LA-UR-24-26124

Title: Improving Atomistic Simulations With Machine Learning

Author(s): Allen, Alice Elisabeth Anastasia

Intended for: Seminar at Heidelberg Institute for Theoretical Studies

Issued: 2024-06-24

Los Alamos
NATIONAL LABORATORY

Improving Atomistic Simulations With Machine Learning

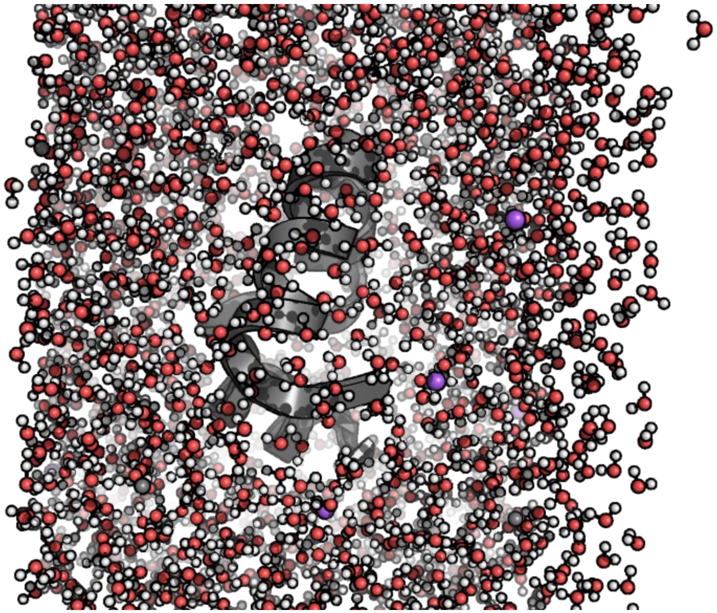
Alice Allen

Los Alamos NATIONAL LABORATORY

Center for Nonlinear Studies

Introduction

What are atomistic simulations and interatomic potentials?



How does machine learning fit in?

How can interatomic potentials be improved?

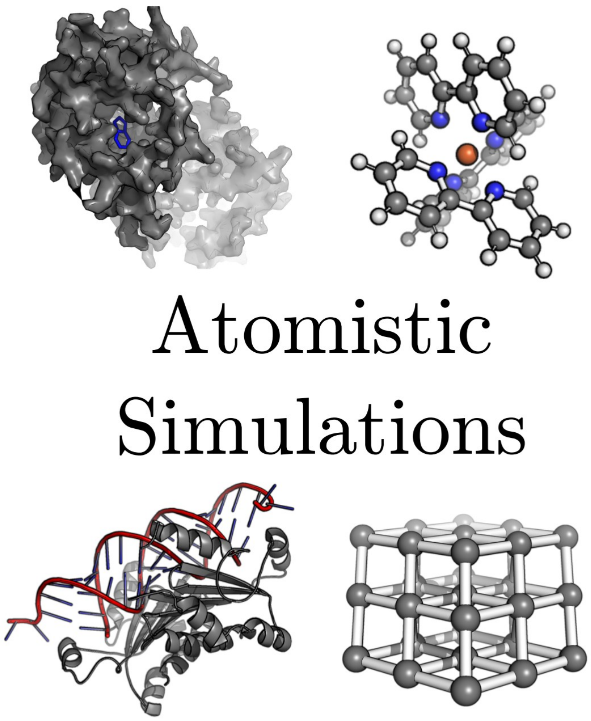
Atomistic Simulation

Atomistic simulations can be used to determine material and chemical properties

Simulating how atoms and molecules behave in a given system

For example:

- Understand reactive systems
- Exploring phases of materials



Protein binding to a drug

Transition Metal Complex

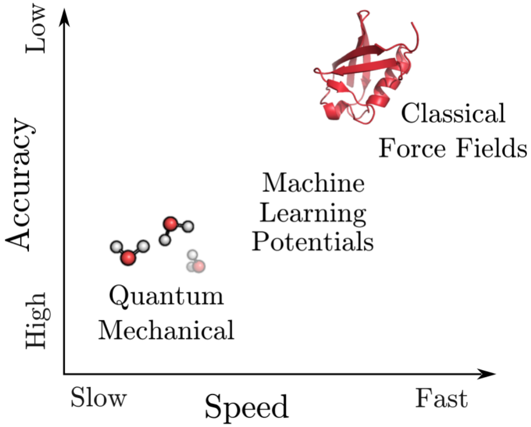
Atomistic Simulations

Biological

Material

Interatomic Potentials

Potentials provide forces and energies



Trade off between speed and accuracy

Classical models use well defined functional form

$$E(\mathbf{x}) = \sum_{\text{bonds}} k_b(r - r_e)^2 + \sum_{\text{angle}} k_\theta(\theta - \theta_e)^2 + \sum_{\text{dihedrals}} k_\chi(1 + \cos n\chi - \delta) + \sum_{\text{impropers}} k_{imp}(\phi - \phi_0)^2 + \sum_{\substack{\textit{nonbonded}\\ \textit{pairs}}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \frac{q_i q_j}{\epsilon_l r_{ij}}$$

How can interatomic potentials be improved?

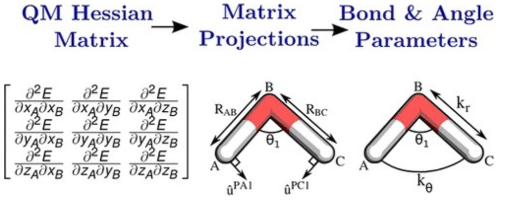
Classical Examples

$$
\begin{aligned}
E(\mathbf{x}) \; = \; & \sum_{\text{bonds}} k_b (r - r_e)^2 + \sum_{\text{angle}} k_\theta (\theta - \theta_e)^2 \\
+ & \sum_{\text{dihedrals}} k_\chi (1 + \cos n\chi - \delta) + \sum_{\text{impropers}} k_{imp} (\phi - \phi_0)^2 \\
& + \sum_{\substack{nonbonded \\ pairs}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] + \frac{q_i q_j}{\epsilon_l r_{ij}}
\end{aligned}
$$

How can classical potentials be improved?

Improve Parameters

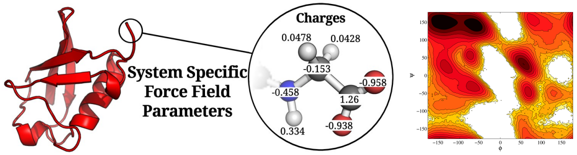
Better bond and angle terms



QM Hessian Matrix
Matrix Projections
Bond & Angle Parameters
$\frac{\partial^2 E}{\partial x_A \partial x_B}$ $\frac{\partial^2 E}{\partial x_A \partial y_B}$ $\frac{\partial^2 E}{\partial x_A \partial z_B}$
$\frac{\partial^2 E}{\partial y_A \partial x_B}$ $\frac{\partial^2 E}{\partial y_A \partial y_B}$ $\frac{\partial^2 E}{\partial y_A \partial z_B}$
$\frac{\partial^2 E}{\partial z_A \partial x_B}$ $\frac{\partial^2 E}{\partial z_A \partial y_B}$ $\frac{\partial^2 E}{\partial z_A \partial z_B}$
B
R_{AB}
R_{BC}
θ_1
A
C
$\hat{u}^{PA1}$
$\hat{u}^{PC1}$
B
k_r
θ_1
A
C
k_θ

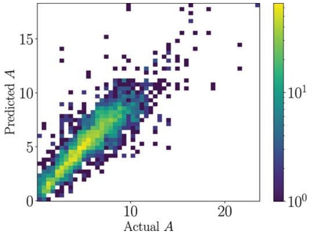
Harmonic force constants for molecular mechanics force fields via Hessian matrix projection, AEA Allen, MC Payne, DJ Cole, 2018, JCTC

System specific parameters



Development and validation of the quantum mechanical bespoke protein force field, AEA Allen, MJ Robertson, MC Payne, DJ Cole, 2019, ACS Omega

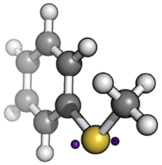
ML of Parameters

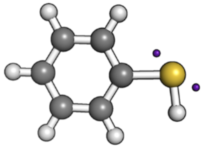


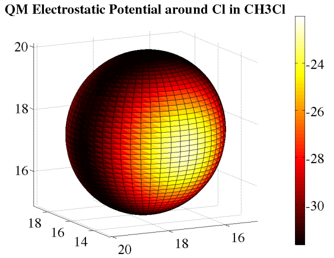
Toward transferable empirical valence bonds: Making classical force fields reactive, AEA Allen, Gabor Csanyi;, J. Chem. Phys, 2024

Improve Functional Form

Improve description of electrostatics

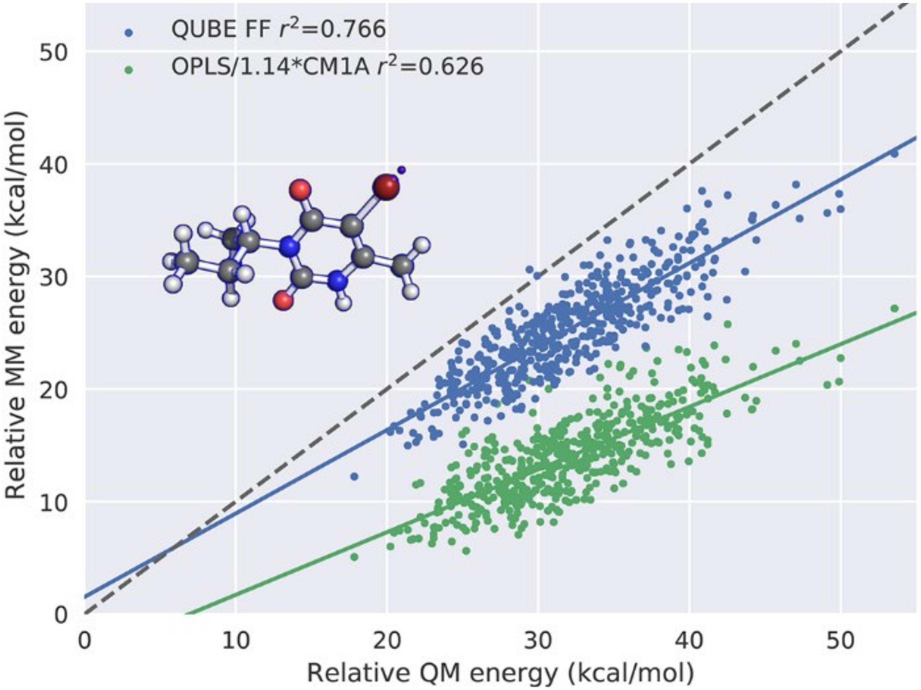






QUBEKit: Automating the derivation of force field parameters from quantum mechanics, JT Horton, AEA Allen, LS Dodda, DJ Cole, 2019, Journal of chemical information and modeling

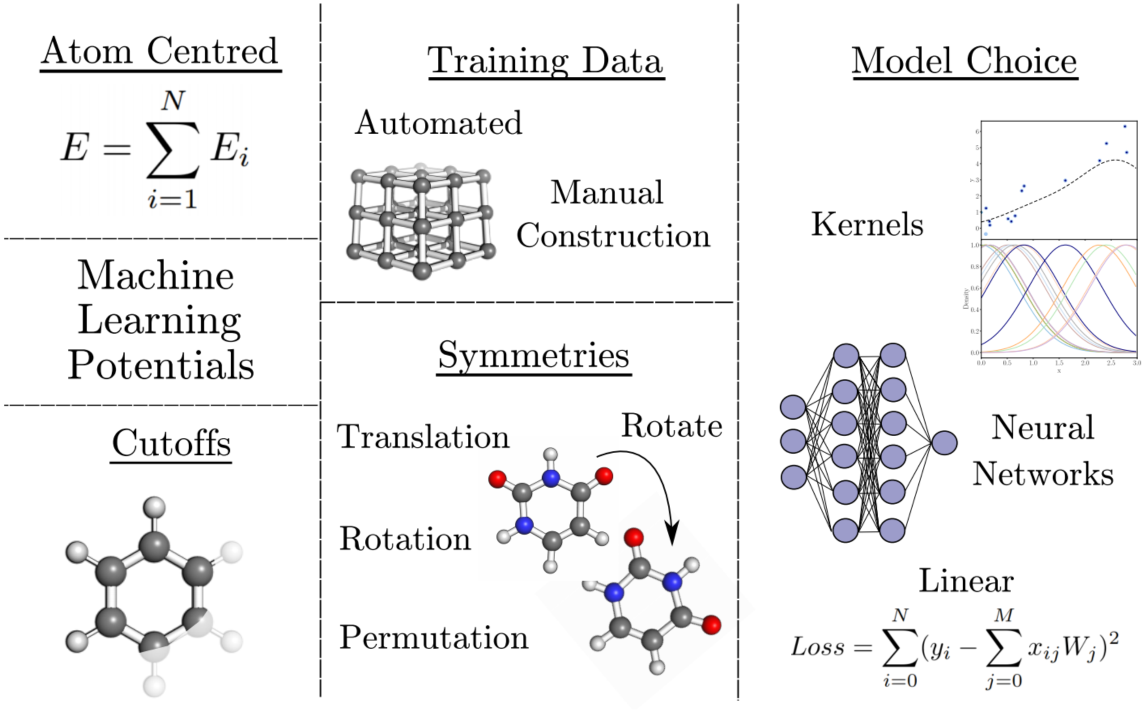
Limitations of Classical Force Fields



QUBEKit: Automating the derivation of force field parameters from quantum mechanics, JT Horton, AEA Allen, LS Dodda, DJ Cole, 2019, Journal of chemical information and modeling

How can interatomic potentials be improved?

