Performance analysis of Mondrian Conformal Prediction for the top 10 targets in the ExCAPE dataset

Khuong An Nguyen ME@KHUONG.UK

Abstract

This report summaries the performance of Mondrian Inductive Conformal Prediction on the top 10 largest targets in the ExCAPE dataset.

1. Overview of all 526 targets

Out of 526 targets in the ExCAPE database, only the top 150 have more than 100,000 compounds. The number of compounds drops below 5,000 beyond the top 200 (see Figure 1). From a machine learning perspective, all targets beyond this threshold would not be of interest, as we do not have enough training examples.

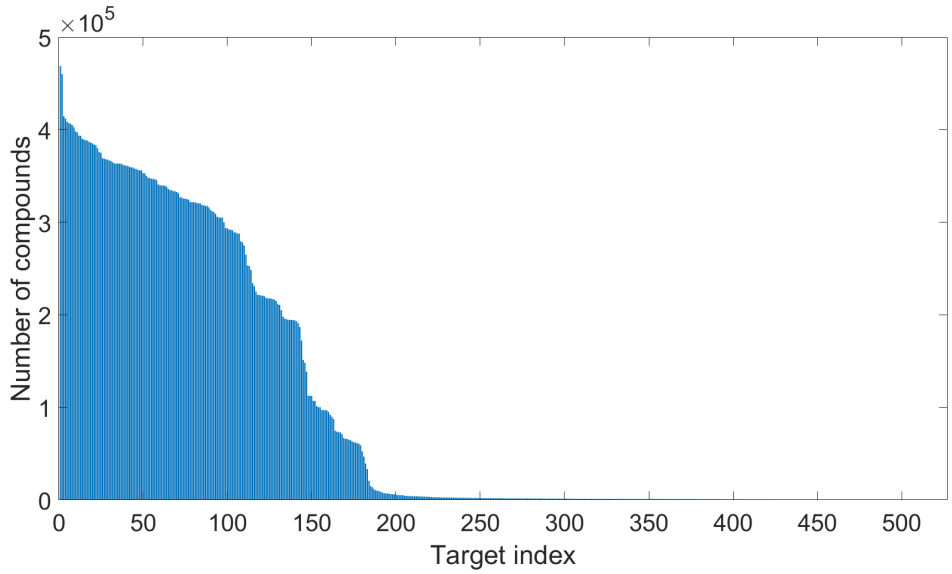


Figure 1: Number of compounds across all 526 targets. Only the top 150 have more than 100,000 compounds.

Having a closer look at each target, another interesting feature is the high imbalance between the active and inactive classes (see Table 1). The majority of targets have just around 1% of active compounds.

Table 1: Top 10 targets with most compounds. Targets with more than 5% of active compounds are highlighted in bold.

Target	Total compounds	Active compounds	% Active
IDH1	468,798	6,091	1.3%
CFTR	459,914	1,166	0.25%
NFE2L2	414,510	28,230	**6.8%**
GLP1R	412,076	33,311	**8.1%**
DRD3	408,536	3,381	0.83%
SMAD3	407,094	21,995	**5.4%**
APOBEC3G	406,341	4,040	1%
GLS	404,997	2,870	0.71%
TARDBP	402,570	10,765	2.67%
PTH1R	398,103	5,738	1.44%

2. Mondrian Inductive Conformal Prediction

The reason Mondrian Inductive Conformal Prediction (MICP) is used in this report, is because of the highly imbalanced nature of the dataset. Standard Conformal Prediction does not guarantee validity amongst classes. This means the error rate for the active class may be much higher in the end, which is compensated by much lower error rate for the inactive class.

In principal, MCIP splits the dataset into categories and set a separate significance level for each category. When calculating the p-values, we have to adjust the condition to pick the conditional label as follows. A more detailed treatment of MCIP can be found in our first deliverable.

$$p(y) = \frac{|\{i = h+1, \ldots, l+1 : y_i = y, \alpha_i \geq \alpha_{l+1}\}|}{|\{i = h+1, \ldots, l+1 : y_i = y\}|} \tag{1}$$

3. Performance results

In this section, we present the results of applying MICP for the top 10 largest targets in the ExCAPE dataset.

