FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Machine Learning for Crystal Structure Prediction

Coffee Talk, 23rd April 2018

Marcel F. Langer

Crystal Structure Prediction

One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition.

– John Maddox

Maddox, *Nature* **335**, 201 (1988)

Crystal Structure Prediction

One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition. [...] ***One would have thought that, by now, it should be possible to equip a sufficiently large computer with a sufficiently large program, type in the formula of the chemical and obtain, as output, the atomic coordinates of the atoms in that unit cell.***

– John Maddox

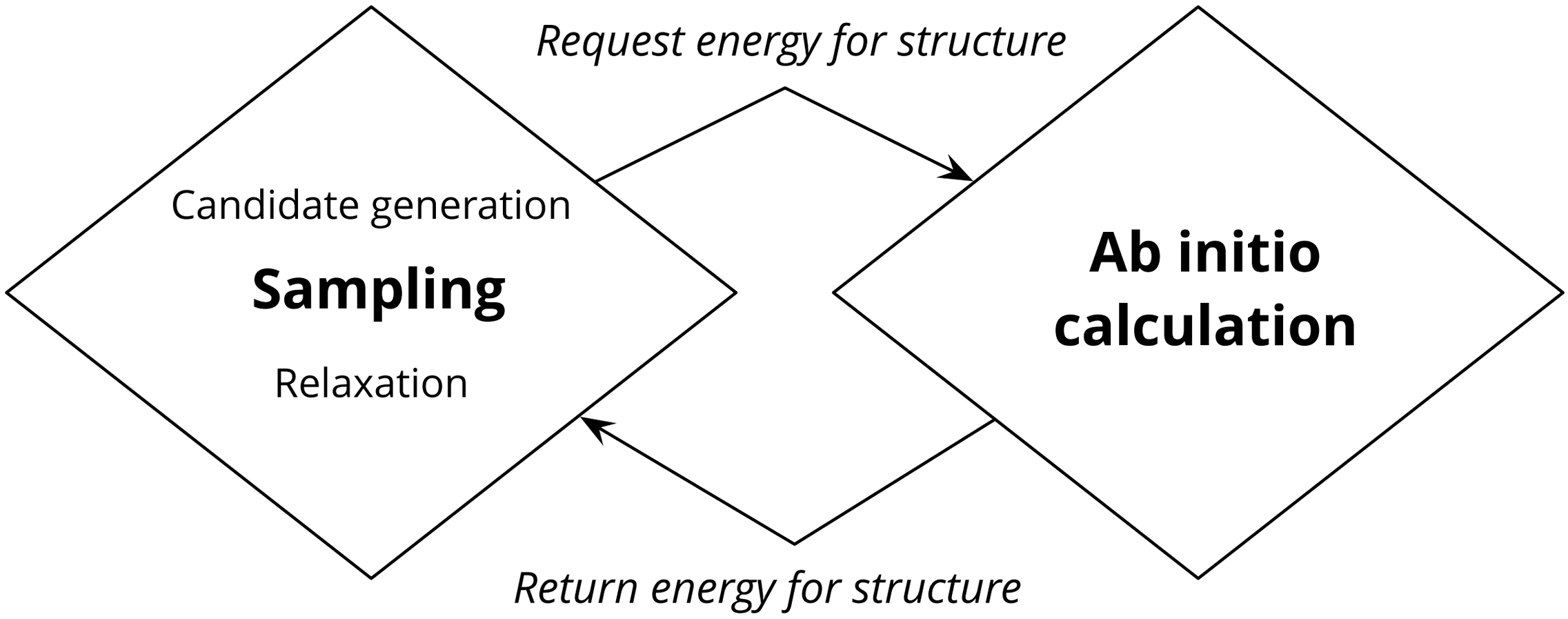
Maddox, *Nature* **335**, 201 (1988)

Crystal Structure Prediction

- Stability in general: Minimum of the free energy, depends on thermodynamic conditions