Machine Learning



Atom Centred

$$E = \sum_{i=1}^{N} E_i$$

Machine Learning Potentials

Cutoffs

Training Data

Automated

Manual Construction

Symmetries

Translation

Rotation

Permutation

Rotate

Model Choice

Kernels

Neural Networks

Linear

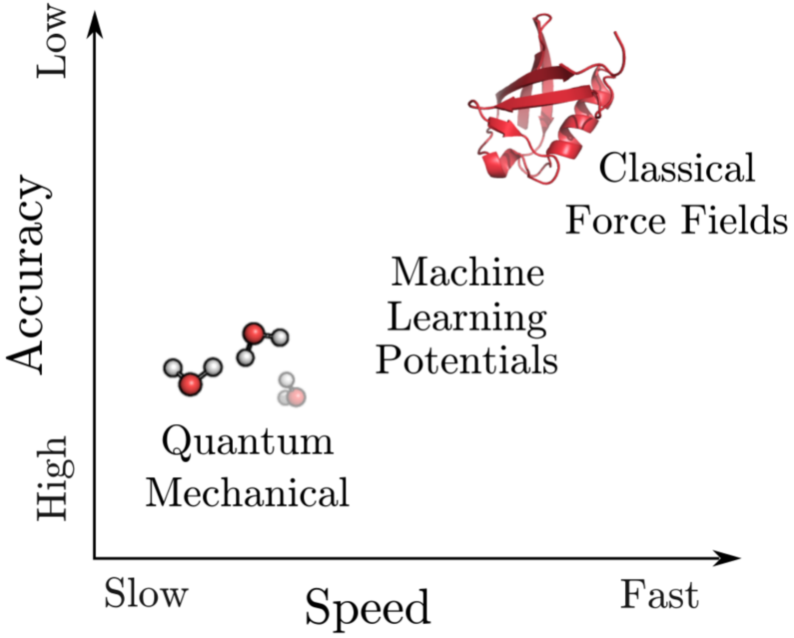
$$Loss = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2$$

Why do we need Machine Learning Potentials?

Potentials can provide forces and energies

Problem suits machine learning as the potential energy surface is highly complex and requires a flexible functional form

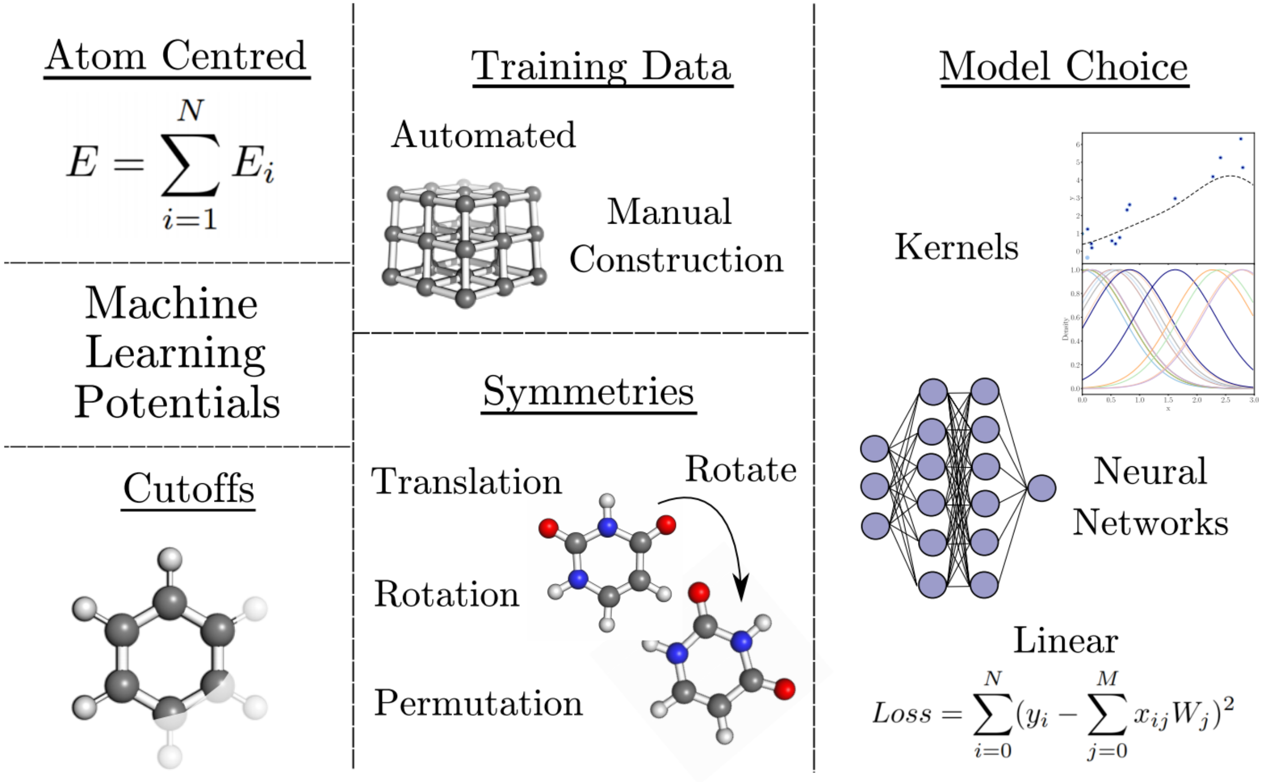
Lots of data can be supplied for the problem



Machine Learning Potentials

The flexible functional form allows for increased accuracy.

Multiple considerations for these forms of models.



Atom Centred

$$E = \sum_{i=1}^{N} E_i$$

Machine Learning Potentials

Cutoffs

Training Data

Automated

Manual Construction

Symmetries

Translation

Rotation

Permutation

Rotate

Model Choice

Kernels

Neural Networks

Linear

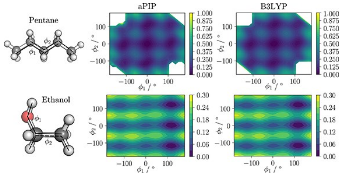
$$Loss = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2$$

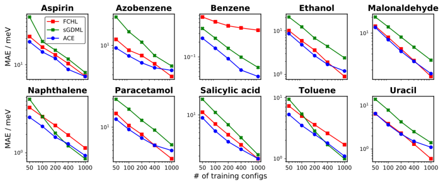
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

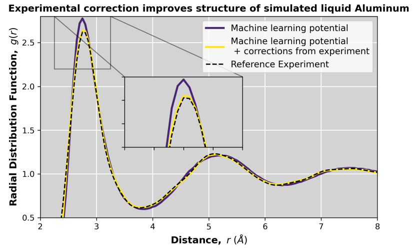
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





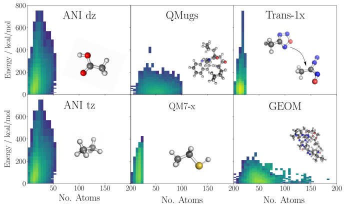
Training Data

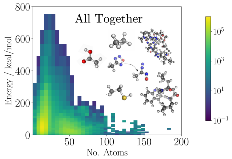
Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



Fitting Methods

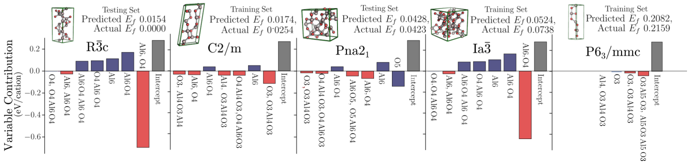
Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024





Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022

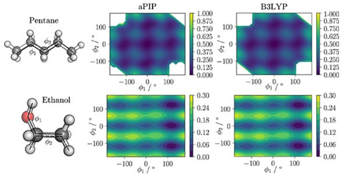


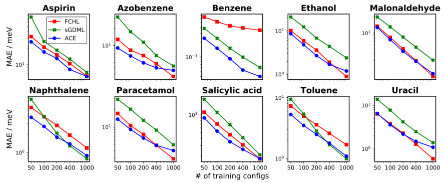
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

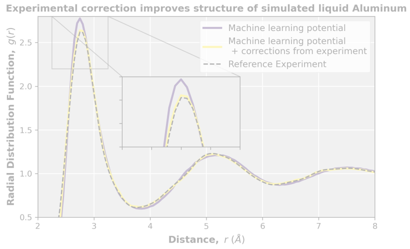
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





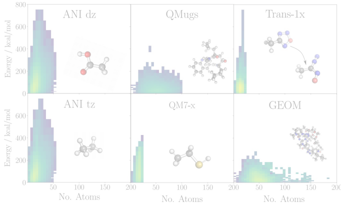
Training Data

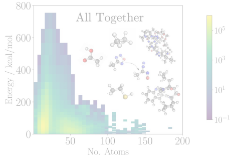
Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



Fitting Methods

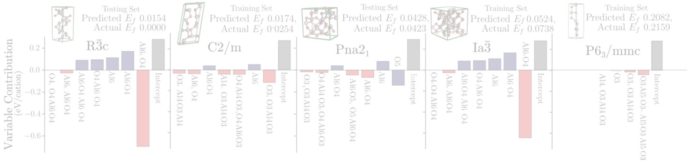
Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024





Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022

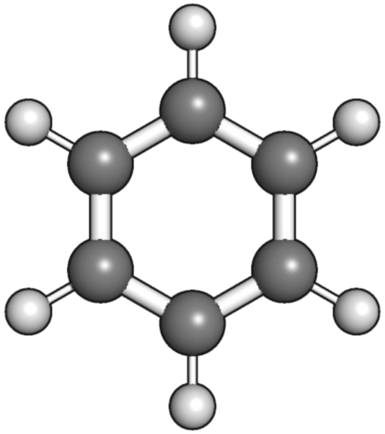


Architecture

- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

    - Rigid translations

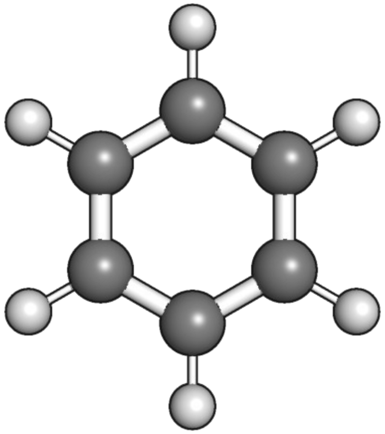


Architecture

- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

    - Rigid translations



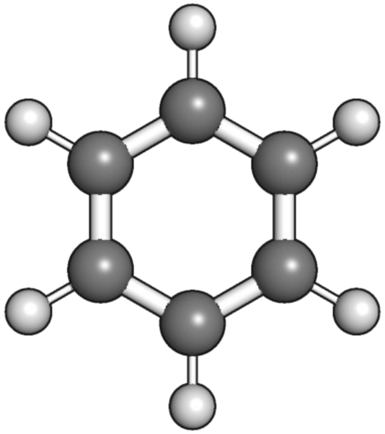
Architecture

- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

    - Rigid translations

    - Rotations and reflections



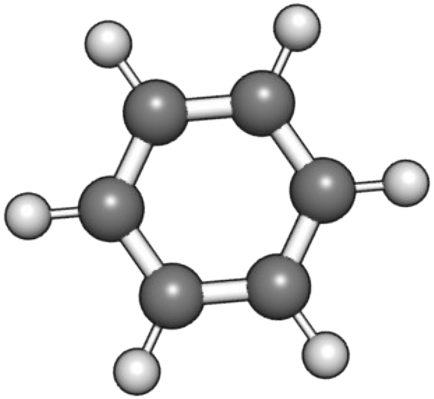
Architecture

- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

    - Rigid translations

    - Rotations and reflections



Architecture

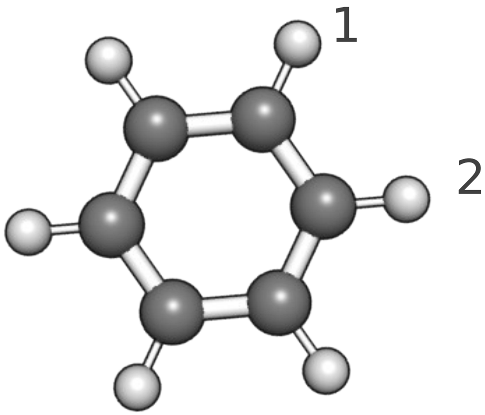
- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

  - Rigid translations

  - Rotations and reflections

  - Permutation of like atoms



Architecture

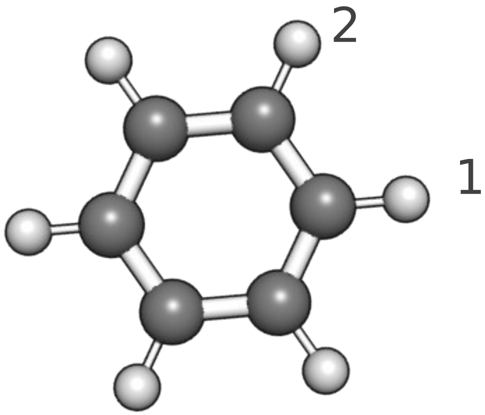
- Interatomic potentials are a regression problem with a regression model and representation chosen

- The representation chosen must satisfy the following properties:

  - Rigid translations

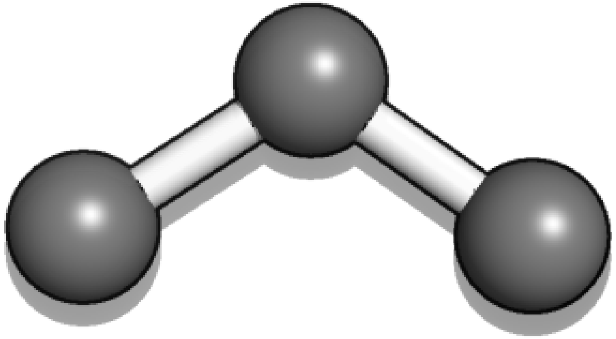
  - Rotations and reflections

  - Permutation of like atoms



Architecture - aPIPs

- Three identical atoms example for PIPs:



Rotational Invariance

Permutational Invariance

$$E(R_1, R_2, R_3) = E(r_{12}, r_{13}, r_{23}) = E(I_1, I_2, I_3)$$

$$I_1 = r_{12} + r_{13} + r_{23}$$

$$I_2 = r_{12}r_{13} + r_{12}r_{23} + r_{13}r_{23}$$

$$I_3 = r_{12}r_{13}r_{23}$$

$$E(R_1, R_2, R_3) = P(I_1, I_2, I_3) = \sum_k c_k I_1^{k_1} I_2^{k_2} I_3^{k_3}$$

- These ideas can then be extended to more atoms and to the non-identical atoms case.

Architecture - aPIPs

- PIPs are a polynomial construction with the underlying invariances present in the basis used.

