

## ML models for computed band gap of HOIPs

Huan Tran, Georgia Institute of Technology  
This notebook is part of [V.N. Tuoc, Nga, T.T. Nguyen, V. Sharma, and T.D. Huan, *Probabilistic deep learning approach for targeted hybrid organic-inorganic perovskites*, *Physical Review Materials* 5, 125402 (2021); <https://doi.org/10.1103/PhysRevMaterials.5.125402>], and is also an example of matSLM toolkit. Results obtained here can be found in this work.

The original (raw) dataset containing the computed band gap of 1,346 atomic structures predicted for 192 chemical compositions of hybrid organic-inorganic perovskites (HOIPs) is available at [C. Kim, T.D. Huan, S. Krishnan, and R. Ramprasad, *Scientific Data* 8, e01757 (2021); <https://www.nature.com/articles/data201757>]. Here, three fingerprinted versions of S1, S2, and S3, which are based on Gaussian Process Regression, fully connected Neural Net, and Probability Neural Net. Computations performed using matSLM toolkit, available at <https://github.com/huantm/matSLM.git>.

Among 5 models developed, M5 demonstrates a reasonable way to handle the aleatoric uncertainty in deep learning of materials data. More details on this topic can be found in "Probabilistic deep learning approach for targeted hybrid organic-inorganic perovskites", the reference mentioned above.

### 1. Download data

Three (fingerprinted) datasets (S1, S2, and S3) used for the work will be obtained. In fact, S2 has 2 versions, one with selector and one not.

```
In [1]: from matSLM.data import Datasets
data = Datasets()
S1 = fp_hoops.S1_1dest
S2a = fp_hoops.S2a_2dest
S2b = fp_hoops.S2b_1dest
S3 = fp_hoops.S3_4tfp
```

data.load\_dataset()  
matSLM version 3.0

Load (fingerprinted) datasets  
Data saved in fp\_hoops.S1\_1dest.csv.gz  
Data saved in fp\_hoops.S2a\_2dest.csv.gz  
Data saved in fp\_hoops.S2b\_1dest.csv.gz  
Data saved in fp\_hoops.S3\_4tfp.csv.gz

### 2. Obtained datasets parameters

```
In [2]: # data parameters for learning
n_train = 0.9 # % for training, 10% for testing
x_scaling = 'minmax' # method for x_train/test
y_scaling = 'minmax' # method for y scaling
```

# Dict of data parameters
data1\_params = {
 'data\_file': 'fp\_hoops.S1\_1dest.csv.gz',
 'nfold\_cv': 5,
 'id\_col': 'ID',
 'y\_cols': ['Ymean'],
 'comment\_cols': [],
 'x\_scaling': 'yScaling',
 'y\_scaling': 'yScaling',
 'x\_sampling': 'sampling',
 'n\_trains': n\_trains,
}

data2b\_params = {
 'data\_file': 'fp\_hoops.S2a\_2dest.csv.gz',
 'nfold\_cv': 5,
 'id\_col': 'ID',
 'y\_cols': ['Ymean', 'Ystd'],
 'comment\_cols': [],
 'x\_scaling': 'yScaling',
 'y\_scaling': 'yScaling',
 'x\_sampling': 'sampling',
 'n\_trains': n\_trains,
}

data3\_params = {
 'data\_file': 'fp\_hoops.S3\_4tfp.csv.gz',
 'nfold\_cv': 5,
 'id\_col': 'ID',
 'y\_cols': ['prop\_value'],
 'comment\_cols': ['hid'],
 'x\_scaling': 'yScaling',
 'y\_scaling': 'yScaling',
 'x\_sampling': 'sampling',
 'n\_trains': n\_trains,
}

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.052365 0.06944 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 3.055918 0.051227 0.053613
cv, rmse\_train, rmse\_test, rmse\_opt = 4.054392 0.053427

Training model w/ cross validation
cv, rmse\_train, rmse\_test, rmse\_opt = 0.0524103 0.084918 0.084918
cv, rmse\_train, rmse\_test, rmse\_opt = 1.052310 0.056313 0.056313
cv, rmse\_train, rmse\_test, rmse\_opt = 2.052310 0.056313 0.0563