XACS
Xiamen Atomistic Computing Suite
XACScloud.com

MLatom.com



AI-enhanced computational chemistry

Pavlo O. Dral
MLatom@XACS, Xiamen University
dral@xmu.edu.cn

27 May 2024

Pavlo O. Dral

dr-dral.com

Lecture 6

Is your data sufficient for ML? Getting data

Pavlo O. Dral

Task 6.1.

Investigate how the performance of the ML model changes when you train it on H_2 PES with:

1. 40 randomly sampled points
2. 20 points chosen using farthest-point sampling

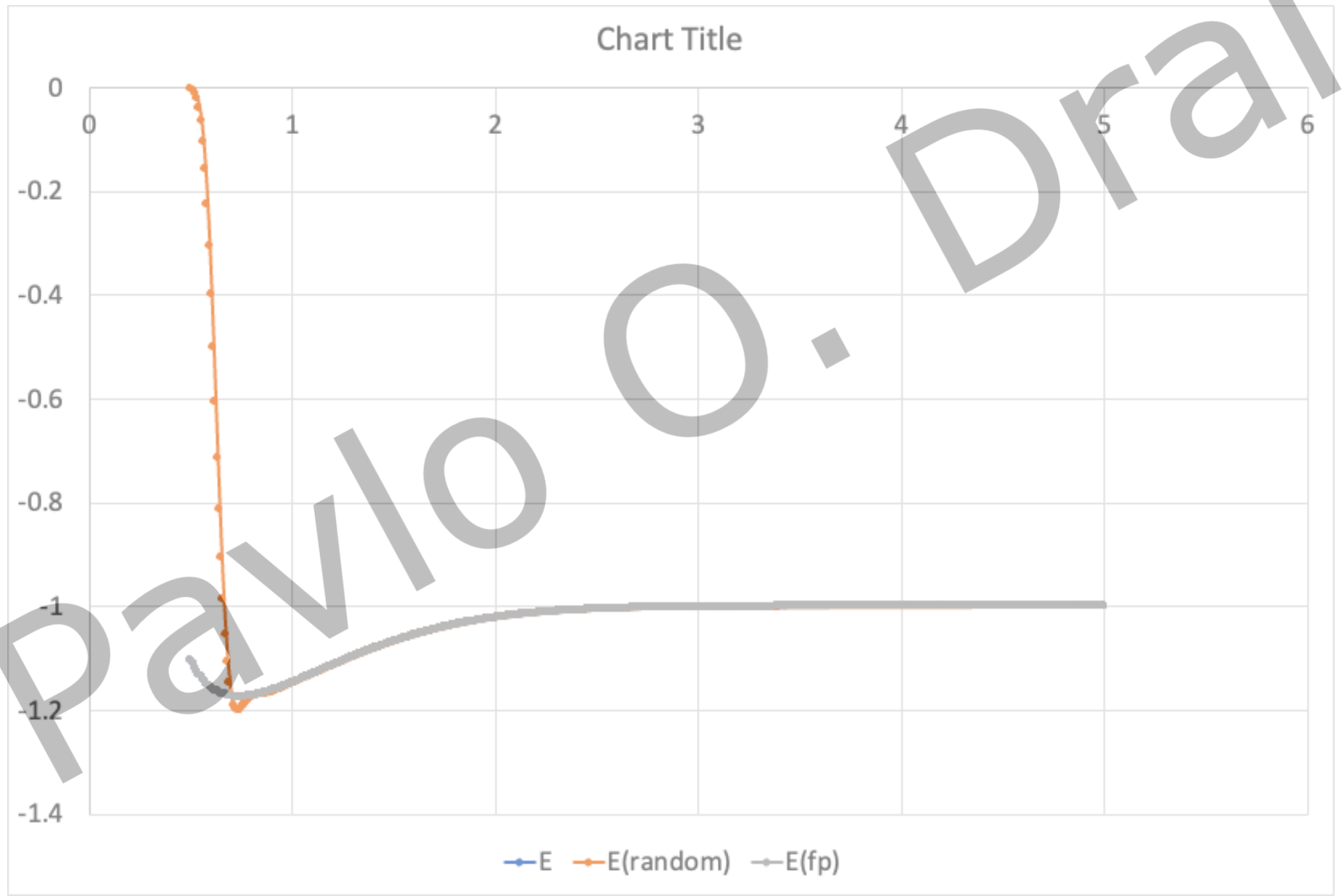
Compare both models':

1. Test RMSE.
2. Performance on the potential energy curve.

Input file:

```
# Uncomment and adjust the lines below as needed:
# sampling=random
# sampling=farthest-point
# Ntrain=                 # Number of training points
estAccMLmodel             # Specify the task for MLatom
XYZfile=h2.xyz            # File with XYZ geometries
Yfile=E_FCI_451.dat       # The file with FCI energies
YestFile=E_est.dat        # File with ML estimated energies
MLmodelType=KREG          # Specify the model type
Nsubtrain=0.9             # request 90% of the training points to be used for validation
sigma=opt
lgSigmaL=-4               # Lower bound of log2(sigma)
lambda=opt
```

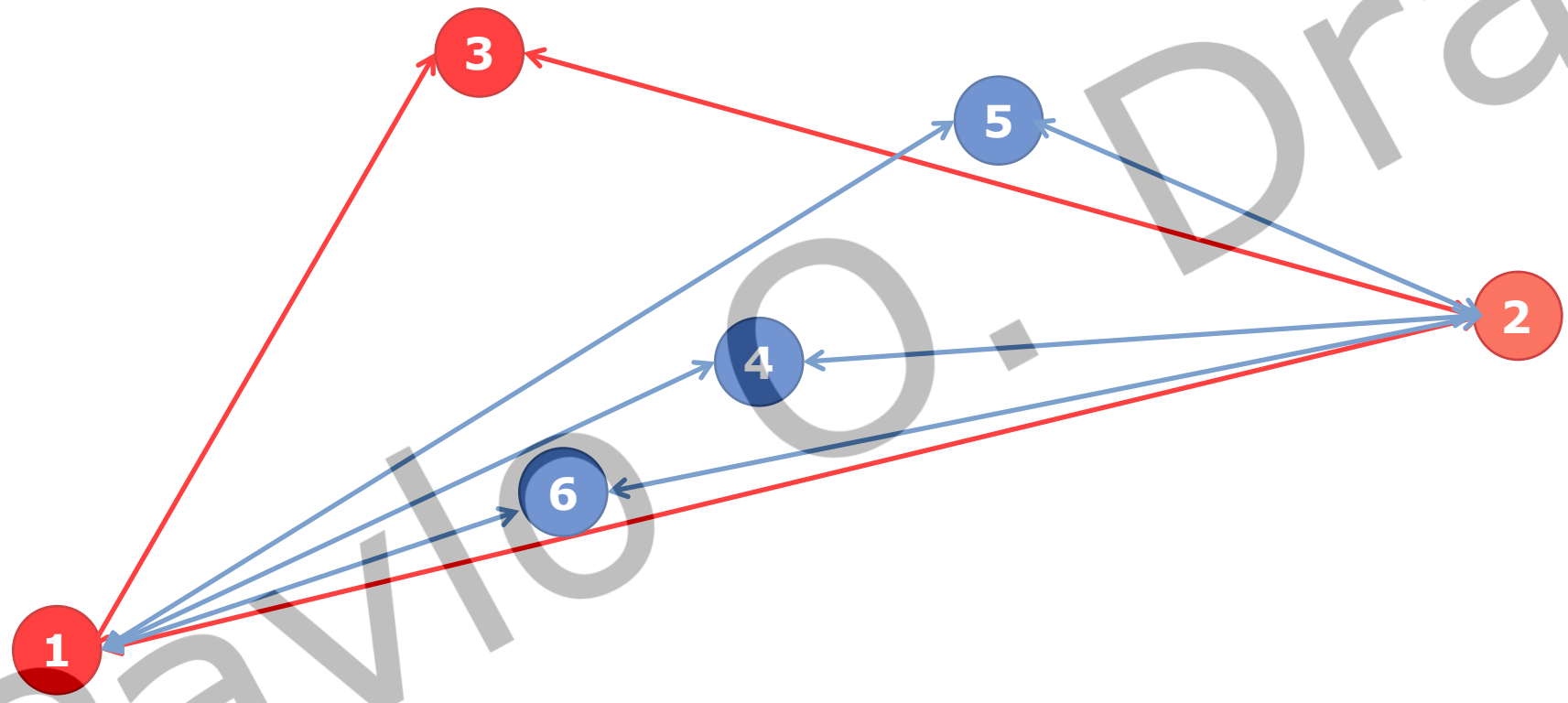
Pavlo O. Dral



Pavlo O. Dral

Farthest-point sampling

Distances in two dimensional space



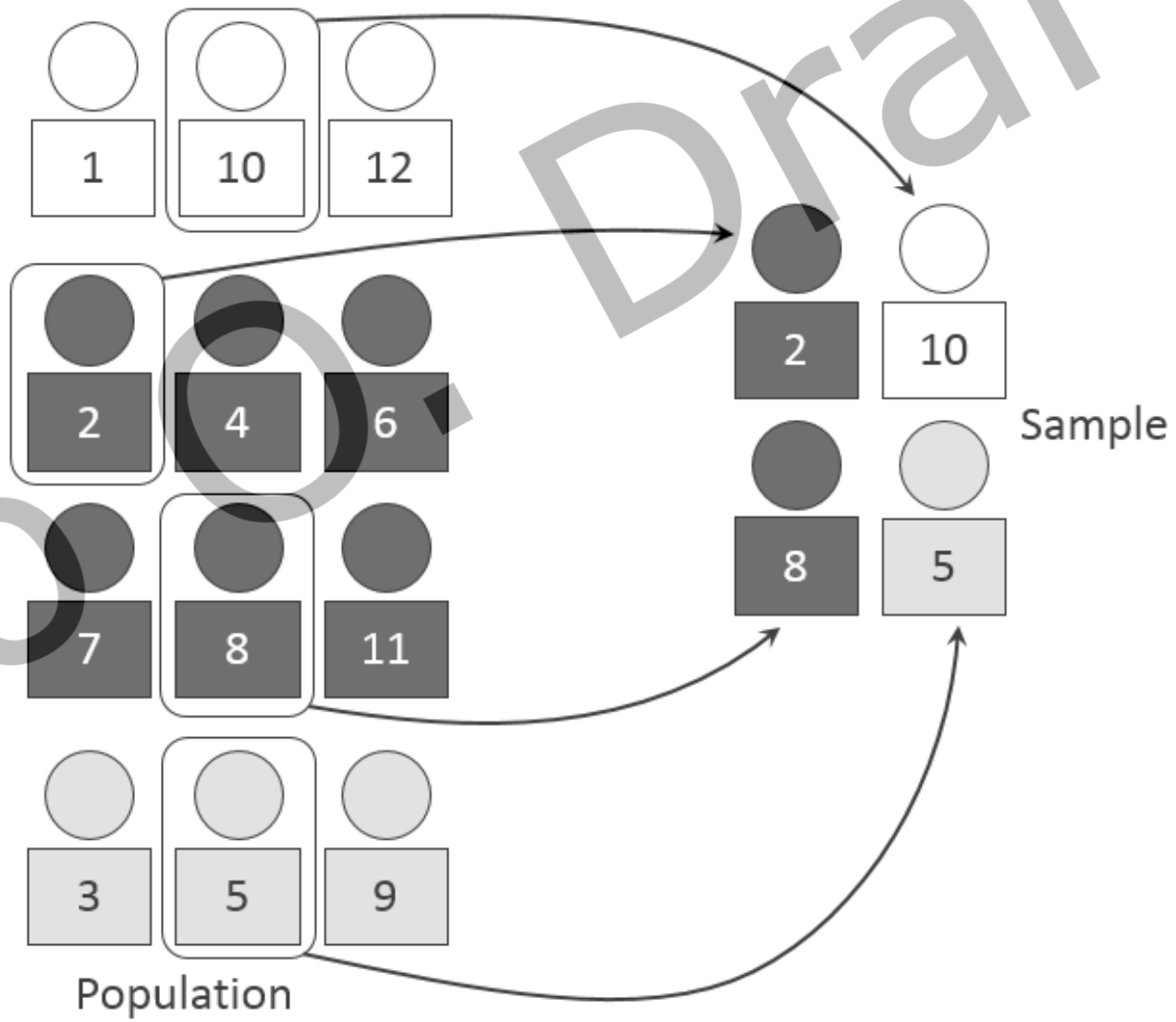
$$D(\mathbf{M}_a, \mathbf{M}_i) = \sqrt{\sum_{\beta}^{N_{\mathbf{M}}} \left(\mathbf{M}_{i,\beta} - \mathbf{M}_{a,\beta}\right)^2}$$

