**Supporting information:**

**Application of Bioactivity Profile-Based Fingerprints for Building Machine Learning Models**

Noé Sturm,1, \* Jiangming Sun,1 Yves Vandriessche,2 Andreas Mayr,3 Günter Klambauer,3 Lars Carlsson,4 Ola Engkvist,1 and Hongming Chen1, \*

*1Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Pepparedsleden 1, 43153 Mölndal, Sweden*

*2Intel Corporation, Data Center Group, Veldkant 31, 2550 Kontich, Belgium.*

*3**LIT AI Lab & Institute for Machine Learning, Johannes Kepler University Linz, Altenbergerstr 69, 4040 Linz, Austria*

*4Quantitative Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Pepparedsleden 1, 43153 Mölndal, Sweden*

**Corresponding authors**

Noé Sturm: [noejoseph.sturm@astrazeneca.com](mailto:noejoseph.sturm@astrazeneca.com)

Hongming Chen: [Hongming.Chen@astrazeneca.com](mailto:Hongming.Chen@astrazeneca.com)

**SMILES standardization procedure:**

*(adapted from ExCAPE-db, with permission)*

Standardisation was performed with ambitcli version 3.0.2. The ambicli tool includes a number of chemical structure processing options (fragments splitting, isotopes removal, handling implicit hydrogens, stereochemistry, InChI generation, SMILES generation, structure transformation via SMIRKS, tautomers generation, neutralisation). The structure standardisation was performed in two steps, as follows:

Read SMILES from AstraZeneca repository, perform standardisation, write output in tabular TXT format. Ambitcli tool was executed with the following options:

**$java -Xmx1536m -jar ambitcli-3.0.2-7433.jar**

-a standardize

-m post

-d page=0 -d pagesize=-1

–d tag\_smiles=AMBIT\_SMILES

-d tag\_inchi=AMBIT\_InChI

-d tag\_inchikey=AMBIT\_InChIKey

–d tautomers=true

-d splitfragments=true

-d implicith=true

-d smilescanonical=false

-d smiles=true

-d inchi=true

-d neutralise=true

-d isotopes=true

-d tag\_tokeep=PUBCHEM\_CID

-i <input>.smiles

-o standardized\_smiles.txt

where <input> corresponds to the original data retrieved from Astrazeneca internal repository.

**ECFP generation:**

Read standardized smiles and calculate fingerprints. The following options were used, which created the ECFP with binary bit values, the ECFP with counts and the raw ECFP with hash keys.