- aPIPs retains the polynomial construction

$$E = \sum_{i=1}^{N} E_i$$

- But moves to an atom centred model

- Uses a body-ordered approached:

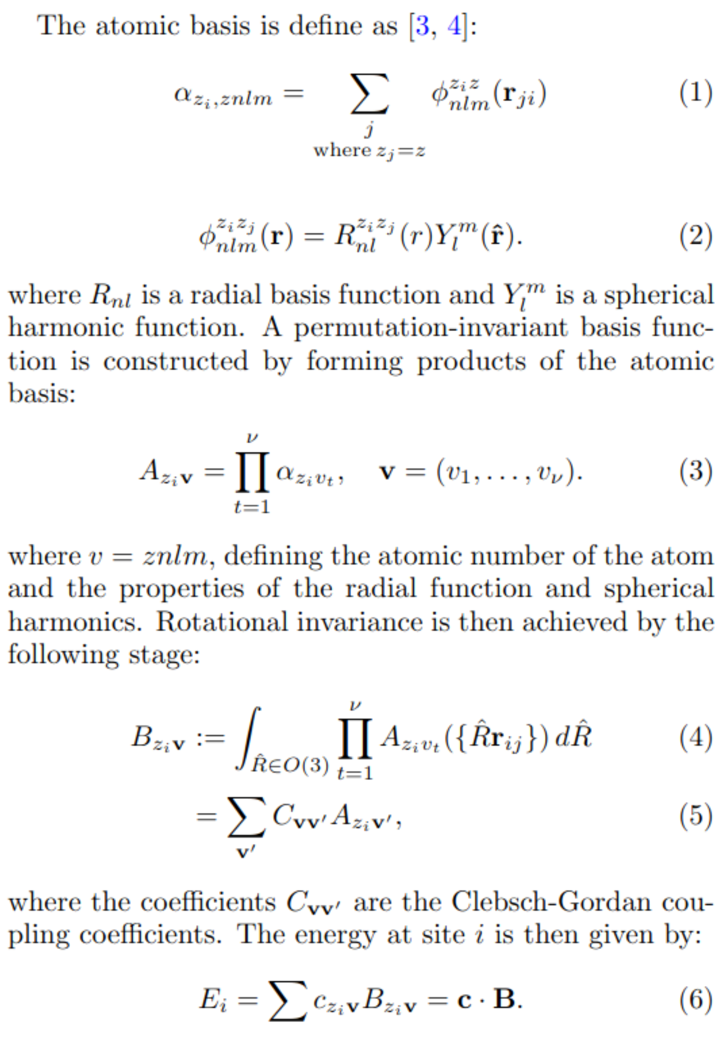
$$P(I_1, I_2, I_3) = \sum_k c_k I_1^{k_1} I_2^{k_2} I_3^{k_3}$$

$$E(\mathbf{R}) = \sum_i E_1 + \sum_{i<j} E_2(r_{ij}) + \sum_{i<j<k} E_3(r_{ij}, r_{ik}, r_{jk}) + \sum_{i<j<k<l} E_4(r_{ij}, r_{ik}, r_{il}, r_{jk}, r_{jl}, r_{kl}) + \ldots$$

Cas van der Oord, Geneviève Dusson, Gábor Csányi and Christoph Ortner, Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials, (2020). Machine Learning: Science and Technology

Allen, Alice E A, Dusson, Geneviève, Ortner, Christoph, Csányi, Gábor, Atomic permutationally invariant polynomials for fitting molecular force fields, (2021). Machine Learning: Science and Technology

Architecture - ACE



The atomic basis is define as [3, 4]:

$$\alpha_{z_i,znlm} = \sum_{\substack{j \\ \text{where } z_j = z}} \phi^{z_i z}_{nlm}(\mathbf{r}_{ji}) \qquad (1)$$

$$\phi^{z_i z_j}_{nlm}(\mathbf{r}) = R^{z_i z_j}_{nl}(r) Y^m_l(\hat{\mathbf{r}}). \qquad (2)$$

where R_{nl} is a radial basis function and Y^m_l is a spherical harmonic function. A permutation-invariant basis function is constructed by forming products of the atomic basis:

$$A_{z_i \mathbf{v}} = \prod_{t=1}^{\nu} \alpha_{z_i v_t}, \quad \mathbf{v} = (v_1, \ldots, v_\nu). \qquad (3)$$

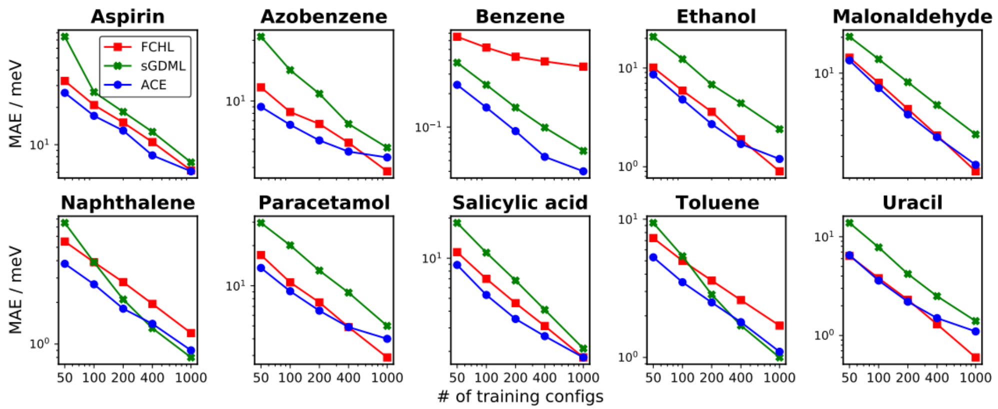
where $v = znlm$, defining the atomic number of the atom and the properties of the radial function and spherical harmonics. Rotational invariance is then achieved by the following stage:

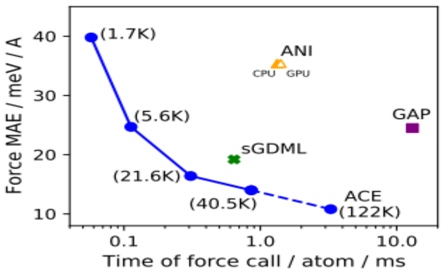
$$B_{z_i \mathbf{v}} := \int_{\hat{R} \in O(3)} \prod_{t=1}^{\nu} A_{z_i v_t}(\{\hat{R}\mathbf{r}_{ij}\})\, d\hat{R} \qquad (4)$$

$$= \sum_{\mathbf{v}'} C_{\mathbf{v}\mathbf{v}'} A_{z_i \mathbf{v}'}, \qquad (5)$$

where the coefficients $C_{\mathbf{v}\mathbf{v}'}$ are the Clebsch-Gordan coupling coefficients. The energy at site i is then given by:

$$E_i = \sum c_{z_i \mathbf{v}} B_{z_i \mathbf{v}} = \mathbf{c} \cdot \mathbf{B}. \qquad (6)$$

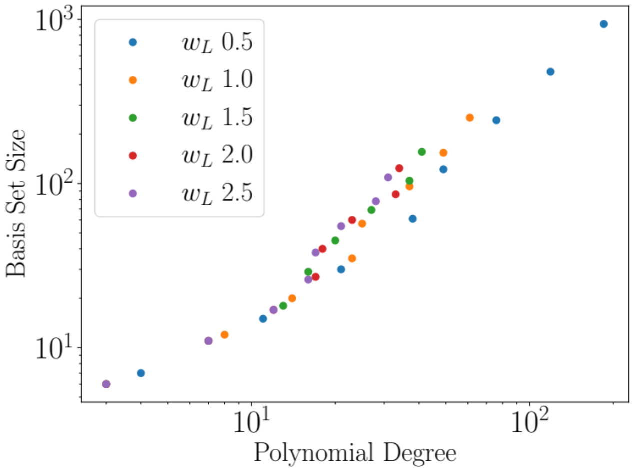


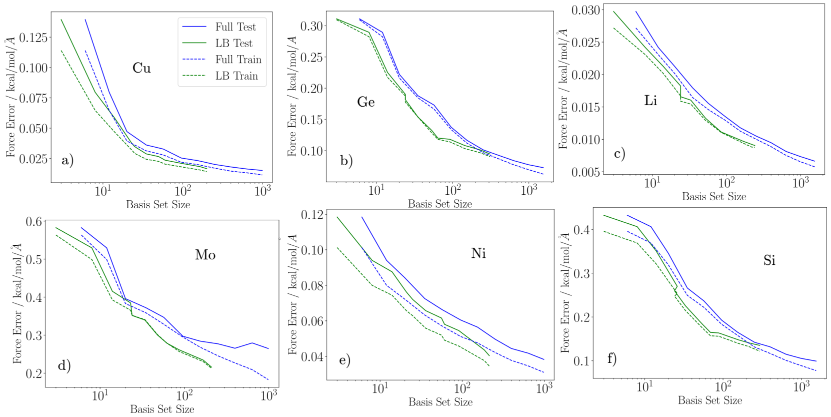


Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE. JCTC, 2022

Architecture - ACE

- Construct matrix of $\frac{\partial B_i}{\partial \alpha_j}$ for one atom (either from training dataset or assigned random neighbours). The matrix $\frac{\partial B_i}{\partial \alpha_j}$ is calculated numerically.
- Initialize a matrix with columns equivalent to the number of atomic basis functions ($\alpha_{z_i,znlm}$) and no rows.
- For each B_i until the utmost number of B_i terms is reached: append the row vector of the partial derivatives of the subsequent invariant into the matrix if it augments the rank of the matrix. From this, constructing an index of B_i components that increase rank.
- Repeat this N times, constructing an index of all B_i components that increase rank for any datapoint present.



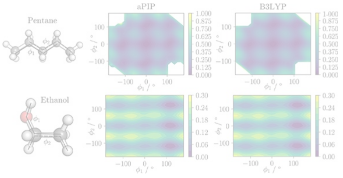


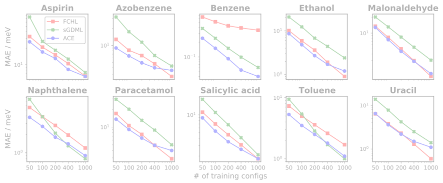
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

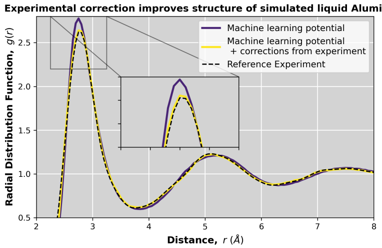
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





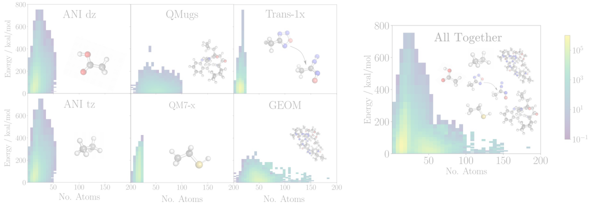
Training Data

Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



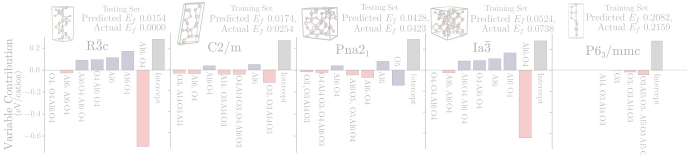
Fitting Methods

Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024



Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022

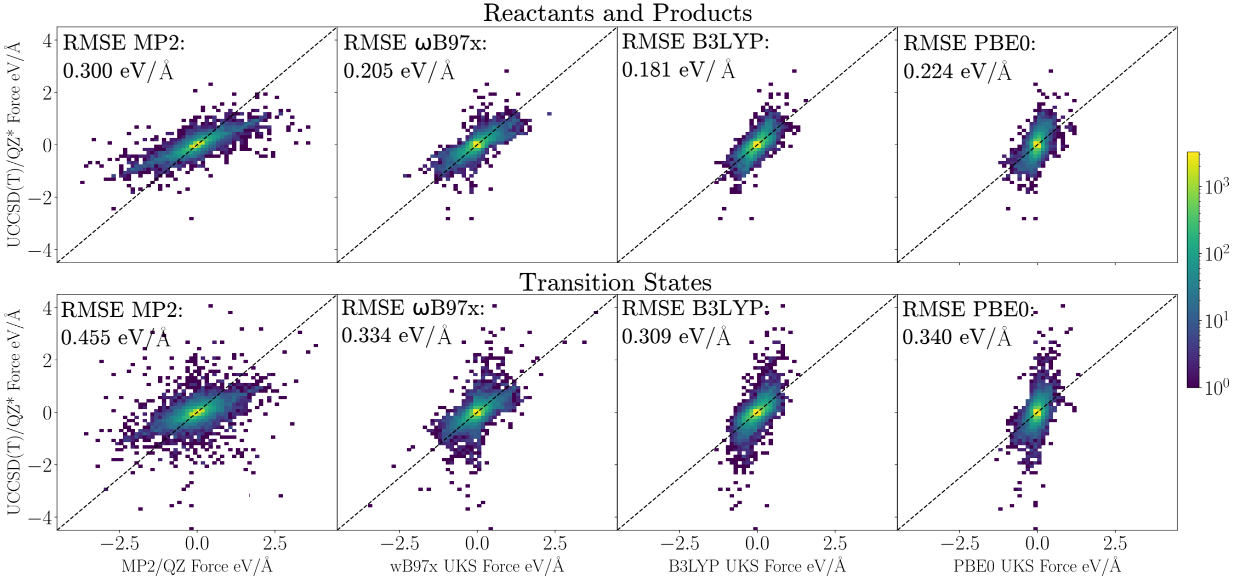


Fitting Data – QM Data

- Improving QM datasets can include improving the level of theory used

- We have been building a unrestricted CCSD(T) dataset for gas phase reactions

- Energies and forces are included



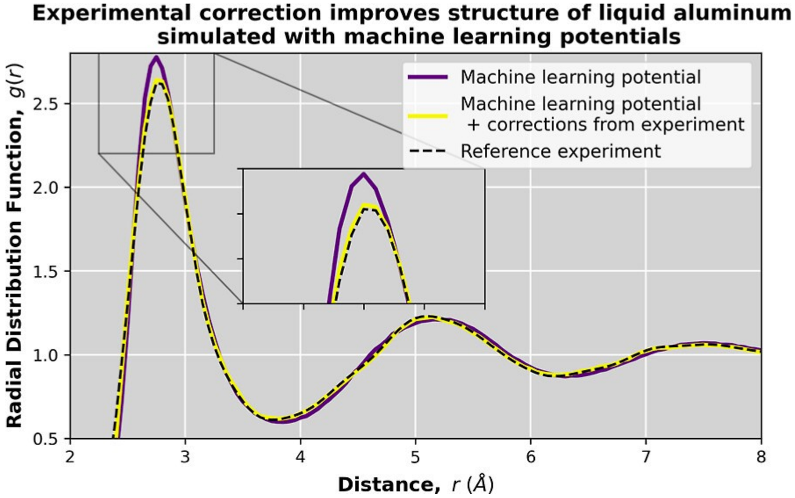
Fitting Data - Experimental

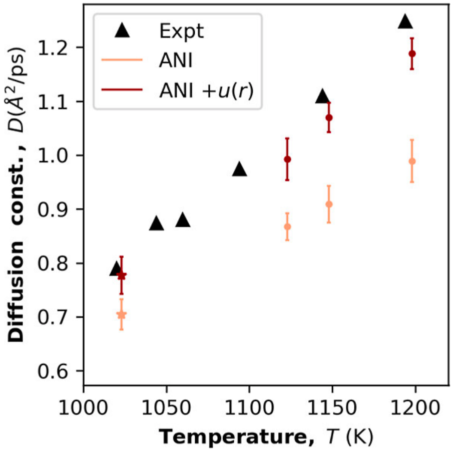
- QM datasets have fundamental limitations

- How can experimental data be incorporated into ML models?