Pavlo O. Dral

Model Selection

Random sampling for model selection is not always a good idea

Sometimes, **stratification** is preferable



Pavlo O. Dral

Figure by Dan Kernler [CC BY-SA 4.0], from Wikimedia Commons

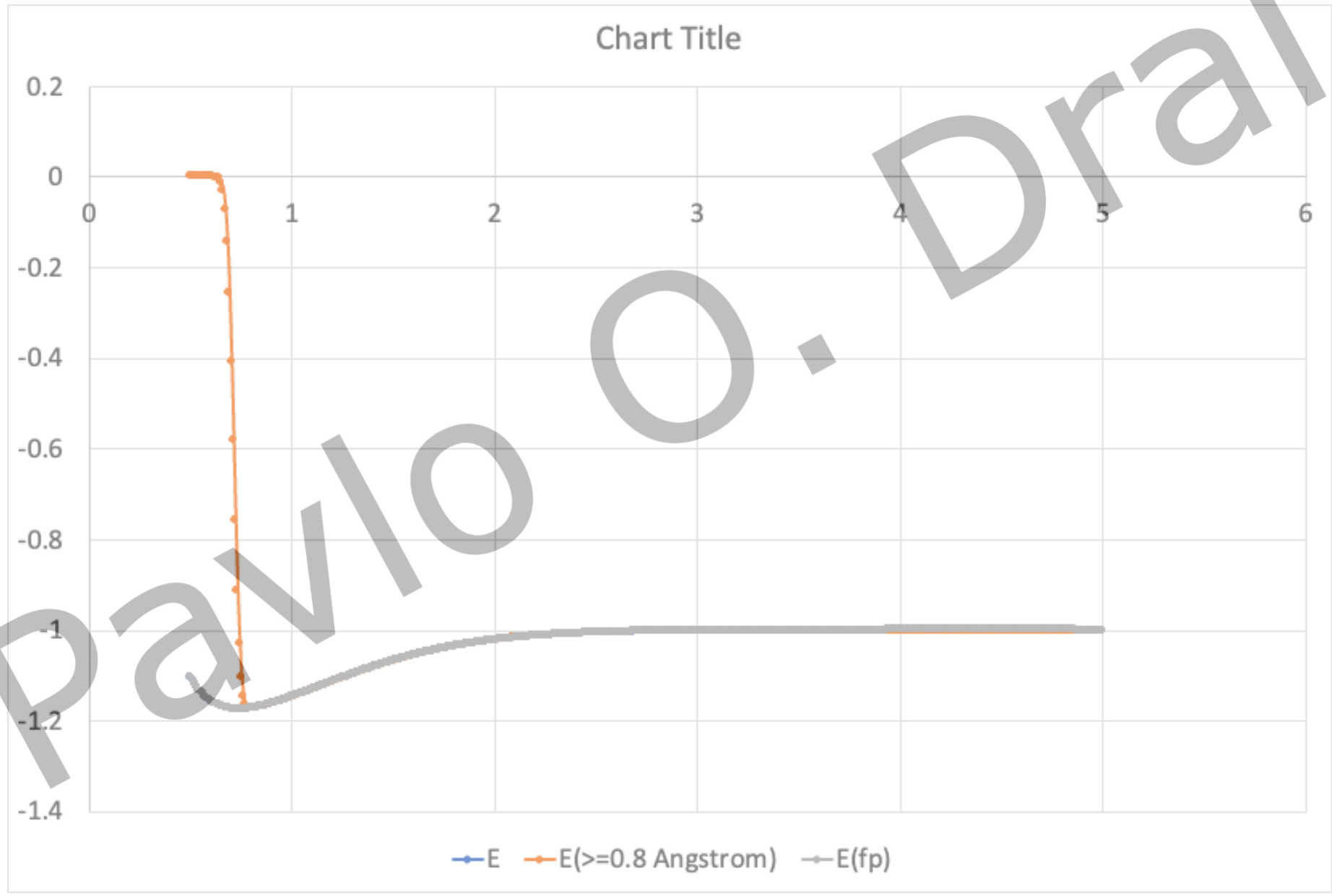
Task 6.2.

Train the KREG model on the H_2 data set for **interatomic distances larger or equal 0.8 Angstrom**. Calculate energies with this KREG model for all internuclear distances from 0.5 to 5 Angstrom. See the previous tasks how to train and predict.

Plot the potential energy curve:

1. How does this KREG model perform?
2. How many training points did you use?
3. Compared to the KREG models you trained previously (e.g., in the previous task)?

Pavlo O. Dral

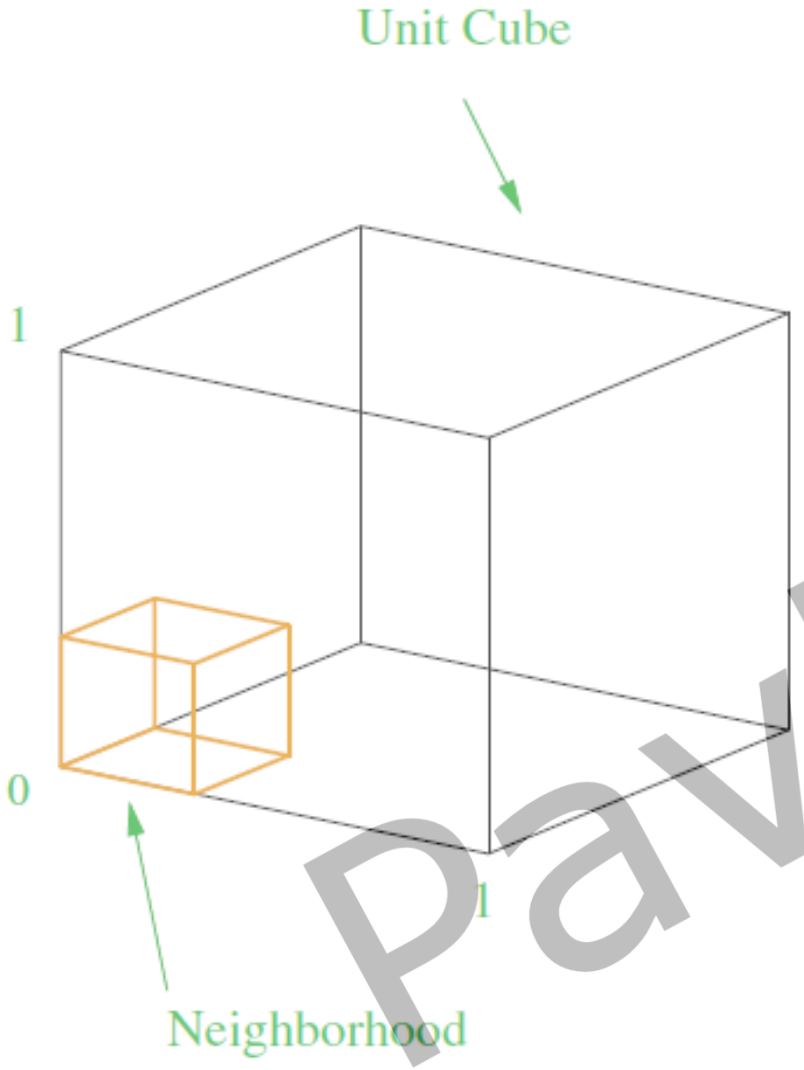


Pavlo O. Dral

How would you proceed to create good models when you are limited to small data?

