MACHINE LEARNING AND ACTIVE EXPLORATION OF CHEMICAL SPACE

K. Gubaev, E. Podryabinkin, I. Novikov, A. Shapeev

Skolkovo Institute of Science and Technology

1

Task

Numerical experiments

Obtain properties of matter (e.g. formation energy, polarizability, phase stability) which are difficult/expensive to get experimentally or are unobservable at all.

Tool

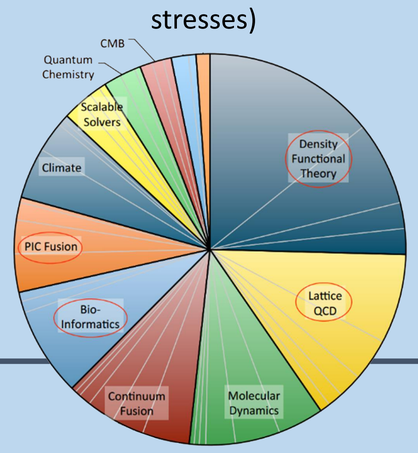
Molecular dynamics

Used for modelling of physical processes on the atomic level. In principle can provide a lot of valuable information.

Problem

Computational cost

Requires computationally demanding quantum mechanical calculations (energies, forces, stresses)



Solution

Machine Learning Interatomic Potentials

To construct a surrogate machine learning model predicting quantum mechanical data. We want to fit $E^{\mathrm{qm}}(\boldsymbol{X})$ with $E(\boldsymbol{X})$, which means being able to predict energy E for a given atomic configuration $\boldsymbol{X}$, as close as possible to the quantum data $E^{\mathrm{qm}}(\boldsymbol{X})$. This results in minimizing the following functional:

$$F = \sum_i \left|E\left(x^{(i)}\right) - E^{\mathrm{qm}}\left(x^{(i)}\right)\right|^2 + (\text{forces}) + \ \ldots$$

It requires:

- Regression model to learn
- Reference data (training set)
- Optimization algorithm

2

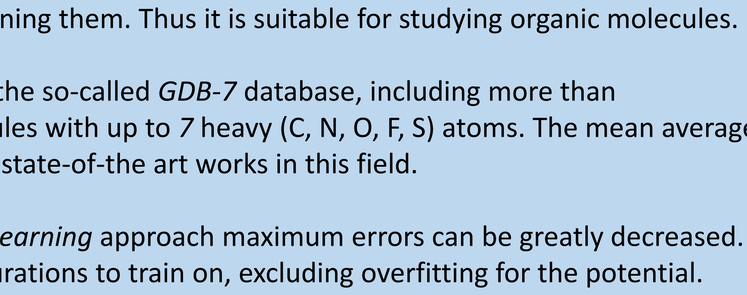
Regression model

Moment Tensor Potentials

Common approach is to introduce locality of interatomic interactions. It works for most cases, when electrostatic forces can be neglected, and gives us energy partitioning $E = \sum_i V(r_{i1}, r_{i2}, \ldots)$, where each atom interacts only with its neighborhood in some cutoff radius (~ 6 Å)

The problem is to find a good V, employing some parameterized functional form and then choose optimal parameters which minimize F function.

$\boldsymbol{u} = \{r_{i1}, \ldots, r_{in}\}$ – atomic environment



M^α – inertia tensor, invariant to translation, rotation and permutation of equivalent atoms

Examples of descriptors:

$$M_i^0(\boldsymbol{u}) = r_{i1} + \ldots + r_{in}$$
$$M_i^1(\boldsymbol{u}) = \boldsymbol{r}_{i1} + \ldots + \boldsymbol{r}_{in}$$
$$M_i^2(\boldsymbol{u}) = \boldsymbol{r}_{i1} \otimes \boldsymbol{r}_{i1} + \ldots + \boldsymbol{r}_{in} \otimes \boldsymbol{r}_{in}$$

Examples of basic functions:

$$B^0(\boldsymbol{u}) = M^0(\boldsymbol{u})$$
$$B^1(\boldsymbol{u}) = M^1(\boldsymbol{u}) \cdot M^1(\boldsymbol{u})$$
$$B^2(\boldsymbol{u}) = M^0(\boldsymbol{u})\left(M^2(\boldsymbol{u}) : M^2(\boldsymbol{u})\right)$$

$V(\boldsymbol{u}; \theta) = \sum_\alpha \theta^\alpha B^\alpha(\boldsymbol{u})$, where $B^\alpha(\boldsymbol{u})$ are all different contractions of $M^\alpha(\boldsymbol{u})$ yielding a scalar. Varying the set of $\{\alpha\}$ we can switch the potential. $B^\alpha(\boldsymbol{u})$ is complete basis, see [1] for proofs.