- Investigated using a pair potential correction to an existing MLIP using radial distribution function data

- Improvements then seen in diffusion coefficients when correction present





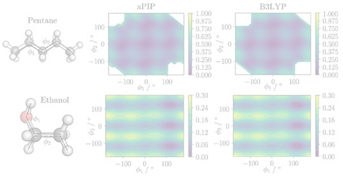
Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024

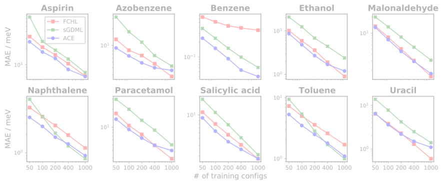
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

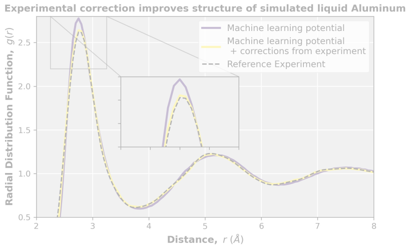
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





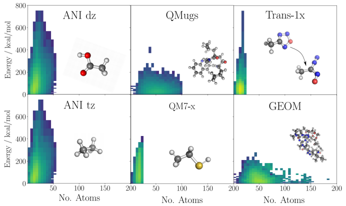
Training Data

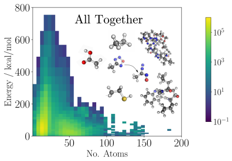
Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



Fitting Methods

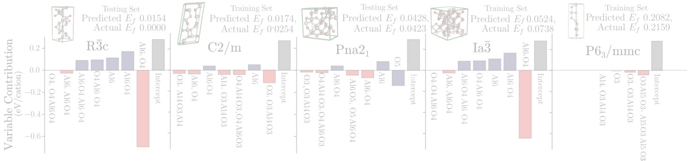
Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024





Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022

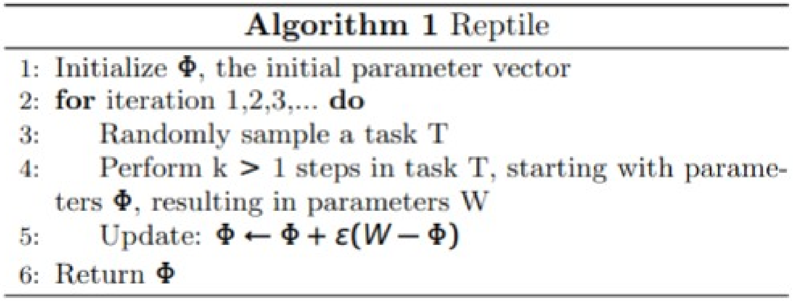


Fitting Methods – Meta Learning

- There is an abundance of datasets containing quantum mechanical calculations for molecular and material systems.

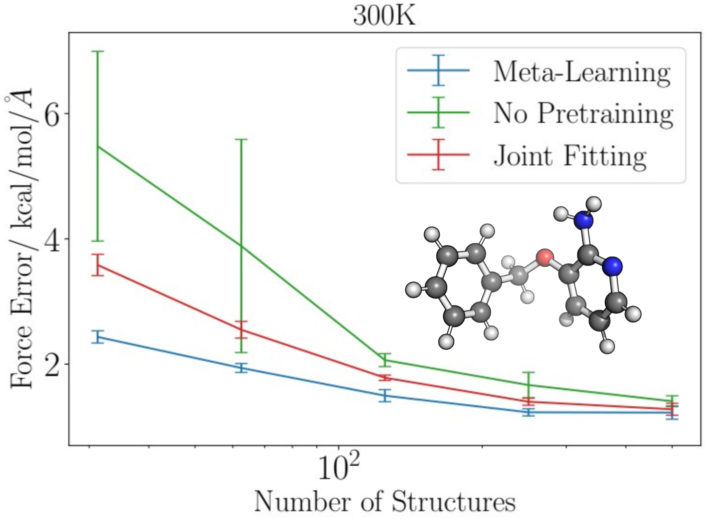
- However, using the information from different datasets together remains a challenge due to the varying levels of theory employed.

- We have shown that meta-learning can be used to pre-train models to multiple datasets.



Algorithm 1 Reptile

```
Initialize $\Phi$, the initial parameter vector
for iteration 1,2,3,... do
    Randomly sample a task T
    Perform k > 1 steps in task T, starting with parameters $\Phi$, resulting in parameters W
    Update: $\Phi \leftarrow \Phi + \varepsilon(W - \Phi)$
Return $\Phi$
```

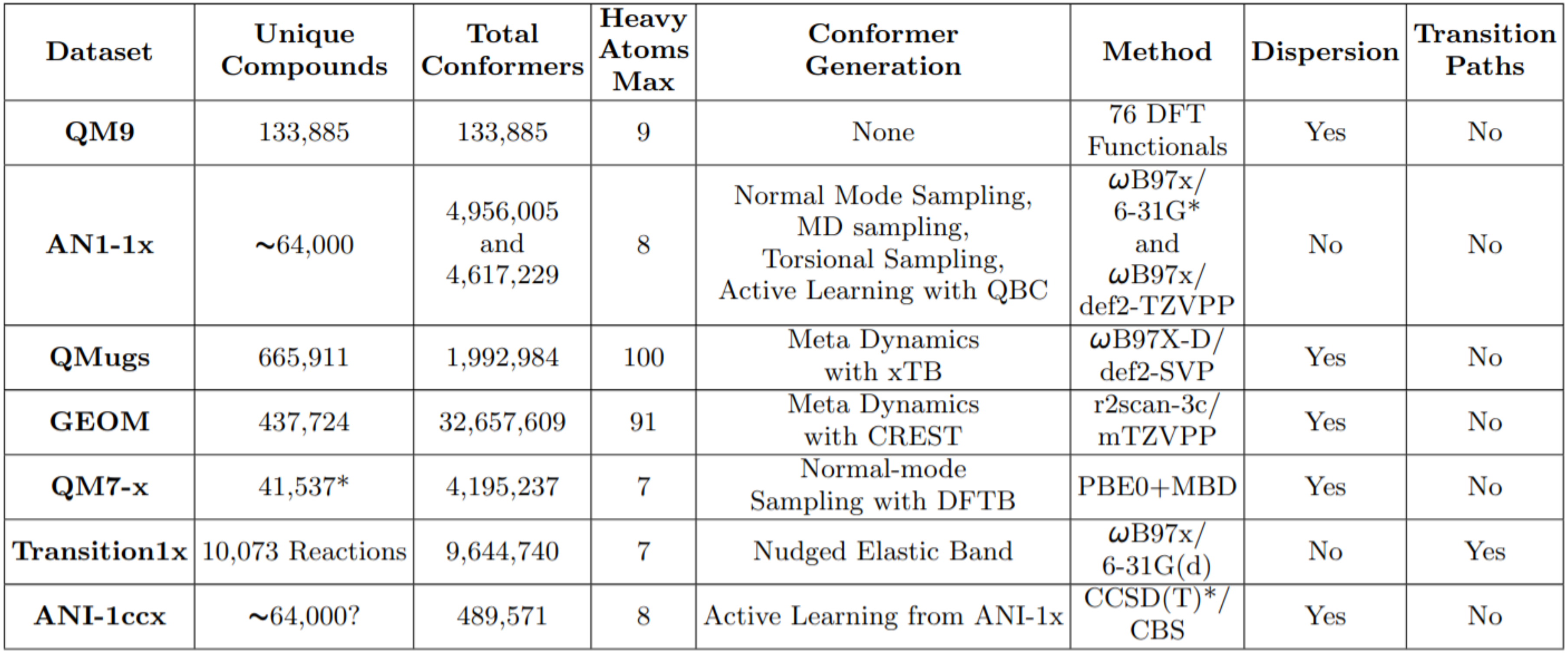


Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024

Problem

How do we fit multiple datasets together that use different QM approximations?

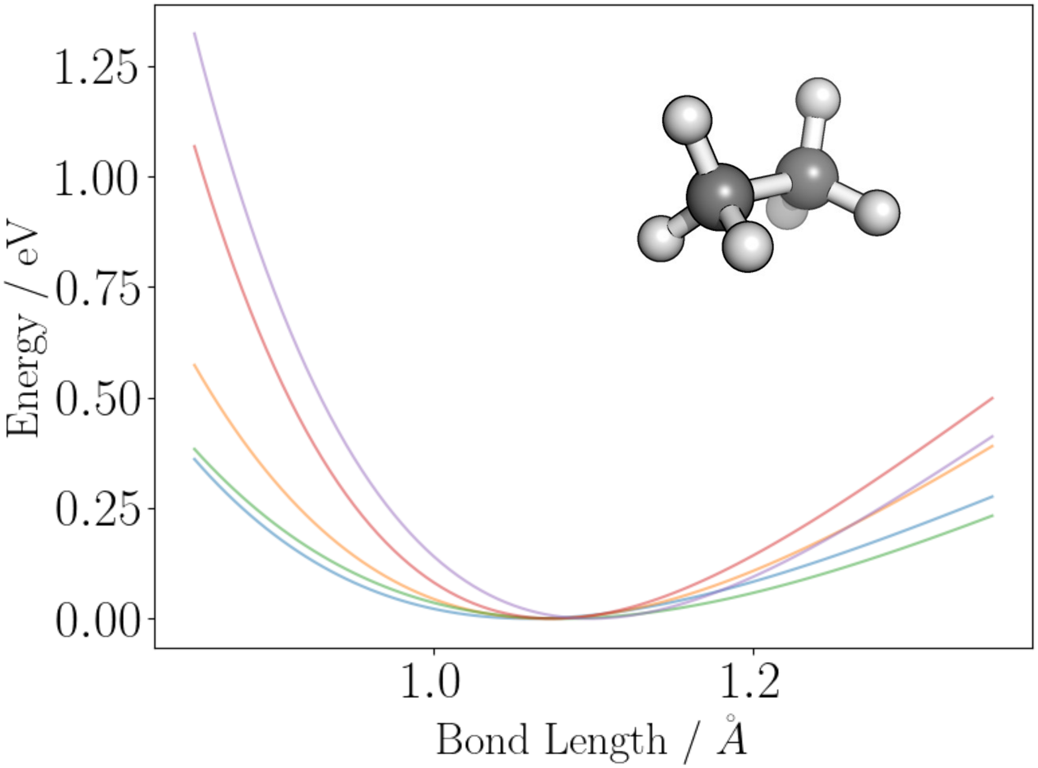
Multiple Existing Datasets Available



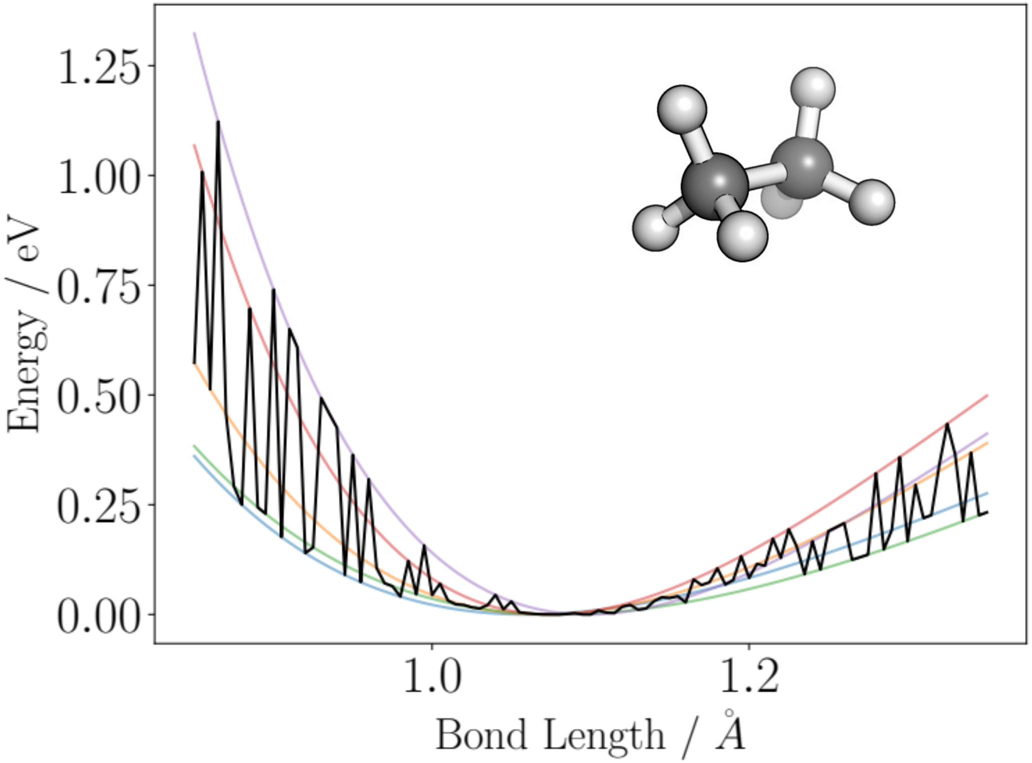
Dataset	Unique Compounds	Total Conformers	Heavy Atoms Max	Conformer Generation	Method	Dispersion	Transition Paths
QM9	133,885	133,885	9	None	76 DFT Functionals	Yes	No
AN1-1x	~64,000	4,956,005 and 4,617,229	8	Normal Mode Sampling, MD sampling, Torsional Sampling, Active Learning with QBC	ωB97x/ 6-31G* and ωB97x/ def2-TZVPP	No	No
QMugs	665,911	1,992,984	100	Meta Dynamics with xTB	ωB97X-D/ def2-SVP	Yes	No
GEOM	437,724	32,657,609	91	Meta Dynamics with CREST	r2scan-3c/ mTZVPP	Yes	No
QM7-x	41,537*	4,195,237	7	Normal-mode Sampling with DFTB	PBE0+MBD	Yes	No
Transition1x	10,073 Reactions	9,644,740	7	Nudged Elastic Band	ωB97x/ 6-31G(d)	No	Yes
ANI-1ccx	~64,000?	489,571	8	Active Learning from ANI-1x	CCSD(T)*/ CBS	Yes	No

Over 90 million calculations included, with more than 200 different levels of theory.

