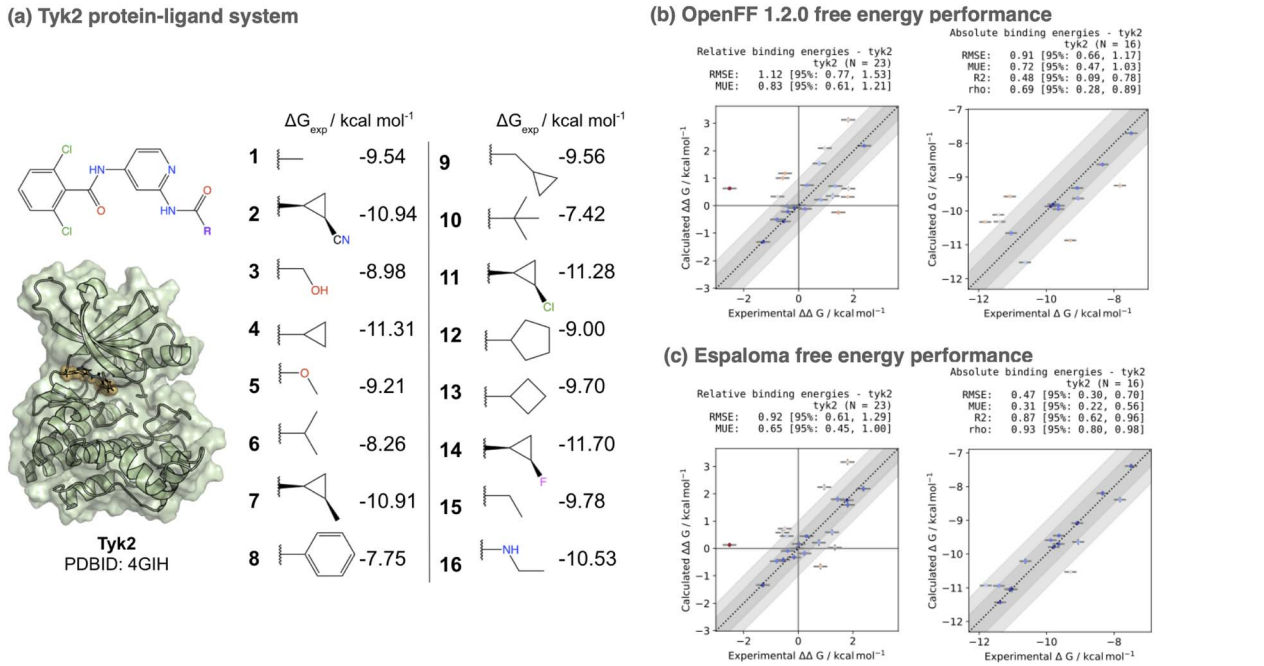


Fig. 6 Espaloma small molecule parameters can be used for accurate protein-ligand alchemical free energy calculations. (a) The Tyk2 congeneric ligand benchmark series taken from the Schrödinger JACS benchmark set[100] is challenging for both commercial force fields (OPLS2.1 achieves a ΔG RMSE of 0.93 $\pm$ 0.12 kcal mol^{-1}[100]) and public force fields (GAFF 1.8 achieves a ΔG RMSE of 1.13 kcal mol^{-1}, and $\Delta\Delta G$ RMSE of 1.27 kcal mol^{-1}[101]). We show the X-ray structure used for all free energy calculations as well as 2D structures of all ligands in the benchmark set, along with their experimental binding free energies. This congeneric series from ref. 100 was selected from ref. 102 where experimental errors in K_i are reported to have $\delta K_i/K_i < 0.3$, yielding $\delta\Delta G \approx 0.18$ kcal mol^{-1} and $\delta\Delta\Delta G \approx 0.25$ kcal mol^{-1}. Here, we used the perses[103] relative free energy calculation tool, based on OpenMM,[17] to assess the accuracy of espaloma on this dataset. (b) The Open Force Field ("Parsley") openff-1.2.0 small molecule force field achieves an absolute free energy (ΔG) RMSE of 0.91 [95% CI: 0.66, 1.17] kcal mol^{-1} on this set. (c) The espaloma-0.2.2 model for predicting valence parameters and partial charges—trained jointly on the same OpenFF Gen2 Optimization dataset used for openff-1.2.0 as well as the PepConf dataset to reproduce quantum chemical energies and AM1-BCC charges—achieves a lower error of 0.48 [95% CI: 0.30, 0.73] on this set, despite having never been trained on any molecules in this set.

8 Discussion

Here, we have demonstrated that graph neural networks not only have the capacity to reproduce legacy atom type classification, but they are sufficiently expressive to fit a traditional molecular mechanics force field and generalize it to new molecules, as well as learn entirely new force fields directly from quantum chemical energies and experimental measurements. The neural framework presented here also affords the modularity to easily experiment with the inclusion of additional potential energy terms, functional forms, or parameter classes, while making it easy to rapidly refit the entire force field afterwards.

8.1 Espaloma enables a wide variety of applications

Espaloma enables a wide variety of applications in the realm of molecular simulation: While many force field packages use complex, difficult to maintain, non-portable custom typing engines,[27,114–116] simply generating examples is sufficient to train espaloma to reproduce this typing, translating it into a model that is easy to extend by providing more quantum chemical training data. Some force fields have traditionally been typed by hand, making them difficult to automate;[116] espaloma can in principle learn to generalize from these examples, provided care is taken to avoid overfitting during training. As we have shown here, espaloma also provides a convenient way to rapidly build new force fields directly from quantum chemical data.