- Common simplification: Minimum of *formation* energy

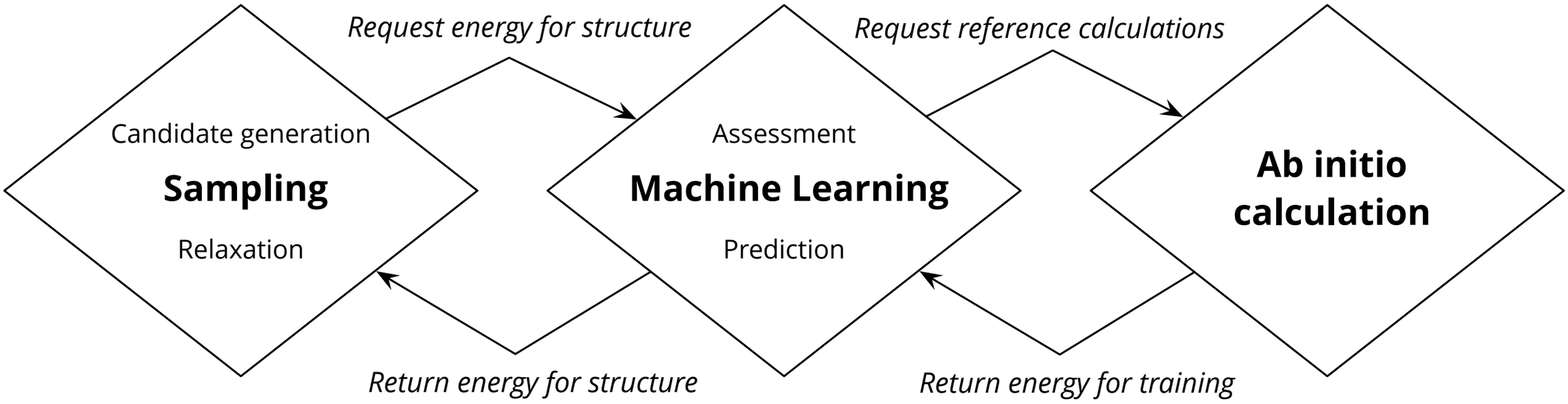
Crystal Structure Prediction



Overview:

Modern Methods of Crystal Structure Prediction (ed: Oganov) (Wiley, 2011)

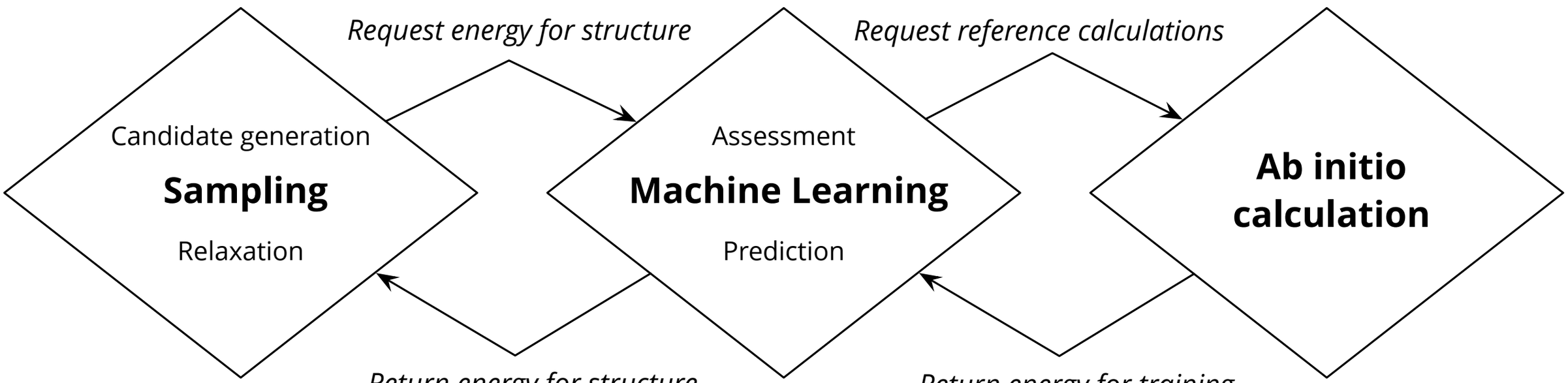
Approach



Application:

Search for transparent conducting oxides

Requirements

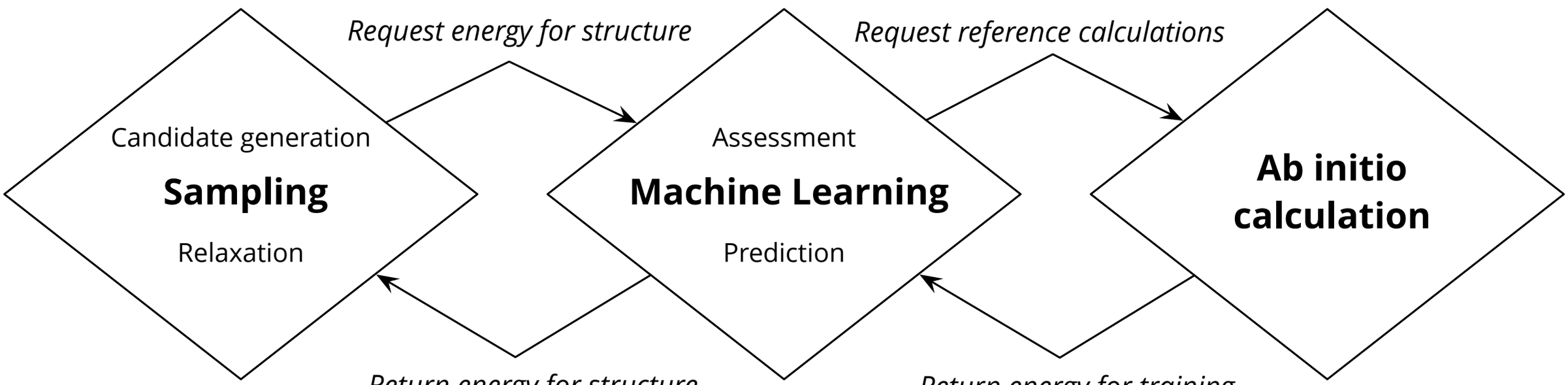


- Simple
- Parallel
- Batched

- Accurate
- Sample efficient
- Quantifiable uncertainty

- Automated
- Consistent

Choices



- Simulated annealing with Monte-Carlo moves
- Nested sampling

- **MBTR +**
- **KRR? +**
- **…?**

- DFT (FHI-aims (PBEsol[1]) + high-aims)