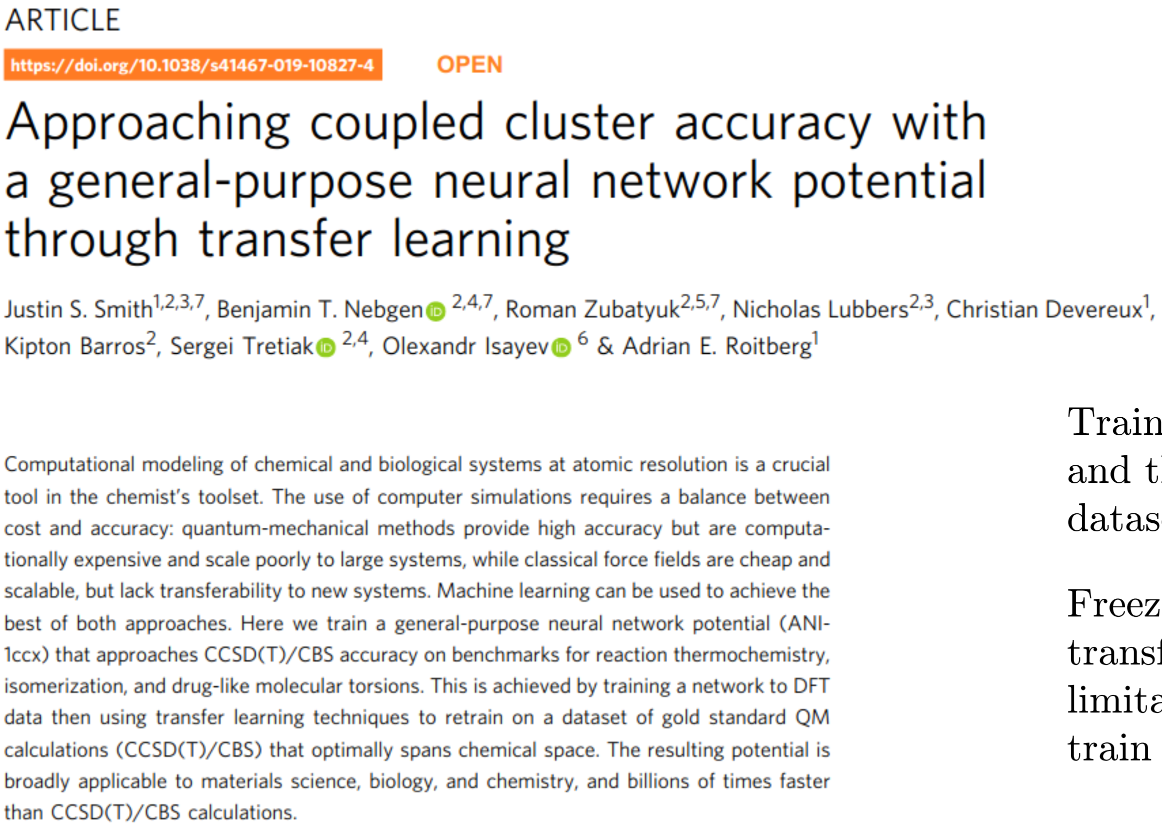
Quantum Mechanical Calculations can be carried out with different approximations



Quantum Mechanical Calculations can be carried out with different approximations



Transfer Learning



ARTICLE

https://doi.org/10.1038/s41467-019-10827-4 OPEN

Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning

Justin S. Smith[1,2,3,7], Benjamin T. Nebgen[2,4,7], Roman Zubatyuk[2,5,7], Nicholas Lubbers[2,3], Christian Devereux[1], Kipton Barros[2], Sergei Tretiak[2,4], Olexandr Isayev[6] & Adrian E. Roitberg[1]

Computational modeling of chemical and biological systems at atomic resolution is a crucial tool in the chemist's toolset. The use of computer simulations requires a balance between cost and accuracy: quantum-mechanical methods provide high accuracy but are computationally expensive and scale poorly to large systems, while classical force fields are cheap and scalable, but lack transferability to new systems. Machine learning can be used to achieve the best of both approaches. Here we train a general-purpose neural network potential (ANI-1ccx) that approaches CCSD(T)/CBS accuracy on benchmarks for reaction thermochemistry, isomerization, and drug-like molecular torsions. This is achieved by training a network to DFT data then using transfer learning techniques to retrain on a dataset of gold standard QM calculations (CCSD(T)/CBS) that optimally spans chemical space. The resulting potential is broadly applicable to materials science, biology, and chemistry, and billions of times faster than CCSD(T)/CBS calculations.

Trained ANI potential to DFT and then retrained to CCSD(T) dataset with frozen parameters.

Freezing parameters and other transfer learning techniques have limitations and can't readily train to multiple datasets.

Problem: How could we extend this to multiple datasets/level of theory?

Meta-learning

The problem:

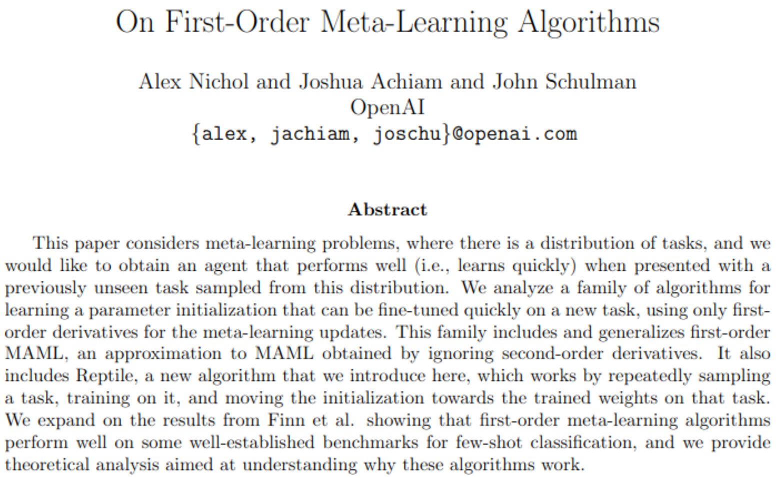
How can we combine multiple QM datasets?

The solution

Meta-learning trains models in such a way that the solution can easily be refit to a new task

Not about finding a model that works well for one task

But finding a model that can be easily retrained to new tasks with limited data



On First-Order Meta-Learning Algorithms

Alex Nichol and Joshua Achiam and John Schulman
OpenAI
{alex, jachiam, joschu}@openai.com

Abstract

This paper considers meta-learning problems, where there is a distribution of tasks, and we would like to obtain an agent that performs well (i.e., learns quickly) when presented with a previously unseen task sampled from this distribution. We analyze a family of algorithms for learning a parameter initialization that can be fine-tuned quickly on a new task, using only first-order derivatives for the meta-learning updates. This family includes and generalizes first-order MAML, an approximation to MAML obtained by ignoring second-order derivatives. It also includes Reptile, a new algorithm that we introduce here, which works by repeatedly sampling a task, training on it, and moving the initialization towards the trained weights on that task. We expand on the results from Finn et al. showing that first-order meta-learning algorithms perform well on some well-established benchmarks for few-shot classification, and we provide theoretical analysis aimed at understanding why these algorithms work.

Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

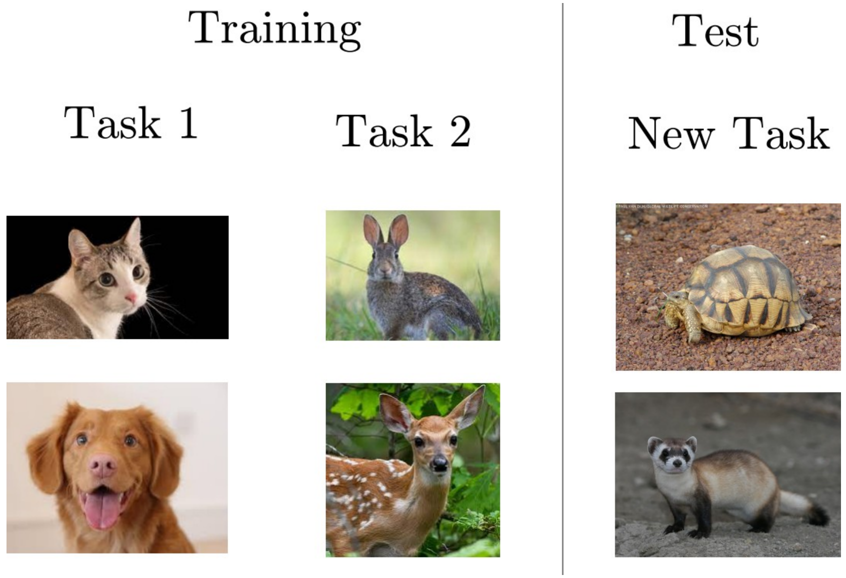
Meta-learning

What is meta-learning?

Trying to train models in such a way that the solution can easily be refit to a new task

Not about finding a model that works well for one task

But finding a model that can be easily retrained to new tasks with limited data



Training

Task 1

Task 2

Test

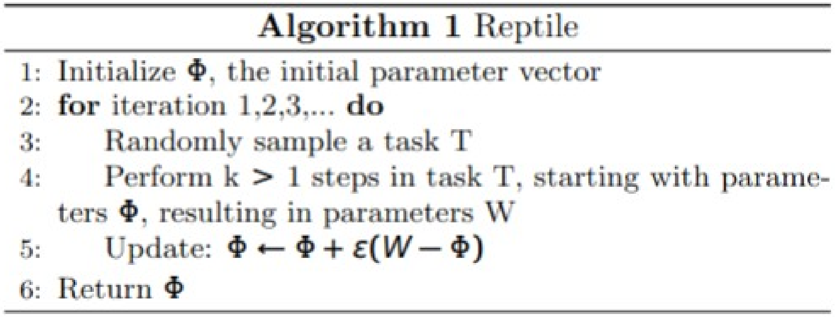
New Task

Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

Meta-learning

Previously implemented the Reptile meta-learning algorithm.

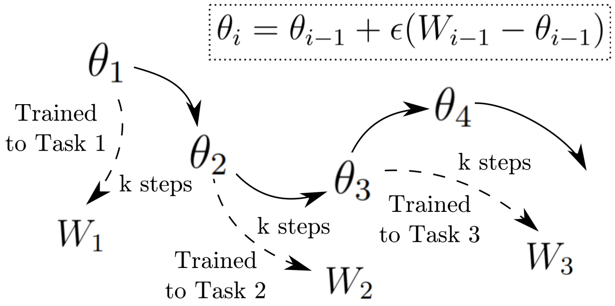
This divides a problem into multiple tasks and then performs multiple optimization steps on individual tasks to find a model that generalizes well to new tasks.



Algorithm 1 Reptile

```
Initialize $\Phi$, the initial parameter vector
for iteration 1,2,3,... do
    Randomly sample a task T
    Perform k > 1 steps in task T, starting with parameters $\Phi$, resulting in parameters W
    Update: $\Phi \leftarrow \Phi + \varepsilon(W - \Phi)$
Return $\Phi$
```

Reptile Algorithm



Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

Reptile

We have implemented the Reptile meta-learning algorithm.

This divides a problem into multiple tasks and then performs multiple optimization steps on individual tasks to find a model that generalizes well to new tasks.

NOT:

Trying to find parameters that are good for one specific task

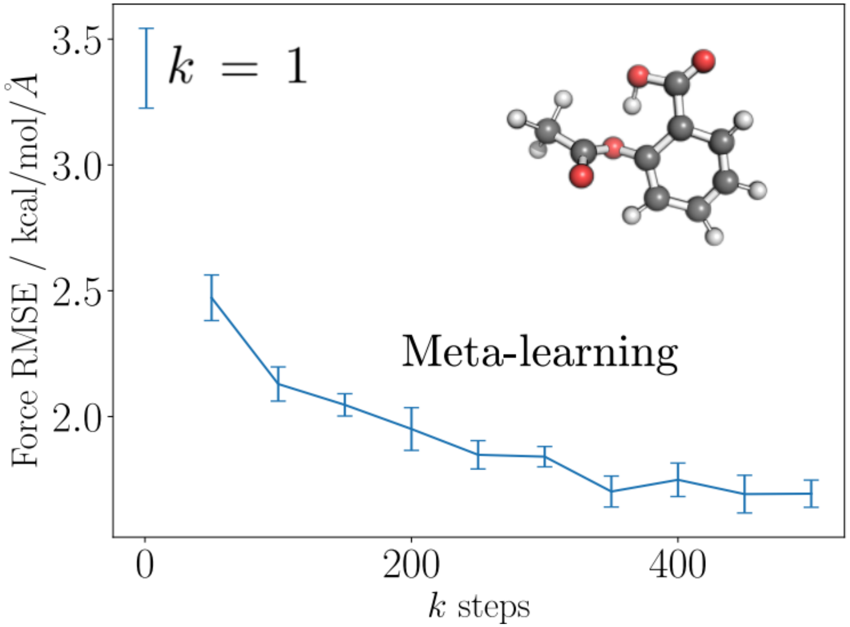
Trying to find parameters that work for all tasks at once

Aspirin

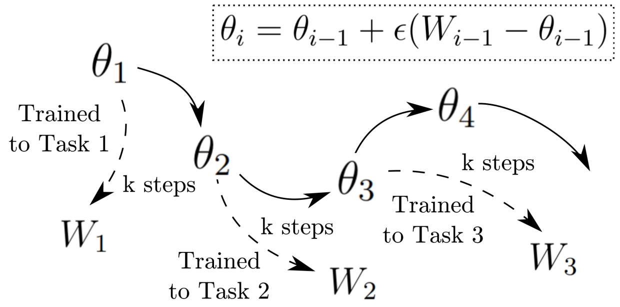
Aspirin was pre-trained to three datasets from 300K, 600K and 900K MD simulations

Three levels of theory were used to pre-train the potential

The potential was then retrained to 400 structures at MP2.



