
DSTAR
Manual

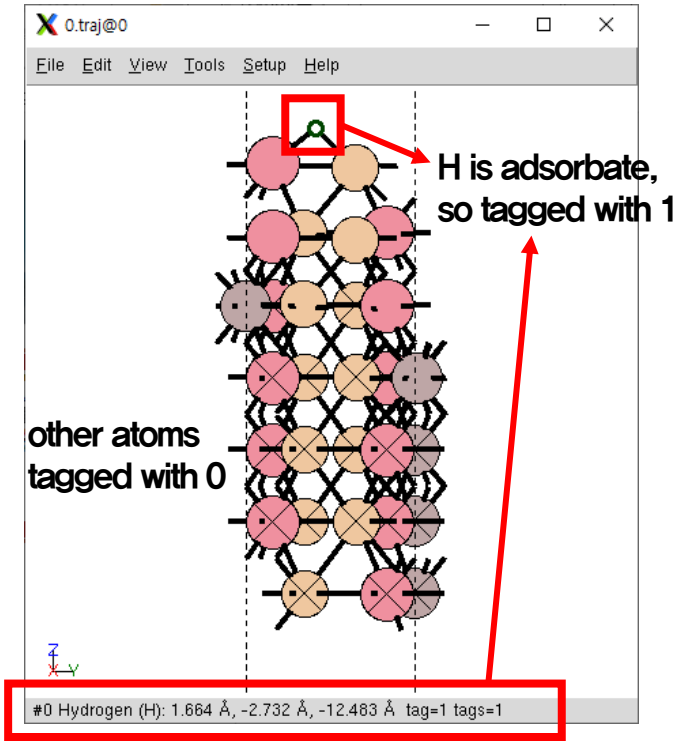
Index

1. Making Input Trajectory

1) Save surfaces with tagged adsorbates as trajectory file format to ~/DSTAR/atoms/ for training

```
/home/ahrehd0506/git/DSTAR/atoms
(base) [ahrehd0506@haber atoms]$ ls
0.traj   11.traj  13.traj  15.traj  17.traj  19.traj  20.traj  2.traj  4.traj  6.traj  8.traj
10.traj  12.traj  14.traj  16.traj  18.traj  1.traj   21.traj  3.traj  5.traj  7.traj  9.traj
```

※ Adsorption atoms should be tagged with non-zero integer, and the other surface atoms are tagged with zero.



2. Making Target Dataframe

1) Generate Dataframe with name and target columns.
It doesn't matter if there are other columns

2) name column indicates name (id) of input surface trajectory file and target column indicates corresponding target value (In our study, binding energy)

surface named 12.traj with target value 4

	name	target
0	1	1
1	2	2
2	3	3
3	4	4
4	5	5
5	6	6
6	7	7
7	8	8
8	9	9
9	0	1
10	10	2
11	11	3
12	12	4
13	13	5
14	14	6
15	15	7
16	16	8
17	17	9
18	18	1
19	19	2
20	20	3
21	21	4

3) Save filled Dataframe named target.csv to ~/DSTAR/atoms/

```
(base) [ahrehd0506@haber atoms]$ pwd
/home/ahrehd0506/git/DSTAR/atoms
(base) [ahrehd0506@haber atoms]$ ls
0.traj   11.traj  13.traj  15.traj  17.traj  19.traj  20.traj  2.traj  4.traj  6.traj  8.traj  target.csv
10.traj  12.traj  14.traj  16.traj  18.traj  1.traj   21.traj  3.traj  5.traj  7.traj  9.traj
```

