**Supplementary Information**

# Multi-objective active machine learning rapidly improves structure-activity models and reveals new protein-protein interaction inhibitors

**Table of contents**

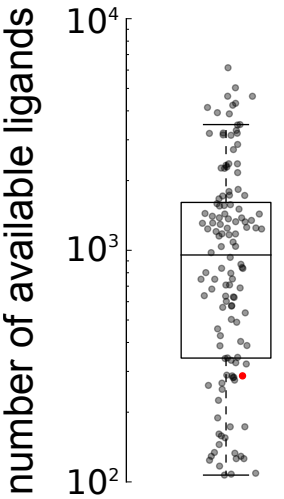



Figure S1: Distribution of numbers of all available ligands for family A GPCRs after pre-processing ChEMBL19 data. The data point for CXCR4 is shown in red.

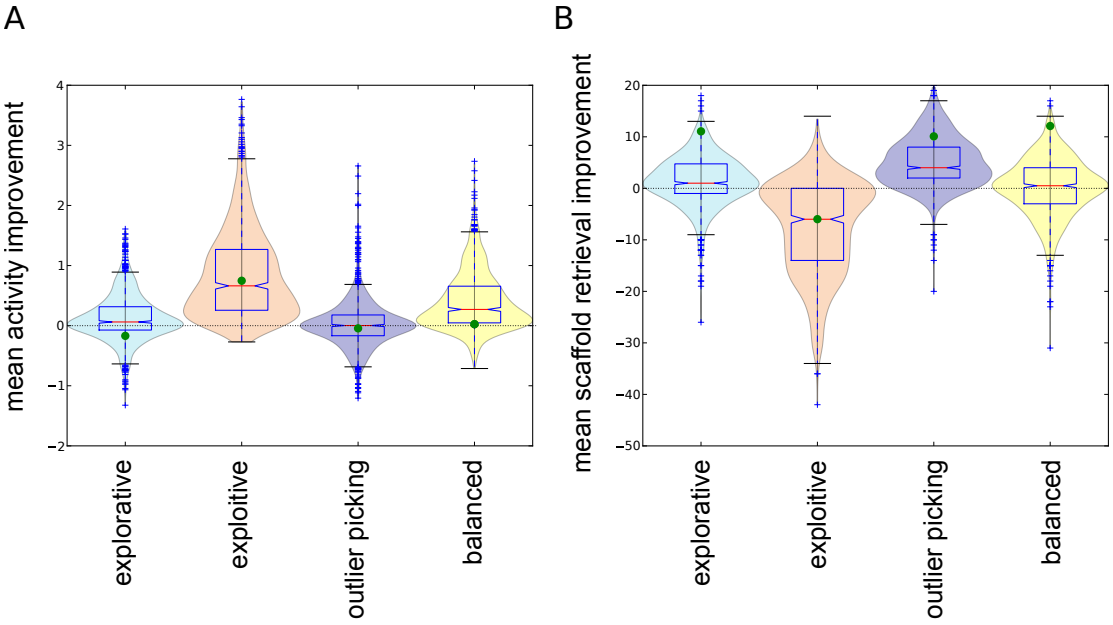



Figure S2: Retrospective evaluation of balanced active learning strategy on all ChEMBL16 targets. Initial training data was selected according to "time-split" criterion using the year of publication for each compound and adapting the threshold to include 33% of the whole dataset. Active learning and random molecule picking was run for 50 iterations per dataset. Active learning was performed using different selection functions: maximum uncertainty (light blue, "explorative"), maximum prediction (orange, "exploitive"), maximum random forest dissimilarity to the training data (violet, "outlier picking"), or an average value of all the strategies (yellow, "balanced"). Performance was evaluated as the difference in average activity of the 50 selected compounds (**A**) or the number of retrieved atomistic molecular scaffolds among those compounds (**B**) compared to random picking on the same data. The values for CXCR4 are shown with a green dot. The balanced strategy apparently provides a compromise in terms of finding active compounds better than explorative or outlier picking, while at the same timing detecting more novel compounds compared to a purely exploitive strategy.

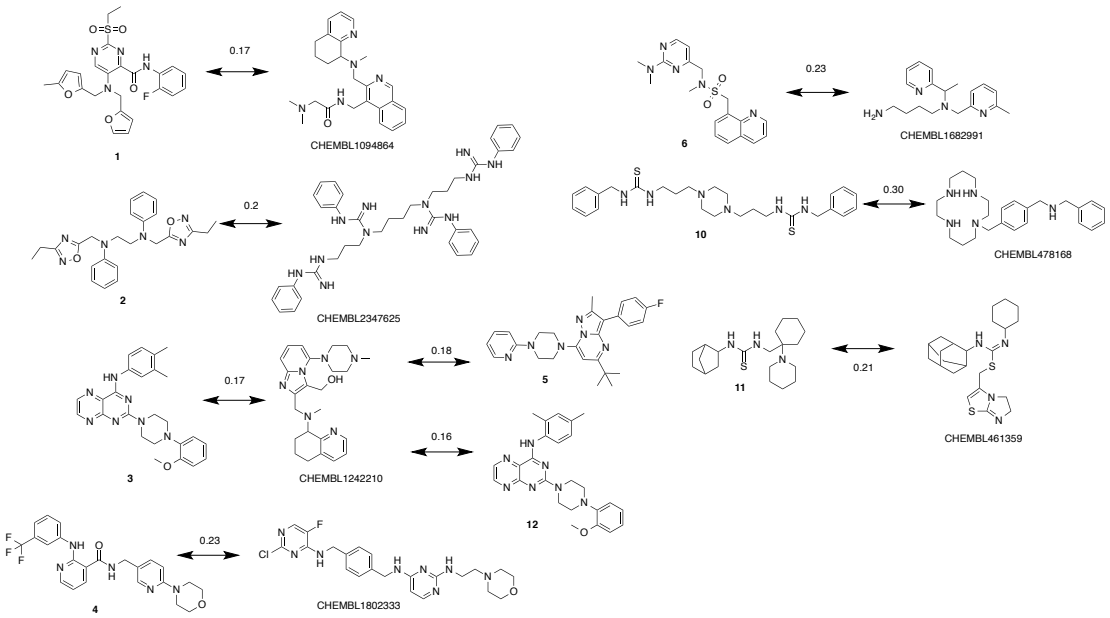



Figure S3: Nearest neighbors in the ChEMBL19 training data of selected hits. Compound similarity was calculated using the Tanimoto coefficient on Morgan fingerprints (RDKit, *radius* = 2, 2048 bit).