8.2 Modern machine learning frameworks offer flexibility in fitting potentials

The flexibility afforded by modern machine learning frameworks solves a long-standing problem in molecular simulation in which it is extremely difficult to assess whether a new functional form might lead to significant benefits in modeling multiple properties of interest. While efforts such as the Open Force Field Initiative aim to streamline the process of refitting force fields,[104] the ease of refitting models in machine learning frameworks makes it extremely easy to experiment with new functional forms: Modern automatic differentiation in these frameworks means that only the potential need be implemented, and gradients are automatically computed.

This enables a wide variety of exploration: Simple improvements could be widely implemented in current molecular simulation packages including adjusting the 1-4 Lennard-Jones and electrostatics scaling parameters, producing 1-4 interaction parameters that override Lennard-Jones combining rules,

exploring different Lennard-Jones combining rules,[117] changing the van der Waals treatment to alternative functional forms (such as Buckingham exp-6 [118] or Halgren potentials[57]), and refitting force fields for various non-bonded treatments (such as PME[119] and reaction field electrostatics[120]). Many simulation packages provide support for Class II molecular mechanics force fields,[30,31] which include additional coupling terms that can drastically reduce errors in modeling quantum chemical energies at essentially no meaningful impact on cost due to the $\mathcal{O}(N)$ number of these terms; simple extensions to espaloma's architecture can easily predict the parameters for these coupling terms from additional symmetry-preserving features.

More radical potential explorations could involve assessing different algebraic functional forms—modern simulation packages such as OpenMM have the ability to automatically differentiate and compile symbolic algebraic expressions to produce optimized force kernels for simulation on fast GPUs.[17,98] Excitingly, the simplicity of incorporating a new generation of quantum machine learning (QML) potentials[121]—such as ANI[10,13] and SchNet[14]—means that it will be easy to explore hybrid potentials that combine the flexibility of QML potentials at short range with the accuracy of physical forces at long range.[122]

8.3 Espaloma can enable modular loss functions and regularization

The ease at which the loss function can be augmented with additional terms enables the addition of other classes of loss terms to the loss function. For example, one of the molecules considered in the Tyk2 : inhibitor system included a cyano group which proved to be slightly unstable with hydrogen mass repartitioning at 4 fs timesteps. The loss function could either be augmented to regularize parameters to increase stability (penalizing short vibrational periods) or to include other data classes (such as Hessians and/or torsion drive data) to improve fits to particular aspects. While this will require tuning of the weighting of different loss classes, these parameters can be selected automatically *via* cross-validation strategies.

8.4 Espaloma can enable Bayesian force field parameterization and model uncertainty quantification

While much of the history of molecular simulation has focused on quantifying the impact of statistical uncertainty,[123–125] critical studies over the last decade[126–135] have improved our ability to quantify and propagate predictive uncertainty in molecular mechanics force fields by quantifying contributions from model uncertainty—which is frequently the major source of predictive uncertainty in applications of interest. While most attention has been focused on the continuous parameters of the force field model with fixed model form, some progress has been made in discrete model selection among candidate model forms.[136–138]

It remains an open problem to rigorously quantify uncertainty in other important parts of the model definition—especially in the definitions of atom-types. These "chemical perception" definitions can involve very large spaces of discrete choices, and crucially influence the behavior of a generalizable molecular mechanics model.[22,139]

An important benefit of the present approach is that it reduces the mixed continuous-discrete task of "being Bayesian about atom-types" to the more familiar task of "being Bayesian about neural network weights." Bayesian treatment of neural networks—while also intractable—has been the focus of productive study and methodological innovation for decades.[140]

We anticipate that Bayesian extensions of this work will enable more comprehensive treatment of predictive uncertainty in molecular mechanics force fields.

8.5 Ensuring full chemical equivalence is nontrivial

In the current experiments, espaloma used a set of atom features (one-hot encoded element, hybridization, aromaticity, formal charge, and membership in rings of various sizes) easily computed using a cheminformatics toolkit; no bond features were used (see Detailed Methods in ESI†). While this provided excellent performance, the non-uniqueness of formal charge assignment (obvious in molecules such as guanidinium, where resonance forms locate the formal charge on different atoms) does not guarantee the assigned parameters will respect chemical equivalence (a form of invariance) in cases where these atom properties are not unique. Ensuring full chemical equivalence would require modifications to this strategy, such as omission of non-unique features (which may require additional data or pre-training to learn equivalent chemical information), averaging of the output of one or more stages over equivalent resonance forms, or architectures such as transformers that more fully encode chemical equivalence.

8.6 Future directions: espaloma for free alchemical parameters

While we used Espaloma to generate parameters for the real physical endstates of an alchemical free energy calculation in this world, we note it is also possible to introduce dependence of these parameters on a global alchemical parameters to generate parameters for alchemical intermediate states as well. More complex loss functions could minimize the thermodynamic length along the alchemical coördinate to provide an efficient way to interpolate alchemical parameters.[141,142]

Disclosures

JDC 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.

Disclaimers

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability

The software package as well as all data is distributed open source at https://github.com/choderalab/espaloma.

Author contributions

Conceptualization: J. F., Y. W., J. D. C.; data curation: Y. W., J. F., J. E. H.; formal analysis: Y. W.; funding acquisition: J. D. C.; investigation: Y. W., J. F.; methodology: Y. W., J. F.; project administration: J. D. C.; resources: J. D. C.; software: Y. W., J. F., B. K., D. R., I. Z., I. P., M. H., H. E. B. M.; supervision: J. D. C.; visualization: Y. W.; writing – original draft: Y. W.; writing – review & editing: Y. W., J. D. C., J. F., B. K., J. E. H., K. T.