3. Generate Fingerprint and Train ML Model

1) Run `python train.py` in ~/DSTAR/

2) Training is successful if displayed as below image.

```
Namespace(algo='total', atom_path='./atoms/', convert_only=False, data_path='./data/fp.csv', desire_range=0.1, desire_target=-0.67, load_data=False, model_path='./model/', subs_path='./subs/fp/', test=False, test_ratio=0.2)

Initiate Conversion Atoms to Fingerprint...
100%|| 23/23 [00:03<00:00,  6.27it/s]
Successfully Generate Fingerprint!

2022-07-13 11:33:29,004 - INFO - Start Gradient Boosting Regression
2022-07-13 11:33:30,608 - INFO - Start Kernel Ridge Regression
2022-07-13 11:33:30,836 - INFO - Start ElasticNet Regression
/home/ahrehd0506/miniconda3/lib/python3.8/site-packages/sklearn/linear_model/_coordinate_descent.py:645: ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations, check the scale of the features or consider increasing regularisation. Duality gap: 3.875e+00, tolerance: 4.991e-02
  model = cd_fast.enet_coordinate_descent(
2022-07-13 11:33:30,924 - INFO - Start Support Vector Regression
2022-07-13 11:33:30,962 - INFO - Start Gaussian Process Regression
/home/ahrehd0506/miniconda3/lib/python3.8/site-packages/sklearn/gaussian_process/kernels.py:430: ConvergenceWarning: The optimal value found for dimension 0 of parameter k1__constant_value is close to the specified upper bound 100000.0. Increasing the bound and calling fit again may find a better value.
  warnings.warn(
/home/ahrehd0506/miniconda3/lib/python3.8/site-packages/sklearn/gaussian_process/kernels.py:420: ConvergenceWarning: The optimal value found for dimension 0 of parameter k2__length_scale is close to the specified lower bound 1e-05. Decreasing the bound and calling fit again may find a better value.
  warnings.warn(
==================================================

Best performance model : Kernel Ridge Regressor
MAE : 5.066529515330492
RMSE : 6.360076086100101

==================================================
```

3) Train_results.csv and Test_results.csv will be generated in ~/DSTAR/.
They have name, target, pred column, indicating surface file id, true target value and predicted target value, respectively

4) Fingerprint file named fp.csv will be generated in ~/DSTAR/data/. It contains surface informations converted into active motif–based fingerprint.

	name	FNN	Same	Sub	target
2	0	{'Co': 1, 'Si': 1}	{'Co': 5, 'Si': 3}	{'Al': 2}	1

surface named 0.traj converted into the number of elements in FNN / SNN_same / SNN_sub site

※ Can change file name and path by running with --data-path {your_data_path} argument (Default : ./data/fp.csv)

```
python main.py --data-path /data/test.csv
```

5) ML model and Scaler used in training will be saved as ~/DSTAR/model/{current–date}/model.pkl & scaler.pkl

```
(base) [ahrehd0506@haber model]$ ls
2022-06-01/  2022-06-02/  2022-06-03/  2022-07-13/
(base) [ahrehd0506@haber model]$ cd 2022-07-13/
(base) [ahrehd0506@haber 2022-07-13]$ ls
model.pkl  scaler.pkl
```

3 – 1. Only Generation or Training

1) Only generate fingerprint by running with --convert-only argument (Default : False)

```
(base) [ahrehd0506@haber DSTAR]$ python train.py --convert-only
Namespace(algo='total', atom_path='./atoms/', convert_only=True,
th='./subs/fp/', test=False, test_ratio=0.2)

Initiate Conversion Atoms to Fingerprint...
100%|
Successfully Generate Fingerprint!
```

2) Only train model from existed fingerprint by running with --load-data --data-path {your_data_path} argument (data–path Default : ./data/fp.csv)

```
(base) [ahrehd0506@haber DSTAR]$ python train.py --load-data --data-path ./data/test.csv
Namespace(algo='total', atom_path='./atoms/', convert_only=False, data_path='./data/test.
path='./subs/fp/', test=False, test_ratio=0.2)

Load Motif From test.csv...
Initiate Conversion Motifs to Fingerprint...
22it [00:00, 148.06it/s]
Successfully Generate Fingerprint!

2022-07-13 13:44:51,304 - INFO - Start Gradient Boosting Regression
2022-07-13 13:44:53,923 - INFO - Start Kernel Ridge Regression
2022-07-13 13:44:53,943 - INFO - Start ElasticNet Regression
```

3 – 2. Other Arguments

1) Change input trajectory files path for training by running with
`--atom-path {your_atom_path}` argument (Default : ./atoms/)

2) Adjust the train / test ratio by running with
`--test-ratio {test_ratio}` argument (Default : 0.2)

3) Change ML algorithm by running with
`--algo {algorithm}` argument (Default : Use all and choose the best)

Available Algorithms : GBR / KRR / ELN / SVR / GRP

4. Fingerprint Substitution

1) Substitute elements in the fingerprint file generated from previous process

2) Run `python subs.py` in ~/DSTAR/. Successful if displayed as below image

```
(base) [ahrehd0506@haber DSTAR]$ python subs.py
Namespace(bi_only=False, data_path='./data/fp.csv', get_bi=False, subs_path='./subs/fp/', subs_type='comb')

Initiate Active Motif Subsititution
100%|
Successfully Generate New Active Motifs!!
```