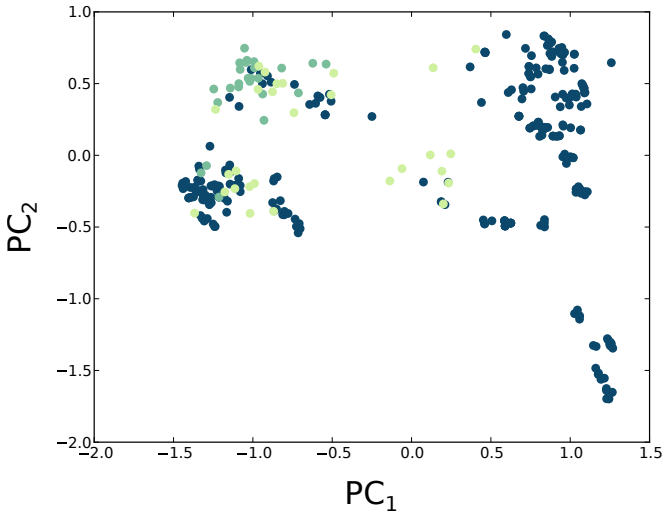



Figure S4: Positions of selected compounds during active learning iterations and initial training compounds in important feature space. This feature space consists of the first two principle components of the normalized descriptor values for the most important features (*cf.* Figure 3B). ChEMBL training data is shown in dark blue, the first active learning selection in light blue and the second active learning selection in light green.

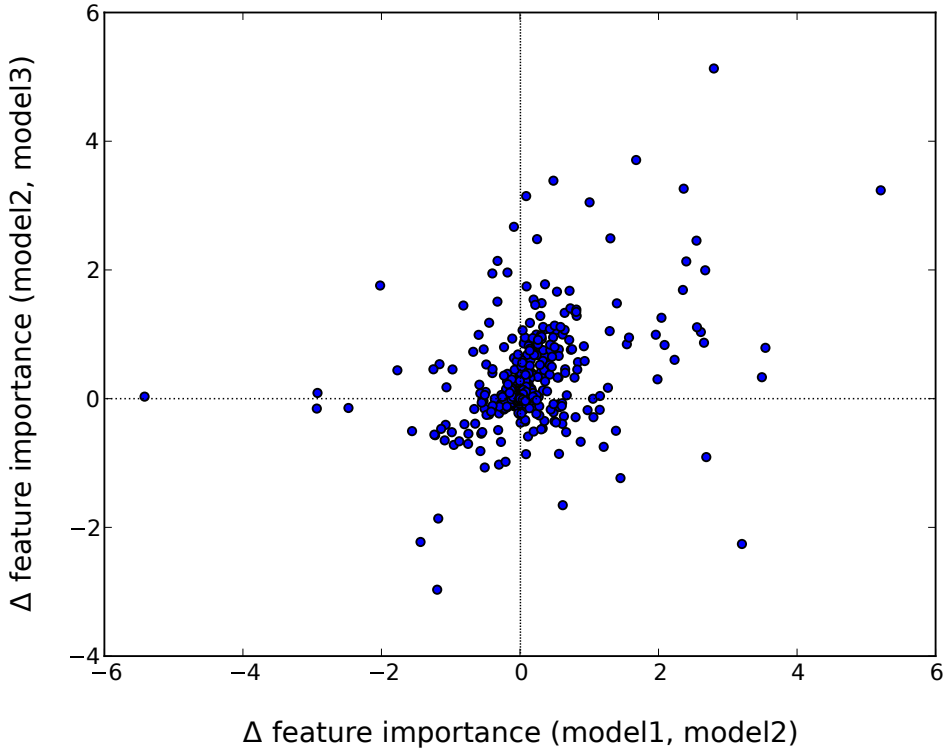



Figure S5: Comparison of change of model importance for individual features (descriptors). Every dot corresponds to a single feature and its position reflects the change in model importance after the first or the second prospective iteration of active learning. Features are visible that steadily increased (quadrant I) or decreased (quadrant III) in importance. More interestingly, other features were re-discovered as important (quadrant II) or neglected (quadrant IV) during the second prospective iteration.

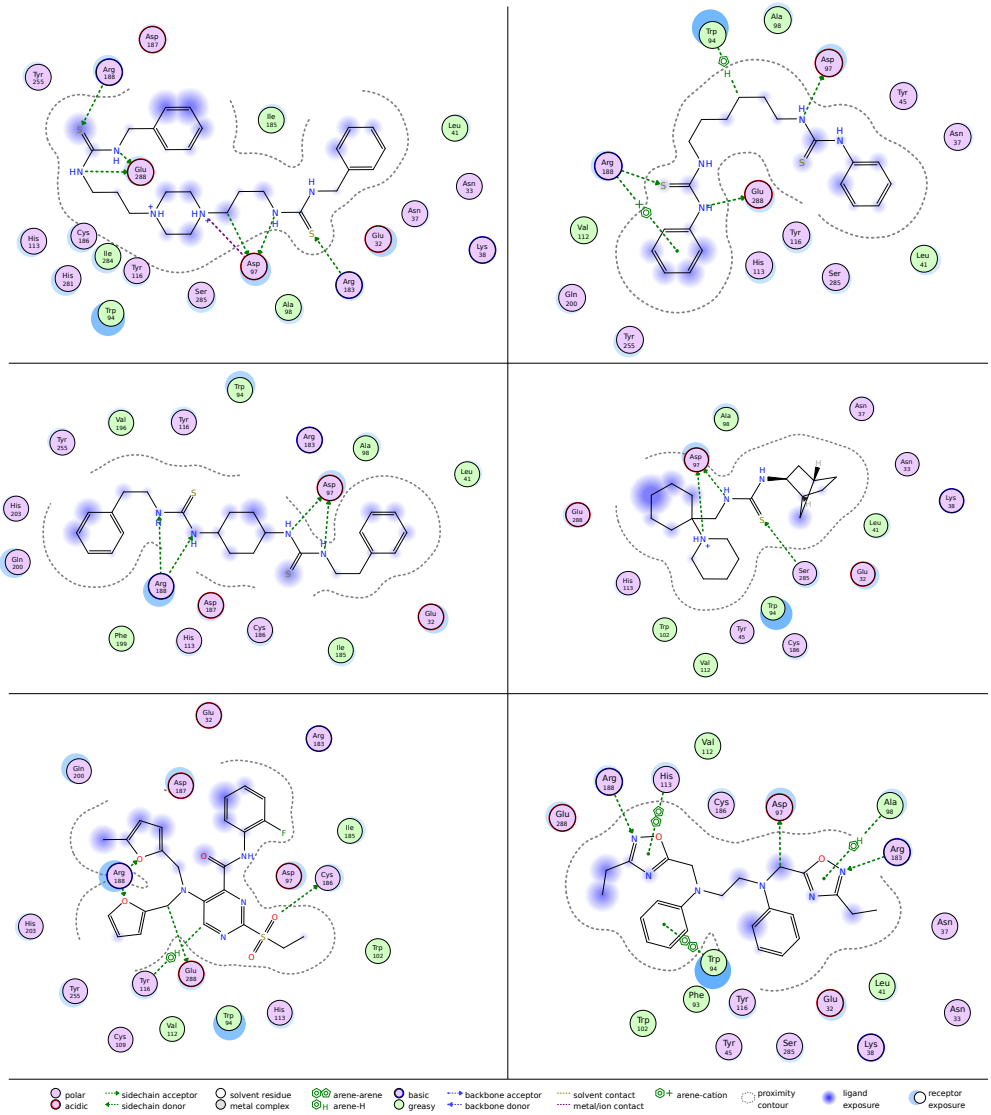



Figure S6: Ligand-receptor interaction diagrams for selected hits generated through GOLD (5.1) docking and visualized in MOE (2011.10).