Conflicts of interest

The authors declare no conflicts of interest, but highlight all sources of funding in Disclosures in the interest of full transparency.

Acknowledgements

The authors thank reviewers and editors for their constructive feedback. The authors wish to thank Yutong Zhao (0000-0003-0090-7801) for helpful feedback while troubleshooting implementations of molecular mechanics models automatic differentiation packages, David L. Mobley (0000-0002-1083-5533) for feedback about issues with improper torsion models and other important considerations, Joshua T. Horton (0000-0001-8694-7200) for insightful comments especially regarding fitting and assessing torsions. The authors wish to thank Christian L. Mueller (0000-0002-3821-7083). We thank OpenEye Scientific Software to provide us with the free academic license. Y. W. acknowledges support from NSF CHI-1904822 and the Sloan Kettering Institute. J. F. acknowledges support from NSF CHE-1738979 and the Sloan Kettering Institute. J. D. C. acknowledges support from NIH grant P30 CA008748, NIH grant R01 GM121505, NIH grant R01 GM132386, NSF CHI-1904822, and the Sloan Kettering Institute. D. A. R. acknowledges support from a Molecular Sciences Software Institute Seed Fellowship.

Notes and references

¶ Note that this discrete type assignment layer is only used to address the question of how well the continuous embeddings approximate discrete types, and is not used in subsequent experiments that utilize the standard espaloma architecture (Fig. 1).

‖ Traditional force fields group bonds, angles, and torsions simply by their composing ordered groups of atoms. For instance, the first bond type in GAFF 1.81 (ref. 27) is defined by the types hw-ow[54], which is equivalent to ow-hw due to mirror symmetry in identifying bonds. Angles and torsions have similar symmetries that must be accounted for when enumerating the atoms or matching valence types. Note that Amber does not uniquely specify equivariant improper torsion orderings—see footnote a of Table 3 of (ref. 22) for details.

** Here, we use the threefold improper formulation used by the Open Force Field "Parsley" generation force fields, which avoids the ambiguity associated with selecting a single arbitrary improper torsion from a set of four atoms involved in the torsion (ref. 53).

†† In ESI Section L, we prove that this form is sufficiently expressive to assign unique valence types.

1 J. W. Ponder and D. A. Case, Force fields for protein simulations, *Advances in protein chemistry*, Elsevier, 2003, vol 66, pp. 27–85.
2 D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, E. M. Alan and H. J. C. Berendsen, Gromacs: fast, flexible, and free, *J. Comput. Chem.*, 2005, **26**(16), 1701–1718.
3 D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, The amber biomolecular simulation programs, *J. Comput. Chem.*, 2005, **26**(16), 1668–1688.
4 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, Scalable molecular dynamics with namd, *J. Comput. Chem.*, 2005, **26**(16), 1781–1802.
5 M. E. Tuckerman, *Free Energy Calculations. Theory and applications in chemistry and biology*, ed. C. Chipot, Springer Series in Chemical Physics, 86, 2007.
6 C. C. Wang, G. Pilania, S. A. Boggs, S. Kumar, C. Breneman and R. Ramprasad, Computational strategies for polymer dielectrics design, *Polymer*, 2014, **55**(4), 979–988, DOI: 10.1016/j.polymer.2013.12.069, ISSN 0032-3861, http://www.sciencedirect.com/science/article/pii/S0032386113011889, http://www.sciencedirect.com/science/article/pii/S0032386113011889.
7 C. Li and A. Strachan, Molecular scale simulations on thermoset polymers: A review, *J. Polym. Sci., Part B: Polym. Phys.*, 2015, **53**(2), 103–122.
8 H. Sun, Z. Jin, C. Yang, R. L. C. Akkermans, H. StruanRobertson, N. A. Spenley, S. Miller and S. M. Todd, Compass ii: extended coverage for polymer and drug-like molecule databases, *J. Mol. Model.*, 2016, **22**(2), 47.
9 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.*, Opls3: a force field providing broad coverage of drug-like small molecules and proteins, *J. Chem. Theory Comput.*, 2016, **12**(1), 281–296.
10 J. S. Smith, O. Isayev and A. E. Roitberg, Ani-1: an extensible neural network potential with dft accuracy at force field computational cost, *Chem. Sci.*, 2017, **8**(4), 3192–3203.
11 J. S. Smith, N. Ben, N. Lubbers, O. Isayev and A. E. Roitberg, Less is more: Sampling chemical space with active learning, *J. Chem. Phys.*, 2018, **148**(24), 241733.

12 J. S. Smith, B. T. Nebgen, Z. Roman, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, *Nat. Commun.*, 2019, **10**(1), 1–8.

13 C. Devereux, J. S. Smith, K. KateDavis, K. Barros, Z. Roman, O. Isayev and A. E. Roitberg, Extending the applicability of the ani deep learning molecular potential to sulfur and halogens, *J. Chem. Theory Comput.*, 2020, **16**(7), 4192–4202.

14 K. T. Schütt, H. E Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, Schnet–a deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**(24), 241722.

15 S. Batzner, T. E Smidt, L. Sun, J. P. Mailoa, M. Kornbluth, N. Molinari, and B. Kozinsky, Se (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, arXiv preprint arXiv:2101.03164, 2021.