1 - Shapeev, A. V. (2016). Moment Tensor Potentials: a class of systematically improvable interatomic potentials. Multiscale Modeling & Simulation, 14(3), 1153-1173.

3

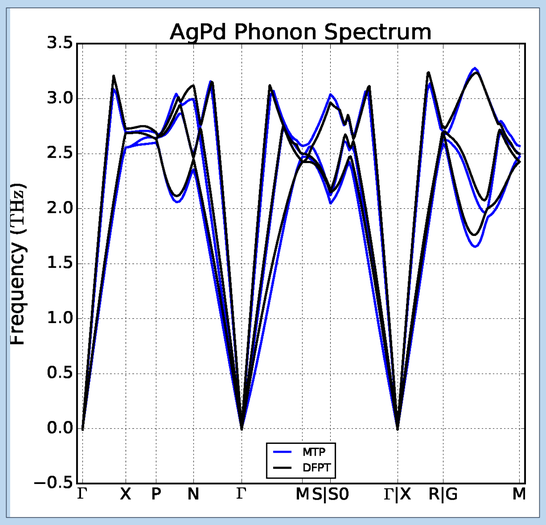
Application 1

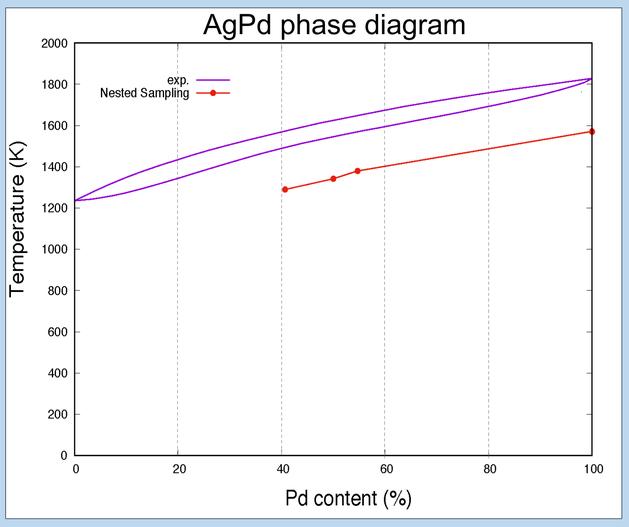
Metallic alloys

The phonon spectrum and a part of the phase diagram for the AgPd system, calculated with MTP potential. Note that fitted DFT data itself may contain errors in melting temperature. The complexity of system processing with MTP potentials scales as N, where N is number of atoms. For comparison, complexity of DFT calculation scales as N^3.

Binary MTP potential was trained on liquid and deformed solid AgPd configurations calculated via VASP DFT package.

MTP potential can be integrated into LAMMPS for carrying on MD simulations. In this case, MTP works as a black-box providing energies/forces/stresses for incoming configurations.





4

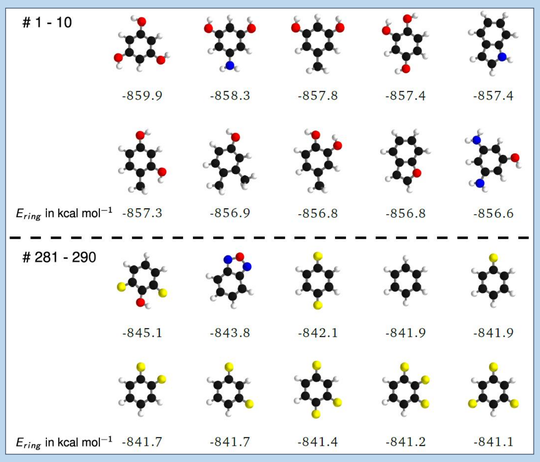
Application 2

Organic molecules

MTP is a reactive potential, which means it can create and destroy interatomic bonds without explicitly defining them. Thus it is suitable for studying organic molecules.

We've fitted the so-called *GDB-7* database, including more than *7,000* molecules with up to *7* heavy (C, N, O, F, S) atoms. The mean average errors are listed and compared to state-of-the art works in this field.

With *Active Learning* approach maximum errors can be greatly decreased. It selects the optimal set of configurations to train on, excluding overfitting for the potential.



