Symbolic Regression in Materials Science: Discovering Interatomic Potentials from Data

Bogdan Burlacu, Michael Kommenda, Gabriel Kronberger, Stephan Winkler, Michael Affenzeller

Josef Ressel Centre for Symbolic Regression
Heuristic and Evolutionary Algorithms Laboratory
University of Applied Sciences Upper Austria

SymReg
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION

HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Materials Science

Highly relevant, interdisciplinary field:

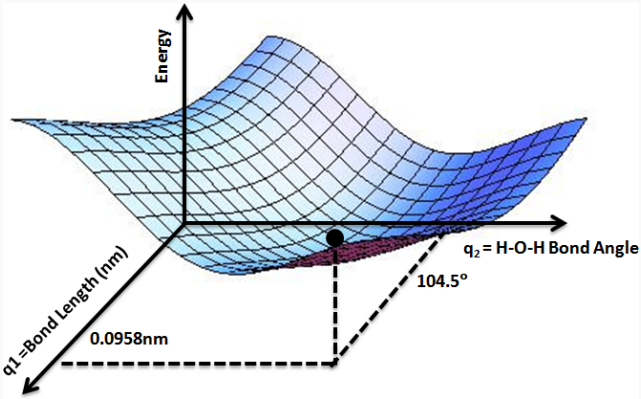
- physics — chemistry — engineering — **machine learning**

Computational modeling is nowadays the main way of studying new materials:

An accurate model of interatomic interactions is required.

Potential Energy Surface (PES)

Describe the relationship between an atomic system's potential energy and the geometry of its atoms.



Potential Energy Surface (PES)

Used in Molecular Dynamics and Monte Carlo simulations.

A simulation is as good as its atomic interaction model:

- First principles models: very slow, very accurate (e.g. Density Functional Theory)
- Classical (empirical) models: very fast, not very accurate
- Semi-empirical: trade-off, combination of empirical and *ab initio*
- Machine learning: best of both worlds

ML-based Potentials – Requirements

Speed

To enable simulations at larger scale and longer time frames.

Accuracy

Of crucial importance, must be close to *ab initio* methods.

Generality and transferability

Models should not be restricted to specific types of atomic configurations.

Complexity and data requirements

Simple structures, few parameters, ability to train from small data.

Ab initio data expensive to acquire.

ML-based Potentials – Physical soundness

Physically meaningful predictions

- ML-based models must exhibit the same invariant behavior as the true PES.
- Must respect all physical properties, symmetries, invariances.
- Interpretability highly desired.
- Ability to provide physical insight.

Physical properties

- Conservation of total energy
- Roto-translational invariance (linear and angular momentum)
- Permutational invariance (atoms with the same nuclear charge can be exchanged)

ML-based Potentials

Data-driven approaches

- neural networks (most popular)
- polynomial fitting
- moment tensor potentials (linear combination of polynomial basis functions)
- Gaussian processes
- support vector machines

Symbolic Regression

- more amenable to the integration of physical knowledge
- not requiring *a priori* knowledge or predefined functional forms
- successful in discovering simple forms of potentials
- smaller number of parameters, better efficiency
- arguably more interpretable by experts

Atomic Interation Models

In molecular dynamics simulations, the system's potential energy is typically decomposed into a set of independent m-body interactions that are a function of each particle's position, $\mathbf{r}$.

$$E = \sum_{\langle i,j \rangle} g(\mathbf{r}_i, \mathbf{r}_j) \qquad \text{two-body}$$

$$E = \sum_{\langle i,j \rangle} g(\mathbf{r}_i, \mathbf{r}_j) + \sum_{\langle i,j,k \rangle} h(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) \qquad \text{three-body}$$

The functions g and h can be all kinds of empirical or semi-empirical functions.

ML enables the automatic discovery of such functional forms using *ab initio* training data.