16 Y. Han, Z. Wang, Z. Wei, J. Liu and J. Li, Machine learning builds full-QM precision protein force fields in seconds, *Briefings Bioinf.*, 2021, **22**(6), 05, DOI: **10.1093/bib/bbab158**, ISSN 1477-4054.

17 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.*, Openmm 7: Rapid development of high performance algorithms for molecular dynamics, *PLoS Comput. Biol.*, 2017, **13**(7), e1005659.

18 M. J. Harvey, G. Giupponi and G. De Fabritiis, Acemd: accelerating biomolecular dynamics in the microsecond time scale, *J. Chem. Theory Comput.*, 2009, **5**(6), 1632–1639.

19 R. Salomon-Ferrer, A. W. Gotz, P. Duncan, S. Le Grand and R. C. Walker, Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald, *J. Chem. Theory Comput.*, 2013, **9**(9), 3878–3888.

20 S. S. Schoenholz and E. D. Cubuk, *Jax, m.d.: End-to-end differentiable, hardware accelerated, molecular dynamics in pure python*, 2019.

21 W. Wang, A. Simon and R. Gómez-Bombarelli, *Differentiable molecular simulations for control and learning*, 2020.

22 D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. V. T. L. Chodera, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, *et al.*, Escaping atom types in force fields using direct chemical perception, *J. Chem. Theory Comput.*, 2018, **14**(11), 6076–6092.

23 L.-P. Wang, J. Chen and T. Van Voorhis, Systematic parametrization of polarizable force fields from quantum chemistry data, *J. Chem. Theory Comput.*, 2012, **9**(1), 452–460, DOI: **10.1021/ct300826t**.

24 L.-P. Wang, T. J. Martinez and V. S. Pande, Building force fields: An automatic, systematic, and reproducible approach, *J. Phys. Chem. Lett.*, 2014, **5**(11), 1885, DOI: **10.1021/jz500737m**.

25 Y. Qiu, P. S. Nerenberg, T. Head-Gordon and L.-P. Wang, Systematic optimization of water models using liquid/vapor surface tension data, *J. Phys. Chem. B*, 2019, **123**(32), 7061–7073, DOI: **10.1021/acs.jpcb.9b05455**.

26 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**(9), 1157–1174.

27 J. Wang, W. Wang, P. A. Kollman and D. A. Case, Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graphics Modell.*, 2006, **25**(2), 247–260.

28 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method, *J. Comput. Chem.*, 2000, **21**(2), 132–146.

29 A. Jakalian, D. B. Jack and C. I. Bayly, Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation, *J. Comput. Chem.*, 2002, **23**(16), 1623–1641.

30 J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, S. E. Carl and A. T. Hagler, Derivation of class ii force fields. i. methodology and quantum force field for the alkyl functional group and alkane molecules, *J. Comput. Chem.*, 1994, **15**(2), 162–182.

31 Mj J. Hwang, T. P. Stockfisch and A. T. Hagler, Derivation of class ii force fields. 2. derivation and characterization of a class ii force field, cff93, for the alkyl functional group and alkane molecules, *J. Am. Chem. Soc.*, 1994, **116**(6), 2515–2525.

32 P. Adam, S. Gross, F. Massa, L. Adam, B. James, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, R. Martin, A. Tejani, S. Chilamkurthy, B. Steiner, F. Lu, J. Bai and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox and R. Garnett, Curran Associates, Inc., 2019, 32, pp. 8024–8035, **http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf**.

33 M. Abadi, A. Agarwal, B. Paul, E. Brevdo, Z. Chen, C. Craig, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vincent, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, W. Martin, W. Martin, Y. Yuan and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, **https://www.tensorflow.org/.Softwareavailablefromtensorflow.org**.

34 B. James, F. Roy, P. Hawkins, M. J. C. L. Johnson, D. Maclaurin, N. George, P. Adam, J. VanderPlas, S. Wanderman-Milne and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, 2018, **http://github.com/google/jax**.

35 K. Xu, W. Hu, J. Leskovec and S. Jegelka, How powerful are graph neural networks?, arXiv preprint arXiv:1810.00826, 2018.