Reptile Algorithm


$\theta_i = \theta_{i-1} + \epsilon(W_{i-1} - \theta_{i-1})$

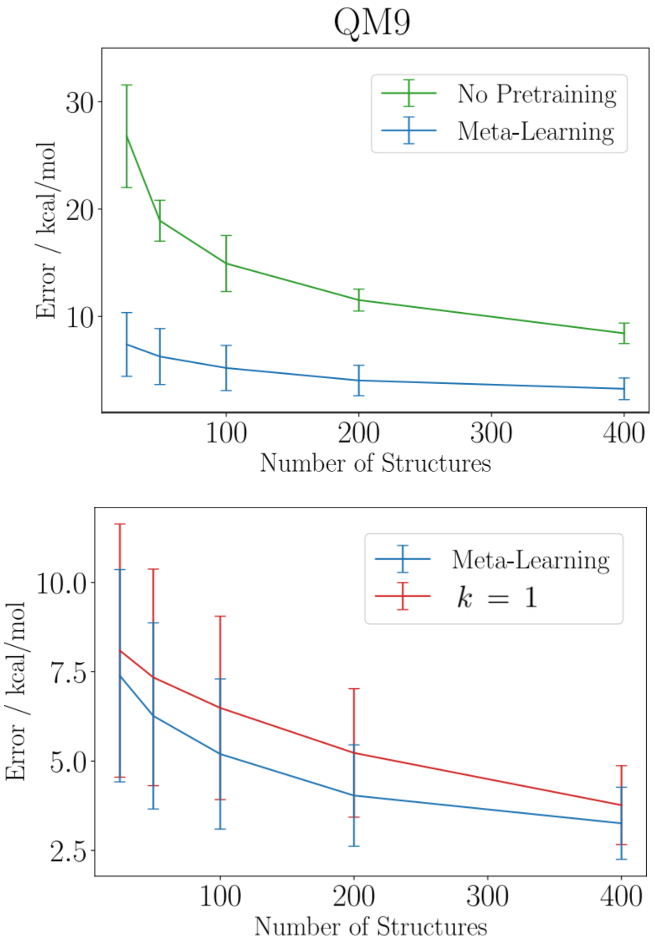
QM9

Meta-learning potential was fit to 150 different levels of theory from the QM9 dataset

Refitting was then performed on 15 different levels of theory that had not previous been encountered

The results of this are shown to the right.

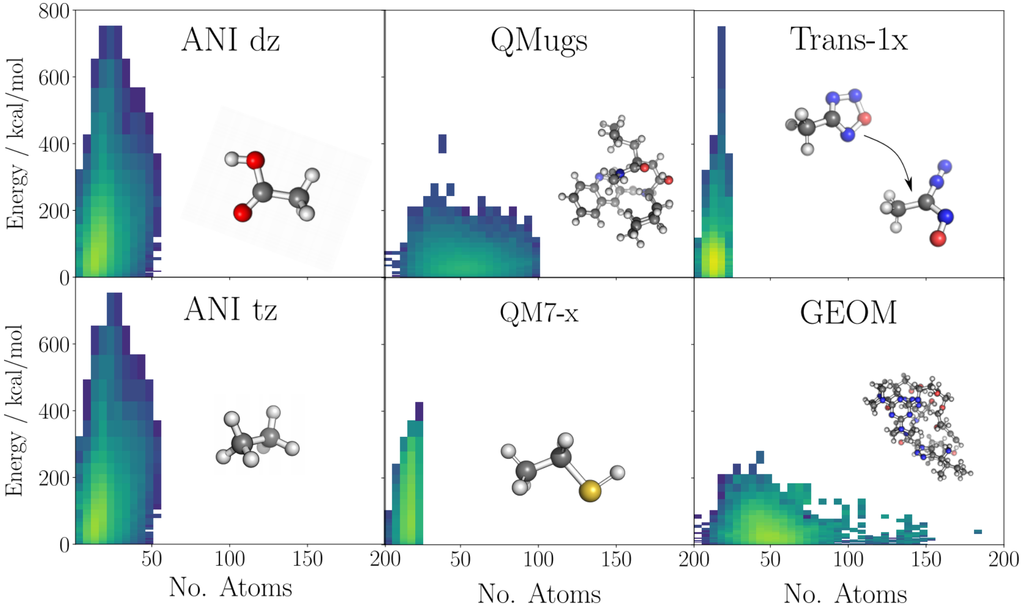
S. Nandi, T. Vegge, and A. Bhowmik, ChemRxiv (2022), 10.26434/chemrxiv-2022-fs70n.

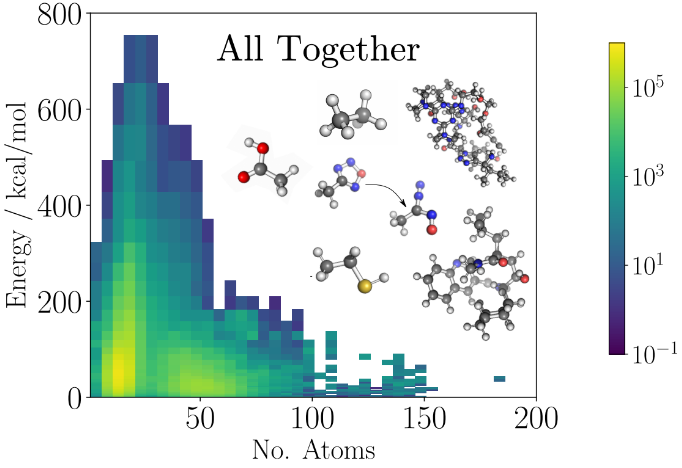


Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

Combining Datasets

Previously implemented the Reptile meta-learning algorithm.



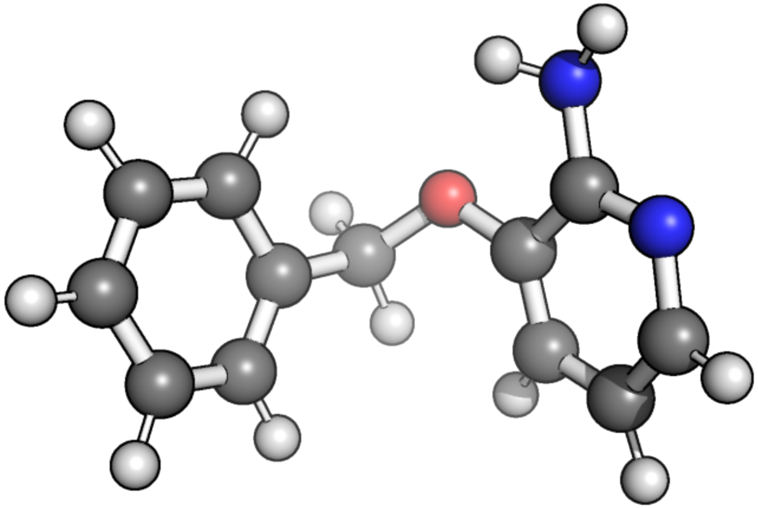


Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

3BPA

Start by investigating the potential for a single molecule

3BPA – shown to the right

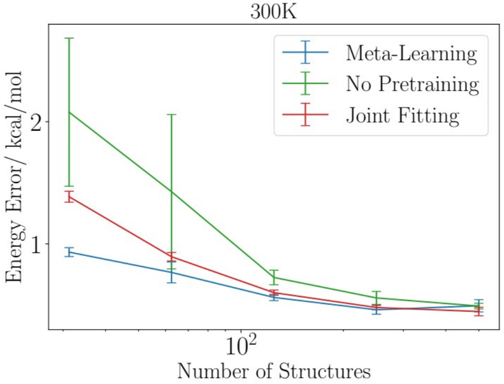


Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., (2021). JCTC.

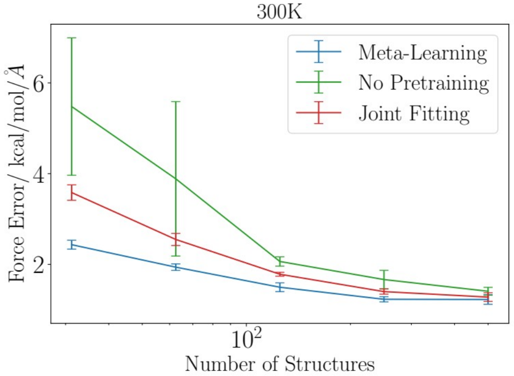
Meta-learning

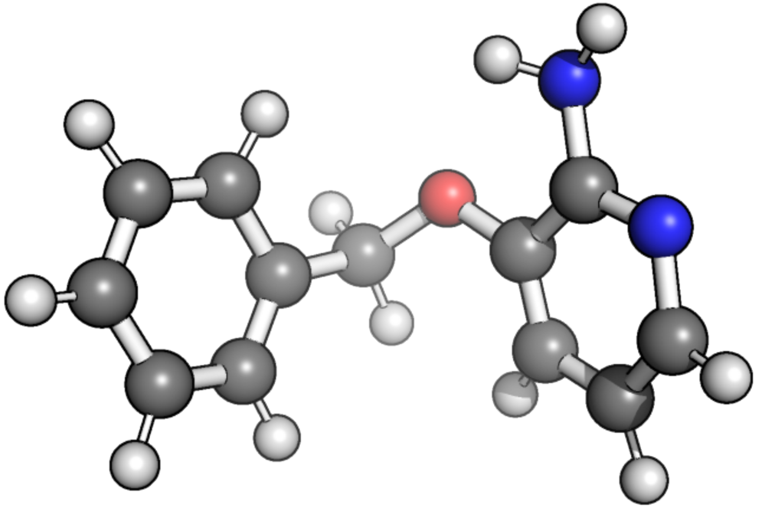
300K

Energy



Force

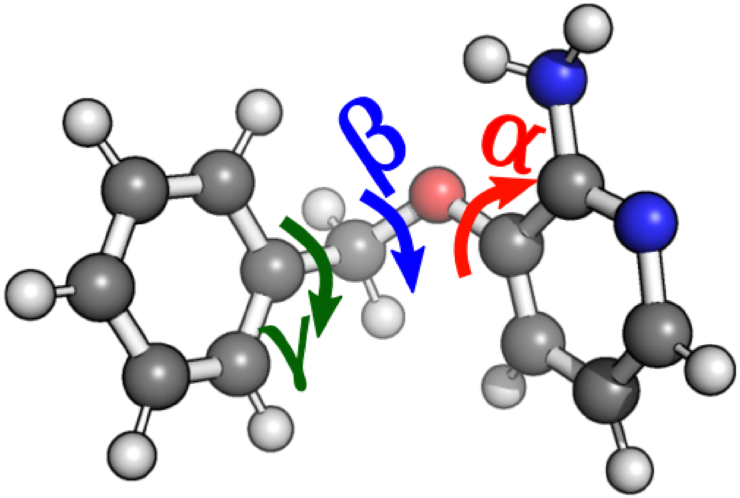




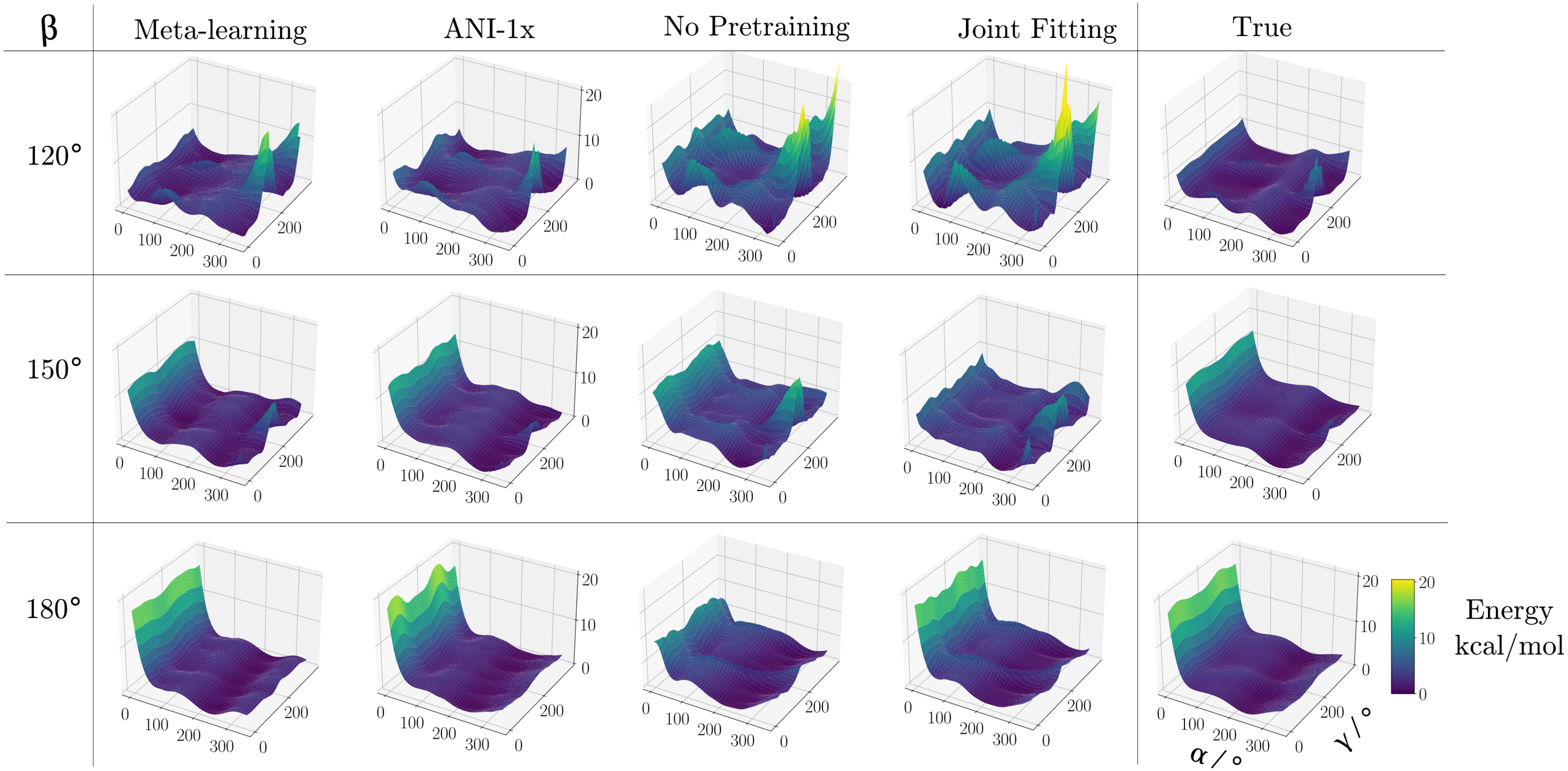
Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