[1]Perdew et al., *Physical Review Letters* **100**, 136406 (2008)

[2]Huo & Rupp, *arXiv* 1704.06439 (2017)

Currently…

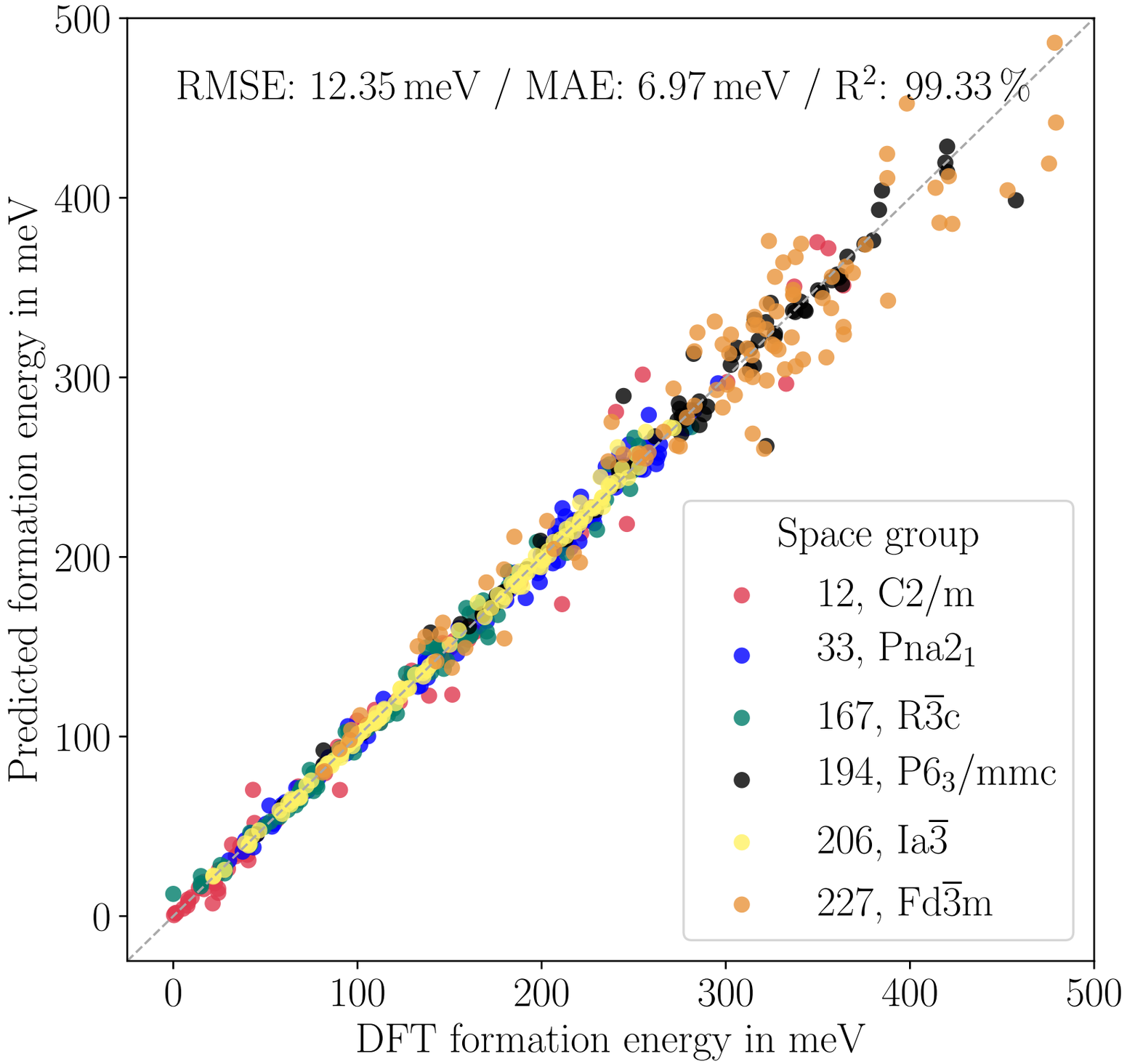
- Identify the ML model to use
- … by comparing on a realistic dataset
- … using automated infrastructure