**$java -Xmx1536m -jar ambitcli-3.0.2-7433.jar**

-a fingerprint

-m post

-d page=0 -d pagesize=-1

–d fpclass=CircularFingerprinter

-d fplen=1024

-d fptype=ECFP6

-d tag\_tokeep=AMBIT\_InChIKey

-d inputtag\_smiles=AMBIT\_SMILES

–d inputtag\_inchikey=AMBIT\_InChIKey

-d inputtag\_inchi=AMBIT\_InChI

-d write\_count=true

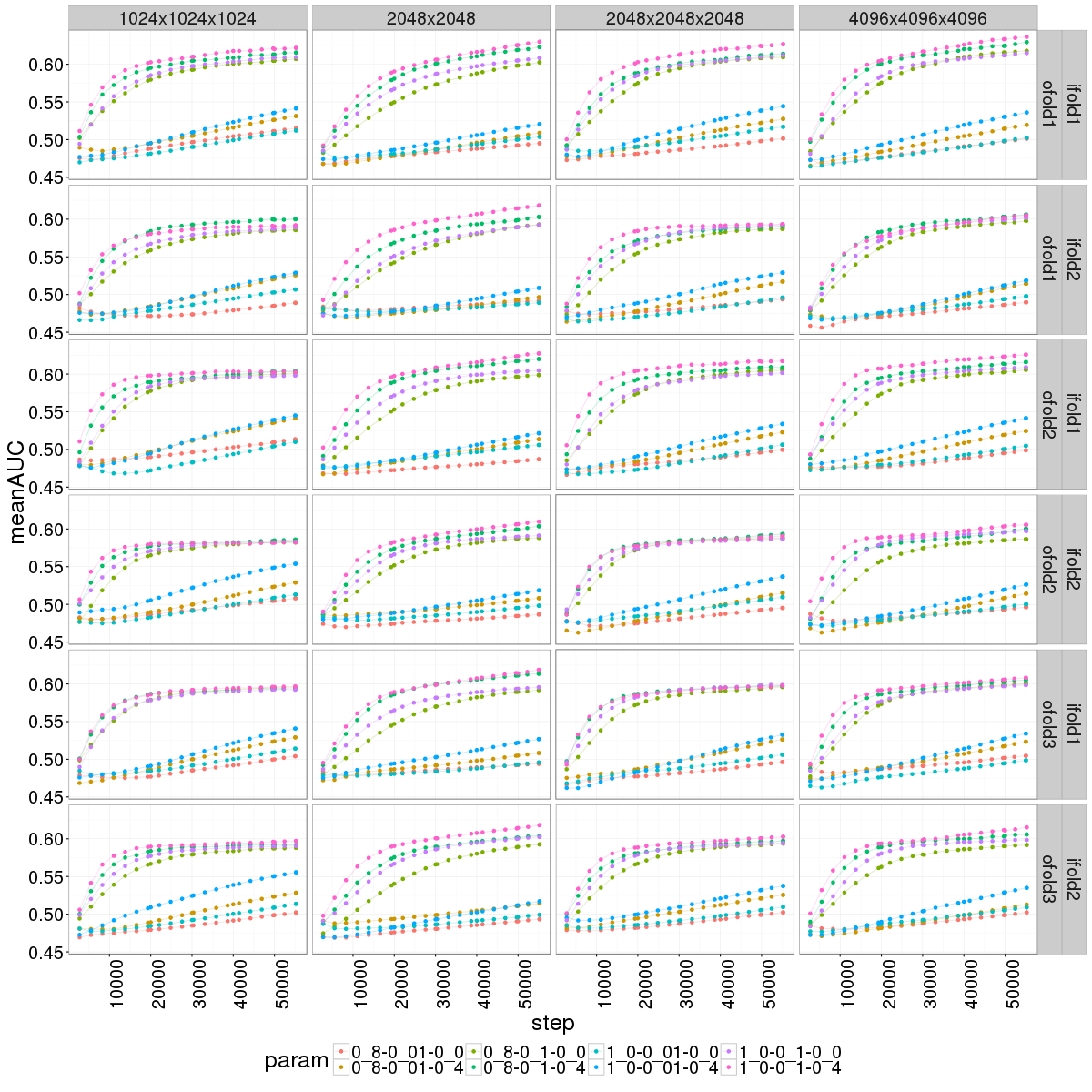
-d write\_raw=true

-i standardized\_smiles.txt

-o ecfp6.txt



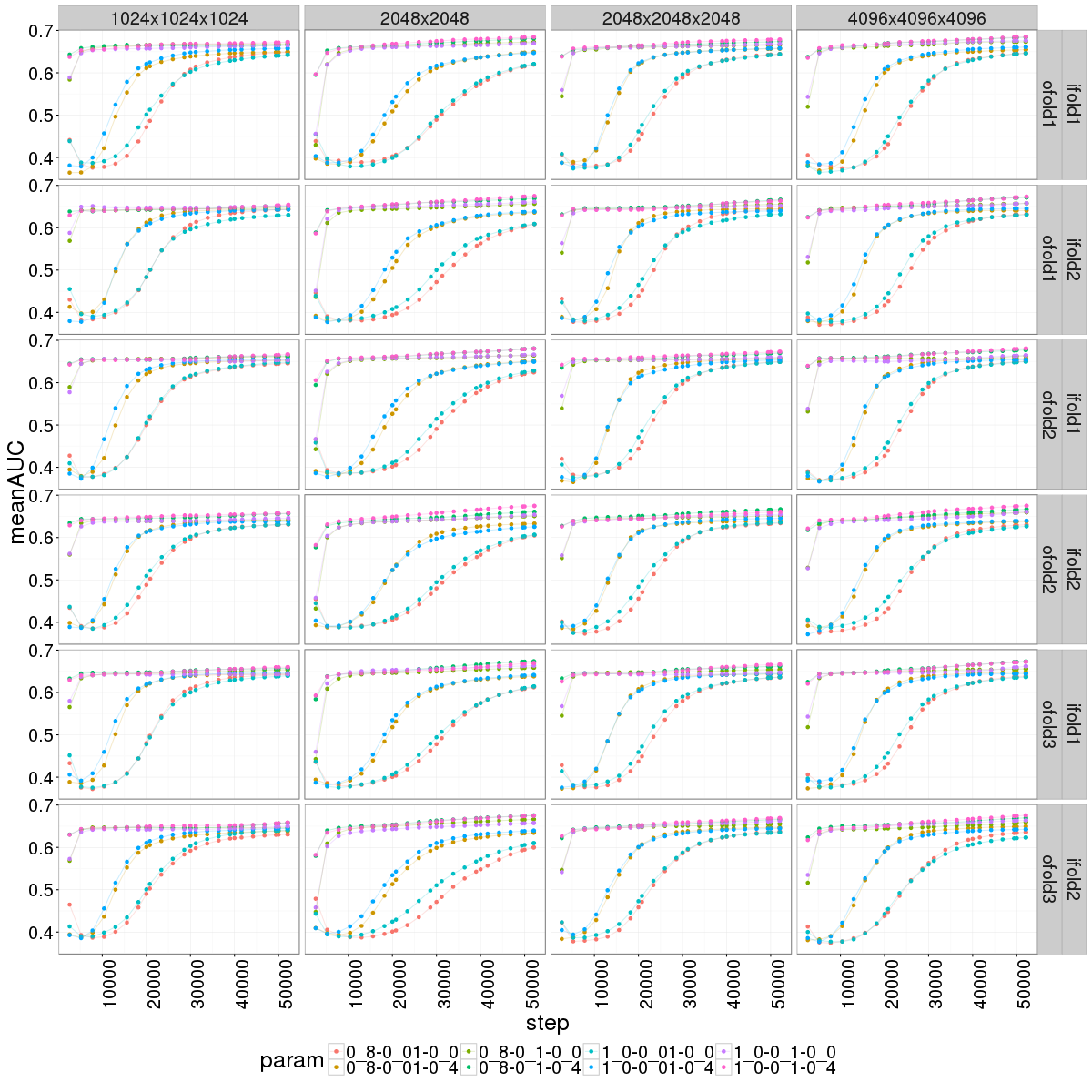
**Figure S1. Hyperparameter search and model evaluation workflow.** The dataset is split in three blocks, A, B and C. During the hyperparameter search phase, a model is trained upon the first block, validated on the second block while leave the third block out. The hyperparameter search is a systematic grid search exploring 32 different hyperparameters. Each hyperparameter set is used to train two models, one with each block left in. Given that this procedure is repeated three times (i.e. once for each left out fold), we obtain six performance estimates per hyperparameter. Each searched hyperparameter is then ranked according to the average performance over the six estimates. The best set of hyperparameter is then used to train three models as in a standard 3-Cross fold validation procedure. Final evaluation of the model is given by the average performance over the 3 folds.



**Figure S2. Inner loop of nested-cross validation learning curves of DNN with ECFP.** Each panel represents the learning curves of one of the four network architecture trained with one training set and evaluated on one validation set. The grey top banner describes the hidden layers layouts. There are eight curves per panel, each of them represents one of the searched hyperparameter set. Y-axis represent mean ROC-AUC averaged over the 131 *external* targets. X-axis represents the steps at which checkpoint models were saved. The grey annotation on the right represents indices of folds used for training and validation.

**Table S1. Ranking of the network hyperparameters obtained with ECFP at the final checkpoint of the training in the inner loop of nested-CV.** Input dropout rate is the rate of input nodes kept. The layout describes the arrangement of the hidden layers with the number of nodes per layer. M represents momentum. MEAN ROC-AUC is averaged over the 131 *external* targets not encoded in the HTSFP. STDEV represents the standard deviation of MEAN ROC-AUC values over the six validations of the nested cross validation.