3BPA

Next lets look at the 2D torsional energy scans

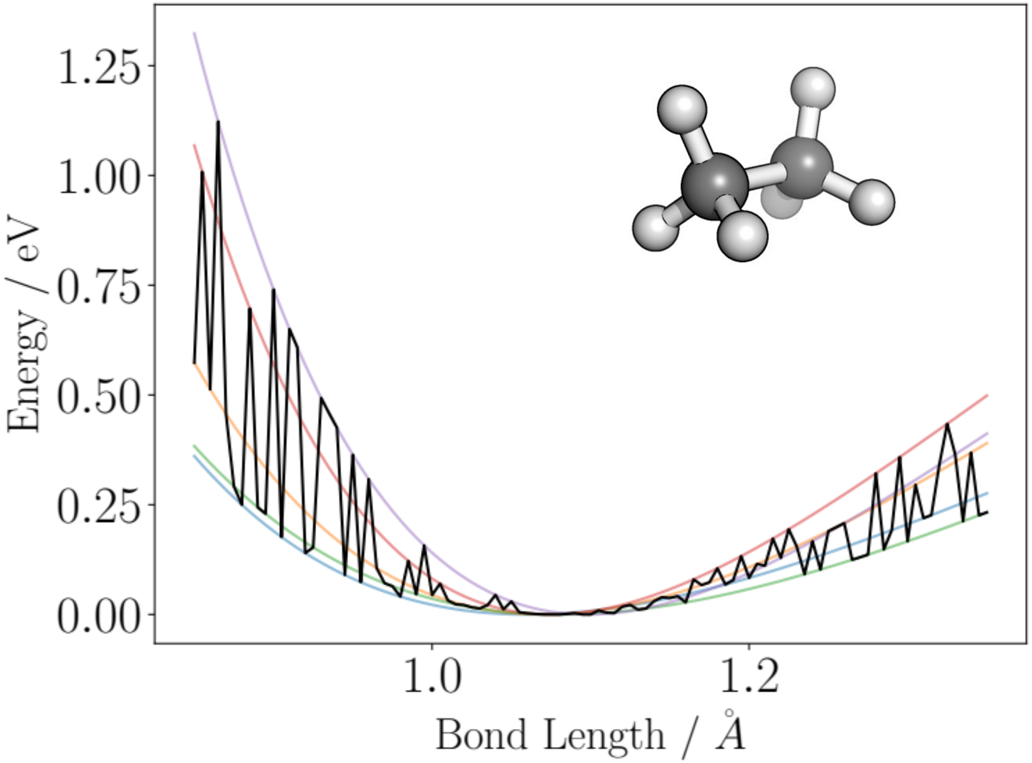


Meta-learning



Learning Together: Towards foundation models for machine learning interatomic potentials with meta-learning, Alice E. A. Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, Kipton Barros, npj Computational Materials, 2024

Quantum Mechanical Calculations can be carried out with different approximations

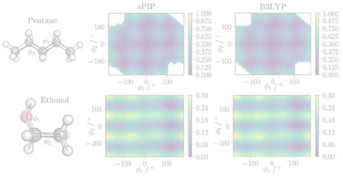


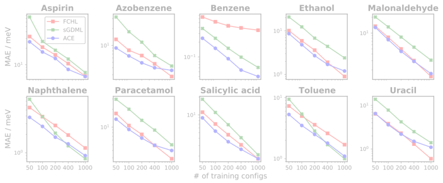
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

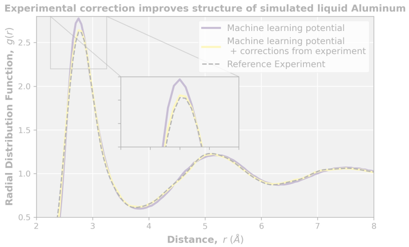
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





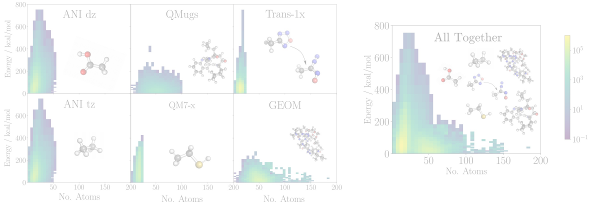
Training Data

Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



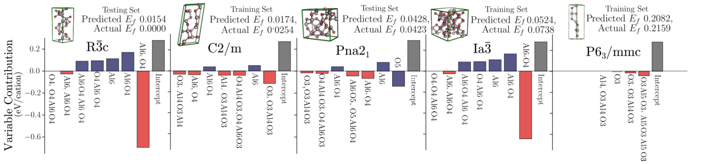
Fitting Methods

Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024



Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022



Interpretability

- Machine learning models can provide fast and accurate predictions of material properties but often lack transparency.
- Interpretability techniques can be used with black box solutions, or alternatively, models can be created that are directly interpretable.
- We revisited several works and demonstrate that simple linear combinations of nonlinear basis functions can be created, which have comparable accuracy to the kernel and neural network approaches originally used.

npj | Computational Materials www.nature.com/npjcompumats

ARTICLE OPEN

Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition

Christopher Sutton[1*], Luca M. Ghiringhelli[1*], Takenori Yamamoto[2], Yury Lysogorskiy[3], Lars Blumenthal[4,5], Thomas Hammerschmidt[3], Jacek R. Golebiowski[4,5], Xiangyue Liu[1], Angelo Ziletti[1] and Matthias Scheffler[1]

A public data-analytics competition was organized by the Novel Materials Discovery (NOMAD) Centre of Excellence and hosted by the online platform Kaggle by using a dataset of 3,000 $(Al_xGa_yIn_{1-x-y})_2O_3$ compounds. Its aim was to identify the best machine-learning (ML) model for the prediction of two key physical properties that are relevant for optoelectronic applications: the electronic bandgap energy and the crystalline formation energy. Here, we present a summary of the top-three ranked ML approaches. The first-place solution was based on a crystal-graph representation that is novel for the ML of properties of materials. The second-place model combined many candidate descriptors from a set of compositional, atomic-environment-based, and average structural properties with the light gradient-boosting machine regression model. The third-place model employed the smooth overlap of atomic position representation with a neural network. The Pearson correlation among the prediction errors of nine ML models (obtained by combining the top-three ranked representations with all three employed regression models) was examined by using the Pearson correlation to gain insight into whether the representation or the regression model determines the overall model performance. Ensembling relatively decorrelated models (based on the Pearson correlation) leads to an even higher

/s41524-019-0239-3

PRL **117,** 135502 (2016) PHYSICAL REVIEW LETTERS week ending 23 SEPTEMBER 2016

Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals

Felix A. Faber,[1] Alexander Lindmaa,[2] O. Anatole von Lilienfeld,[1,3,*] and Rickard Armiento[2,†]
[1]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, 4056 Basel, Switzerland*
[2]*Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden*
[3]*General Chemistry, Free University of Brussels, Pleinlaan 2, 1050 Brussels, Belgium*
(Received 24 August 2015; published 20 September 2016)

Elpasolite is the predominant quaternary crystal structure ($AlNaK_2F_6$ prototype) reported in the Inorganic Crystal Structure Database. We develop a machine learning model to calculate density functional theory quality formation energies of all ~2×10^6 pristine ABC_2D_6 elpasolite crystals that can be made up from main-group elements (up to bismuth). Our model's accuracy can be improved systematically, reaching a mean absolute error of 0.1 eV/atom for a training set consisting of 10×10^3 crystals. Important bonding trends are revealed: fluoride is best suited to fit the coordination of the D site, which lowers the formation energy whereas the opposite is found for carbon. The bonding contribution of the elements A and B is very small on average. Low formation energies result from A and B being late elements from group II, C being a late (group I) element, and D being fluoride. Out of 2×10^6 crystals, 90 unique structures are predicted to be on the convex hull—among which is $NFAl_2Ca_6$, with a peculiar stoichiom oxidation state for Al.

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE View Article Online View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2017, **8**, 5137

Predicting electronic structure properties of transition metal complexes with neural networks†

Jon Paul Janet and Heather J. Kulik*

High-throughput computational screening has emerged as a critical component of materials discovery. Direct density functional theory (DFT) simulation of inorganic materials and molecular transition metal complexes is often used to describe subtle trends in inorganic bonding and spin-state ordering, but these calculations are computationally costly and properties are sensitive to the exchange–correlation functional employed. To begin to overcome these challenges, we trained artificial neural networks (ANNs) to predict quantum-mechanically-derived properties, including spin-state ordering, sensitivity to Hartree–Fock exchange, and spin-state specific bond lengths in transition metal complexes. Our ANN is trained on a small set of inorganic-chemistry-appropriate empirical inputs that are both maximally transferable and do not require precise three-dimensional structural information for prediction. Using these descriptors, our ANN predicts spin-state splittings of single-site transition metal complexes (*i.e.*, Cr–Ni) at arbitrary amounts of Hartree–Fock exchange to within 3 kcal mol^{-1} accuracy of DFT calculations. Our exchange-sensitivity ANN enables improved predictions on a diverse test set of experimentally-characterized transition metal complexes by extrapolation from semi-local DFT to hybrid DFT. The ANN also outperforms other machine learning models (*i.e.*, support vector regression and kernel ridge regression), demonstrating particularly improved performance in transferability, as measured by prediction errors on the diverse test set. We establish the value of new uncertainty quantification tools to estimate ANN prediction uncertainty in computational chemistry, and we provide additional heuristics for identification of when a compound of interest is likely to be poorly predicted by the ANN. The ANNs developed in this work provide a strategy for screening transition metal complexes both with direct ANN prediction and with improved structure generation for validation with first principles simulation.

Received 19th March 2017
Accepted 9th May 2017

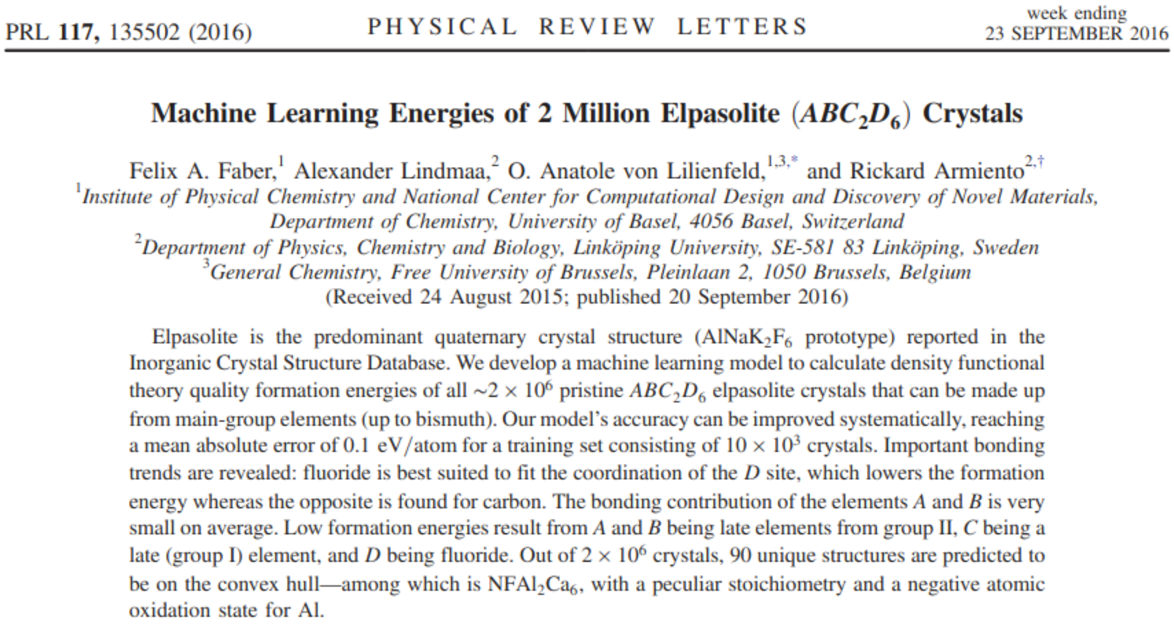
DOI: 10.1039/c7sc01247k

rsc.li/chemical-science

Machine learning of material properties: Predictive and interpretable multilinear models, Allen and Tkatchenko, Sci. Adv. 8, eabm7185 (2022)

Interpretability

- A predictive model for the formation energy of 10,000 elpasolite structures (ABC_2D_6 in the *Fm3m* space group) was produced using kernel ridge regression.



PRL **117,** 135502 (2016) PHYSICAL REVIEW LETTERS week ending 23 SEPTEMBER 2016

Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals

Felix A. Faber,[1] Alexander Lindmaa,[2] O. Anatole von Lilienfeld,[1,3,*] and Rickard Armiento[2,†]

[1]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, 4056 Basel, Switzerland*

[2]*Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden*

[3]*General Chemistry, Free University of Brussels, Pleinlaan 2, 1050 Brussels, Belgium*

(Received 24 August 2015; published 20 September 2016)

Elpasolite is the predominant quaternary crystal structure ($AlNaK_2F_6$ prototype) reported in the Inorganic Crystal Structure Database. We develop a machine learning model to calculate density functional theory quality formation energies of all ~2×10^6 pristine ABC_2D_6 elpasolite crystals that can be made up from main-group elements (up to bismuth). Our model's accuracy can be improved systematically, reaching a mean absolute error of 0.1 eV/atom for a training set consisting of 10×10^3 crystals. Important bonding trends are revealed: fluoride is best suited to fit the coordination of the *D* site, which lowers the formation energy whereas the opposite is found for carbon. The bonding contribution of the elements *A* and *B* is very small on average. Low formation energies result from *A* and *B* being late elements from group II, *C* being a late (group I) element, and *D* being fluoride. Out of 2×10^6 crystals, 90 unique structures are predicted to be on the convex hull—among which is $NFAl_2Ca_6$, with a peculiar stoichiometry and a negative atomic oxidation state for Al.

