Predicting GPR40 Agonists with A Deep Learning-Based Ensemble Model

Jiamin Yang,1 Chen Jiang,1 Jing Chen,1 Lu-Ping Qin\*1 and Gang Cheng\*1

1School of Pharmaceutical Sciences, Zhejiang Chinese Medical University, Hangzhou 310053, People’s Republic of China. E-mail:lpqin@zcmu.edu.cn; gangcheng@zcmu.edu.cn

Content: Supplementary **Fig. 1** shows the detail of the process of data processing, including converting to canonical SMILES, labeling, merging datasets, checking SMILES, removing reduplicative SMILES, remove salt. Supplementary **Fig. 2** and **3** show the frameworks of GPR40 agonists and non-agonists. When the area of the framework in the picture is bigger, the framework is more important. Supplementary **Table 1** show the process of roughly reducing 25 features to 10 features in Layer 1. The optimization process of molecular features in layer 1 are shown in Supplementary **Table** **2**, **3**, **4,** **5** and **6**. Finally, supplementary **Table 6** and **7** including the hyperparameter values of different models.



Supplementary **Fig. 1** The process of data processing



**Supplementary Fig. 2** The frameworks of GPR40 agonists



**Supplementary Fig. 3** The frameworks of GPR40 non-agonists

**Supplementary Table** 1. Roughly reducing 25 features to 10 features in Layer 1

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | model (algorithm and feature) | AUC-ROC | SD |
| 1 | FCNN+MAP4 | 0.9314 | 0.0005 |
| 2 | FCNN+AtomPairFP | 0.9261 | 0.0006 |
| 3 | FCNN+Topology | 0.9164 | 0.0006 |
| 4 | FCNN+MorganFP | 0.9159 | 0.0008 |
| 5 | FCNN+MOE | 0.9108 | 0.0045 |
| 6 | FCNN+Kappa | 0.9089 | 0.0009 |
| 7 | FCNN+TorsionFP | 0.9087 | 0.0009 |
| 8 | FCNN+MHFP6 | 0.9081 | 0.0053 |
| 9 | FCNN+Connectivity | 0.9075 | 0.0002 |
| 10 | FCNN+Charge | 0.9071 | 0.0001 |
| 11 | FCNN+Estate | 0.9011 | 0.0025 |
| 12 | FCNN+PharmacoErGFP | 0.8999 | 0.0003 |
| 13 | FCNN+RDkitFP | 0.8945 | 0.0037 |
| 14 | FCNN+PharmacoPFP | 0.8928 | 0.0021 |
| 15 | FCNN+InfoContent | 0.8883 | 0.0028 |
| 16 | FCNN+Property | 0.8868 | 0.0013 |
| 17 | FCNN+Constitution | 0.8746 | 0.0022 |
| 18 | FCNN+Matrix | 0.8740 | 0.0205 |
| 19 | FCNN+MACCSFP | 0.8615 | 0.0013 |
| 20 | FCNN+PubChemFP | 0.8610 | 0.0017 |
| 21 | FCNN+Path | 0.8571 | 0.0047 |
| 22 | FCNN+AvalonFP | 0.8558 | 0.0037 |
| 23 | FCNN+EstateFP | 0.8460 | 0.0013 |
| 24 | FCNN+Autocorr | 0.8149 | 0.0145 |
| 25 | FCNN+Fragment | 0.7956 | 0.0008 |

SD: Standard deviation. These data points are same to the AUC-ROC of models based on FCNN in Figure 3.