Dataset

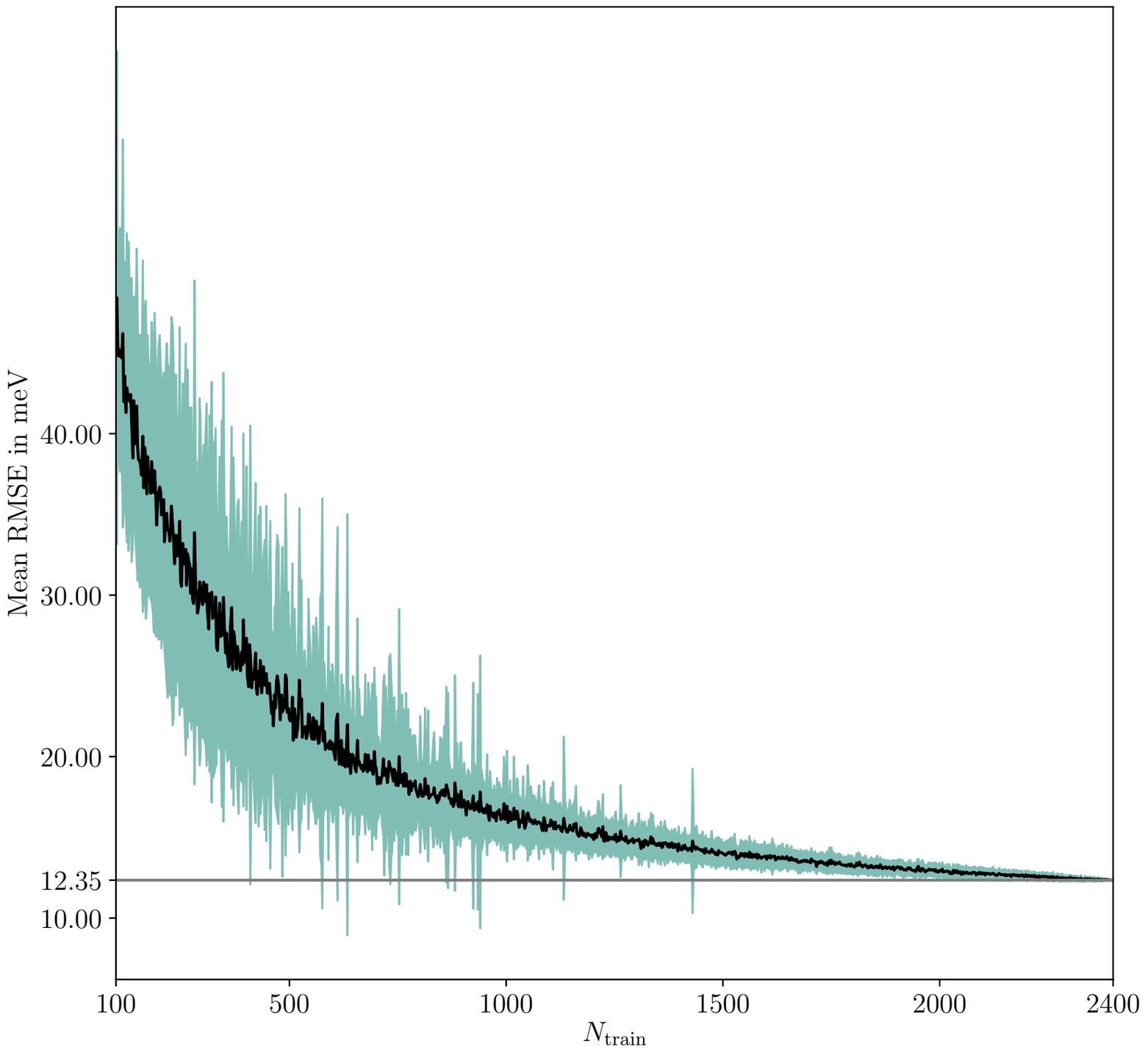
- Kaggle competition dataset:
  - 2400+600 TCOs with formula $(Al_xGa_yIn_{1-x-y})_{2N}O_{3N}$
    - 6 space groups, 10 to 80 atoms in unit cell
    - Idealised geometries with Vegard's rule[1]
- Now: *Relaxed* geometries

[1]Vegard, *Zeitschrift für Physik* **5**, 17–26 (1921)