36 Y. Wang, J. Fass, C. D. Stern, K. Luo and J. Chodera, Graph nets for partial charge prediction, 2019, arXiv preprint arXiv:1909.07903.
37 T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, CoRR, abs/1609, 029072016http://arxiv.org/abs/1609.02907.
38 P. W. Battaglia, B. H. Jessica, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, S. Adam, F. Ryan, *et al.*, Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261, 2018.
39 J. Du, S. Zhang, G. Wu, M. Jose, F. Moura and S. Kar, Topology Adaptive Graph Convolutional Networks, arXiv:1710.10370 [cs, stat], 2018.
40 F. Wu, T. Zhang, A. Holanda de Souza Jr, C. Fifty, Y. Tao and K. Q. Weinberger, Simplifying graph convolutional networks, arXiv preprint arXiv:1902.07153, 2019.
41 M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, *et al.*, Deep graph library: A graph-centric, highly-performant package for graph neural networks, 2019, arXiv preprint arXiv:1909.01315.
42 Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein and J. M. Solomon, Dynamic graph cnn for learning on point clouds, *ACM Trans. Graph.*, 2019, **38**(5), 1–12.
43 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, Convolutional networks on graphs for learning molecular fingerprints, *Advances in neural information processing systems*, 2015, pp. 2224–2232.
44 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Neural message passing for quantum chemistry, *International conference on machine learning*, PMLR, 2017, pp. 1263–1272.
45 C. D. Stern, C. I. Bayly, D. G. A. Smith, J. Fass, L.-P. Wang, D. L. Mobley and J. D. Chodera, Capturing non-local through-bond effects when fragmenting molecules for quantum chemical torsion scans, *bioRxiv*, 2020.
46 T. W. Ko, J. A. Finkler, S. Goedecker and J. Behler, A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer, *Nat. Commun.*, 2021, **12**(1), 1–11.
47 K. Vanommeslaeghe, E. P. Raman and A. D. MacKerell Jr, Automation of the charmm general force field (cgenff) ii: assignment of bonded parameters and partial atomic charges, *J. Chem. Inf. Model.*, 2012, **52**(12), 3155–3168.
48 K. Vanommeslaeghe, E. P. Raman and A. D. MacKerell Jr, Automation of the charmm general force field (cgenff) ii: assignment of bonded parameters and partial atomic charges, *J. Chem. Inf. Model.*, 2012, **52**(12), 3155–3168.
49 B. Weisfeiler and A. Leman, *The reduction of a graph to canonical form and the algebra which appears therein*, 1968.
50 J. J. Irwin and K. BrianShoichet, Zinc- a free database of commercially available compounds for virtual screening, *J. Chem. Inf. Model.*, 2005, **45**(1), 177–182.
51 *An informal amber small molecule force field: parm@frosst*, 2010. http://www.ccl.net/cca/data/parm_at_Frosst/.
52 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**(9), 1157–1174, DOI: 10.1002/jcc.20035, https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035.
53 Y. Qiu, D. Smith, B. Simon, H. Jang, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. Stern, A. Rizzi, *et al.*, Development and benchmarking of open force field v1.0.0, the parsley small molecule force field, *J. Chem. Theory Comput.*, 2021, **17**(10), 6262–6280, DOI: 10.33774/chemrxiv-2021-l070l-v4.
54 gaff-1.81https://github.com/openmm/openmmforcefields/blob/0.9.0/amber/gaff/dat/gaff-1.81.dat#L87.
55 R. L. Murphy, B. Srinivasan, V. A. Rao and B. Ribeiro, Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs, CoRR, abs/1811.01900http://arxiv.org/abs/1811.01900, 2018.
56 H. Jang, J. Maat, Y. Qiu, G. Daniel.A. Smith, B. Simon, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, X. Lucas, *et al. openforcefield/openforcefields: Version 1.2.0 "Parsley" update*. 2020. DOI: 10.5281/zenodo.3872244.
57 T. A. Halgren, The representation of van der waals (vdw) interactions in molecular mechanics force fields: potential form, combination rules, and vdw parameters, *J. Am. Chem. Soc.*, 1992, **114**(20), 7827–7843.
58 C. M. Baker, P. E. M. Lopes, X. Zhu, B. Roux and A. D. MacKerell Jr, Accurate calculation of hydration free energies using pair-specific lennard-jones parameters in the charmm drude polarizable force field, *J. Chem. Theory Comput.*, 2010, **6**(4), 1181–1198.
59 P. Chatterjee, M. Y. Sengul, A. Kumar and A. D. MacKerell, Harnessing deep learning for optimization of lennard-jones parameters for the polarizable classical drude oscillator force field, *J. Chem. Theory Comput.*, 2022, **18**(4), 2388–2407, DOI: 10.1021/acs.jctc.2c00115. PMID: 35362975.
60 P. Ren and J. W. Ponder, Consistent treatment of inter-and intramolecular polarization in molecular mechanics calculations, *J. Comput. Chem.*, 2002, **23**(16), 1497–1506.
61 J. A. Lemkul, J. Huang, B. Roux and A. D. MacKerell Jr, An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications, *Chem. Rev.*, 2016, **116**(9), 4983–5013.
62 I. Leven and T. Head-Gordon, C-gem: Coarse-grained electron model for predicting the electrostatic potential in molecules, *J. Phys. Chem. Lett.*, 2019, **10**(21), 6820–6826.
63 J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, S. E. Carl and A. T. Hagler, Derivation of class ii force fields. i. methodology and quantum force field for the alkyl functional group and alkane molecules, *J. Comput. Chem.*, 1994, **15**(2), 162–182.
64 M. J. Hwang, T. P. Stockfisch and A. T. Hagler, Derivation of class ii force fields. 2. derivation and characterization of a class ii force field, cff93, for the alkyl functional group and alkane molecules, *J. Am. Chem. Soc.*, 1994, **116**(6), 2515–2525.
65 J. R. Maple, M.-J. Hwang, T. P. Stockfisch and A. T. Hagler, Derivation of class ii force fields. iii. characterization of

a quantum force field for alkanes, *Isr. J. Chem.*, 1994, **34**(2), 195–231.

66 T. Gokey, *Openff sandbox cho phalkethoh v1.0*, 2020, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-09-18-OpenFF-Sandbox-CHO-PhAlkEthOH.

67 C. C. Bannan and D. Mobley, *ChemPer: An Open Source Tool for Automatically Generating SMIRKS Patterns*, 2019, 6, DOI: 10.26434/chemrxiv.8304578.v1, https://chemrxiv.org/articles/preprint/ChemPer_An_Open_Source_Tool_for_Automatically_Generating_SMIRKS_Patterns/8304578.

68 Openff phenyl dataset, 2020 https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-10-06-OpenFF-Phenyl-Set.

69 J. Wagner, D. L. Mobley, M. Thompson, J. Chodera, C. Bannan, A. Rizzi, T. Gokey, D. Dotson, J. Rodríguez-Guerra, C. Zanette, B. Pavan, C. Bayly, A. Josh, J. H. Mitchell, N. M. Lim, V. Lim, S. Sasmal, L. Wang, A. Dalke, S. Boothroyd, I. Pulido, D. Smith, L.-P. Wang and Y. Zhao, *openforcefield/openff-toolkit: 0.10.0 Improvements for force field fitting*, 2021, DOI: 10.5281/zenodo.5153946.