※ Can change fingerprint file for substitution by running with `--data-path {your_data_path}` argument (Default : ./data/fp.csv)

3) Substituted fingerprints files will be saved in ~/DSTAR/subs/{your_subs_path}/ (Default : ~/DSTAR/subs/fp/)

```
(base) [ahrehd0506@haber fp]$ ls
Ag_Ag.csv  Al_Al.csv  As_Au.csv  Au_Cu.csv  Co_In.csv  Cr_Os.csv  Cu_Ru.csv  Fe_W.csv   Ge_Os.csv
Ag_Al.csv  Al_As.csv  As_Co.csv  Au_Fe.csv  Co_Ir.csv  Cr_Pb.csv  Cu_Sb.csv  Fe_Zn.csv  Ge_Pb.csv
```

4 – 1. Elements & Substitution Type Selection

1) Can change elements for substitution by modifying el_set_A / el_set_B list in Line 31~34 from ~/DSTAR/sub.py

```
def subs(args):
    el_set_A =['Ag','Al','As','Au','Co','Cr','Cu','Fe','Ga','Ge','In','Ir','Mn','Mo',
               'Ni','Os','Pb','Pd','Pt','Re','Rh','Ru','Sb','Se','Si','Sn','Ti','V',
               'W','Zn']
    el_set_B = []
```

default elements in el_set_A (30 Transition Metals)
role of el_set_B is described below

2) Select substitution type by running with **--subs-type {subs_type}** argument (Default : comb)

- comb : Combination N elements in el_set_A (${}_N\Pi_2$)
- prod : Product between el_set_A and el_set_B
- ex) If you want substitute with combination between Metal + Chalcogen, modify like below image and use prod

```
el_set_A =['Ag','Al','As','Au','Co','Cr','Cu','Fe','Ga','Ge','In','Ir','Mn','Mo',
           'Ni','Os','Pb','Pd','Pt','Re','Rh','Ru','Sb','Se','Si','Sn','Ti','V',
           'W','Zn']
el_set_B = ['O','S','Se','Te']
```

4 – 2. Other Arguments

1) Change substituted fingerprint save path by running with
`--subs-path {your_subs_path}` argument (Default : ./subs/fp/)

2) Substitute without unary materials ($_NC_2$) by running with
`--bi-only` argument (Default : False)

3) Extract original fingerprint only with binary materials by running with
`--get-bi` argument (Default : False)

(Automatically substitute only binary without this arg.)

Extracted fingerprint only with binary materials saved as
{your_data_path}/{data_name}_binary.csv (Default : ./data/fp_binary.csv)

5. Target Prediction from Substituted Fingerprint

1) Run `python train.py --test --model-path {your_model_path}` in ~/DSTAR/.
Successful if displayed as below image.

```
(base) [ahrehd0506@haber DSTAR]$ python train.py --test --model-path ./model/2022-07-13/
Namespace(algo='total', atom_path='./atoms/', convert_only=False, data_path='./data/fp.cs
3/', subs_path='./subs/fp/', test=True, test_ratio=0.2)

Load Model 2022-07-13
Initiate Screening...
100%|
Successfully Predict Ideal Surface Density!!
```

※ model path argument must be included

※ Can change substituted fingerprint files path by running with `--subs-path {your_subs_path}` argument (Default : ./subs/fp/)

2) pred column (predicted target value will be added in fingerprint files in {your_subs_path}

	name	FNN	Same	Sub	target	pred
0	5	{'Ag': 1}	{'Ag': 2}	{'Ag': 1}	0	4
1	5	{'Ag': 1}	{'Ag': 2}	{'Ag': 1}	0	4

There are two same ids because both A_xB_y / B_xA_y are substituted from one material

4) dens.csv file will be generated in ~/DSTAR/, which contains elements and density column.
desnity indicates the ratio of surfaces with predicted target value within desire target range. –〉 higher, the better activity.

※ Set desire target for ideal surface by running with `--desire-target {target_value}` argument
(Default : – 0.67 , Ideal CO binding energy for CO2RR)

※ Set target error range for ideal surface by running with `--desire-range {range_value}` argument
(Default : ± 0.1)

A	B
elements	density
Ga_Ge	0
Fe_Pd	0
As_Ru	0
Re_Zn	0
Fe_Ti	0
Fe_Se	0
Ga_Pb	0
Fe_Fe	0
Au_Re	0
Cr_Mn	0
Ag_Ga	0