SR-based Potentials

Successful rediscovery of simple potentials (Morse, Lennard-Jones, …)

Ongoing issues: extrapolation, generalization, transferability

- directed search
- hard complexity limits (length, depth), simple primitive set
- physically-augmented fitness – weigh down high-energy points
- inclusion of derivative information ($\mathrm{F} = -\nabla_{\mathbf{r}}\mathrm{E}$)

GP "at scale"

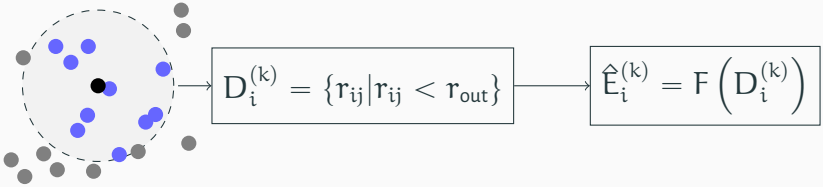
- parallel tempering
- hierarchical fair competition
- co-evolution of populations and operator probabilities
- local search (coefficient tuning)

Modeling Interatomic Potentials

Training data from molecular dynamics simulations (relatively scarce).

Format: atomic configuration → energy (+ forces, stresses)

Snapshot k **Aggregate distances** **Estimate energy**



Modeling Interatomic Potentials

Approach

Include $\sum$ symbol in the primitive set (**Hernandez et al., 2019**).

$$E = \sum_{\langle i,j \rangle} g(\mathbf{r}_i, \mathbf{r}_j) \qquad \text{two-body}$$

Three input dimensions:

- number of training snapshots N
- number of atoms M
- number of pairwise distances L (depends on cut-off radius)

Modeling Interatomic Potentials

Extending *Operon* with $\sum$ symbols:

- nested $\sum$ symbols behave as the identity function
- if not under a $\sum$ symbol, r acts as a constant $c = 1$
- evaluation is performed using a nested interpreter that aggregates the distances

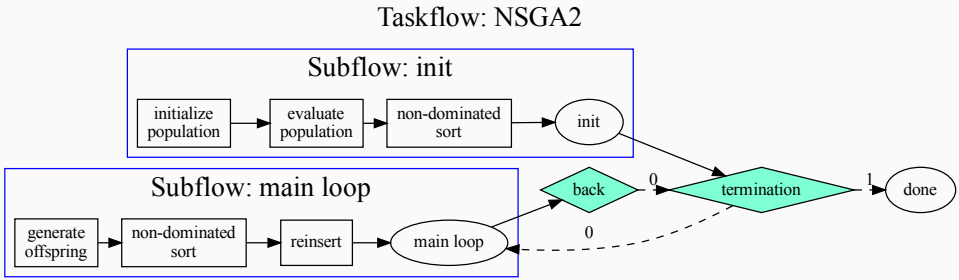
NSGA2 algorithm

- ε-dominance to ensure more even distribution of solutions over Pareto fronts
- duplicate solutions penalized with worst Pareto rank
- objectives: accuracy (R^2 coefficient) + parsimony (model length)

https://github.com/heal-research/operon
https://github.com/foolnotion/atomic-potentials

Modeling Interatomic Potentials

Fine-grained parallelism



Modeling Interatomic Potentials

population size	10,000 individuals
tree limits	max length 20, max depth 10
crossover probability	100%
crossover operator	subtree crossover
mutation probability	25%
mutation operator	uniformly chosen from: · subtree removal/insertion/replacement · change function symbol · change variable name · additive one point leaf mutation ($v = v + \mathcal{N}(0, 1)$) · discrete point leaf mutation
selection operator	crowded tournament selection, group size = 17
objectives	Pearson R^2 and model length
evaluation budget	10^8 fitness evaluations