MAE	Models				
Property	MTP	MTP, active	BAML[1]	SOAP[2]	DTNN[3]
atomization energy, [kcal/mol]	**0.55**	**0.43**	1.15	0.92	1.04
atomization energy max.err, [kcal/mol]	**20**	**3.79**			
Polarizability, [Å^3]	**0.04**		0.07	0.05	-
HOMO_gw, [eV]	0.12		**0.1**	0.12	-
LUMO_gw, [eV]	0.12		**0.11**	0.12	-

1 - Huang, B., & von Lilienfeld, O. A. (2016) The Journal of Chemical Physics.
2 - De, S., Bartók, A. P., Csányi, G., & Ceriotti, M. (2016). Physical Chemistry Chemical Physics
3 - Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Nature Communications

5

Improving transferability

The Active Learning approach

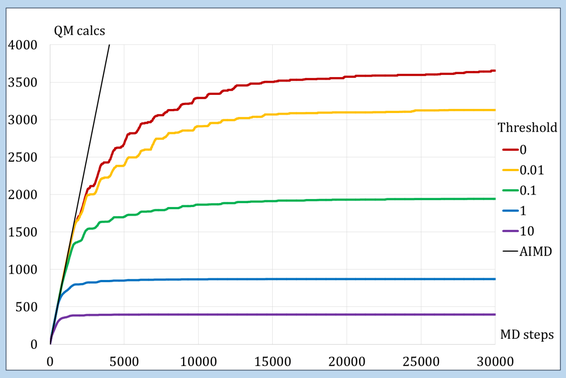
The main idea is to predict, whether our potential will extrapolate on a given configuration and if yes, include it in the training set.

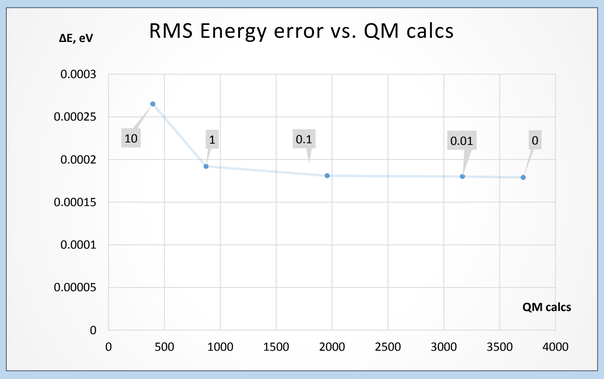
This makes *Learning on the Fly* possible: we run MD with our potential, calling slow QM calculations only if extrapolating configuration occurs. This significantly reduces computation time and almost preserves accuracy.

QM calculations are launched only when configuration occurred in MD is "new" to our potential.

With time, the potential learns itself and requires less and less quantum data, which boosts the whole MD significantly.

Below you can see number of QM calculations with MD step and accuracy for different thresholds, which means allowed degree of extrapolation.





6

Desired impact

Software allowing fast & accurate MD

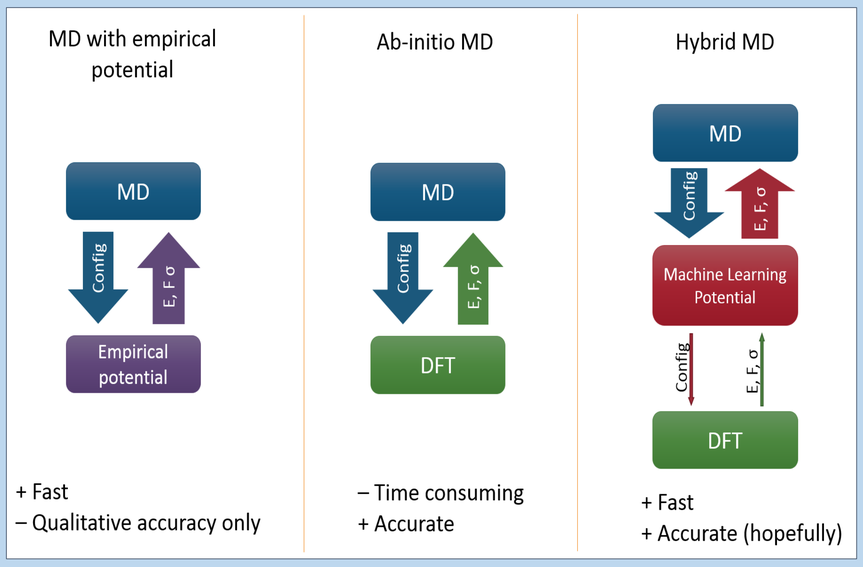
Ab-initio MD is too slow.

MD with empirical potentials is too inaccurate.

MD with our Machine Learning Potential is fast & precise.

Fully automated usage is possible, no additional expertise required!

Our package launches DFT when needed, otherwise provides much faster response.



Contact: konstantin.gubaev@skoltech.ru