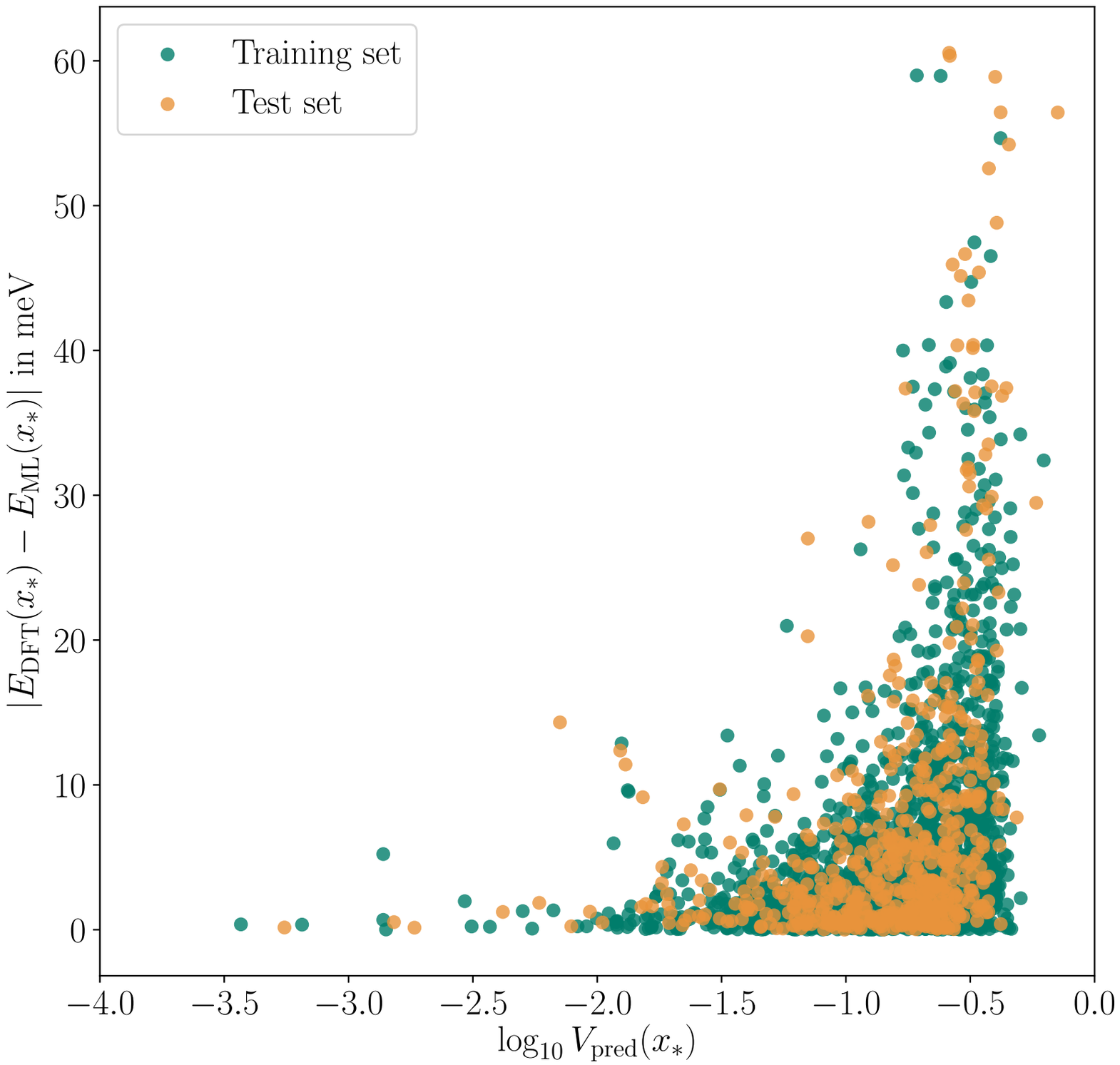
Predictions



Learning curve



Predictive Variance



Next steps

- Uncertainty quantification for KRR
- (Deep) neural networks
- Extended dataset
- Nested sampling with MBTR+KRR (but no uncertainty)