**Supplementary Table** 2. Removing the molecular features (in layer 1) from 10 features to 9 features

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | Features in layer 1 | AUC-ROC | SD |
| 1 | 10\_features\* | 0.9450 | 0.0006 |
| 2 | 10\_features\_remove\_AtomPairFP | 0.9316 | 0.0026 |
| 3 | 10\_features\_remove\_Connectivity | 0.9436 | 0.0029 |
| 4 | 10\_features\_remove\_Charge | 0.9437 | 0.0001 |
| 5 | 10\_features\_remove\_Kappa | 0.9444 | 0.0008 |
| 6 | 10\_features\_remove\_MAP4 | 0.9418 | 0.0002 |
| 7 | 10\_features\_remove\_MHFP6 | **0.9466** | 0.0012 |
| 8 | 10\_features\_remove\_ MOE | 0.9438 | 0.0013 |
| 9 | 10\_features\_remove\_MorganFP | 0.9432 | 0.0003 |
| 10 | 10\_features\_remove\_Topology | 0.9440 | 0.0013 |
| 11 | 10\_features\_remove\_TorsionFP | 0.9459 | 0.0009 |

\* including AtomPairFP, Connectivity, Charge, Kappa, MAP4, MHFP6, MOE, MorganFP, Topology and TorsionFP. SD: Standard deviation.

**Supplementary Table** 3. Removing the molecular features (in layer 1) from 9 features to 8 features

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | Features in layer 1 | AUC-ROC | SD |
| 1 | 9\_features\* | 0.9467 | 0.0018 |
| 2 | 9\_features\_remove\_AtomPairFP | 0.9301 | 0.0010 |
| 3 | 9\_features\_remove\_Connectivity | 0.9465 | 0.0008 |
| 4 | 9\_features\_remove\_Charge | 0.9460 | 0.0017 |
| 5 | 9\_features\_remove\_Kappa | 0.9476 | 0.0022 |
| 6 | 9\_features\_remove\_MAP4 | 0.9437 | 0.0002 |
| 7 | 9\_features\_remove\_MOE | 0.9458 | 0.0018 |
| 8 | 9\_features\_remove\_MorganFP | 0.9461 | 0.0009 |
| 9 | 9\_features\_remove\_Topology | 0.9470 | 0.0011 |
| 10 | 9\_features\_remove\_TorsionFP | **0.9486** | 0.0011 |

\* including AtomPairFP, Connectivity, Charge, Kappa, MAP4, MOE, MorganFP, Topology and TorsionFP. SD: Standard deviation.

**Supplementary Table** 4. Removing the molecular features (in layer 1) from 8 features to 7 features

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | Features in layer 1 | AUC-ROC | SD |
| 1 | 8\_features\* | 0.9483 | 0.0010 |
| 2 | 8\_features\_remove\_AtomPairFP | 0.9297 | 0.0031 |
| 3 | 8\_features\_remove\_Connectivity | **0.9495** | 0.0020 |
| 4 | 8\_features\_remove\_Charge | 0.9479 | 0.0011 |
| 5 | 8\_features\_remove\_Kappa | 0.9488 | 0.0015 |
| 6 | 8\_features\_remove\_MAP4 | 0.9444 | 0.0007 |
| 7 | 8\_features\_remove\_MOE | 0.9482 | 0.0029 |
| 8 | 8\_features\_remove\_MorganFP | 0.9481 | 0.0024 |
| 9 | 8\_features\_remove\_Topology | 0.9488 | 0.0014 |

\* including AtomPairFP, Connectivity, Charge, Kappa, MAP4, MOE, MorganFP and Topology. SD: Standard deviation.

**Supplementary Table** 5. Removing the molecular features (in layer 1) from 7 features to 6 features

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | Features in layer 1 | AUC-ROC | SD |
| 1 | 7\_features\* | 0.9486 | 0.0013 |
| 2 | 7\_features\_remove\_AtomPairFP | 0.9282 | 0.0033 |
| 3 | 7\_features\_remove\_Charge | 0.9485 | 0.0021 |
| 4 | 7\_features\_remove\_Kappa | **0.9497** | 0.0029 |
| 5 | 7\_features\_remove\_MAP4 | 0.9457 | 0.0003 |
| 6 | 7\_features\_remove\_MOE | 0.9488 | 0.0027 |
| 7 | 7\_features\_remove\_MorganFP | 0.9483 | 0.0016 |
| 8 | 7\_features\_remove\_Topology | 0.9496 | 0.0031 |

\* including AtomPairFP, Charge, Kappa, MAP4, MOE, MorganFP and Topology. SD: Standard deviation.

