Electronic Supplementary Material (ESI) for RSC Advances.
This journal is © The Royal Society of Chemistry 2019

# Supporting Information for "Chemical space exploration guided by deep neural networks"

Dmitry Karlov,[†] Sergey Sosnin,[†] Igor V. Tetko,[‡] and Maxim V. Fedorov[*,†]

[†]*Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow 143026, Russia*

[‡]*Helmholtz Zentrum München - Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg Germany*

E-mail: m.fedorov@skoltech.ru

## The architecture of our encoding ANN

The structure of our neural network is presented in *Table 1.* To train our network we used Adam optimizer with $learning\_rate = 10^{-5}$ Neural networks training was performed using PyTorch 0.4 [https://pytorch.org/] with NVIDIA GeForce GTX 1080 Ti (Driver Version 390.42, CUDA V8.0.61). The parameters provided the best performance are listed in *Table 2.* The grid search space are given in *Table 3.*

Table 1: The architecture of the used neural network

| Layer | Neurons | Batch Normalization |
|---|---|---|
| Input | 2048 | Yes |
| 1 | 1024 | Yes |
| 2 | 1024 | Yes |
| 3 | 1024 | Yes |
| Output | 2 | No |

## Results

## ML methods parameters

Table 2: The optimal hyperparameters of classifires both found by grid search optimization and default

| Descriptor set | ML method | Parameters | |
|---|---|---|---|
| | | GPCR ligands | NR ligands |
| ECFP6 descriptors | kNN | n_neighbours = 24 | n_neighbours = 9 |
| | | weights = all_equal | weights = all_equal |
| | SVM | C = 0.015625 | C = 0.01562 |
| | | kernel = linear | kernel = linear |
| | XGBoost | learning_rate = 0.05 | learning_rate = 0.05 |
| | | l2 = 0.01 | l2 = 1. |
| | | max_depth = 3 | max_depth = 3 |
| | Random forest | n_estimators = 10 | n_estimators = 100 |
| | | max_features = all | max_features = sqrt(all) |
| | | min_sample_leaf = 10 | min_sample_leaf = 100 |
| pTSNE mapping | kNN | n_neighbours = 24 | n_neighbours = 9 |
| | | weights = all_equal | weights = all_equal |
| | SVM | C = 64 | C = 0.25 |
| | | kernel = polynomial | Kernel = rbf |
| | | gamma = 0.001 | gamma = 0.003 |
| | XGBoost | learning_rate = 0.05 | learning_rate = 0.05 |
| | | l2 = 0.001 | l2 = 0.01 |
| | | max_depth = 3 | max_depth = 3 |
| | Random forest | n_estimators = 10 | n_estimators = 300 |
| | | max_features = all | max_features = auto |
| | | min_sample_leaf = 10 | min_sample_leaf = 10 |
| PCA mapping | kNN | n_neighbours = 24 | n_neighbours =9 |
| | | weights = all_equal | weights = all_equal |
| | SVM | l2 = 0.015625 | l2 = 0.015625 |
| | | kernel = linear | kernel = linear |
| | XGBoost | learning_rate = 0.05 | learning_rate = 0.05 |
| | | l2 = 0.1 | l2 = 1. |
| | | max_depth = 3 | max_depth = 3 |
| | Random forest | n_estimators = 10 | n_estimators = 13 |
| | | max_features = all | max_features = sqrt(all) |
| | | min_sample_leaf = 10 | min_sample_leaf = 100 |
| MDS mapping | kNN | n_neighbours = 24 | n_neighbours =9 |
| | | weights = all_equal | weights = all_equal |
| | SVM | l2 = 0.015625 | l2 = 0.015625 |
| | | kernel = linear | kernel = linear |
| | XGBoost | learning_rate = 0.05 | learning_rate = 0.05 |
| | | l2 = 1. | l2 = 1. |
| | | max_depth = 4 | max_depth = 3 |
| | Random forest | n_estimators = 3000 | n_estimators = 300 |
| | | max_features = log2(all) | max_features = all |
| | | min_sample_leaf = 100 | min_sample_leaf = 30 |

Table 3: Sets of parameters for grid search procedure

| Method | Parameter | Values |
|---|---|---|
| kNN | Number of neigbors to consider | 1, 3, 9, 12, 15, 18, 21, 24, 27, 30 |
| | Distance metric | manhattan |
| | weights for neghbours | all equal, inverse distance |
| RF | Number of estimators | 10, 30, 100, 300, 1000, 3000 |
| | maximum number of features | all, sqrt(all), log2(all) |
| | minimum number of samples in leafs | 10, 30, 100, 300 |
| SVM | constant at L2 penalty | 0.015625, 0.0625, 0.25, 1, 4, 16, 64, 256, 1024 |
| | kernel type | linear, rbf, polynomial of degree 3 |
| | kernel coefficient (rbf, polinomial) | 0.01, 0.003, 0.001, 0.0003, 0.0001, 0.00003, 0.00001 |
| XGBoost | booster | gbtree |
| | learning rate | 0.05, 0.1, 0.15, 0.2, 0.25, 0.3 |
| | max depth | 3, 4, 5, 6, 7, 8, 9 |
| | L2 penalty | 0.001, 0.01, 0.1, 1, 10 |