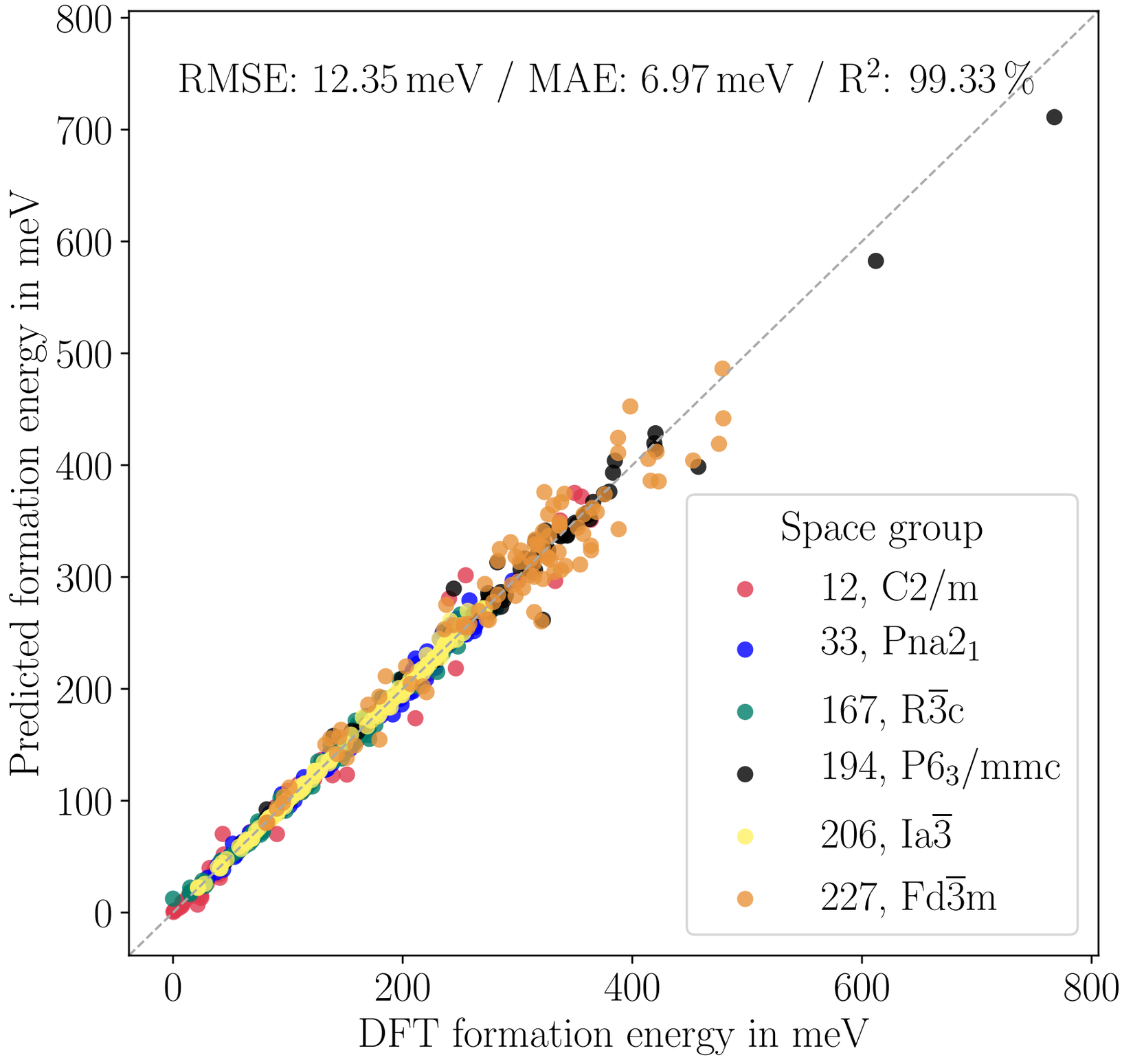
Thank you

- Matthias Rupp
- Christopher Sutton
- Xiangyue Liu

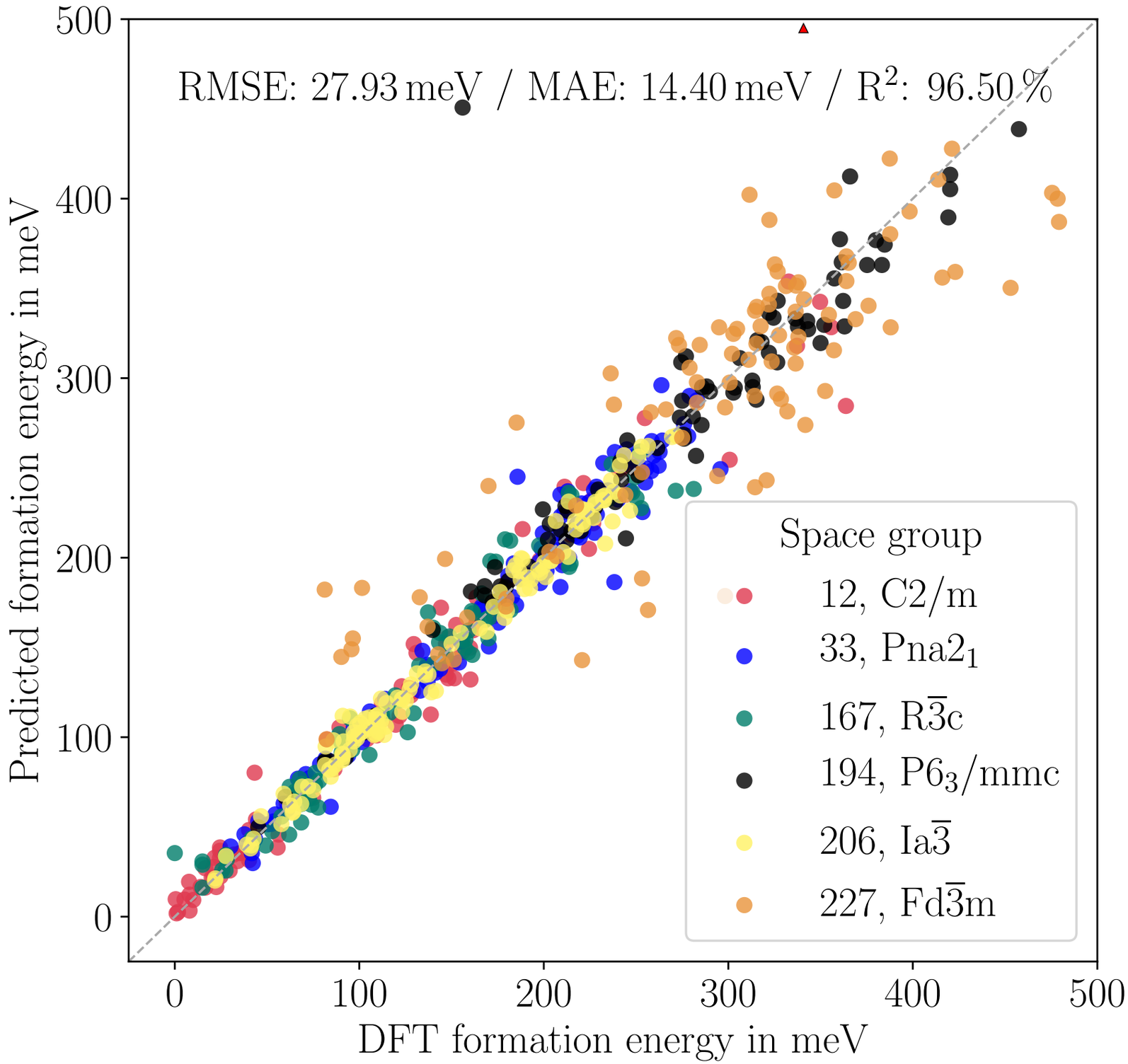
Questions!