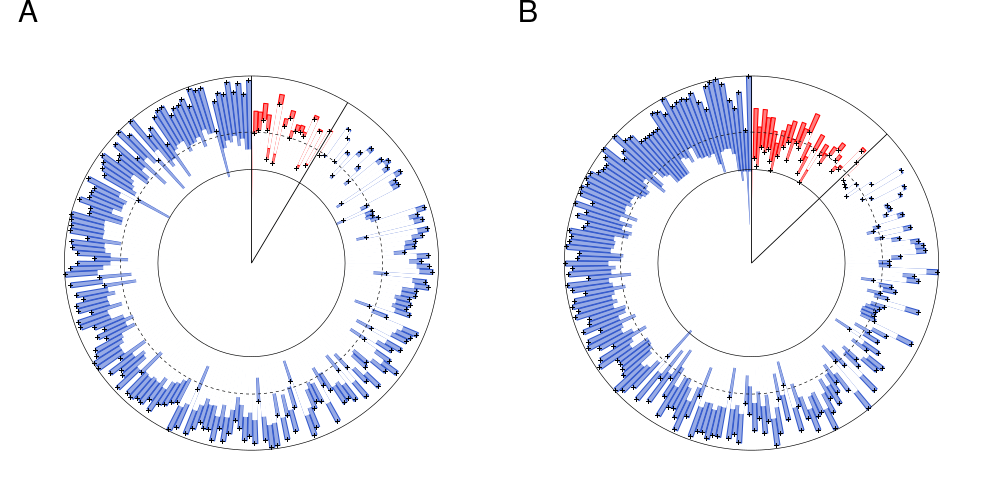
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | layout | INPUT DROPOUT | LEARNING RATE | m | mean  ROC-AUC | stdev |
| 1 | 2048x2048 | 1.0 | 0.1 | 0.4 | 0,6207 | 0,0078 |
| 2 | 4096x4096x4096 | 1.0 | 0.1 | 0.4 | 0,6190 | 0,0142 |
| 3 | 4096x4096x4096 | 0.8 | 0.1 | 0.4 | 0,6128 | 0,0131 |
| 4 | 2048x2048 | 0.8 | 0.1 | 0.4 | 0,6100 | 0,0099 |
| 5 | 2048x2048x2048 | 1.0 | 0.1 | 0.4 | 0,6071 | 0,0158 |
| 6 | 4096x4096x4096 | 1.0 | 0.1 | 0.0 | 0,6043 | 0,0084 |
| 7 | 1024x1024x1024 | 1.0 | 0.1 | 0.4 | 0,6020 | 0,0162 |
| 8 | 2048x2048x2048 | 0.8 | 0.1 | 0.4 | 0,6011 | 0,0102 |
| 9 | 2048x2048 | 1.0 | 0.1 | 0.0 | 0,6009 | 0,0076 |
| 10 | 4096x4096x4096 | 0.8 | 0.1 | 0.0 | 0,6006 | 0,0138 |
| 11 | 1024x1024x1024 | 0.8 | 0.1 | 0.4 | 0,6001 | 0,0127 |
| 12 | 2048x2048x2048 | 1.0 | 0.1 | 0.0 | 0,5986 | 0,0107 |
| 13 | 2048x2048x2048 | 0.8 | 0.1 | 0.0 | 0,5970 | 0,0100 |
| 14 | 1024x1024x1024 | 1.0 | 0.1 | 0.0 | 0,5953 | 0,0112 |
| 15 | 2048x2048 | 0.8 | 0.1 | 0.0 | 0,5952 | 0,0059 |
| 16 | 1024x1024x1024 | 0.8 | 0.1 | 0.0 | 0,5930 | 0,0111 |
| 17 | 1024x1024x1024 | 1.0 | 0.01 | 0.4 | 0,5461 | 0,0108 |
| 18 | 2048x2048x2048 | 1.0 | 0.01 | 0.4 | 0,5382 | 0,0056 |
| 19 | 4096x4096x4096 | 1.0 | 0.01 | 0.4 | 0,5311 | 0,0072 |
| 20 | 1024x1024x1024 | 0.8 | 0.01 | 0.4 | 0,5290 | 0,0021 |
| 21 | 2048x2048x2048 | 0.8 | 0.01 | 0.4 | 0,5231 | 0,0054 |
| 22 | 2048x2048 | 1.0 | 0.01 | 0.4 | 0,5171 | 0,0043 |
| 23 | 4096x4096x4096 | 0.8 | 0.01 | 0.4 | 0,5156 | 0,0032 |
| 24 | 1024x1024x1024 | 1.0 | 0.01 | 0.0 | 0,5119 | 0,0026 |
| 25 | 2048x2048x2048 | 1.0 | 0.01 | 0.0 | 0,5098 | 0,0077 |
| 26 | 2048x2048 | 0.8 | 0.01 | 0.4 | 0,5085 | 0,0065 |
| 27 | 1024x1024x1024 | 0.8 | 0.01 | 0.0 | 0,5052 | 0,0098 |
| 28 | 4096x4096x4096 | 1.0 | 0.01 | 0.0 | 0,5035 | 0,0046 |
| 29 | 2048x2048x2048 | 0.8 | 0.01 | 0.0 | 0,4995 | 0,0039 |
| 30 | 4096x4096x4096 | 0.8 | 0.01 | 0.0 | 0,4989 | 0,0050 |
| 31 | 2048x2048 | 1.0 | 0.01 | 0.0 | 0,4989 | 0,0053 |
| 32 | 2048x2048 | 0.8 | 0.01 | 0.0 | 0,4925 | 0,0031 |