Modeling Interatomic Potentials

Dataset

- Atomic system of 32 Cu atoms.
- 150 snapshots of atomic positions, energies, forces and stresses
- data is shuffled and split evenly between training and test
- additional feature $\mathrm{p} = \mathrm{r}^{-1}$ added to help the search

https://gitlab.com/muellergroup/poet/-/tree/master/poet_run/data/DFT_Cu

Results – summary

ID	Primitive set	Inputs	$\text{MAE}_{\text{train}}$	MAE_{test}	Length	Runtime (s)
A	$\sum, +, -, \times, \div$	r	0.568 ± 0.045	0.602 ± 0.059	32.0	118.52
B	$\sum, +, -, \times, \div$	q	0.518 ± 0.036	0.599 ± 0.069	44.0	142.01
C	$\sum, +, -, \times, \div$	r, q	0.512 ± 0.043	0.595 ± 0.091	42.0	143.69
D	$\sum, +, -, \times$, aq	r	0.498 ± 0.047	0.583 ± 0.060	56.0	165.49
E	$\sum, +, -, \times$, aq	q	0.500 ± 0.066	0.574 ± 0.068	56.5	162.51
F	$\sum, +, -, \times$, aq	r, q	0.493 ± 0.046	0.593 ± 0.060	60.0	169.64
G	$\sum, +, -, \times, \div$, pow	r	0.501 ± 0.042	0.620 ± 0.039	39.0	286.95
H	$\sum, +, -, \times, \div$, pow	q	0.516 ± 0.048	0.604 ± 0.065	46.5	241.25
I	$\sum, +, -, \times, \div$, pow	r, q	0.514 ± 0.051	0.596 ± 0.057	47.0	290.53
J	$\sum, +, -, \times$, aq, pow	r	0.507 ± 0.052	0.608 ± 0.059	47.0	269.26
K	$\sum, +, -, \times$, aq, pow	q	0.489 ± 0.053	0.623 ± 0.085	57.0	244.44
L	$\sum, +, -, \times$, aq, pow	r, q	0.497 ± 0.053	0.594 ± 0.068	57.0	281.86

Results – statistical significance – test performance

	A	B	C	D	E	F	G	H	I	J	K	L
A		↑3e-01	↑4e-01	↑2e-01	**↑7e-03**	↑8e-02	↓9e-01	↓4e-01	↑1e-01	↓5e-01	↓9e-01	↑7e-02
B	↓3e-01		↑9e-01	↑1e+00	↑2e-01	↑6e-01	↓2e-01	↓7e-01	↑6e-01	↓8e-02	↓2e-01	↑4e-01
C	↓4e-01	↓9e-01		↑1e+00	↑2e-01	↑5e-01	↓4e-01	↓8e-01	↓8e-01	↓1e-01	↓3e-01	↑6e-01
D	↓2e-01	↓1e+00	↓1e+00		↑1e-01	↓6e-01	↓3e-01	↓8e-01	↓9e-01	**↓4e-02**	↓3e-01	↓6e-01
E	**↓7e-03**	↓2e-01	↓2e-01	↓1e-01		↓4e-01	**↓3e-03**	↓8e-02	↓3e-01	**↓1e-03**	↓1e-02	↓4e-01
F	↓8e-02	↓6e-01	↓5e-01	↑6e-01	↑4e-01		↓5e-02	↓4e-01	↓7e-01	↓1e-02	↓8e-02	↓9e-01
G	↑9e-01	↑2e-01	↑4e-01	↑3e-01	**↑3e-03**	↑5e-02		↑3e-01	**↑3e-02**	↑5e-01	↓7e-01	**↑2e-02**
H	↑4e-01	↑7e-01	↑8e-01	↑8e-01	↑8e-02	↑4e-01	↓3e-01		↑4e-01	↓1e-01	↓3e-01	↑3e-01
I	↓1e-01	↓6e-01	↑8e-01	↑9e-01	↑3e-01	↑7e-01	**↓3e-02**	↓4e-01		**↓2e-02**	↓1e-01	↑7e-01
J	↑5e-01	↑8e-02	↑1e-01	**↑4e-02**	**↑1e-03**	↑1e-02	↓5e-01	↑1e-01	**↑2e-02**		↓7e-01	**↑7e-03**
K	↑9e-01	↑2e-01	↑3e-01	↑3e-01	**↑1e-02**	↑8e-02	↑7e-01	↑3e-01	↑1e-01	↑7e-01		↑7e-02
L	↓7e-02	↓4e-01	↓6e-01	↑6e-01	↑4e-01	↑9e-01	**↓2e-02**	↓3e-01	↓7e-01	**↓7e-03**	↓7e-02	