Bonus Content!

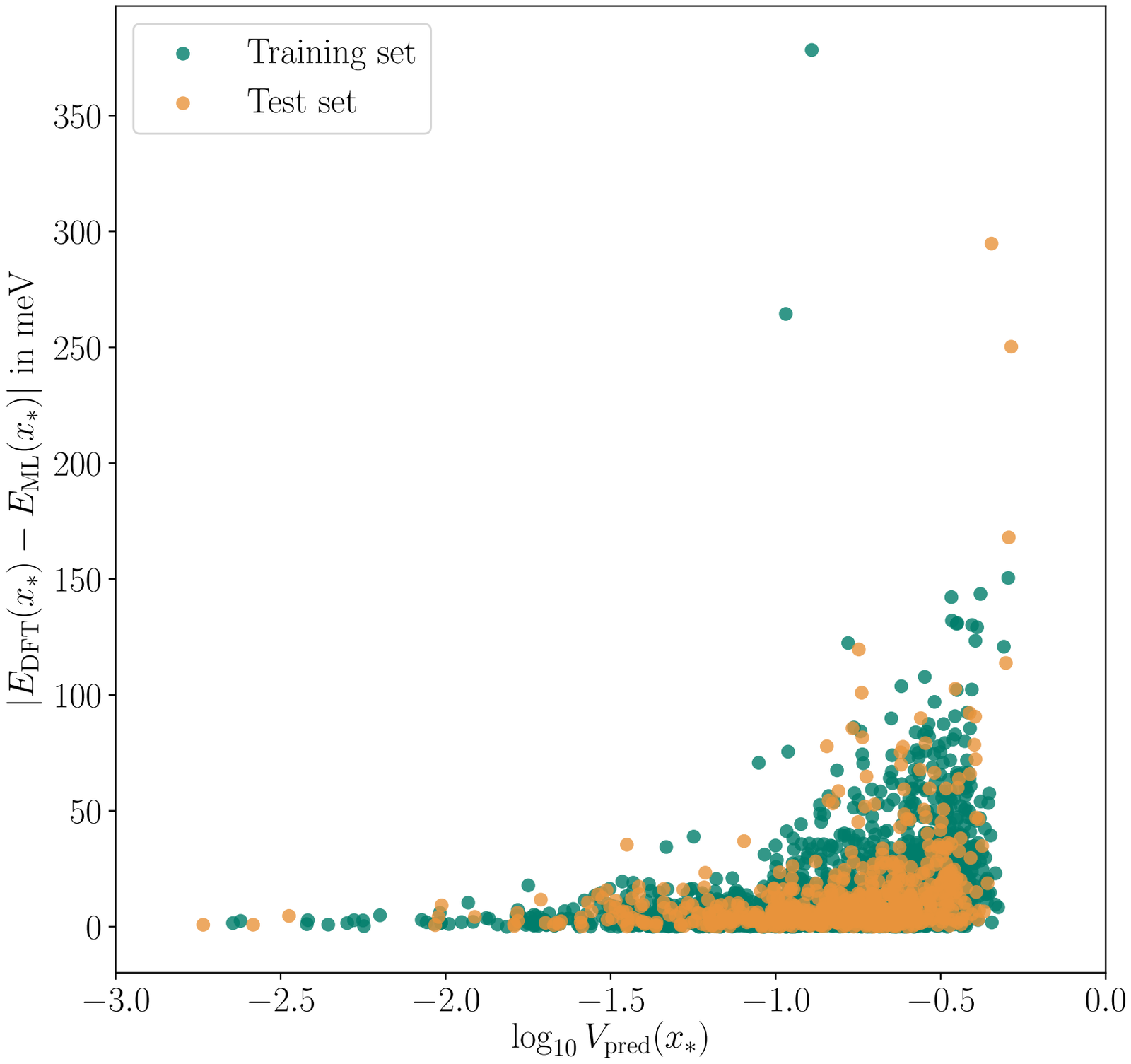
Predictions (Outliers)



Predictions (Kaggle)