****

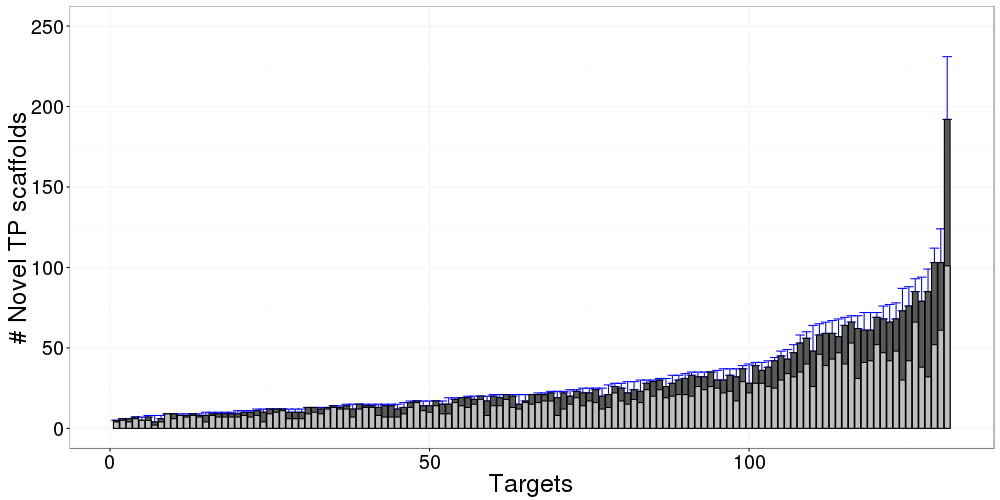
**Figure S3. Nested-cross validation learning curves of DNN with HTSFP.** Each panel represents the learning curves of one of the four network architecture as trained with one training set and evaluated on one validation set. The grey top banner describes the hidden layers layouts. There are eight curves per panel. each of them represents one of the searched hyperparameter set. Y-axis represent mean ROC-AUC averaged over the 131 *external* targets. X-axis represents the steps at which checkpoint models were saved. The grey annotation on the right represents indices of folds used for training and validation.