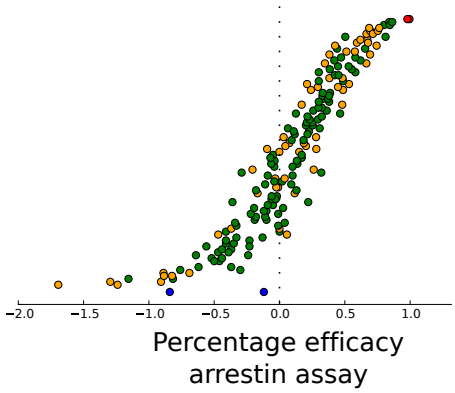



Figure S7: Comparison of observed “percentage efficacy” values in the CXCR4 arrestin assay. The positive control is shown on top in red, an agonist control at 400 nM CXCL12 is shown at the bottom in blue. Screening measurements are shown in pairs as technical duplicates, sorted according to the mean efficacy value. The positioning of the data points in “y” direction is arbitrary and for better visibility only. The dots are colored according to the selection strategy, with the learning iterations in green and the exploitive iterations in orange.

Table S1: Parameter optimization for balanced active learning model on CXCR4 time series data. Compounds were sorted according to publication year and the first 33% served as training data. The remaining 66% were randomly split into learning set and external test set. Active learning was run for 50 iterations. The parameters are the weights of the weighted average for the selection function, balancing the influence of the predicted affinity ($w_1$), the uncertainty about that prediction ($w_2$), and the random forest outlier measure calculated for that compound ($w_3$). After the active learning model was trained, we calculated four different evaluation criteria: (i) the reduction of the mean squared error (MSE) on a randomly selected test set (ii) the number of scaffolds investigated (scaffold count SC) (iii) the average affinity of the picked compounds (AF) (iv) the area under the learning curve (ALC). The selected parameter set and the associated model quality is shown in bold.

| tested parameters | | | model quality | | | |
|---|---|---|---|---|---|---|
| $w_1$ | $w_2$ | $w_3$ | MSE | SC | AF | ALC |
| 0.00 | 0.00 | 1.00 | 1.58 | 90.00 | 7.37 | 0.09 |
| 0.00 | 1.00 | 0.00 | 1.63 | 88.00 | 7.36 | 0.16 |
| 0.00 | 0.71 | 0.71 | 1.62 | 91.00 | 7.37 | 0.18 |
| 0.00 | 0.45 | 0.89 | 1.62 | 91.00 | 7.37 | 0.18 |
| 0.00 | 0.24 | 0.97 | 1.62 | 91.00 | 7.37 | 0.18 |
| 0.00 | 0.89 | 0.45 | 1.63 | 92.00 | 7.33 | 0.17 |
| 0.00 | 0.97 | 0.24 | 1.63 | 91.00 | 7.36 | 0.17 |
| 1.00 | 0.00 | 0.00 | 0.70 | 86.00 | 7.38 | 0.13 |
| 0.71 | 0.00 | 0.71 | 1.04 | 91.00 | 7.34 | 0.13 |
| 0.45 | 0.00 | 0.89 | 1.04 | 91.00 | 7.34 | 0.13 |
| 0.24 | 0.00 | 0.97 | 1.04 | 91.00 | 7.34 | 0.13 |
| 0.71 | 0.71 | 0.00 | 1.64 | 90.00 | 7.38 | 0.17 |
| 0.58 | 0.58 | 0.58 | 1.62 | 91.00 | 7.40 | 0.20 |
| 0.41 | 0.41 | 0.82 | 1.62 | 91.00 | 7.40 | 0.20 |
| 0.24 | 0.24 | 0.94 | 1.62 | 91.00 | 7.40 | 0.20 |
| 0.45 | 0.89 | 0.00 | 1.63 | 92.00 | 7.43 | 0.17 |
| 0.41 | 0.82 | 0.41 | 1.63 | 92.00 | 7.38 | 0.19 |
| 0.33 | 0.67 | 0.67 | 1.62 | 92.00 | 7.41 | 0.19 |
| 0.22 | 0.44 | 0.87 | 1.62 | 92.00 | 7.41 | 0.19 |
| 0.24 | 0.97 | 0.00 | 1.65 | 91.00 | 7.41 | 0.18 |
| 0.24 | 0.94 | 0.24 | 1.63 | 91.00 | 7.40 | 0.21 |
| 0.22 | 0.87 | 0.44 | 1.64 | 91.00 | 7.40 | 0.21 |
| 0.17 | 0.70 | 0.70 | 1.61 | 91.00 | 7.35 | 0.22 |
| 0.89 | 0.00 | 0.45 | 1.04 | 91.00 | 7.34 | 0.13 |
| 0.89 | 0.45 | 0.00 | 1.63 | 90.00 | 7.47 | 0.17 |
| **0.82** | **0.41** | **0.41** | **1.59** | **92.00** | **7.56** | **0.17** |
| 0.67 | 0.33 | 0.67 | 1.59 | 92.00 | 7.56 | 0.17 |
| 0.44 | 0.22 | 0.87 | 1.59 | 92.00 | 7.56 | 0.17 |
| 0.67 | 0.67 | 0.33 | 1.63 | 90.00 | 7.35 | 0.20 |
| 0.44 | 0.87 | 0.22 | 1.66 | 88.00 | 7.37 | 0.18 |
| 0.97 | 0.00 | 0.24 | 1.04 | 91.00 | 7.34 | 0.13 |
| 0.97 | 0.24 | 0.00 | 1.43 | 87.00 | 7.64 | 0.15 |
| 0.94 | 0.24 | 0.24 | 1.49 | 88.00 | 7.50 | 0.15 |
| 0.87 | 0.22 | 0.44 | 1.49 | 88.00 | 7.50 | 0.15 |
| 0.70 | 0.17 | 0.70 | 1.49 | 88.00 | 7.50 | 0.15 |
| 0.87 | 0.44 | 0.22 | 1.59 | 92.00 | 7.56 | 0.17 |
| 0.70 | 0.70 | 0.17 | 1.66 | 91.00 | 7.40 | 0.19 |

Table S2: Comparison of compounds selected with the initial machine learning model trained on ChEMBL19 training data. Compounds were either selected naïvely according to their highest score ("without batch selection") or were re-scored after every selection using the random forest similarity metric ("with batch selection").