Predictive variance (Kaggle)



```
##### k12_wei_1-best0 #####
Standard loss:
   Trained model on kgrlx (property fecreal) and predicted kgrlx_test (property fe).
=> RMSE=0.01235  MAE=0.00697 R2=99.32643
Log loss:
   Trained model on kgrlx (property fecreal) and predicted kgrlx_test (property fe).
=> RMSE=0.00946  MAE=0.00557 R2=99.42410
Standard loss:
   Trained model on kgrlx (property fecreal) and predicted kgrlx_test (property fepa).
=> RMSE=0.00494  MAE=0.00279 R2=99.32643
Training set loss:
   Trained model on kgrlx (property fecreal) and predicted kgrlx (property fe).
=> RMSE=0.00793  MAE=0.00440 R2=99.70990
Training set loss (same property):
   Trained model on kgrlx (property fecreal) and predicted kgrlx (property fecreal).
=> RMSE=0.13102  MAE=0.08175 R2=99.90032
No fd3m Standard loss:
   Trained model on kgrlx-nofd3m (property fecreal) and predicted kgrlx_test-nofd3m (property fe).
=> RMSE=0.00922  MAE=0.00410 R2=99.55190
No fd3m Log loss:
   Trained model on kgrlx-nofd3m (property fecreal) and predicted kgrlx_test-nofd3m (property fe).
=> RMSE=0.00730  MAE=0.00336 R2=99.59717
No fd3m Standard loss:
   Trained model on kgrlx-nofd3m (property fecreal) and predicted kgrlx_test-nofd3m (property fepa).
=> RMSE=0.00369  MAE=0.00164 R2=99.55190
No fd3m Training set loss:
   Trained model on kgrlx-nofd3m (property fecreal) and predicted kgrlx-nofd3m (property fe).
=> RMSE=0.00341  MAE=0.00208 R2=99.93288
No fd3m Training set loss (same property):
   Trained model on kgrlx-nofd3m (property fecreal) and predicted kgrlx-nofd3m (property fecreal).
=> RMSE=0.06459  MAE=0.04583 R2=99.97876
```

```
###### k12_wei_1-best0 ######
Model during autotune run

### DATA ###
SON([('property', 'fecreal')])

### KRR ###
{'centering': False,
 'kernelf': ['gaussian', 107.63474115247546],
 'nl': 2.6973983046972182e-06}

### MBTR ###
{'mbtr_1': {'acc': 0.001,
            'aindexf': 'full',
            'corrf': 'identity',
            'd': [-0.5, 1.6333333333333333, 30],
            'distrf': ['normal', 4.0],
            'eindexf': 'full',
            'elems': None,
            'flatten': True,
            'geomf': 'count',
            'k': 1,
            'norm': 13.138686868686868,
            'weightf': ['exp_-1/identity', 171.71858585858587]},
 'mbtr_2': {'acc': 0.001,
            'aindexf': 'noreversals',
            'corrf': 'identity',
            'd': [-0.04, 0.008, 100],
            'distrf': ['normal', 0.015625],
            'eindexf': 'noreversals',
            'elems': None,
            'flatten': True,
            'geomf': '1/distance',
            'k': 2,
            'norm': None,
            'weightf': ['exp_-1/identity^2', 10.10150505050505]}}
```

```
##### k12_run_1-best0 #####
Standard loss:
   Trained model on kaggle (property fecreal) and predicted kaggle_test (property fe).
=> RMSE=0.02793  MAE=0.01440 R2=96.50097
Training set:
   Trained model on kaggle (property fecreal) and predicted kaggle (property fe).
=> RMSE=0.02365  MAE=0.01315 R2=97.38817
Log loss:
   Trained model on kaggle (property fecreal) and predicted kaggle_test (property fe).
=> RMSE=0.02125  MAE=0.01159 R2=97.06138

##### k123_mr #####
Standard loss:
   Trained model on kaggle (property feclog) and predicted kaggle_test (property fe).
=> RMSE=0.03159  MAE=0.01486 R2=95.85761
Training set:
   Trained model on kaggle (property feclog) and predicted kaggle (property fe).
=> RMSE=0.01845  MAE=0.00980 R2=98.41860
Log loss:
   Trained model on kaggle (property feclog) and predicted kaggle_test (property fe).
=> RMSE=0.02312  MAE=0.01176 R2=96.67117
```

```
###### k12_run_1-best0 ######
Model during autotune run

### DATA ###
SON([('property', 'fecreal')])

### KRR ###
{'centering': False,
 'kernelf': ['gaussian', 1505.0215477863412],
 'nl': 1.7973733224146392e-07}

### MBTR ###
{'mbtr_1': {'acc': 0.001,
            'aindexf': 'full',
            'corrf': 'identity',
            'd': [-0.5, 1.6333333333333333, 30],
            'distrf': ['normal', 1448.1546878700494],
            'eindexf': 'full',
            'elems': None,
            'flatten': True,
            'geomf': 'count',
            'k': 1,
            'norm': 23.237676767676767,
            'weightf': 'identity'},
 'mbtr_2': {'acc': 0.001,
            'aindexf': 'noreversals',
            'corrf': 'identity',
            'd': [-0.04, 0.008, 100],
            'distrf': ['normal', 0.0078125],
            'eindexf': 'noreversals',
            'elems': None,
            'flatten': True,
            'geomf': '1/distance',
            'k': 2,
            'norm': None,
            'weightf': 'identity^2'}}
```