Pavlo O. Dral

Curse of Dimensionality



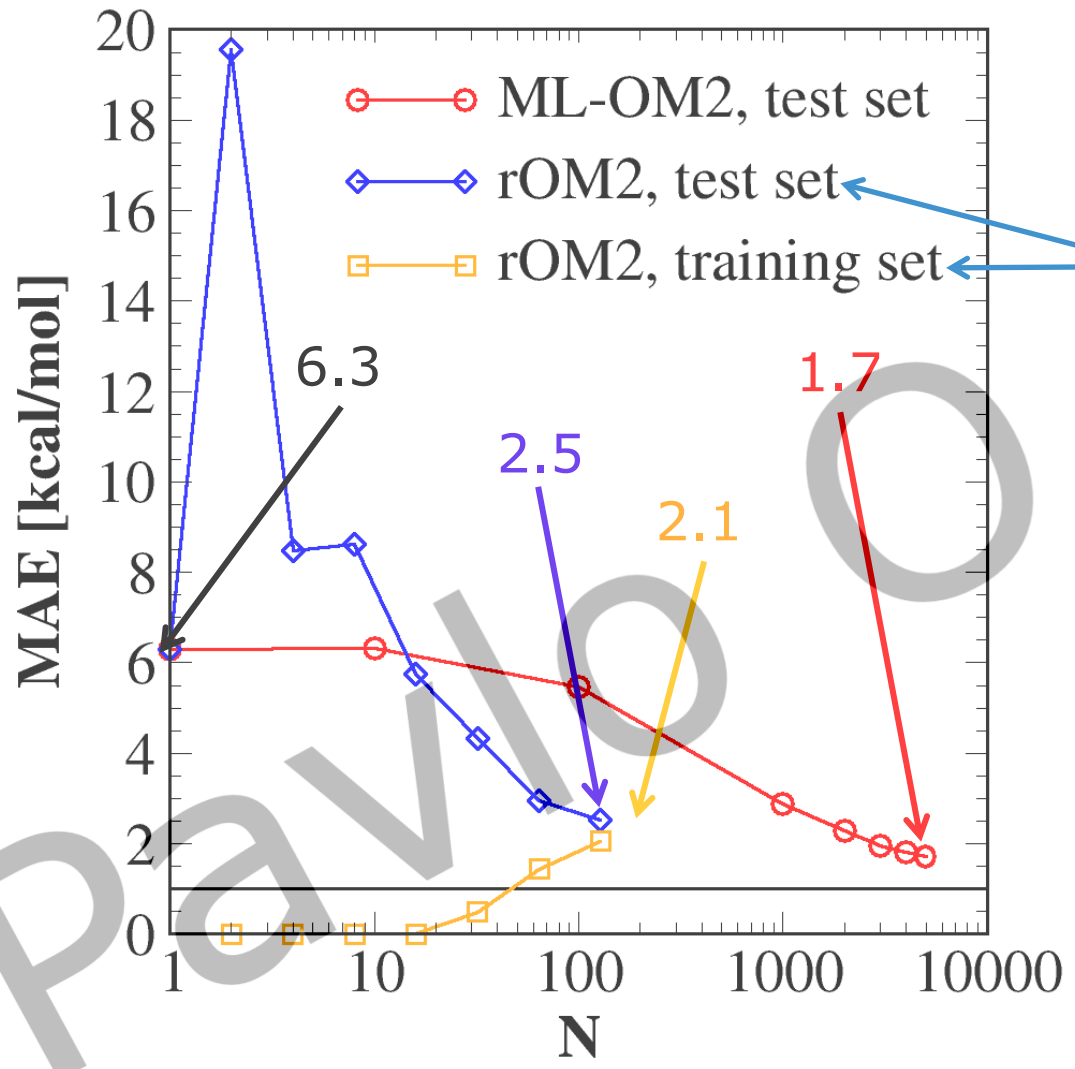
To keep the same sampling density with increasing number *d* of dimensions you need to sample N^d training point.

Thus, in high dimensions you can only afford **very, very sparse sampling**

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed.; Springer-Verlag, **2009**

Pavlo O. Dral

Improving Hamiltonian

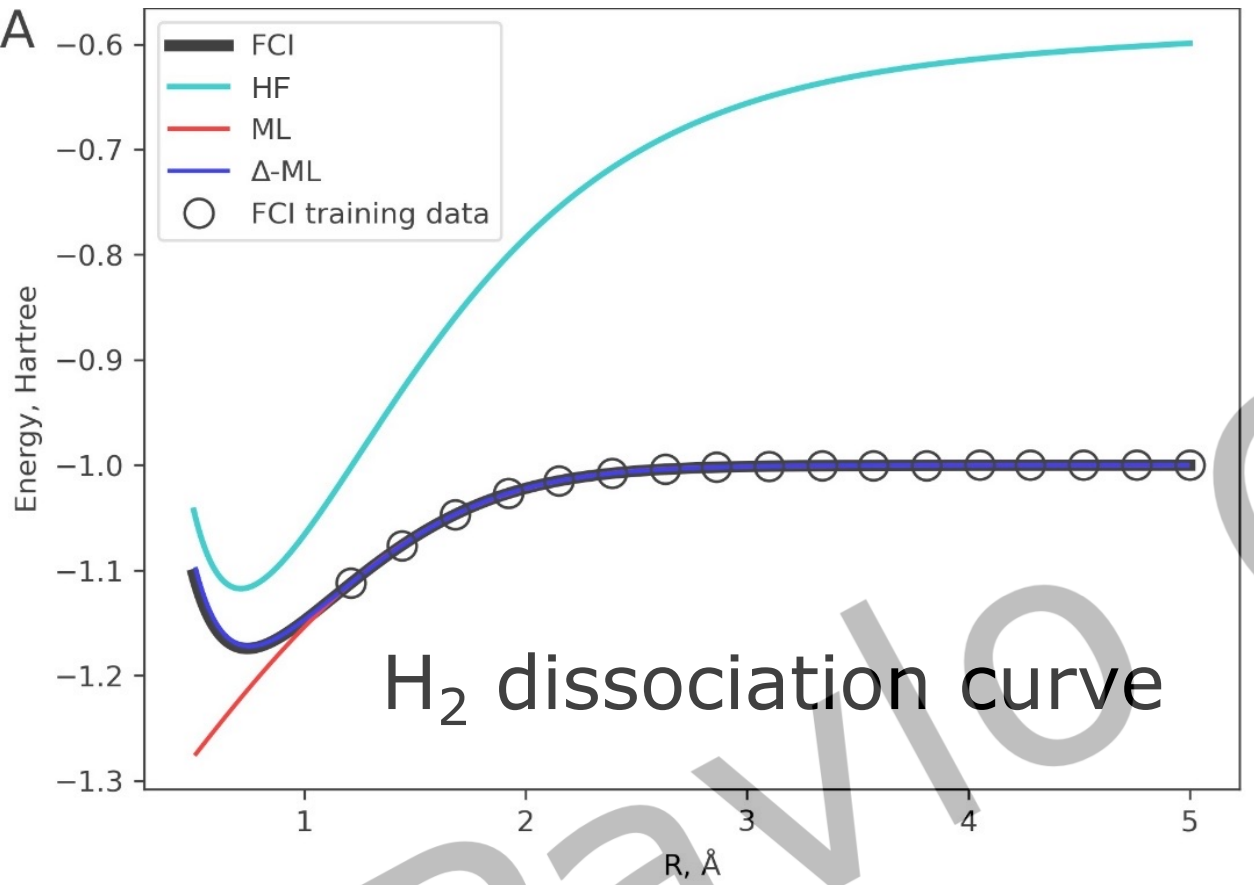


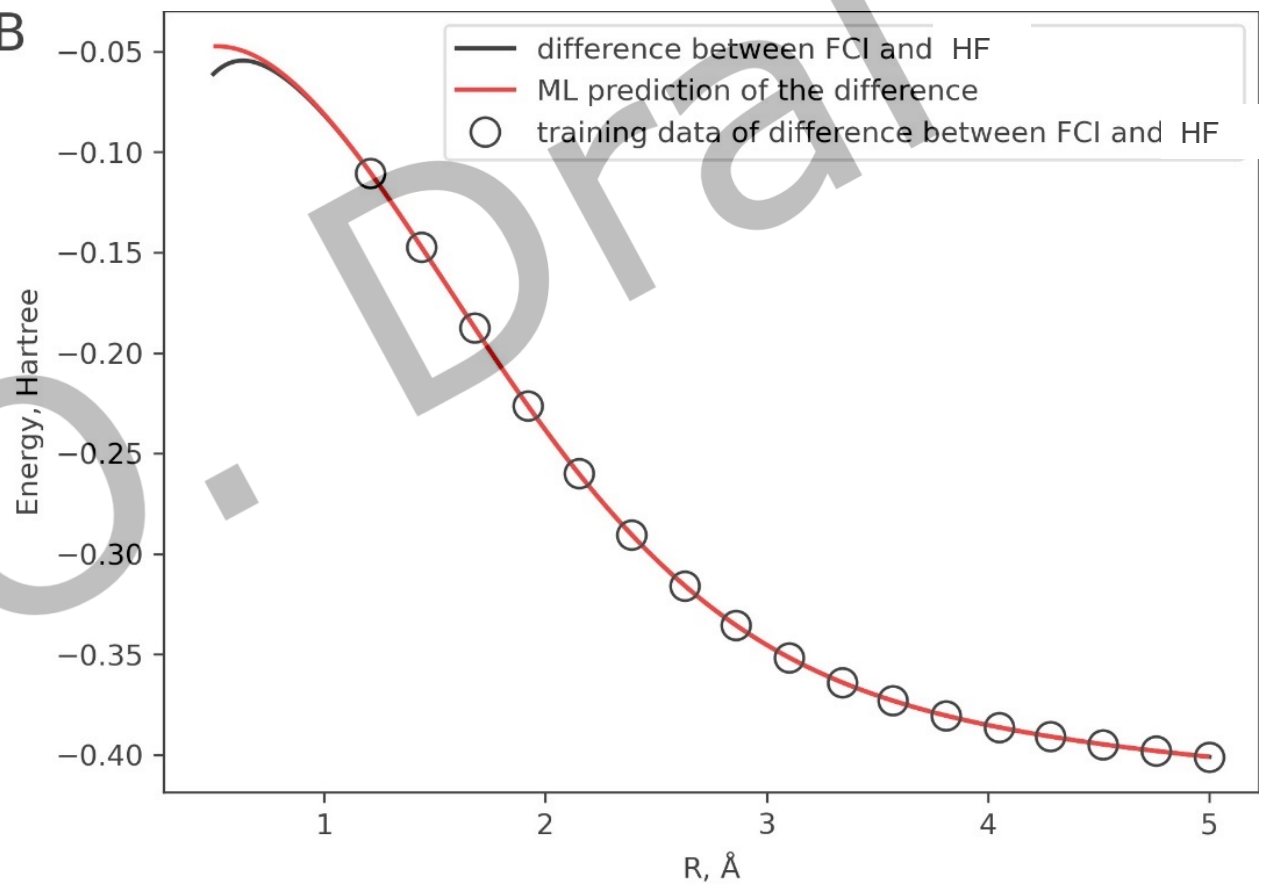
Traditional reparametrization (molecule-specific orbital exponents in the basis set)

Pavlo O. Dral

P. O. Dral, O. A. von Lilienfeld, W. Thiel, *J. Chem. Theory Comput.* **2015**, *11*, 2120

Multi-fidelity data: Δ-learning





Pavlo O. Dral

Δ-learning: R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087

Figure: Pavlo O. Dral, Tetiana Zubatiuk, Bao-Xin Xue, Learning from multiple quantum chemical methods: Δ-learning, transfer learning, co-kriging, and beyond. In *Quantum Chemistry in the Age of Machine Learning*, Pavlo O. Dral, Ed. Elsevier: **2023**. Paperback ISBN: 9780323900492