70 D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. V. T. L. Chodera, N. M. Lim, K. A. Beauchamp, M. R. Shirts, M. K. Gilson, *et al.*, Open force field consortium: Escaping atom types using direct chemical perception with smirnoff v0. 1, *BioRxiv*, 2018, 286542.

71 D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard and T. D. Crawford, *The molssi qcarchive project: An open-source platform to compute, organize, and share quantum chemistry data*, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2020, p. e1491.

72 Openff sandbox gen2 optimization dataset, 2020 https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-1-Roche, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-2-Coverage, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-3-Pfizer-Discrepancy, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-4-eMolecules-Discrepancy, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-03-20-OpenFF-Gen-2-Optimization-Set-5-Bayer.

73 Y. Qiu, D. Smith, B. Simon, H. Jang, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. Stern, A. Rizzi, *et al.*, *Development and benchmarking of open force field v1. 0.0, the parsley small molecule force field*, 2021.

74 Vehicle dataset, 2020, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2019-07-02VEHICLeoptimizationdataset.

75 W. R. Pitt, D. M. Parry, B. G. Perry and C. R. Groom, Heteroaromatic rings of the future, *J. Med. Chem.*, 2009, **52**(9), 2952–2963, DOI: 10.1021/jm801513z. PMID: 19348472.

76 Pepconf dataset, 2020, https://github.com/openforcefield/qca-dataset-submission/tree/master/submissions/2020-10-26-PEPCONF-Optimization.

77 V. K. Prasad, A. O. Roza and A. D. Gino, *Pepconf, a diverse data set of peptide conformational energies*, 2019, DOI: DOI: 10.1038/sdata.2018.310.

78 J. Horton, *openforcefield/openff-qcsubmit: 0.3.1*, 2022, doi: DOI: 10.5281/zenodo.6338096.

79 J. Wagner, M. Thompson, D. Dotson, Hyejang and J. Rodríguez-Guerra, *openforcefield/openforcefields, Version 1.2.1 "Parsley" Update*, 2020, DOI: DOI: 10.5281/zenodo.4021623.

80 D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, R. H. A. W. Goetz, S. Izadi, S. A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, L. Wilson, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York and P. A. Kollman, *Amber*, 2020, 2020.

81 V. Prasad, A. Otero-de La-Roza and G. DiLabio, Pepconf, a diverse data set of peptide conformational energies, *Sci. Data*, 2019, **6**, 180310.

82 Y. Qiu, D. G. A. Smith, C. D. Stern, M. Feng, H. Jang and L.-P. Wang, Driving torsion scans with wavefront propagation, *J. Chem. Phys.*, 2020, **152**, 244116.

83 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.*, 1995, **117**(19), 5179–5197.

84 Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, *et al.*, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *J. Comput. Chem.*, 1999, **24**(16), 20122003.

85 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell Jr, Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles, *J. Chem. Theory Comput.*, 2012, **8**(9), 3257–3273.

86 D. S. Cerutti, J. E. Rice, W. C. Swope and D. A. Case, Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization, *J. Phys. Chem. B*, 2013, **117**(8), 2328–2338.

87 K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case and L. T. Chong, Further along the road less traveled: Amber ff15ipq, an original protein force field

built on a self-consistent physical model, *J. Chem. Theory Comput.*, 2016, **12**(8), 3926–3947.

88 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model, *J. Phys. Chem.*, 1993, **97**(40), 10269–10280.

89 M. Schauperl, S. N. Paul, H. Jang, L.-P. Wang, C. I. Bayly, D. L. Mobley and M. K. Gilson, Non-bonded force field model with advanced restrained electrostatic potential charges (resp2), *Commun. Chem.*, 2020, **3**(1), 1–11.

90 G. A. Khoury, J. P. Thompson, S. James, C. A. Kieslich and C. A. Floudas, Forcefield_ptm: *Ab initio* charge and amber forcefield parameters for frequently occurring post-translational modifications, *J. Chem. Theory Comput.*, 2013, **9**(12), 5653–5674.

91 K. Atz, C. Isert, N. Markus, A. Böcker, J. Jiménez-Luna and G. Schneider, Delta-quantum machine learning for medicinal chemistry, *ChemRxiv*, 2021, DOI: **10.26434/chemrxiv-2021-fz6v7-v2**.

92 P. Bleiziffer, K. Schaller and S. Riniker, Machine learning of partial charges derived from high-quality quantum-mechanical calculations, *J. Chem. Inf. Model.*, 2018, **58**(3), 579–590.

93 N. Lubbers, J. S. Smith and K. Barros, Hierarchical modeling of molecular energies using a deep neural network, *J. Chem. Phys.*, 2018, **148**, 241715.

94 A. E. Sifain, N. Lubbers, B. T. Nebgen, J. S. Smith, A. Y. Lokhov, O. Isayev, A. E. Roitberg, K. Barros and S. Tretiak, Discovering a transferable charge assignment model using machine learning, *J. Phys. Chem. Lett.*, 2018, **9**(16), 4495–4501.

95 A. K. Rappe and W. A. Goddard III, Charge equilibration for molecular dynamics simulations, *J. Phys. Chem.*, 1991, **95**(8), 3358–3363.