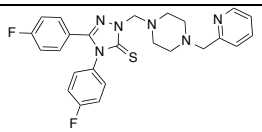
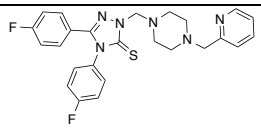
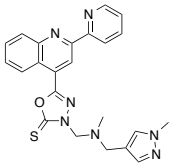
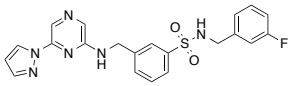
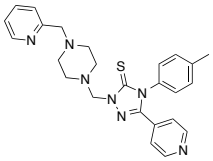
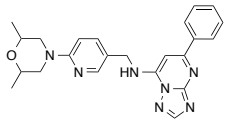
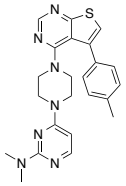
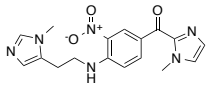
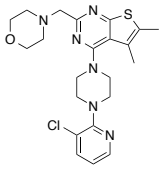
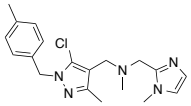
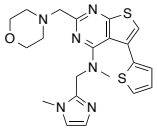
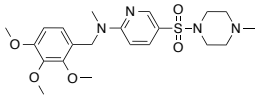
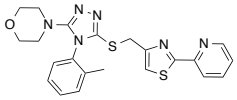
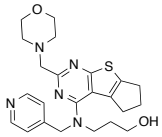
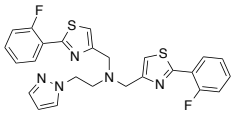
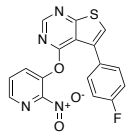
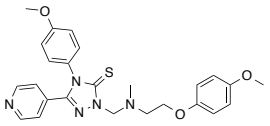
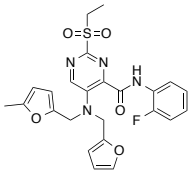
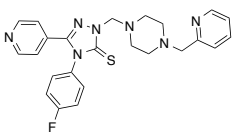
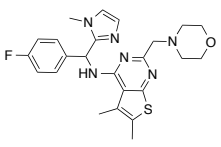
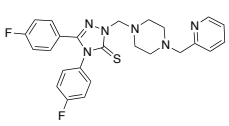
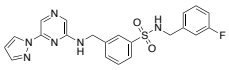
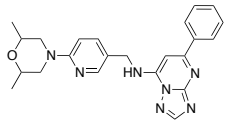
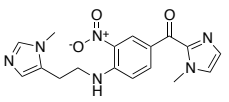
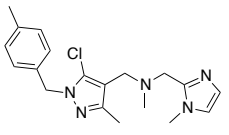
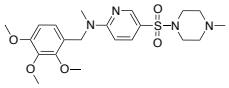
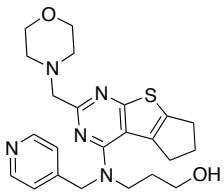
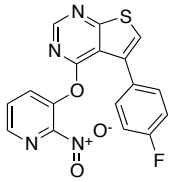
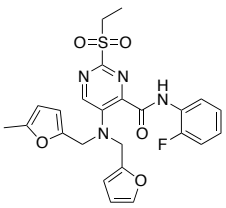
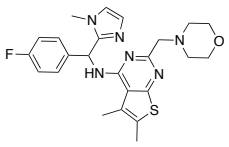
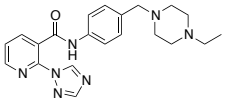
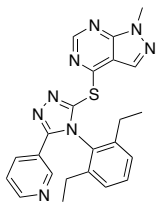
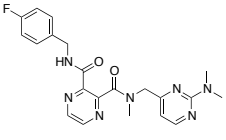
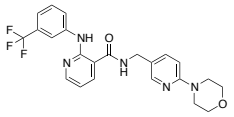
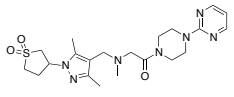
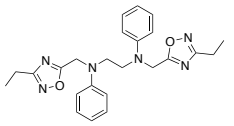
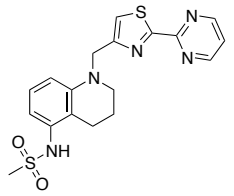
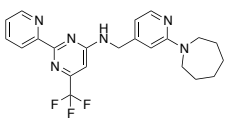
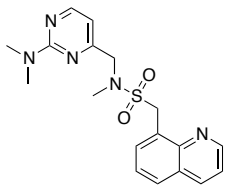
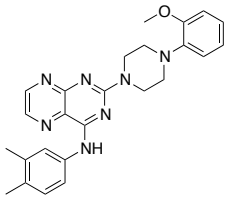
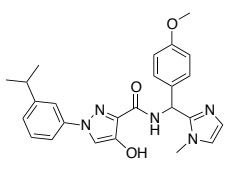
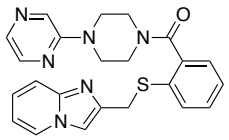
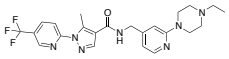
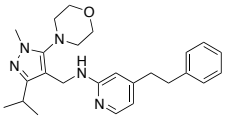
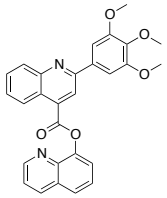
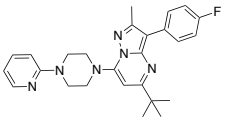
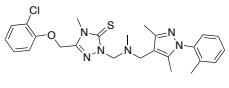
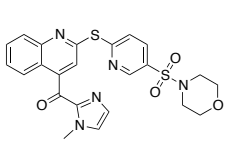
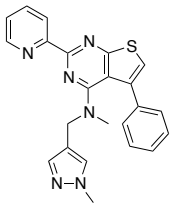
| Rank | without batch selection | with batch selection |
|---|---|---|
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

Table S3: Compounds for the first explorative active learning iteration with measured effect in the CXCR4 arrestin assay. Reported is the Enamine catalogue ID, the chemical structure of the compound, the predicted $pIC_{50}$ for this compound using the initial machine learning model, the predictive uncertainty in terms of the standard deviation of the random forest model, the re-scored random forest similarity during batch selection, the structural similarity towards the initial training data in terms of the Morgan fingerprint (*radius* = 2, 2048 bit, RDKit). Furthermore, we report the initial rank of this compound in the screening list without batch selection. Both replicates of the *in vitro* experiment are given, as well as an approximate $IC_{50}$ value derived from these two measurements using the inverse of the Hill equation.