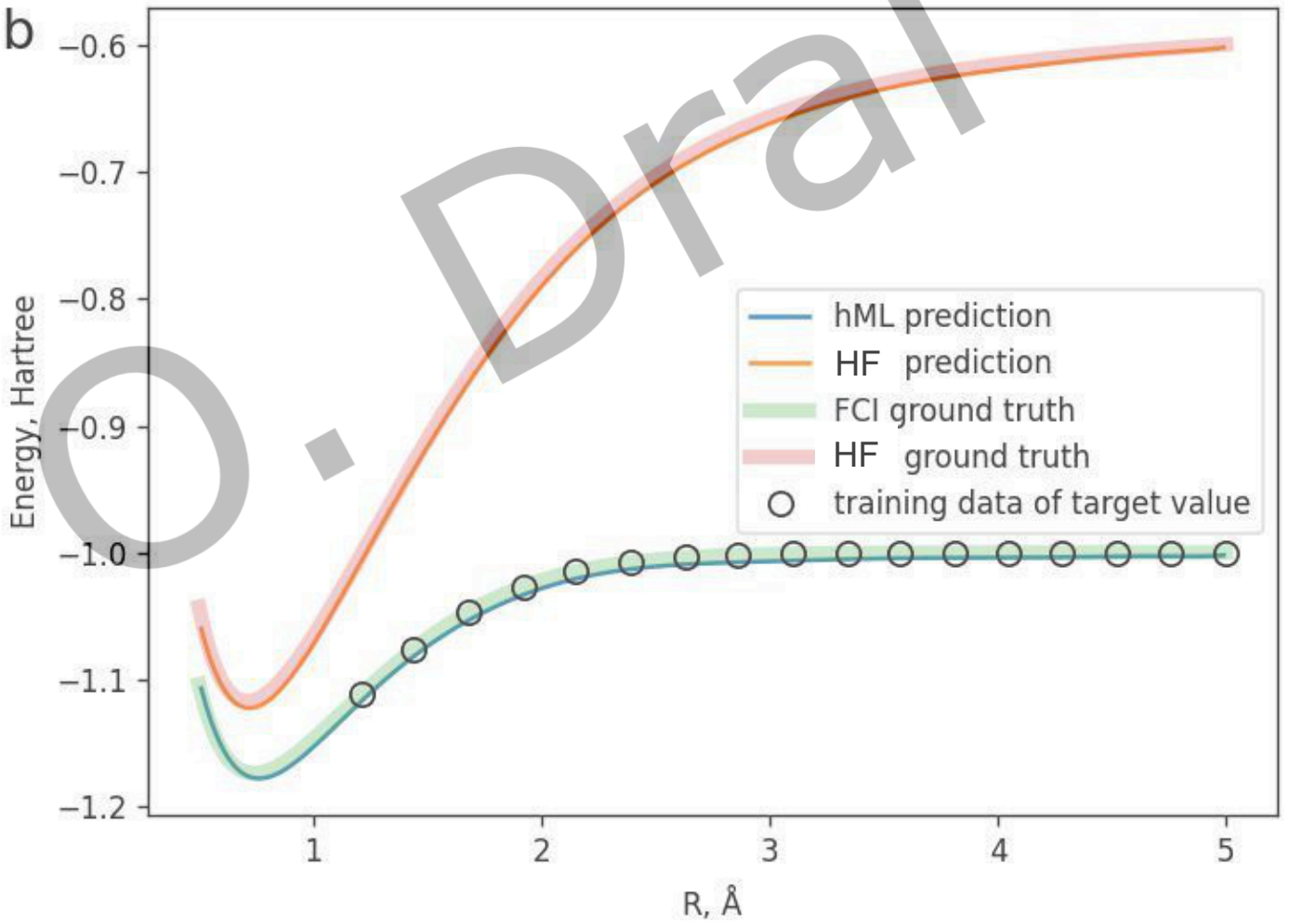
Hierarchical ML (hML): generalized Δ-learning

Automatic procedure to find optimal training points for each delta-model

$$\hat{y}_{\Sigma,i} = \sum_M \hat{y}_{M,i}\left(N_{\mathrm{tr},M}\right)$$

hML for H_2

$$\mathrm{hML} = \Delta_0^{\mathrm{HF}} + \Delta_{\mathrm{HF}}^{\mathrm{FCI}}$$



Hierarchical ML: P. O. Dral, A. Owens, A. Dral, G. Csányi, *J. Chem. Phys.* **2020**, *152*, 204110

Figure: Pavlo O. Dral, Tetiana Zubatiuk, Bao-Xin Xue, Learning from multiple quantum chemical methods: Δ-learning, transfer learning, co-kriging, and beyond. In *Quantum Chemistry in the Age of Machine Learning*, Pavlo O. Dral, Ed. Elsevier: **2023**. Paperback ISBN: 9780323900492

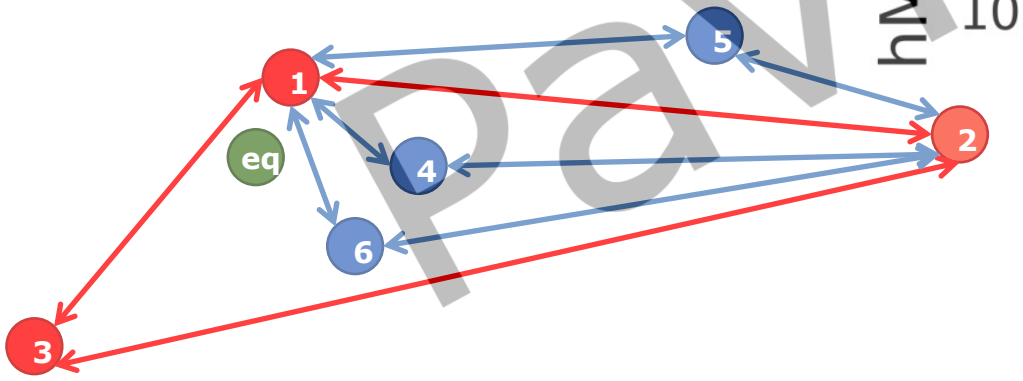
Pavlo O. Dral
dr-dral.com

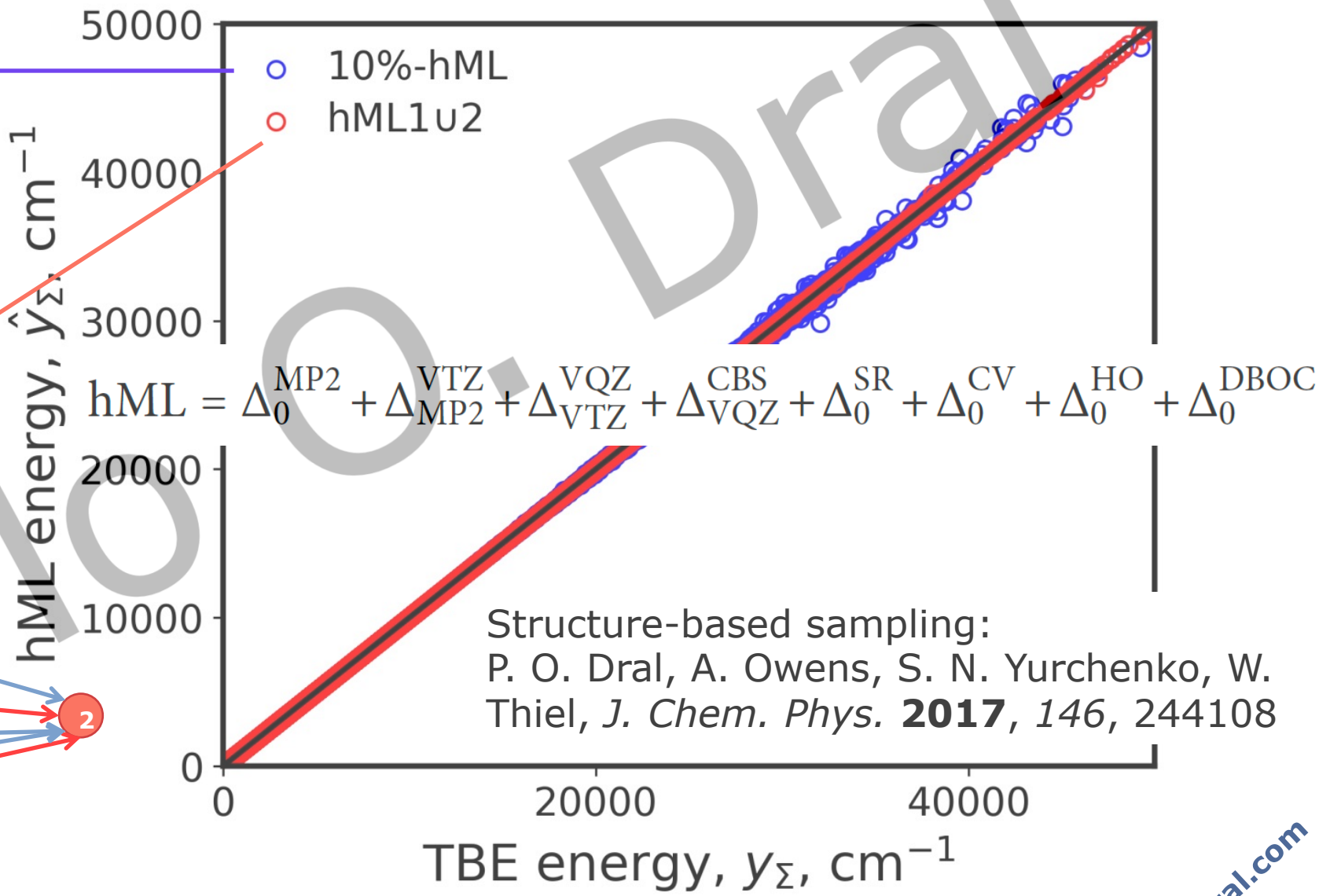
Hierarchical ML for CH_3Cl PES

Pure ML-model trained on **saved 90% of CPU-time**

Weighted RMSE:
3.49 cm⁻¹ = 0.01 kcal/mol

Δ-ML models:
saved 99% of CPU-time
Weighted RMSE:
1.12 cm⁻¹ = 0.003 kcal/mol





$$\mathrm{hML} = \Delta_0^{\mathrm{MP2}} + \Delta_{\mathrm{MP2}}^{\mathrm{VTZ}} + \Delta_{\mathrm{VTZ}}^{\mathrm{VQZ}} + \Delta_{\mathrm{VQZ}}^{\mathrm{CBS}} + \Delta_0^{\mathrm{SR}} + \Delta_0^{\mathrm{CV}} + \Delta_0^{\mathrm{HO}} + \Delta_0^{\mathrm{DBOC}}$$