96 W. J. Mortier, S. K. Ghosh and S. Shankar, Electronegativity-equalization method for the calculation of atomic charges in molecules, *J. Am. Chem. Soc.*, 1986, **108**(15), 4315–4320, DOI: **10.1021/ja00275a013**.

97 M. K. Gilson, H. S. R. Gilson and M. J. Potter, Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms, *J. Chem. Inf. Comput. Sci.*, 1982, **43**(6), 19972003.

98 P. Eastman and V. Pande, Accelerating development and execution speed with just-in-time gpu code generation, *GPU Computing Gems Jade Edition*, Elsevier, 2012, pp. 399–407.

99 L.-P. Wang, T. J. Martinez and V. S. Pande, Building force fields: An automatic, systematic, and reproducible approach, *J. Phys. Chem. Lett.*, 2014, **5**(11), 1885–1891.

100 L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, *et al.*, Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field, *J. Am. Chem. Soc.*, 2015, **137**(7), 2695–2703.

101 L. F. Song, T.-S. Lee, C. Zhu, D. M. York and K. M. Merz Jr, Using amber18 for relative free energy calculations, *J. Chem. Inf. Model.*, 2019, **59**(7), 3128–3135.

102 J. Liang, A. van Abbema, M. Balazs, K. Barrett, B. Leo, B. Wade, C. Chang, D. Delarosa, J. DeVoss, J. Driscoll, *et al.*, Lead optimization of a 4-aminopyridine benzamide scaffold to identify potent, selective, and orally bioavailable tyk2 inhibitors, *J. Med. Chem.*, 2013, **56**(11), 4521–4536.

103 J. Chodera, A. Rizzi, L. Naden, K. Beauchamp, P. Grinaway, J. Fass, A. Wade, B. Rustenburg, G. A. Ross, A. Krämer, H. B. Macdonald, J. Rodríguez-Guerra, M. Henry, A. Simmonett, D. W. H. Swenson, I. Pulido, S. Roet, M. J. Williamson, S. Boothroyd, A. Silveira, and Dominicrufa, *choderalab/openmmtools: Bugfix release v0.21.2*, 2022, DOI: DOI: **10.5281/zenodo.6260174**.

104 Y. Qiu, D. Smith, B. Simon, H. Jang, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. Stern, A. Rizzi, *et al.*, *Development and benchmarking of open force field v1. 0.0, the parsley small molecule force field*, 2020.

105 A. S. J. S. Mey, B. Allen, H. E. B. Macdonald, J. D. Chodera, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, *et al.*, Best practices for alchemical free energy calculations, 2020, arXiv preprint arXiv:2008.03067.

106 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb, *J. Chem. Theory Comput.*, 2015, **11**(8), 3696–3713.

107 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.*, 1983, **79**(2), 926–935.

108 S. Joung and T. E. Cheatham III, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations, *J. Phys. Chem. B*, 2008, **112**(30), 9020–9041.

109 J. Chodera, A. Rizzi, N. Levi, K. Beauchamp, P. Grinaway, J. Fass, A. Wade, B. Rustenburg, A. R. Gregory, A. Krämer, H. B. Macdonald, J. Rodríguez-Guerra, dominicrufa, A. Simmonett, D. W. H. Swenson, M. Henry, S. Roet and A. Silveira, *Choderalab/openmmtools: 0.20.3 Bugfix Release*, Zenodo, 2021.

110 F. D. Hahn and J. Wagner, *openforcefield/protein-ligand-benchmark: 0.1.2 Release to create Zenodo record*, 2021, DOI: DOI: **10.5281/zenodo.4813735**.

111 J. Chodera, R. Wiewiora, C. Stern and P. Eastman, *openmm/openmm-forcefields: Fix GAFF AM1-BCC charging bug for some molecules*, 2020, DOI: DOI: **10.5281/zenodo.3627391**.

112 H. Xu, Optimal measurement network of pairwise differences, *J. Chem. Inf. Model.*, 2019, **59**(11), 4720–4728.

113 H. B. Macdonald, M. Henry, J. Chodera, D. Dotson, W. Glass and I. Pulido, *openforcefield/openff-arsenic: v0.2.1*, 2022, DOI: DOI: **10.5281/zenodo.6210305**.

114 K. Vanommeslaeghe and A. D. MacKerell Jr, Automation of the charmm general force field (cgenff) i: bond perception

and atom typing, *J. Chem. Inf. Model.*, 2012, **52**(12), 3144–3154.

115 K. Vanommeslaeghe, E. P. Raman and A. D. MacKerell Jr, Automation of the charmm general force field (cgenff) ii: assignment of bonded parameters and partial atomic charges, *J. Chem. Inf. Model.*, 2012, **52**(12), 3155–3168.

116 J. D. Yesselman, D. J. Price, J. L. Knight and C. L. Brooks III, Match: An atom-typing toolset for molecular mechanics force fields, *J. Comput. Chem.*, 2012, **33**(2), 189–202.

117 M. Waldman and A. T. Hagler, New combining rules for rare gas van der waals parameters, *J. Comput. Chem.*, 1993, **14**(9), 1077–1084.

118 J. P. Toennies, On the validity of a modified buckingham potential for the rare gas dimers at intermediate distances, *Chem. Phys. Lett.*, 1973, **20**(3), 238–241.

119 T. Darden, D. York and P. Lee, Particle mesh ewald: An nlog (n) method for ewald sums in large systems, *J. Chem. Phys.*, 1993, **98**(12), 10089–10092.