Results – best models

ID	Model
C	$\text{MAE}_{\text{train}} = 0.579, \text{MAE}_{\text{test}} = 0.448$, Absolute rank: 1 $-110.531 - \frac{2929.411 \sum \left(\left(-0.974 + \frac{2.68}{r} \right) \left(0.727 - \frac{2.888}{r} \right) \left(0.727 - \frac{1.747}{r} \right) \right)}{\sum \left(-\frac{0.972}{r(0.899r - 1.815)} \right)}$
E	$\text{MAE}_{\text{train}} = 0.612, \text{MAE}_{\text{test}} = 0.454$, Absolute rank: 2 $\frac{3.037 \left(-\sum \left(\frac{0.211}{r^2} \right) - 2.396 \right)}{\sqrt{\sum^2 \left(\left(-2.409 + \frac{6.254}{r} \right) \left(1.209 - \frac{4.99}{r} \right) \left(1.209 - \frac{2.956}{r} \right) \right) + 1}} - 101.086$
C	$\text{MAE}_{\text{train}} = 0.585, \text{MAE}_{\text{test}} = 0.458$, Absolute rank: 3 $-111.611 + \frac{12327.356 \sum \left(\left(-0.817 + \frac{2.014}{r} \right) \left(0.318 - \frac{1.255}{r} \right) \left(0.706 - \frac{1.913}{r} \right) \right)}{\sum \left(\frac{0.806}{r\left(0.307r - \frac{1.292}{r} \right)} \right)}$
C	$\text{MAE}_{\text{train}} = 0.550, \text{MAE}_{\text{test}} = 0.473$, Absolute rank: 6 $-108.409 + \frac{14618.749 \sum \left(\left(0.555 - \frac{1.538}{r} \right) \left(0.707 - \frac{1.667}{r} \right) \left(0.787 - \frac{3.116}{r} \right) \right)}{\sum \left(\frac{3.142}{r(1.922 - 0.953r)} \right)}$
B	$\text{MAE}_{\text{train}} = 0.549, \text{MAE}_{\text{test}} = 0.475$, Absolute rank: 7 $-109.903 - \frac{82734.094 \sum \left(\left(-0.361 + \frac{1.414}{r} \right) \left(0.527 - \frac{1.433}{r} \right) \left(0.622 - \frac{1.512}{r} \right) \right)}{\sum \left(\frac{0.873}{r\left(-0.339 + \frac{0.686}{r} \right)} \right)}$

Results

Unrestricted/undirected search enabled by $\sum$-symbols.

Relative simple primitive sets provide the ability to discover accurate potentials.

- discovered expressions do not resemble known potentials
- arithmetic-only configurations perform just as well on test data
- no advantage of analytical quotient over unprotected division
- a simple potential function was found

$$-110.531 - \frac{2929.411 \sum \left(\left(-0.974 + \frac{2.68}{r} \right) \left(0.727 - \frac{2.888}{r} \right) \left(0.727 - \frac{1.747}{r} \right) \right)}{\sum \left(-\frac{0.972}{r(0.899r - 1.815)} \right)}$$

Conclusion

Symbolic regression is a powerful approach for obtaining simple and accurate potentials.

Further work is needed to establish the right approach in terms of:

- Primitive set – possibly augmented with new operators
- Emphasis on respecting fundamental physical laws
- Problem-specific extensions and hybridizations
  - multi-target GP
  - shape constraints
  - derivative information
- GP deployment in HPC environments