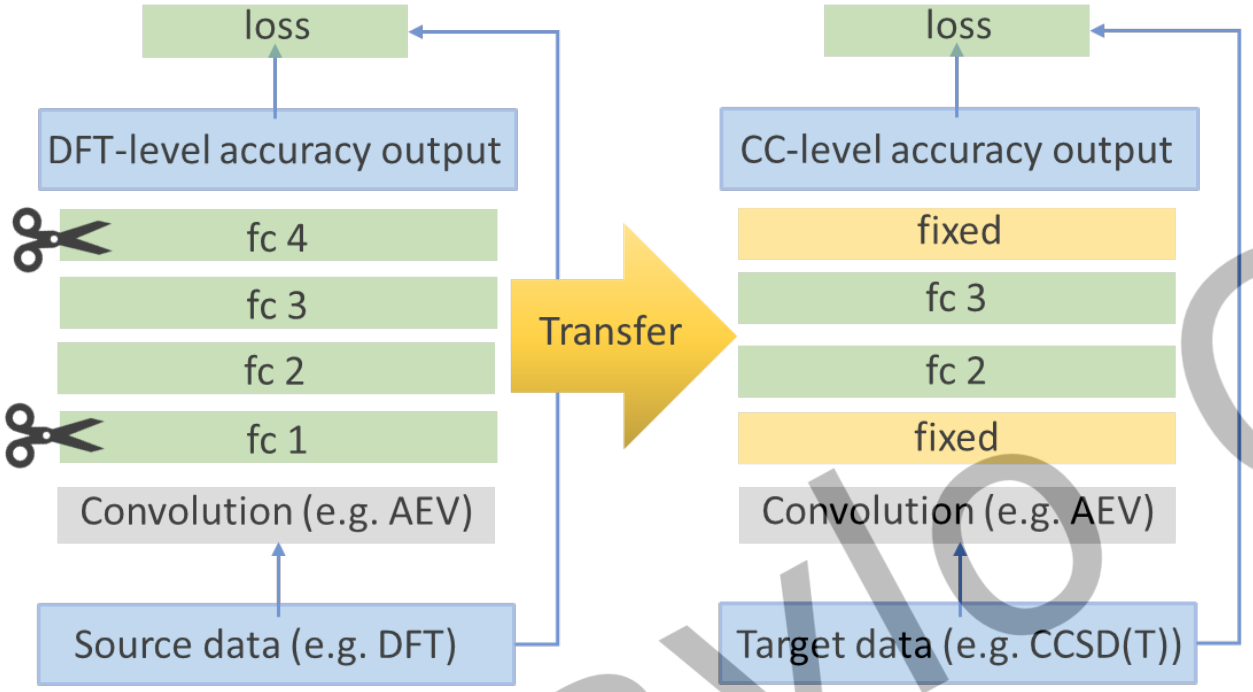
Structure-based sampling:
P. O. Dral, A. Owens, S. N. Yurchenko, W. Thiel, *J. Chem. Phys.* **2017**, *146*, 244108

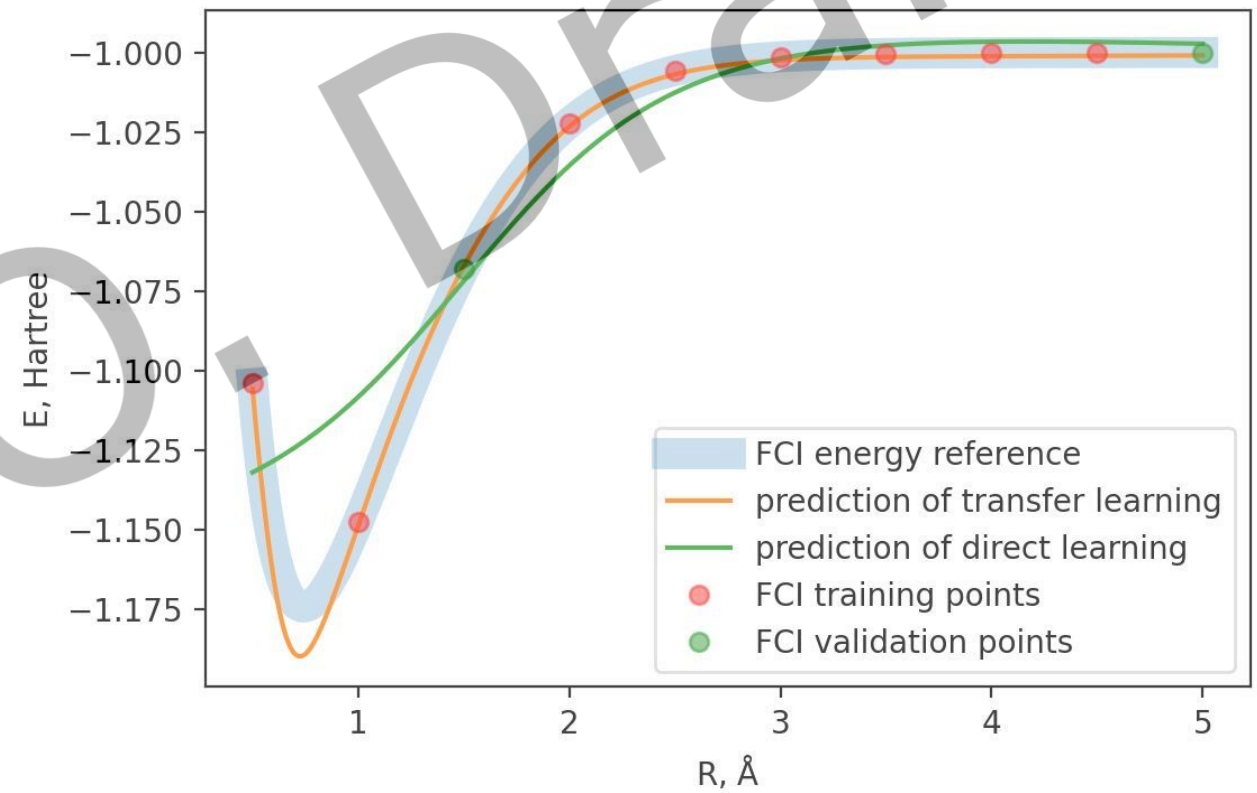
Pavlo O. Dral

P. O. Dral, A. Owens, A. Dral, G. Csányi, *J. Chem. Phys.* **2020**, *152*, 204110

dr-dral.com

Multi-fidelity data: Transfer learning

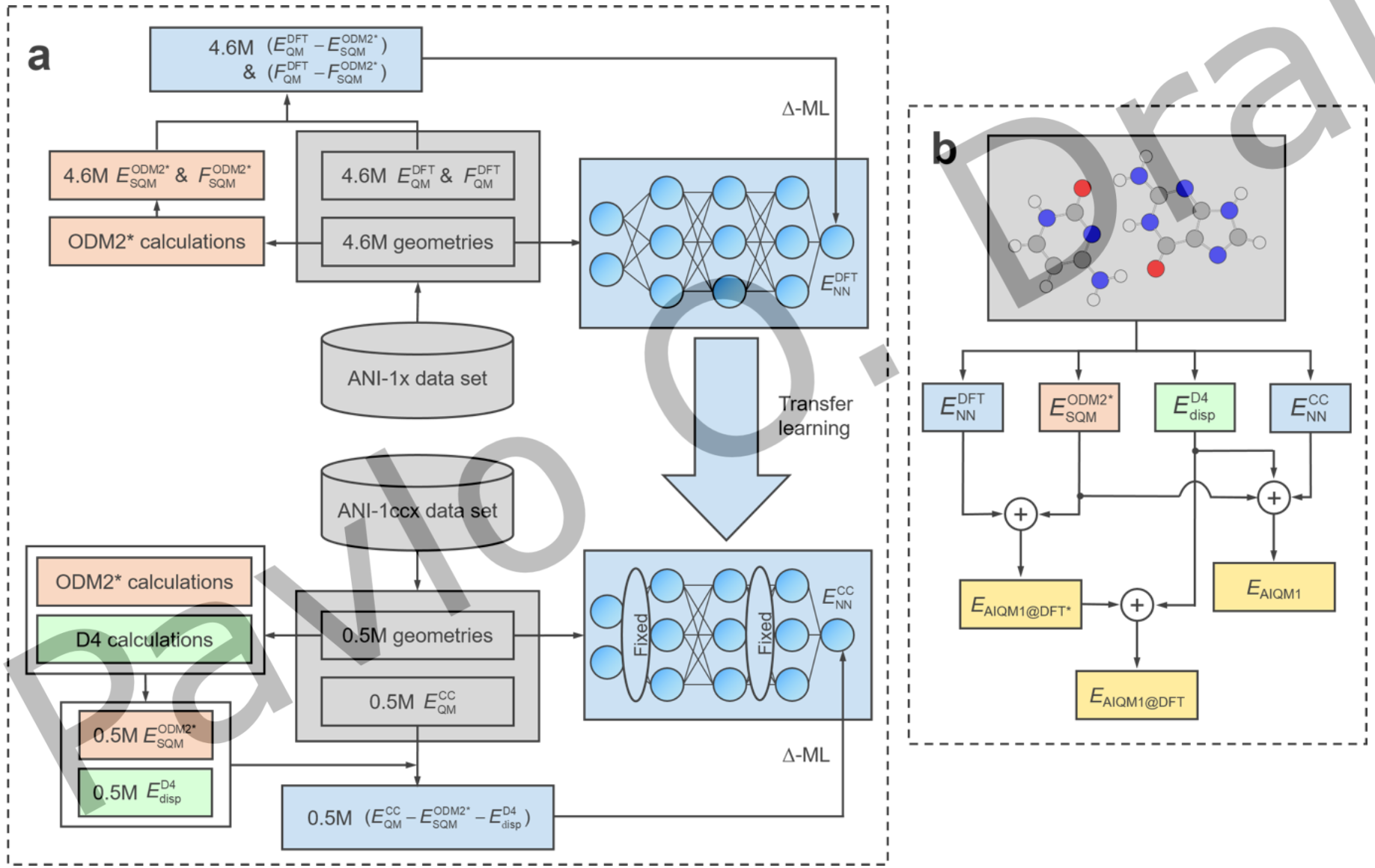




Pavlo O. Dral

Figures: Pavlo O. Dral, Tetiana Zubatiuk, Bao-Xin Xue, Learning from multiple quantum chemical methods: Δ-learning, transfer learning, co-kriging, and beyond. In *Quantum Chemistry in the Age of Machine Learning*, Pavlo O. Dral, Ed. Elsevier: **2023**. Paperback ISBN: 9780323900492

Design



Pavlo O. Dral

P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, P. O. Dral, *Nat. Commun.* **2021**, *12*, 7022

Categories of machine learning potentials

Data ?