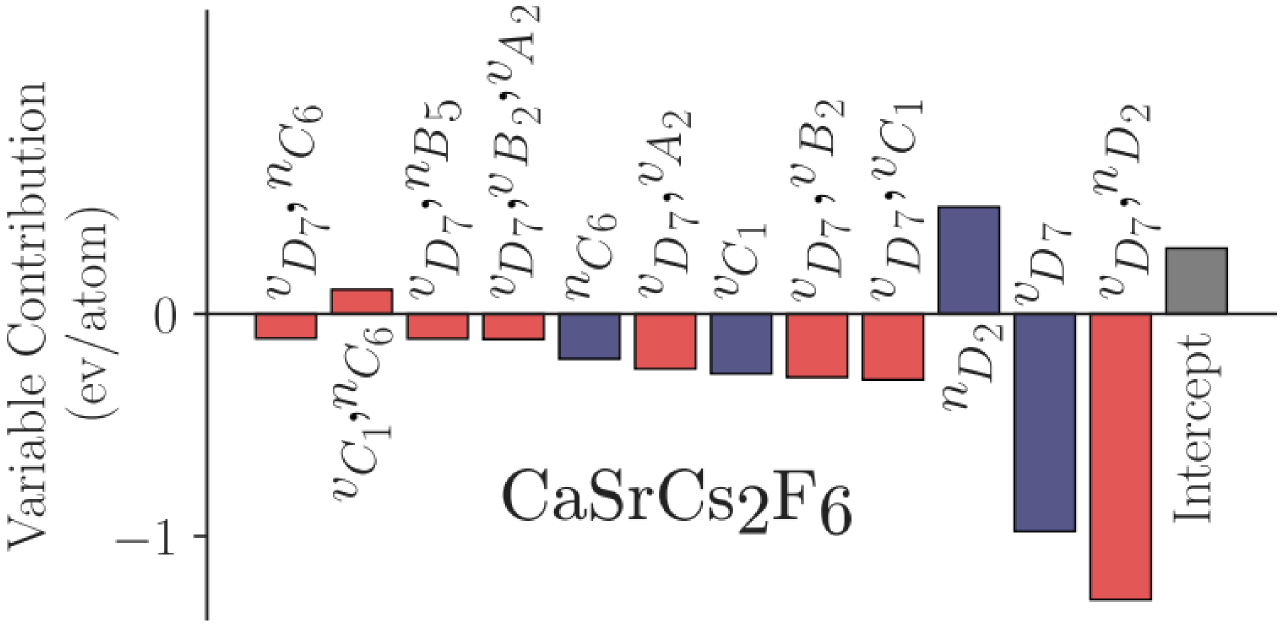
Interpretability

- The features used to describe the structures were the principle quantum number (n) and number of valence electrons (v) at each site A,B,C or D.
- Reach accuracy of 0.11 eV/atom, compared to 0.10 eV/atom for the KRR.
- The accuracy of the DFT data has been stated as between 0.10 – 0.19 eV/atom
- So we have comparable accuracy to KRR and DFT

$$
\begin{aligned}
E(n,v) = \sum_{i} \alpha_i n_i + \sum_{i} \beta_i v_i + \sum_{j<i} \alpha_{ij} n_i n_j + \sum_{j<i} \beta_{ij} v_i v_j \\
+ \sum_{i,j} \gamma_{ij} n_i v_j + \sum_{k<j<i} \alpha_{ijk} n_i n_j n_k + \sum_{k<j<i} \beta_{ijk} v_i v_j v_k \\
+ \sum_{k<i,j} \gamma_{ijk} n_i v_j n_k + \sum_{k<j,i} \lambda_{ijk} n_i v_j v_k \qquad (3)
\end{aligned}
$$

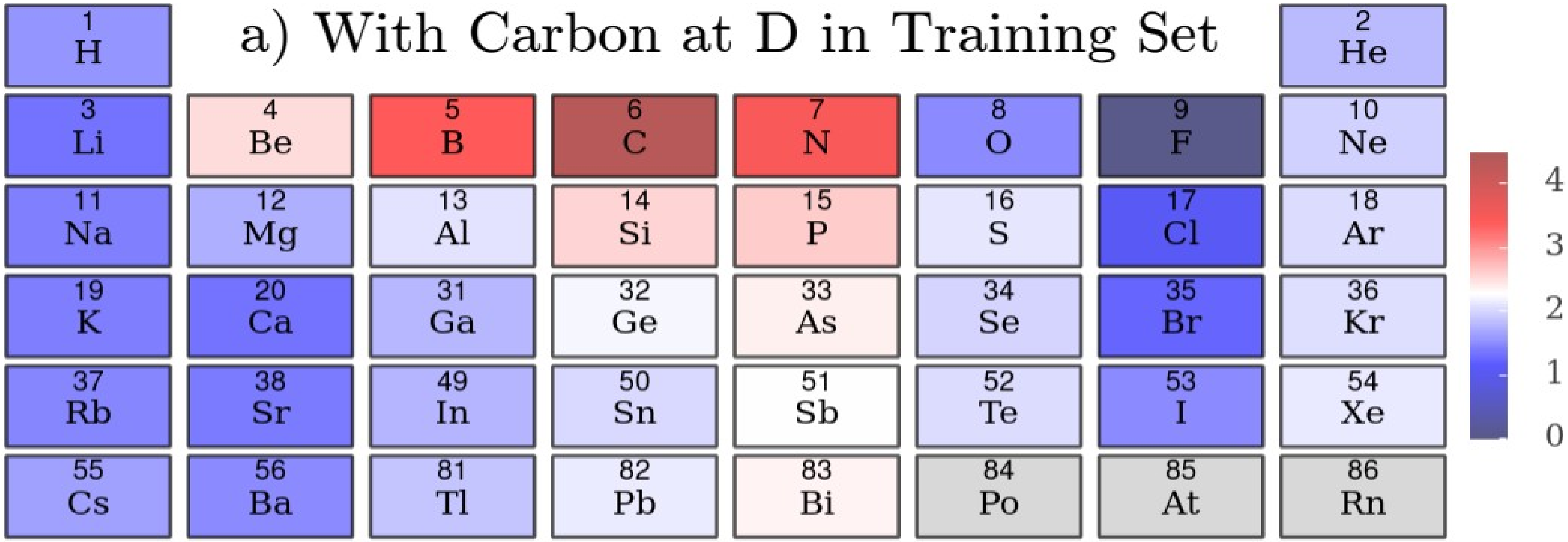
Interpretability

- Search 2 million possible structures
- Individual predictions can be broken down into variable and interaction contributions:



Interpretability

- Contributions to the formation energy at position D can be found directly from the coefficients of the model:

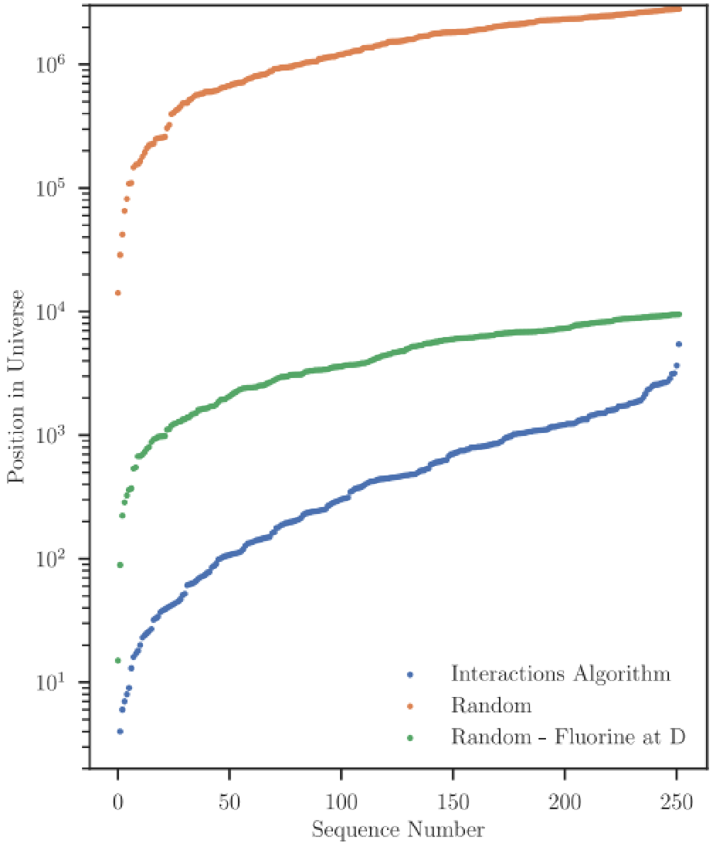


a) With Carbon at D in Training Set

Interpretability

- With the linear model the coefficients can be used to guide predictions.

- If we want low formation energy structures we can focus on this region.



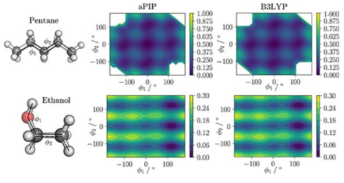
Machine learning of material properties: Predictive and interpretable multilinear models, Allen and Tkatchenko, Sci. Adv. 8, eabm7185 (2022)

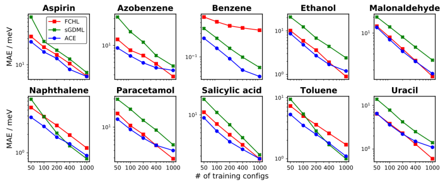
Improving ML Models

Architecture

Atomic permutationally invariant polynomials for fitting molecular force fields. Alice E. A. Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Machine Learning: Science and Technology, 2021

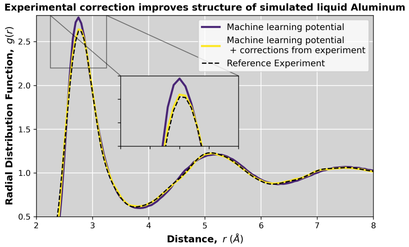
Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE., Kovacs, D. P., van der Oord, C., Kucera, J., Allen, A., Cole, D., Ortner, C., & Csanyi, G., JCTC, 2021





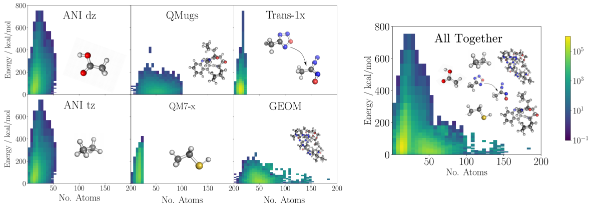
Training Data

Machine learning potentials with Iterative Boltzmann Inversion: training to experiment, Sakib Matin, Alice Allen, Justin S Smith, Nicholas Lubbers, Ryan B Jadrich, Richard A Messerly, Benjamin T Nebgen, Ying Wai Li, Sergei Tretiak, Kipton Barros, JCTC, 2024



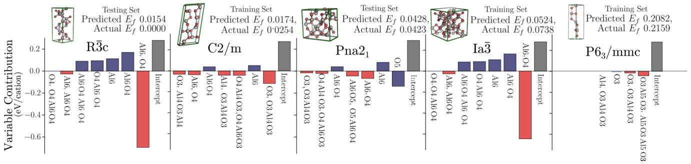
Fitting Methods

Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, AEA Allen, N Lubbers, S Matin, J Smith, R Messerly, S Tretiak, K Barros, npj Computational Materials, 2024



Interpretability

Alice E. A. Allen and Alexandre Tkatchenko, Machine Learning of Material Properties: Predictive and Interpretable Multilinear Models, Science Advances, 2022

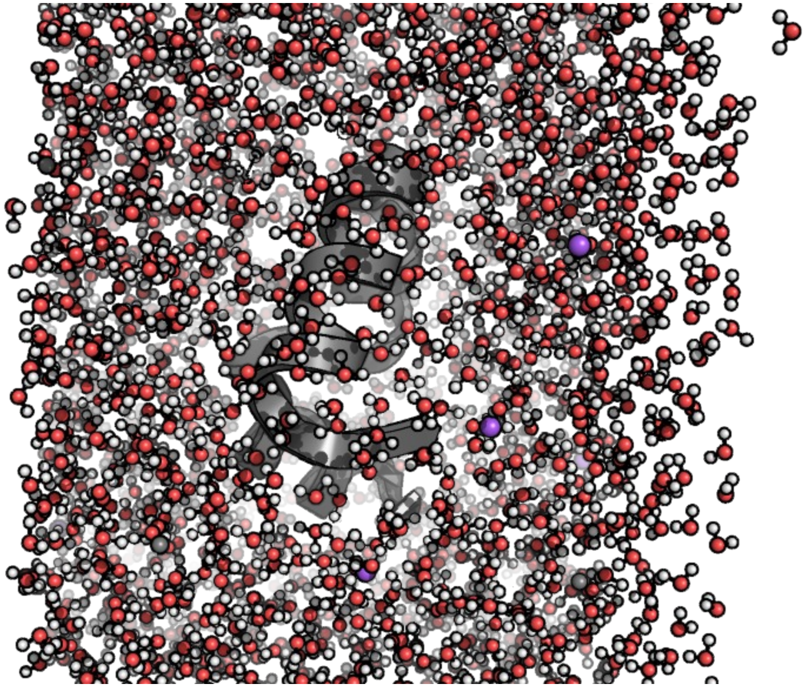


Conclusion

- Atomistic simulations and the description of interactions between atoms and molecules can be greatly improved with ML

- Both building novel machine learning models and exploiting techniques from ML community is important for constructing effective models

- This can help us improve ML through better architecture, fitting data, fitting methods and interpretability



Acknowledgements

Classical Force Fields:

Daniel Cole

Mike Payne

Joshua Horton

Michael Robertson

aPIPs/ACE:

Geneviève Dusson

Gábor Csányi

Cas van der Oord

Christoph Ortner

Dávid Péter Kovács

Emily Shinkle

Roxana Bujack

Meta-learning & Reactive Potentials

Kipton Barros

Nicholas Lubbers

Sakib Matin

Richard Messerly

Ben Nebgen

Sergei Tretiak

Justin Smith

Garnet Chan

Interpretability

Alexandre Tkatchenko

UNIVERSITY OF CAMBRIDGE

Los Alamos NATIONAL LABORATORY

Center for Nonlinear Studies

Questions?