| Rank | EnamineID | Structure | Predicted $pIC_{50}$ | Uncertainty | RF Similarity | $T_C$(Morgan) | Rank before batch | Inhibition (repl. 1) Arrestin@10µM [% of control] | Inhibition (repl. 2) Arrestin@10µM [% of control] | Approx. $IC_{50}$ [µM] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Z223473328 |  | 6.7 | 1.9 | 50 | 0.28 | 1 | 0.33 | 0.37 | 20 |
| 2 | Z646433752 |  | 6.3 | 2.1 | 85 | 0.22 | 17 | 0.29 | 0.23 | 32 |
| 3 | Z645179942 |  | 6.7 | 1.9 | 81 | 0.21 | 102 | 0.36 | 0.26 | 25 |
| 4 | Z1661113969 |  | 6.5 | 2.0 | 69 | 0.16 | 69 | 0.06 | 0.31 | 40 |
| 5 | Z1632770703 |  | 6.5 | 1.9 | 63 | 0.34 | 136 | 0.32 | 0.30 | 25 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Z31133182 |  | 6.2 | 2.1 | 97 | 0.20 | 219 | 0.38 | 0.30 | 20 |
| 7 | Z1329095597 |  | 6.8 | 1.9 | 79 | 0.21 | 88 | 0.32 | 0.21 | 25 |
| 8 | Z31112468 |  | 6.4 | 2.0 | 105 | 0.14 | 324 | 0.42 | 0.45 | 13 |
| 9 | Z1213699875 |  | 6.1 | 2.1 | 88 | 0.17 | 1411 | 0.84 | 0.86 | **2** |
| 10 | Z185386180 |  | 6.5 | 2.0 | 121 | 0.24 | 32 | 0.33 | 0.38 | 20 |
| 11 | Z434893686 |  | 6.5 | 2.0 | 115 | 0.26 | 20 | -0.06 | -0.03 | 1000 |
| 12 | Z229071932 |  | 6.3 | 2.0 | 77 | 0.15 | 1391 | 0.39 | 0.33 | 20 |
| 13 | Z1614492513 |  | 6.7 | 2.0 | 131 | 0.24 | 74 | 0.09 | 0.18 | 63 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | Z236278906 |  | 6.6 | 2.0 | 116 | 0.23 | 834 | 0.66 | 0.44 | **8** |
| 15 | Z106710628 |  | 6.4 | 2.0 | 103 | 0.23 | 41 | 0.14 | -0.06 | 251 |
| 16 | Z1318029062 |  | 6.4 | 1.9 | 84 | 0.20 | 542 | 0.80 | 0.84 | **2** |
| 17 | Z649280220 |  | 6.3 | 1.9 | 75 | 0.22 | 708 | 0.49 | 0.30 | 16 |
| 18 | Z1155806433 |  | 6.5 | 1.9 | 105 | 0.31 | 188 | 0.58 | 0.30 | 13 |
| 19 | Z1587990692 |  | 6.3 | 2.0 | 119 | 0.23 | 487 | 0.47 | 0.58 | **10** |
| 20 | Z1041113954 |  | 6.5 | 2.0 | 119 | 0.17 | 24 | 0.81 | 0.50 | 5 |
| 21 | Z1262346502 |  | 6.3 | 1.9 | 82 | 0.15 | 712 | 0.13 | 0.46 | 25 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 22 | Z317763120 |  | 6.6 | 1.9 | 121 | 0.21 | 16 | 0.23 | 0.11 | 50 |
| 23 | Z955361072 |  | 6.4 | 2.0 | 136 | 0.26 | 315 | -0.04 | 0.13 | 200 |
| 24 | Z1625062616 |  | 6.5 | 1.9 | 78 | 0.27 | 604 | 0.25 | 0.20 | 32 |
| 25 | Z28175954 |  | 6.4 | 2.0 | 110 | 0.22 | 486 | -0.82 | -1.16 | 1000 |
| 26 | Z1172208304 |  | 6.5 | 1.9 | 110 | 0.18 | 40 | 0.62 | 0.42 | **10** |
| 27 | Z872486250 |  | 6.5 | 2.0 | 117 | 0.18 | 108 | 0.03 | -0.19 | 1000 |
| 28 | Z127624426 |  | 6.0 | 2.1 | 128 | 0.18 | 1978 | 0.12 | 0.21 | 50 |
| 29 | Z68342921 |  | 6.7 | 1.9 | 140 | 0.20 | 23 | 0.45 | 0.53 | **10** |

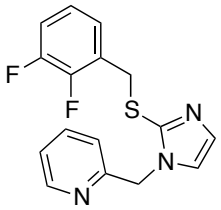
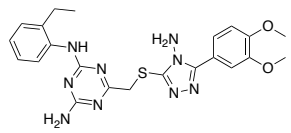
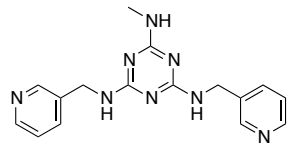
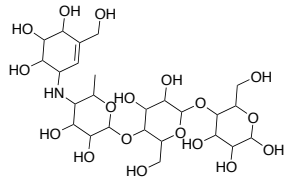
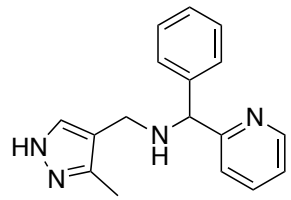
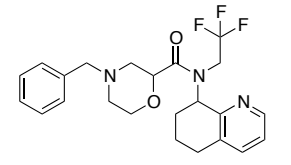
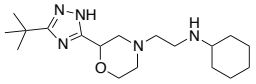
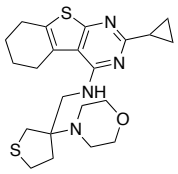
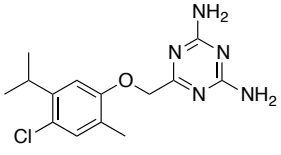
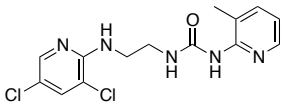
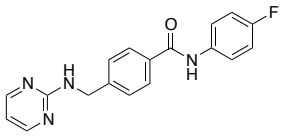
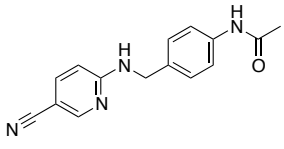
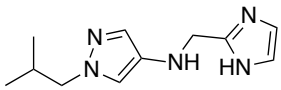
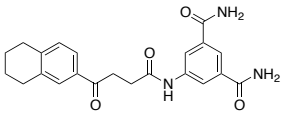
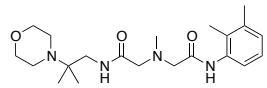
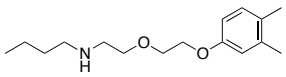
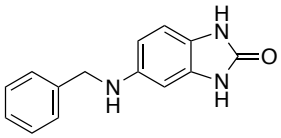
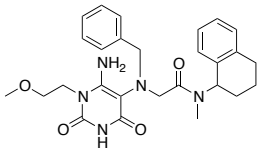
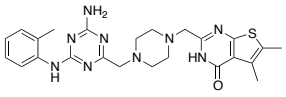
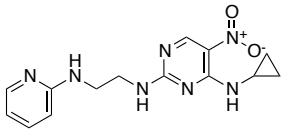
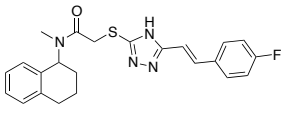
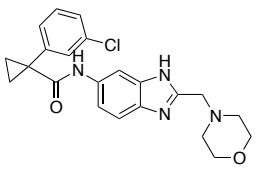
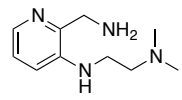
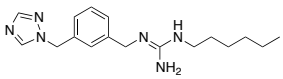
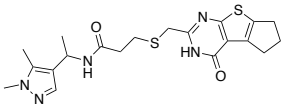
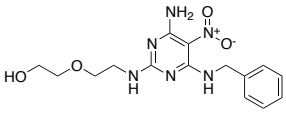
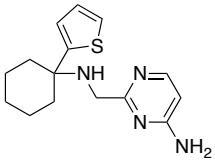
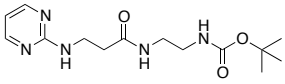
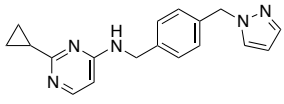
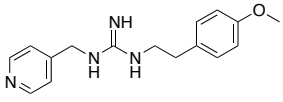
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | Z1190352304 |  | 6.4 | 2.0 | 133 | 0.23 | 1355 | 0.55 | 0.38 | 13 |