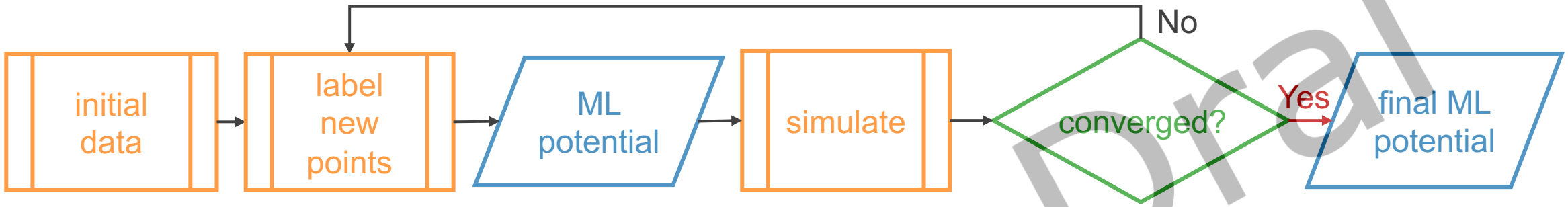


Figure based on M. Pinheiro Jr, F. Ge, N. Ferré, P. O. Dral, M. Barbatti. *Chem. Sci.* **2021**, *12*, 14396–14413