**Supplementary Table** 6. Removing the molecular features (in layer 1) from 6 features to 5 features

|  |  |  |  |
| --- | --- | --- | --- |
| Entry | Features in layer 1 | AUC-ROC | SD |
| 1 | 6\_features | **0.9504** | 0.0022 |
| 2 | 6\_features\_remove\_AtomPairFP | 0.9272 | 0.0046 |
| 3 | 6\_features\_remove\_Charge | 0.9486 | 0.0018 |
| 4 | 6\_features\_remove\_MAP4 | 0.9462 | 0.0014 |
| 5 | 6\_features\_remove\_MOE | 0.9500 | 0.0020 |
| 6 | 6\_features\_remove\_MorganFP | 0.9496 | 0.0014 |
| 7 | 6\_features\_remove\_Topology | 0.9501 | 0.0022 |

\* including AtomPairFP, Charge, MAP4, MOE, MorganFP and Topology. SD: Standard deviation.

**Supplementary** **Table** 7. The hyperparameters of FCNN, RF and XGBoost models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| hyperparameter | FCNN | hyperparameter | RF | hyperparameter | XGBoost |
| layer\_sizes | 800 | bootstrap | TRUE | max\_depth | 5 |
| weight\_init\_stddevs | 0.02 | class\_weight | balanced | learning\_rate | 0.05 |
| bias\_init\_consts | 1 | criterion | gini | n\_estimators | 1000 |
| dropouts | 0.5 | max\_depth | None | gamma | 0 |
| weight\_decay\_penalty | 0.1 | max\_features | auto | min\_child\_weight | 5 |
| weight\_decay\_penalty\_type | l2 | max\_leaf\_nodes | None | max\_delta\_step | 1 |
| batch\_size | 8192 | min\_impurity\_decrease | 0 | subsample | 0.53 |
| learning\_rate | 0.001 | min\_impurity\_split | None | colsample\_bytree | 0.66 |
|  |  | min\_sample\_leaf | 1 | cosample\_bylevel | 1 |
|  |  | min\_sample\_split | 2 | reg\_alpha | 0 |
|  |  | min\_weight\_fraction\_leaf | 0 | reg\_lambda | 1 |
|  |  | n\_estimators | 50 | scale\_pos\_weight | 1 |
|  |  | n\_jobs | -1 | base\_score | 0.5 |
|  |  | oob\_score | False |  |  |
|  |  | random\_state | None |  |  |
|  |  | verbose | 0 |  |  |
|  |  | warm\_start | False |  |  |

**Supplementary** **Table** 8. The hyperparameters of LR and D-MPNN models

|  |  |  |  |
| --- | --- | --- | --- |
| hyperparameter | LR | hyperparameter | D-MPNN |
| C | 1 | activation | ReLU |
| class\_weight | balance | atom\_messages | False |
| dual | False | batch\_size | 50 |
| fit\_intercept | True | bias | False |
| intercept\_scaling | 1 | dataset\_type | classification |
| max\_iter | 100 | depth | 3 |
| multi\_class | ovr | dropout | 0 |
| n\_jobs | -1 | ensemble\_size | 1 |
| penalty | l2 | epochs | 30 |
| random\_state | None | features\_scaling | True |
| solver | liblinear | ffn\_hidden\_size | 300 |
| tol | 0.0001 | ffn\_num\_layers | 2 |
| verbose | 0 | final\_lr | 0.0001 |
| warm\_start | False | hidden\_size | 300 |
|  |  | init\_lr | 0.0001 |
|  |  | max\_lr | 0.001 |
|  |  | split\_sizes | 0.8, 0.1, 0.1 |
|  |  | split\_type | random |
|  |  | warmup\_epochs | 2 |