Table S4: Compounds for the second explorative active learning iteration with measured effect in the CXCR4 arrestin and the CXCR4 cAMP assay. Reported is the Enamine catalogue ID, the chemical structure of the compound, the predicted $pIC_{50}$ for this compound using the second machine learning model, the predictive uncertainty in terms of the standard deviation of the second random forest model, the re-scored random forest similarity during batch selection, the maximal structural similarity towards the initial training data and the first iteration in terms of the Morgan fingerprint (*radius* = 2, 2048 bit, RDKit). Furthermore, we report the initial rank of this compound in the screening list without batch selection. Both replicates of the two *in vitro* experiment are given, as well as an approximate $IC_{50}$ value derived from the two measurement from the cAMP assay using the inverse of the Hill equation.

| Rank | EnamineID | Structure | Predicted $pIC_{50}$ | Uncertainty | RF Similarity | $T_C$(ECFP4) | Rank before batch | Arrestin @10 µM (repl. 1 & repl. 2) [% of control] | cAMP @ 30 µM (repl. 1 & repl. 2) [% of control] | Approx. $IC_{50}$ [µM] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Z25375413 |  | 6.3 | 1.6 | 32 | 0.23 | 1 | 0.17<br>-0.07 | 0.11<br>0.08 | 100 |
| 2 | Z1410217337 |  | 6.1 | 1.8 | 95 | 0.30 | 2 | 0.02<br>-0.04 | 0.48<br>0.37 | 13 |
| 3 | Z1558515434 |  | 5.8 | 1.7 | 27 | 0.10 | 5 | 0.13<br>-0.18 | 0.14<br>0.11 | 79 |
| 4 | Z728822720 |  | 6.2 | 1.8 | 123 | 0.29 | 7 | 0.22<br>-0.36 | 0.11<br>0.03 | 126 |
| 5 | Z1592230569 |  | 5.8 | 1.7 | 35 | 0.38 | 9 | -0.41<br>-0.00 | 0.01<br>0.10 | 158 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Z1268953925 |  | 5.8 | 1.7 | 31 | 0.20 | 28 | -0.05<br>0.17 | 0.24<br>0.18 | 40 |
| 7 | Z951360858 |  | 5.8 | 1.7 | 39 | 0.28 | 27 | 0.25<br>0.21 | 0.19<br>0.11 | 63 |
| 8 | Z107201840 |  | 6.2 | 1.6 | 45 | 0.17 | 16 | -0.23<br>-0.41 | -0.14<br>-0.11 | 1000 |
| 9 | Z513427274 |  | 6.0 | 1.7 | 71 | 0.22 | 33 | -0.10<br>-0.12 | -0.12<br>-0.13 | 1000 |
| 10 | Z1419914482 |  | 6.2 | 1.8 | 165 | 0.36 | 35 | -0.09<br>0.10 | -0.15<br>-0.12 | 1000 |
| 11 | Z246519464 |  | 6.0 | 1.8 | 124 | 0.34 | 43 | -0.02<br>-0.06 | 0.06<br>-0.01 | 316 |
| 12 | Z1537364795 |  | 6.0 | 1.7 | 81 | 0.22 | 13 | -0.22<br>-0.01 | 0.11<br>0.11 | 79 |
| 13 | Z33485836 |  | 6.1 | 1.6 | 31 | 0.23 | 56 | 0.32<br>-0.29 | 0.00<br>0.06 | 316 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | Z105938746 |  | 5.7 | 1.7 | 47 | 0.22 | 20 | -0.64<br>-0.24 | 0.13<br>0.07 | 100 |
| 15 | Z237480070 |  | 5.9 | 1.7 | 74 | 0.21 | 170 | -0.18<br>-0.11 | 0.03<br>0.09 | 158 |
| 16 | Z138720204 |  | 6.2 | 1.8 | 175 | 0.29 | 200 | -0.13<br>-0.33 | 0.01<br>0.01 | 631 |
| 17 | Z354712800 |  | 6.0 | 1.6 | 36 | 0.27 | 90 | -0.37<br>-0.62 | -0.01<br>-0.01 | 1000 |
| 18 | Z53006412 |  | 5.9 | 1.7 | 65 | 0.29 | 6 | -0.34<br>0.04 | -0.13<br>-0.13 | 1000 |
| 19 | Z354279164 |  | 5.9 | 1.7 | 76 | 0.32 | 85 | -0.19<br>-0.10 | -0.09<br>-0.06 | 1000 |
| 20 | Z228563452 |  | 6.1 | 1.6 | 35 | 0.28 | 228 | -0.33<br>-0.01 | -0.14<br>-0.22 | 1000 |
| 21 | Z955589808 |  | 5.8 | 1.7 | 41 | 0.24 | 177 | -0.04<br>0.10 | 0.07<br>0.17 | 79 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 22 | Z1463975251 |  | 6.1 | 1.7 | 107 | 0.30 | 281 | 0.09<br>-0.11 | 0.14<br>0.17 | 50 |
| 23 | Z1127455297 |  | 6.0 | 1.6 | 47 | 0.22 | 116 | -0.05<br>-0.09 | 0.17<br>0.23 | 40 |
| 24 | Z359434674 |  | 5.6 | 1.7 | 43 | 0.22 | 147 | -0.35<br>-0.41 | 0.02<br>0.01 | 631 |
| 25 | Z295462974 |  | 6.1 | 1.6 | 81 | 0.22 | 12 | -0.30<br>-0.76 | 0.27<br>0.25 | 32 |
| 26 | Z1492840408 |  | 6.1 | 1.6 | 78 | 0.20 | 51 | -0.52<br>-0.44 | 0.16<br>0.15 | 50 |
| 27 | Z899987010 |  | 5.8 | 1.7 | 81 | 0.20 | 148 | -0.47<br>-0.39 | 0.21<br>0.12 | 50 |
| 28 | Z1172365327 |  | 5.9 | 1.7 | 80 | 0.35 | 681 | -0.46<br>-0.50 | -0.05<br>-0.12 | 1000 |
| 29 | Z991127312 |  | 6.1 | 1.7 | 107 | 0.23 | 195 | -0.56<br>-0.23 | 0.07<br>0.08 | 126 |

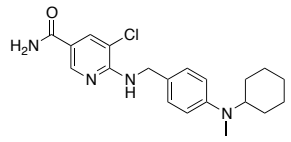
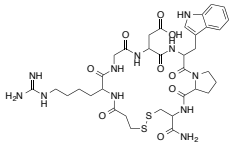
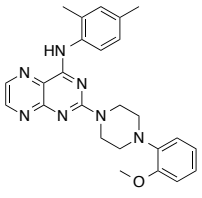
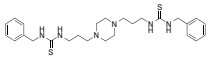
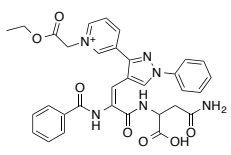
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | Z900157686 |  | 6.0 | 1.6 | 64 | 0.24 | 41 | -0.37<br>-0.44 | -0.05<br>0.03 | 1000 |