Models implemented/interfaced in MLatom
not implemented in MLatom

Pavlo O. Dral

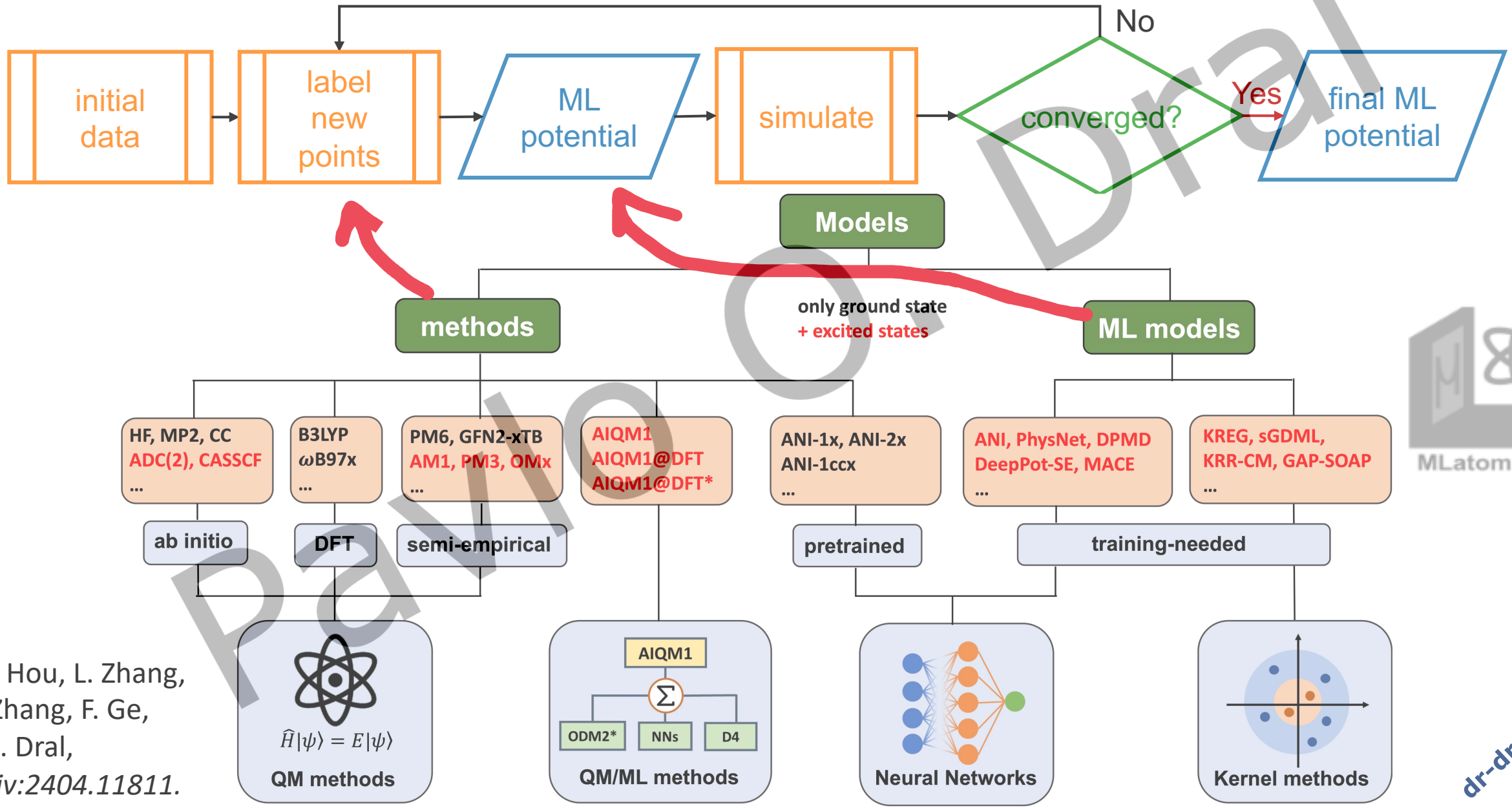
Physics-informed active learning



Pavlo O. Dral

Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral, *arXiv:2404.11811.*

Physics-informed active learning

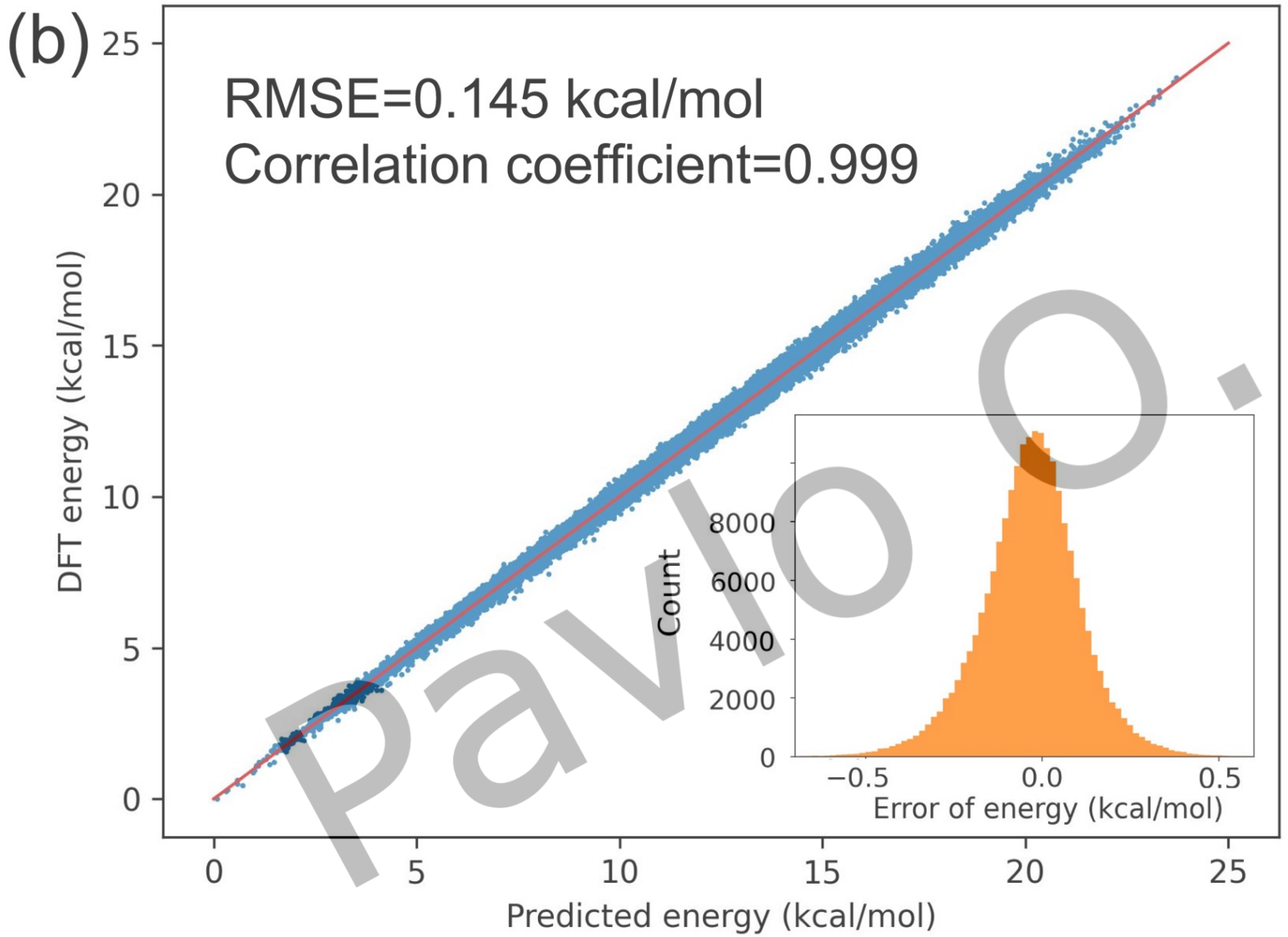


Pavlo O. Dral

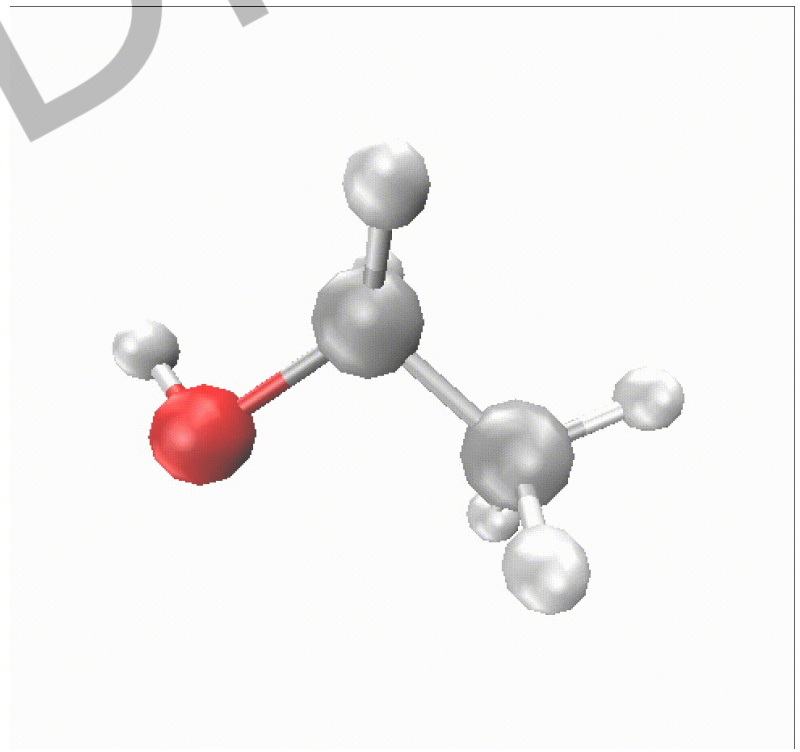
MLatom.com

Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral, *arXiv:2404.11811.*

Accurate PES



< 1000 training points with ANI potential

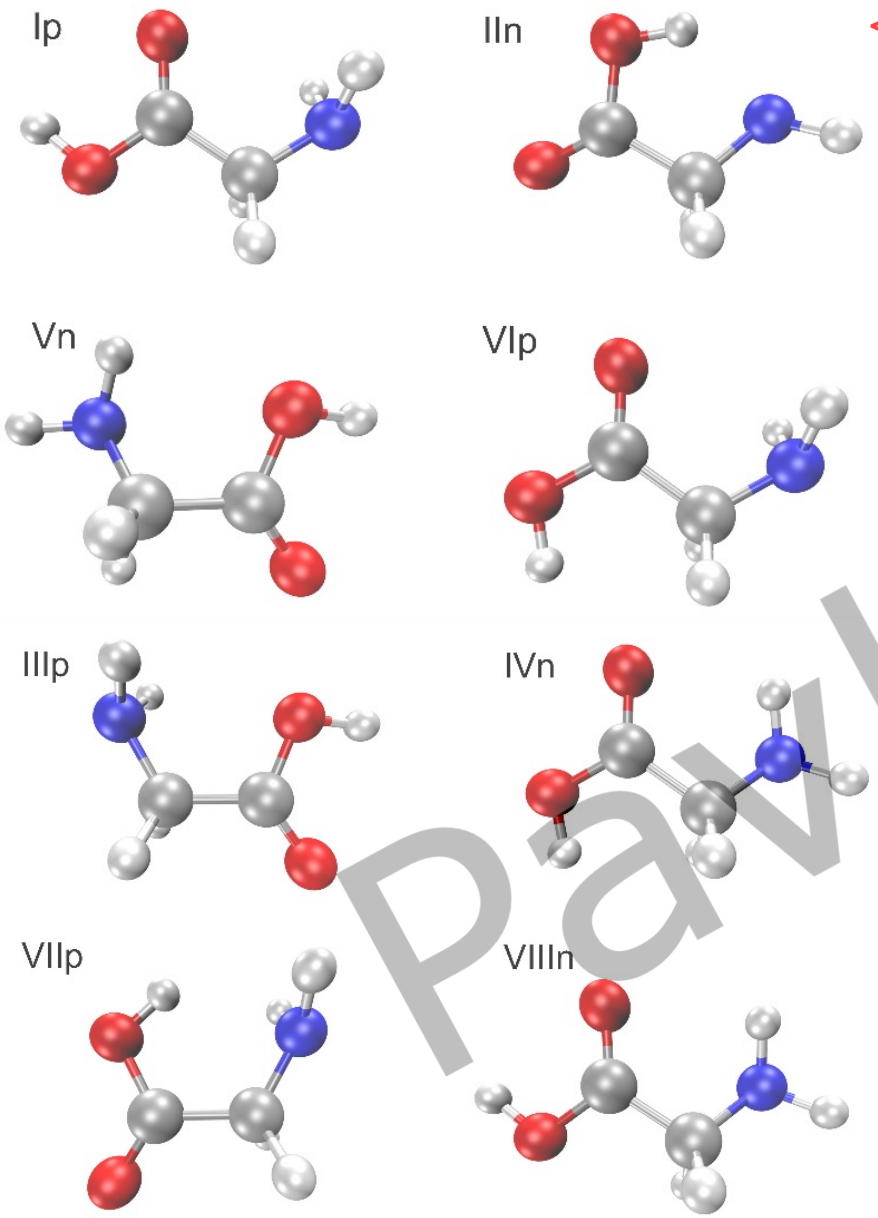


Pavlo O. Dral

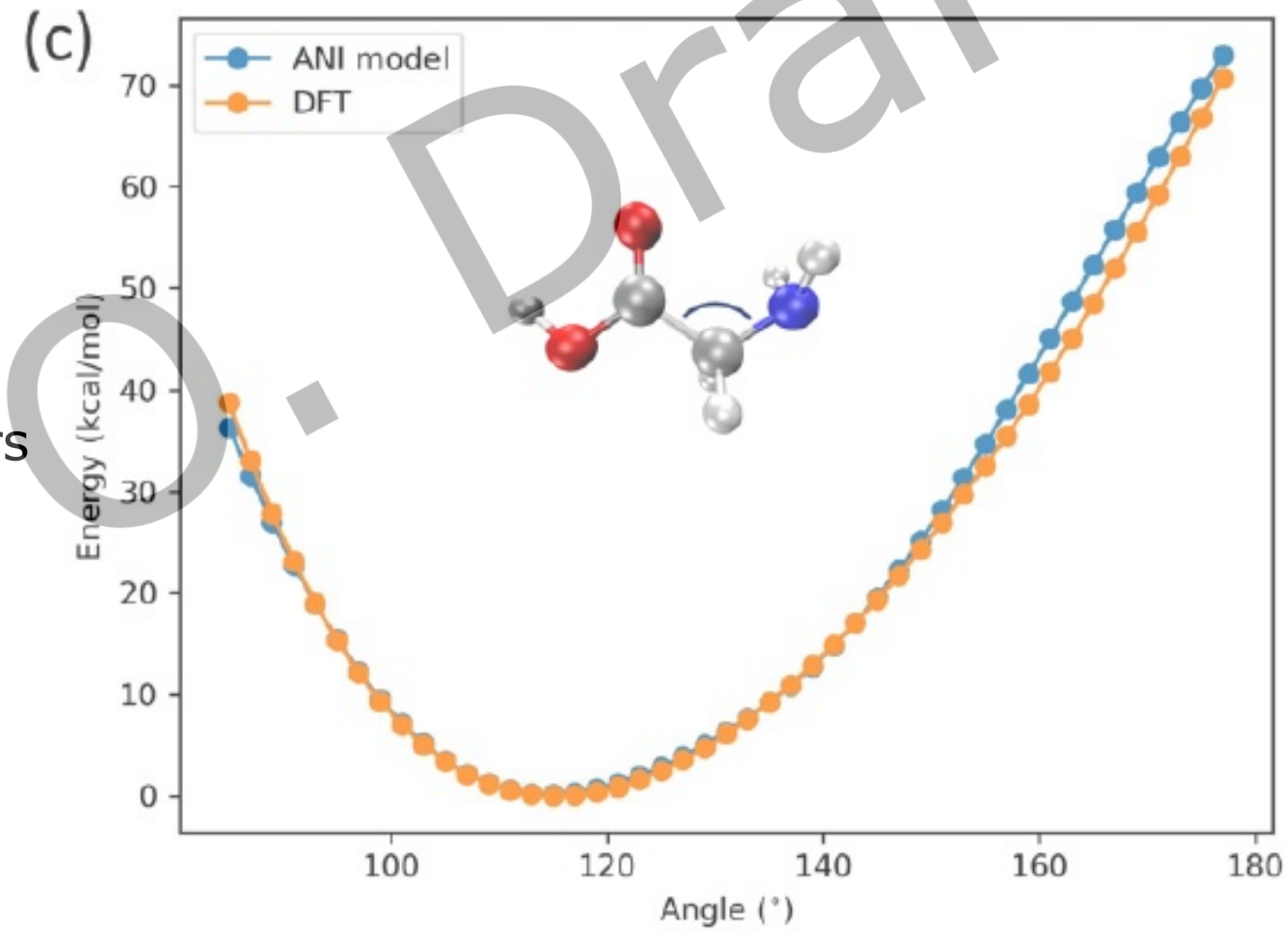
Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral. *arXiv:2404.11811*.

Conformer search through quasi-classical MD

< 2000 training points with ANI potential



Glycine
8
conformers

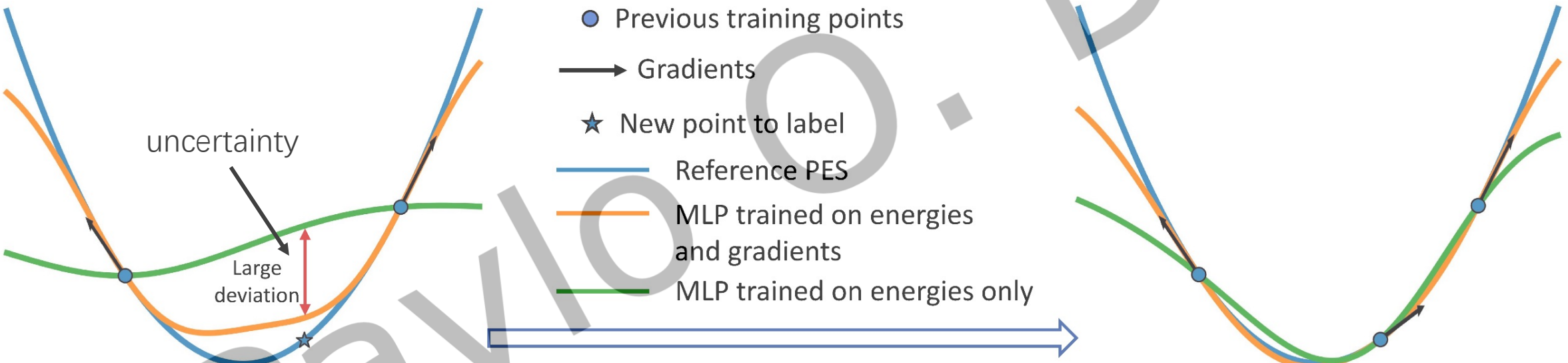


Y.-F. Hou, L. Zhang, Q. Zhang,
F. Ge, P. O. Dral, *arXiv:2404.11811.*

Pavlo O. Dral

Physics-informed active learning

Utilizing different amount of physics-derived information for uncertainty quantification