120 A. Kubincová, S. Riniker and P. H. Hünenberger, Reaction-field electrostatics in molecular dynamics simulations: Development of a conservative scheme compatible with an atomic cutoff, *Phys. Chem. Chem. Phys.*, 2020, **22**(45), 26419–26437.

121 O. A. Von Lilienfeld, Quantum machine learning in chemical compound space, *Angew. Chem., Int. Ed.*, 2018, **57**(16), 4164–4169.

122 O. T. Unke and M. Meuwly, Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges, *J. Chem. Theory Comput.*, 2019, **15**(6), 3678–3693.

123 A. Grossfield and D. M. Zuckerman, Quantifying uncertainty and sampling quality in biomolecular simulations, *Annu. Rep. Comput. Chem.*, 2009, **5**, 23–48.

124 J. D. Chodera, A simple method for automated equilibration detection in molecular simulations, *J. Chem. Theory Comput.*, 2016, **12**(4), 1799–1805.

125 A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. W. Siderius and D. M. Zuckerman, Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1. 0], *Living J. Comp. Mol. Sci.*, 2018, **1**(1), 5067.

126 F. Cailliez and P. Pascal, Statistical approaches to forcefield calibration and prediction uncertainty in molecular simulation, *J. Chem. Phys.*, 2011, **134**(5), 054124, DOI: **10.1063/1.3545069**, ISSN 0021-9606.

127 F. Cailliez, B. Arnaud and P. Pascal, Calibration of forcefields for molecular simulation: Sequential design of computer experiments for building cost-efficient kriging metamodels, *J. Comput. Chem.*, 2014, **35**(2), 130–149, DOI: **10.1002/jcc.23475**, ISSN 1096-987X.

128 P. Angelikopoulos, C. Papadimitriou and P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework, *J. Chem. Phys.*, 2012, **137**(14), 144103, DOI: **10.1063/1.4757266**, ISSN 0021-9606.

129 P. E. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou and P. Koumoutsakos, Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models, *J. Comput. Phys.*, 2015, **284**(1–21), DOI: **10.1016/j.jcp.2014.12.006**), ISSN 0021-9991.

130 B. Cooke and S. C. Schmidler, Statistical Prediction and Molecular Dynamics Simulation, *Biophys. J.*, 2008, **95**(10), 4497–4511, DOI: **10.1529/biophysj.108.131623**, ISSN 0006-3495.

131 X. Zhou and S. M. Foiles. *Uncertainty Quantification and Reduction of Molecular Dynamics Models*, 2017, DOI: DOI: **10.5772/intechopen.68507**.

132 F. Rizzi, H. Najm, B. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson and O. Knio, Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-Field Parameters, *Multiscale Model. Simul.*, 2012, **10**(4), 1460–1492, DOI: **10.1137/110853170**, ISSN 1540-3459.

133 S. Wu, P. Angelikopoulos, C. Papadimitriou, R. Moser and P. Koumoutsakos, A hierarchical Bayesian framework for force field selection in molecular dynamics simulations, *Philos. Trans. R. Soc., A*, 2016, **374**(2060), 20150032, DOI: **10.1098/rsta.2015.0032**, ISSN 1364-503X, 1471-2962.

134 L. Kulakova, G. Arampatzis, P. Angelikopoulos, P. Chatzidoukas, C. Papadimitriou and P. Koumoutsakos, Experimental data over quantum mechanics simulations for inferring the repulsive exponent of the Lennard-Jones potential in Molecular Dynamics, 2017, arXiv:1705.08533 [physics, stat].

135 N. P. Patrone and A. Dienstfrey. Uncertainty Quantification for Molecular Dynamics, 2018, arXiv:1801.02483 [physics].

136 S. Wu, P. Angelikopoulos, G. Tauriello, C. Papadimitriou and P. Koumoutsakos, Fusing heterogeneous data for the calibration of molecular dynamics force fields using hierarchical Bayesian models, *J. Chem. Phys.*, 2016, **145**(24), 244112, DOI: **10.1063/1.4967956**, ISSN 0021-9606.

137 R. A. Messerly, T. A. Knotts and W. V. Wilding, Uncertainty quantification and propagation of errors of the Lennard-Jones 12-6 parameters for n-alkanes, *J. Chem. Phys.*, 2017, **146**(19), 194110, DOI: **10.1063/1.4983406**, ISSN 0021-9606.

138 C. OwenMadin, S. Boothroyd, R. A. Messerly, J. D. Chodera, J. Fass and M. R. Shirts, Bayesian inference-driven model parameterization and model selection for 2CLJQ fluid models, 2021, arXiv:2105.07863 [physics, stat].

139 C. Zanette, C. C. Bannan, C. I. Bayly, J. Fass, M. K. Gilson, M. R. Shirts, J. D. Chodera and D. L. Mobley, Toward Learned Chemical Perception of Force Field Typing Rules, *J. Chem. Theory Comput.*, 2019, **15**(1), 402–423, DOI: **10.1021/acs.jctc.8b00821**, ISSN 1549-9618.

140 M. R. Neal, *Bayesian Learning for Neural Networks, Lecture Notes in Statistics*, Springer-Verlag, New York, 1996, ISBN no. 978-0-387-94724-2, DOI: DOI: **10.1007/978-1-4612-0745-0**.

141 D. K. Shenfeld, H. Xu, M. P. Eastwood, R. O. Dror and D. E. Shaw, Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **80**, 046705, DOI: **10.1103/PhysRevE.80.046705**.

142 G. E. Crooks, Measuring thermodynamic length, *Phys. Rev. Lett.*, 2007, **99**, 100602, DOI: **10.1103/PhysRevLett.99.100602**.