Table S5: Compounds selected in the exploitive, greedy iteration. Reported is the Enamine catalogue ID, the chemical structure of the compound, the predicted $pIC_{50}$ for this compound using the second machine learning model, the predictive uncertainty in terms of the standard deviation of the second random forest model, and their difference which served as the selection criteria. We also report the maximal structural similarity towards the initial training data and the first and second iteration in terms of the Morgan fingerprint (*radius* = 2, 2048 bit, RDKit). Furthermore, we report the rank of this compound in hypothetical screening list generated with the same scoring function but using a model trained only with the initial ChEMBL data to monitor the impact of using the actively trained model. Both replicates of the *in vitro* experiment in the arrestin assay are given, as well as an approximate $IC_{50}$ value derived from the two measurement using the inverse of the Hill equation

| Rank | EnamineID | Structure | Predicted $pIC_{50}$ | Uncertainty | $pIC_{50}$ - Uncertainty | $T_C$(ECFP4) | Initial rank (model 0) | Inhibition (repl. 1) Arrestin @ 30 µM [% of control] | Inhibition (repl. 2) Arrestin @ 30 µM [% of control] | Approx. $IC_{50}$ [µM] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Z1558506262 |  | 6.1 | 1.4 | 4.7 | 0.44 | 8802 | -0.01 | -0.03 | 1000 |
| 2 | Z1041113924 |  | 5.1 | 1.0 | 4.1 | 0.73 | 66703 | 0.66 | 0.35 | 32 |
| 3 | Z45766764 |  | 5.5 | 1.5 | 4.0 | 0.30 | 5151 | 0.59 | 0.68 | 20 |
| 4 | Z56790850 |  | 5.5 | 1.5 | 4.0 | 0.18 | 19015 | -0.37 | 0.00 | 1000 |

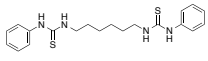
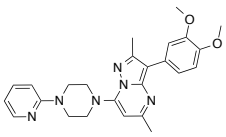
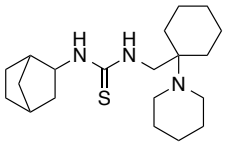
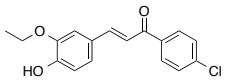
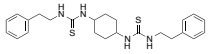
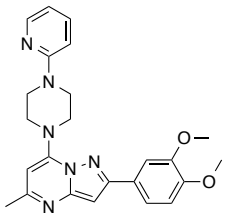
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Z45801934 |  | 5.4 | 1.5 | 4.0 | 0.51 | 25592 | 0.29 | 0.45 | 50 |
| 6 | Z1172231060 |  | 4.9 | 1.0 | 3.9 | 0.59 | 21257 | -1.24 | -1.69 | Ago |
| 7 | Z45904687 |  | 5.5 | 1.6 | 3.9 | 0.21 | 263368 | 0.74 | 0.43 | 20 |
| 8 | Z46033340 |  | 5.5 | 1.6 | 3.9 | 0.75 | 10742 | 0.28 | 0.03 | 158 |
| 9 | Z45831362 |  | 5.4 | 1.6 | 3.9 | 0.35 | 30040 | -1.30 | -0.91 | Ago |
| 10 | Z432102094 |  | 4.9 | 1.0 | 3.9 | 0.44 | 983 | -0.82 | -0.88 | Ago |

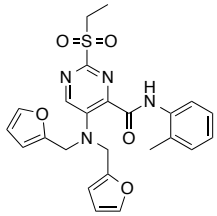
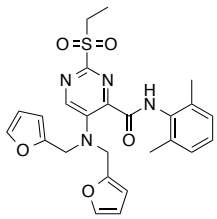
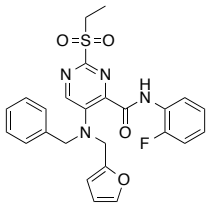
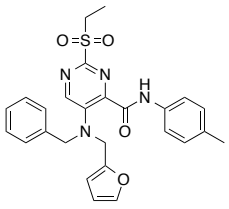
Table S6: Top 15 known CXCR4 antagonists used for the prediction of the most potent identified thiourea compound **10** with their respective compound class annotation. Compounds were identified by determining the leaf that compound 10 is predicted with for every tree of the final random forest model and then determine the training examples that are predicted with the same leaf. The column "# used" indicates the number of trees for which the reported reference was perceived equivalent to compound **10** by being predicted with the same leaf.

| Reference | # used | compound class |
|---|---|---|
| CHEMBL478168 | 45 | Cyclam |
| CHEMBL2170444 | 45 | Diamine |
| CHEMBL2170443 | 45 | Diamine |
| CHEMBL2170299 | 45 | Diamine |
| CHEMBL545532 | 31 | Cyclam |
| CHEMBL2347623 | 30 | Guanidine |
| CHEMBL2347624 | 28 | Guanidine |
| CHEMBL2347627 | 26 | Guanidine |
| CHEMBL543895 | 24 | Cyclam |
| CHEMBL2347626 | 22 | Guanidine |
| CHEMBL1202231 | 22 | Cyclam |
| CHEMBL477121 | 21 | Cyclam |
| CHEMBL516480 | 21 | Isothiourea |
| CHEMBL2347631 | 19 | Guanidine |
| CHEMBL2347625 | 18 | Guanidine |

Table S7: Origin of known ligands used in the prediction of the 10 compounds selected for the greedy iteration. Compounds were identified by determining the leaf that a prospective compound is predicted with for every tree and then determine the training examples that are predicted with the same leaf. For these training examples we determined their origin (original ChEMBL19 data, first active learning iteration, second active learning iteration) and counted them. Note that the number is higher than the number of trees because the out-of-bag examples are also considered. We report the absolute number of times that any reference compound from a specific source was used for the tree-based prediction as well as the relative number normalized by the number of compounds contained in a specific source.

| | **Original** | | **Iteration 1** | | **Iteration 2** | |
|---|---|---|---|---|---|---|
| **Compound** | **absolute** | **relative** | **absolute** | **relative** | **absolute** | **relative** |
| **Z1041113924** | 329 | 1.15 | 786 | **26.20** | 122 | 4.07 |
| **Z1172231060** | 461 | 1.61 | 747 | **24.90** | 123 | 4.10 |
| **Z1558506262** | 1311 | **4.58** | 14 | 0.47 | 52 | 1.73 |
| **Z432102094** | 423 | 1.48 | 759 | **25.30** | 128 | 4.27 |
| **Z45766764** | 1062 | 3.71 | 102 | 3.40 | 131 | **4.37** |
| **Z45801934** | 926 | 3.24 | 91 | 3.03 | 279 | **9.30** |
| **Z45831362** | 1019 | 3.56 | 109 | 3.63 | 189 | **6.30** |
| **Z45904687** | 970 | 3.39 | 66 | 2.20 | 194 | **6.47** |
| **Z46033340** | 729 | 2.55 | 141 | 4.70 | 372 | **12.40** |
| **Z56790850** | 1100 | 3.85 | 142 | 4.73 | 164 | **5.47** |