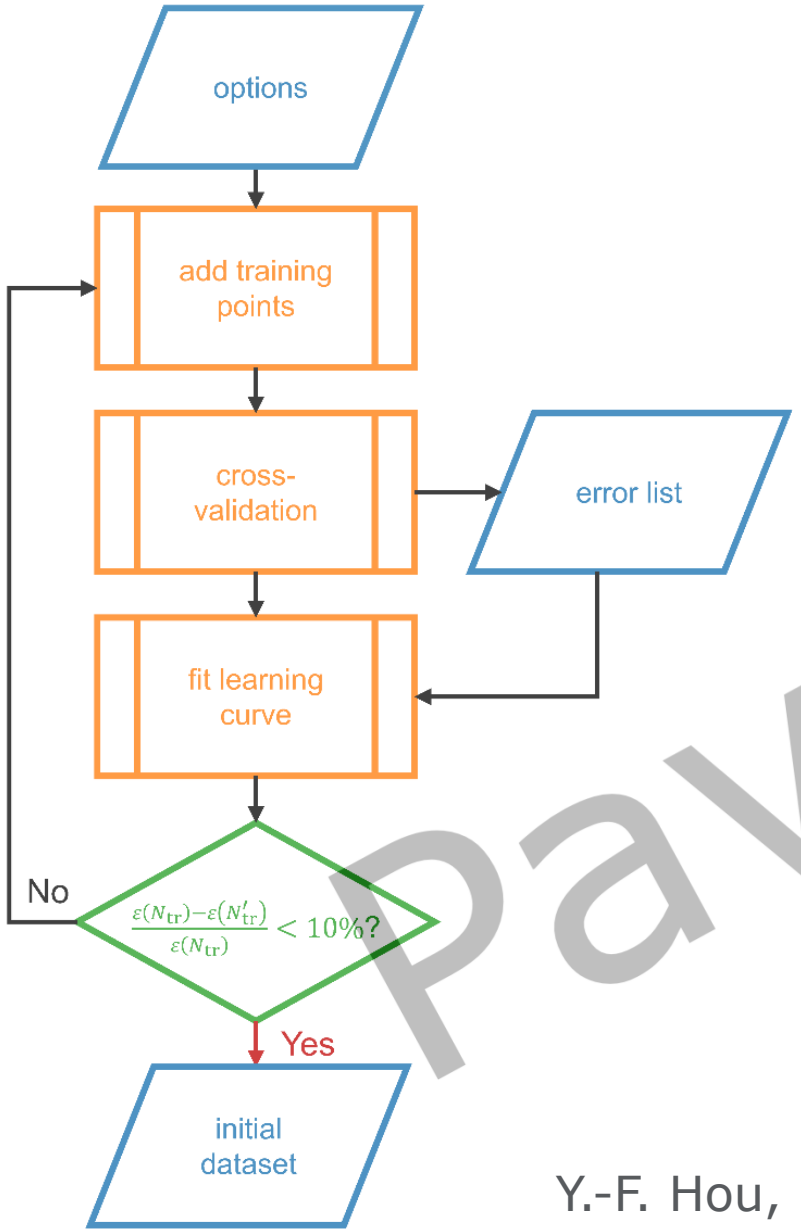


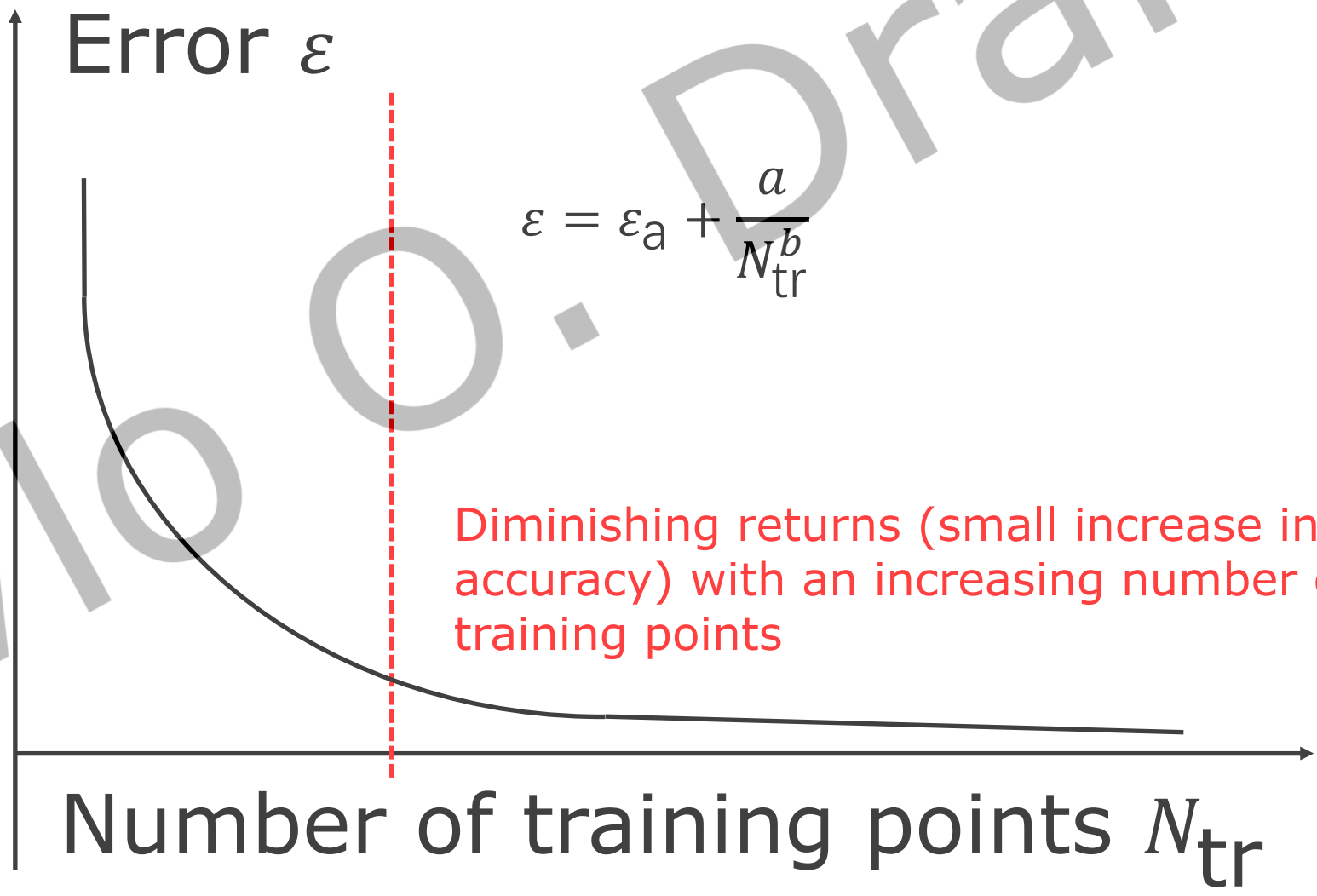
main model (single ML model, used for MD)
auxiliary model (not used for MD)

Pavlo O. Dral

Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral. *arXiv:2404.11811*.

Automatic initial data building

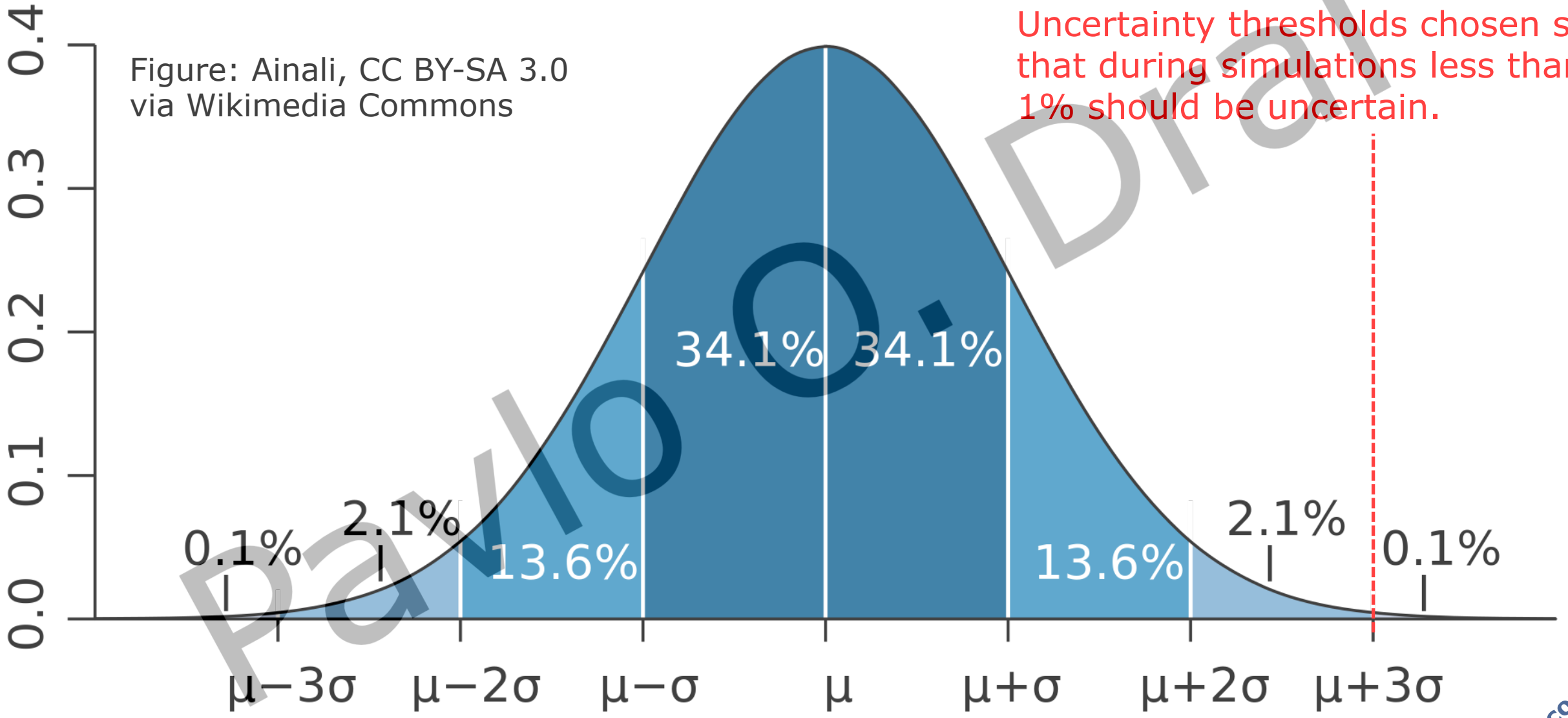




Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral. *arXiv:2404.11811*.

Pavlo O. Dral

Automatic uncertainty quantification



Pavlo O. Dral

Y.-F. Hou, L. Zhang, Q. Zhang, F. Ge, P. O. Dral. *arXiv:2404.11811*.

Note

MLatom enables this by providing end-to-end implementation of all the required steps making it automatic. Active learning is currently only possible to do with MLatom Python API. You can find an example in Jupyter notebook `al_h2.ipynb` (open in Jupyter Lab) which I placed in the folder `compchem_course/task6.5`.

Pavlo O. Dral

Homework 6.1.

Assuming you want to keep the same sampling density as for H_2 (where you needed 20 points for one degree of freedom), calculate the number of training points you would need for:

1. H_2O.
2. Methane molecule.
3. C_{60}.

Hint: for two dimensions, you have a rectangle, if you want to get the same density, you would need 20 times 20 points...

Pavlo O. Dral

Homework 6.2.

Train the KREG model on the differences between the FCI and HF/STO-3G energies **interatomic distances larger or equal 0.8 Angstrom** of H_2. These differences are residual error of HF/STO-3G wrt FCI.

Estimate the residual errors with this KREG delta-learning model for all internuclear distances from 0.5 to 5 Angstrom. Add them back to HF/STO-3G energies to obtain an estimate of FCI energies with such a delta-learning model.

See the previous tasks how to train and predict.

Plot the potential energy curve:

1. How does this delta-learning model perform?
2. Compared to the KREG model you trained previously directly on the FCI data for the same training set (e.g., in the previous task)?

Pavlo O. Dral