**Table S2. Ranking of the network hyperparameters obtained with HTSFP at the final checkpoint of the training in the inner loop of the nested-CV.** Input dropout rate is the rate of input nodes kept. The layout describes the arrangement of the hidden layers with the number of nodes per layer. M represents momentum. MEAN ROC-AUC is averaged over the 131 *external* targets not encoded in the HTSFP. STDEV represents the standard deviation of MEAN ROC-AUC values over the six validations of the nested cross validation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | layout | input dropout | learning rate | m | mean  ROC-AUC | stdev |
| 1 | 4096x4096x4096 | 1.0 | 0.1 | 0.4 | 0.6751 | 0.0030 |
| 2 | 2048x2048 | 1.0 | 0.1 | 0.4 | 0.6748 | 0.0034 |
| 3 | 4096x4096x4096 | 0.8 | 0.1 | 0.4 | 0.6713 | 0.0038 |
| 4 | 2048x2048 | 0.8 | 0.1 | 0.4 | 0.6691 | 0.0076 |
| 5 | 2048x2048x2048 | 0.8 | 0.1 | 0.4 | 0.6663 | 0.0025 |
| 6 | 2048x2048x2048 | 1.0 | 0.1 | 0.4 | 0.6648 | 0.0045 |
| 7 | 4096x4096x4096 | 1.0 | 0.1 | 0.0 | 0.6607 | 0.0026 |
| 8 | 2048x2048 | 1.0 | 0.1 | 0.0 | 0.6599 | 0.0051 |
| 9 | 4096x4096x4096 | 0.8 | 0.1 | 0.0 | 0.6587 | 0.0026 |
| 10 | 1024x1024x1024 | 1.0 | 0.1 | 0.4 | 0.6585 | 0.0046 |
| 11 | 2048x2048 | 0.8 | 0.1 | 0.0 | 0.6563 | 0.0050 |
| 12 | 1024x1024x1024 | 0.8 | 0.1 | 0.4 | 0.6558 | 0.0046 |
| 13 | 2048x2048x2048 | 0.8 | 0.1 | 0.0 | 0.6543 | 0.0028 |
| 14 | 2048x2048x2048 | 1.0 | 0.1 | 0.0 | 0.6535 | 0.0042 |
| 15 | 1024x1024x1024 | 0.8 | 0.1 | 0.0 | 0.6494 | 0.0072 |
| 16 | 1024x1024x1024 | 1.0 | 0.1 | 0.0 | 0.6483 | 0.0038 |
| 17 | 4096x4096x4096 | 1.0 | 0.01 | 0.4 | 0.6449 | 0.0053 |
| 18 | 2048x2048x2048 | 1.0 | 0.01 | 0.4 | 0.6447 | 0.0049 |
| 19 | 2048x2048x2048 | 0.8 | 0.01 | 0.4 | 0.6445 | 0.0061 |
| 20 | 1024x1024x1024 | 0.8 | 0.01 | 0.4 | 0.6443 | 0.0044 |
| 21 | 4096x4096x4096 | 0.8 | 0.01 | 0.4 | 0.6437 | 0.0072 |
| 22 | 1024x1024x1024 | 1.0 | 0.01 | 0.4 | 0.6437 | 0.0062 |
| 23 | 2048x2048x2048 | 0.8 | 0.01 | 0.0 | 0.6404 | 0.0066 |
| 24 | 1024x1024x1024 | 0.8 | 0.01 | 0.0 | 0.6394 | 0.0066 |
| 25 | 2048x2048 | 0.8 | 0.01 | 0.4 | 0.6379 | 0.0069 |
| 26 | 2048x2048x2048 | 1.0 | 0.01 | 0.0 | 0.6368 | 0.0063 |
| 27 | 2048x2048 | 1.0 | 0.01 | 0.4 | 0.6363 | 0.0094 |
| 28 | 4096x4096x4096 | 0.8 | 0.01 | 0.0 | 0.6358 | 0.0080 |
| 29 | 1024x1024x1024 | 1.0 | 0.01 | 0.0 | 0.6352 | 0.0074 |
| 30 | 4096x4096x4096 | 1.0 | 0.01 | 0.0 | 0.6335 | 0.0087 |
| 31 | 2048x2048 | 1.0 | 0.01 | 0.0 | 0.6123 | 0.0084 |
| 32 | 2048x2048 | 0.8 | 0.01 | 0.0 | 0.6108 | 0.0073 |

****

**Figure S4.** Comparison ofROC-AUC performance of HTSFP and ECFP models for the 209 targets *internal* (A: DNN. B: SVM) of the HTSFP assay panel. Bars represent the target-wise difference of ROC-AUC. The cross points on the bars mark the values of HTSFP models and the opposite extremities of the bars mark the performance of ECFP models. Red bars correspond to targets for which ECFP performs better whereas blue bars correspond to targets for which HTSFP performs better. The dashed line represents ROC-AUC value of 0.7. whereas plain lines represent ROC-AUC values of 0.5 and 1 respectively.

****

**Figure S5.** Count of novel true positive scaffolds per target outside the HTSFP. Counts are represented by the heights of the bars. these are not stacked bars. Dark grey bars represent the count of novel molecular scaffolds per target. Light grey bars represent the number of novel topological scaffolds per target. these bars are superimposed to dark bars. Blue error bars represent the total number of true positives per target. Only a small fraction of the true positives is not novel due to structural cluster-based data splitting.