Table S8: Compounds selected in the hit expansion for compound **1**. Reported is the Enamine catalogue ID, the chemical structure of the compound, the predicted $pIC_{50}$ for this compound using the third machine learning model, the predictive uncertainty in terms of the standard deviation of the second random forest model, and the random forest similarity to compound **1** that served as selection criteria. We also report the maximal structural similarity towards the original training data and the first and second iteration in terms of the Morgan fingerprint (*radius* = 2, 2048 bit, RDKit). Furthermore, we report the rank of this compound in hypothetical screening list generated with the greedy function using a model trained only with the initial ChEMBL data to monitor the impact of using the hit expansion approach. Both replicates of the *in vitro* experiment in the arrestin assay are given, as well as an approximate $pIC_{50}$ value derived from the two measurement using the inverse of the Hill equation

| Rank | EnamineID | Structure | RF similarity **1** | Predicted $pIC_{50}$ | Uncertainty | $T_C$(ECFP4) | Initial rank (model 0) | Inhibition (repl. 1) Arrestin@30µM [% of control] | Inhibition (repl. 2) Arrestin@30µM [% of control] | Approx. $IC_{50}$ [µM] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Z1213699877 |  | 346 | 4.5 | 1.3 | 0.75 | 561047 | -0.89 | -0.69 | Ago |
| 2 | Z1213699847 |  | 333 | 4.5 | 1.3 | 0.67 | 790378 | 0.67 | 0.72 | 13 |
| 3 | Z1213699860 |  | 318 | 4.5 | 1.4 | 0.81 | 592996 | 0.49 | 0.24 | 50 |
| 4 | Z1213679310 |  | 250 | 4.3 | 1.3 | 0.64 | 819887 | 0.51 | 0.58 | 25 |

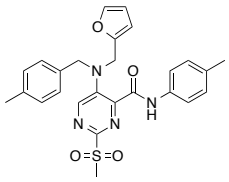
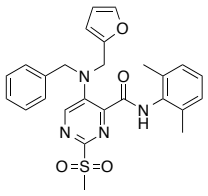
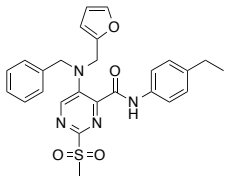
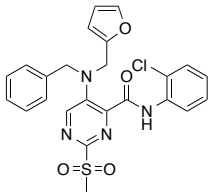
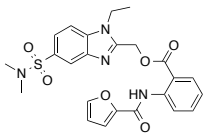
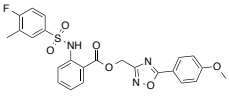
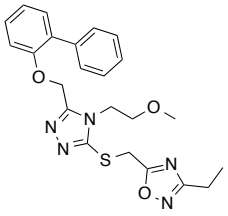
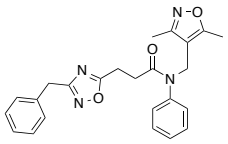
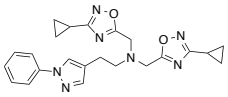
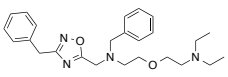
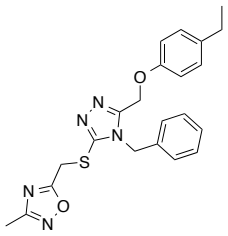
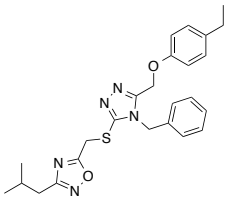
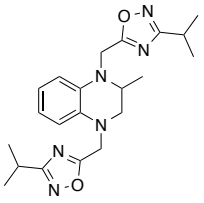
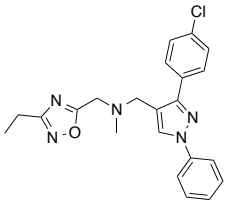
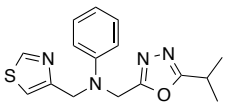
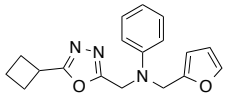
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Z1213680034 |  | 216 | 4.3 | 1.2 | 0.54 | 910377 | 0.38 | 0.69 | 25 |
| 6 | Z1213681397 |  | 209 | 4.4 | 1.3 | 0.54 | 966671 | 0.28 | -0.09 | 316 |
| 7 | Z1213679311 |  | 207 | 4.5 | 1.4 | 0.54 | 552642 | 0.69 | 0.76 | 13 |
| 8 | Z1213680017 |  | 204 | 4.6 | 1.4 | 0.58 | 535945 | 0.38 | 0.48 | 40 |
| 9 | Z54981321 |  | 150 | 4.5 | 1.2 | 0.27 | 1356103 | 0.17 | 0.48 | 63 |
| 10 | Z285623594 |  | 144 | 4.4 | 1.1 | 0.21 | 1395094 | N.T. | N.T. | N.T. |

Table S9: Compounds selected in the hit expansion for compound **2**. Reported is the Enamine catalogue ID, the chemical structure of the compound, the predicted $pIC_{50}$ for this compound using the third machine learning model, the predictive uncertainty in terms of the standard deviation of the third random forest model, and the random forest similarity to compound **2** that served as selection criteria. We also report the maximal structural similarity towards the original training data and the first and second iteration in terms of the Morgan fingerprint (*radius* = 2, 2048 bit, RDKit). Furthermore, we report the rank of this compound in hypothetical screening list generated with the greedy function using a model trained only with the initial ChEMBL data to monitor the impact of using the hit expansion approach. Both replicates of the *in vitro* experiment in the arrestin assay are given, as well as an approximate $pIC_{50}$ value derived from the two measurement using the inverse of the Hill equation.

| Rank | EnamineID | Structure | RF similarity **2** | Predicted $pIC_{50}$ | Uncertainty | $T_C$(ECFP4) | Initial rank (model 0) | Inhibition (repl. 1) Arrestin @ 30 µM [% of control] | Inhibition (repl. 2) Arrestin @ 30 µM [% of control] | Approx. $IC_{50}$ [µM] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Z270217062 |  | 126 | 5.0 | 1.5 | 0.31 | 182794 | 0.26 | -0.21 | 1000 |
| 2 | Z442860018 |  | 125 | 4.5 | 1.4 | 0.35 | 810616 | 0.20 | 0.04 | 200 |
| 3 | Z957336016 |  | 119 | 4.9 | 1.4 | 0.29 | 13772 | 0.07 | 0.19 | 200 |
| 4 | Z1653693772 |  | 117 | 5.2 | 1.6 | 0.43 | 39283 | 0.15 | 0.00 | 398 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Z237664514 |  | 116 | 4.5 | 1.2 | 0.24 | 751272 | 0.53 | 0.21 | 50 |
| 6 | Z243468286 |  | 115 | 4.8 | 1.3 | 0.28 | 827982 | 0.06 | -0.47 | 1000 |
| 7 | Z1318138976 |  | 111 | 4.9 | 1.3 | 0.23 | 101304 | 0.69 | 0.75 | 13 |
| 8 | Z267133972 |  | 103 | 4.8 | 1.3 | 0.42 | 82611 | 0.67 | 0.62 | 16 |
| 9 | Z1171389532 |  | 102 | 4.6 | 1.3 | 0.31 | 113941 | 0.12 | -0.17 | 1000 |
| 10 | Z1225553502 |  | 102 | 4.6 | 1.4 | 0.29 | 132241 | -0.03 | 0.04 | 1000 |