3.1. Testing environments

All experiments were carried out on the IT4I supercomputer located in the Czech Republic. The Salomon cluster consists of 1,008 compute nodes, with 129 TB RAM. Each node is a computer with 24 cores and 128 GB RAM. We used Python 3 to analyse the dataset.

3.2. Underlying algorithms

To analyse the performance of MICP, we will use Linear Support Vector Classification, Gradient Boosted Tree, and k-Nearest Neighbours as underlying algorithms, all provided as standard packages under Python. We then combine the p-values of individual algorithms (i.e. by taking the minimum p-values, maximum, mean, etc.) as described in another paper, giving a total of 11 separate algorithms performed on each target. Their results will be analysed below.

3.3. Comparison of algorithms performance across the top 10 targets

Efficiency-wise, we rank the 11 algorithms based on the F1 score, across the top 10 targets (see Figures 2 and 3). F1 Score is the weighted average of Precision and Recall, in which 1 represents perfect precision and recall and 0 otherwise. It takes both false positives and false negatives into account. F1 score is useful than just accuracy, especially for imbalanced class distribution. The following conclusions are drawn:

- F1 score when predicting active compounds is low, less than 0.1, for highly imbalanced dataset (e.g. CFTR - 0.25% of active, GLS - 0.71% of active).
- F1 score when predicting inactive compounds is always higher than active ones, because of having more training examples.

- Fisher and combination of p-values using the mean performed better overall when predicting active compounds.
- Fisher performed better overall when predicting inactive compounds.
- The average precision of Fisher was the highest amongst all algorithms.

Validity-wise, except Fisher and combination of p-values by taking the minimum, MICP remains valid for all algorithms.

3.4. Detailed results for CFTR target

With 459,914 compounds, CFTR is the second biggest target. However, it is one of the most imbalanced targets in the top 30, with just 0.25% (1,166) active compounds.

The confusion matrices, F1 score for precise predictions, and ranking precision for CFTR are described below in Tables 2, 3 and 4. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher also returned the highest F1 score when predicting inactive compounds. Whereas, Fisher ECDF and combination of p-values by taking the mean were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.5. Detailed results for NFE2L2 target

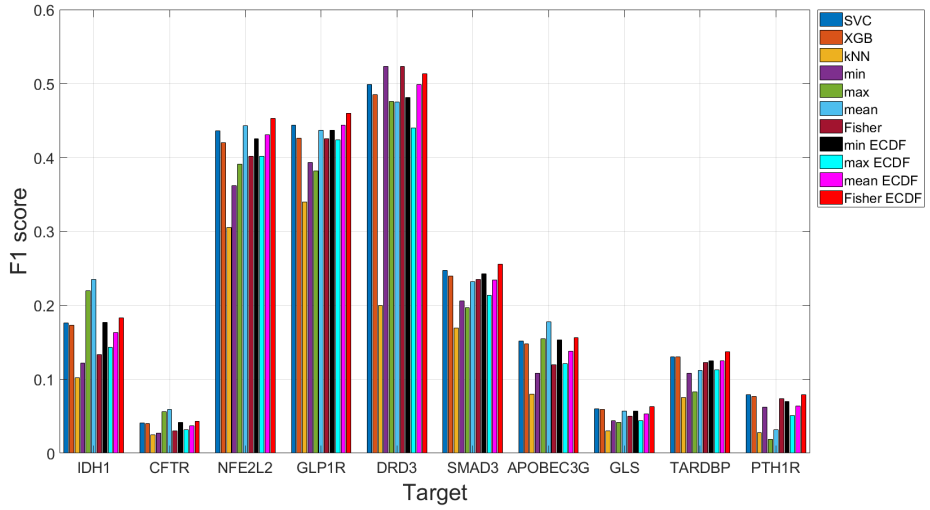
NFE2L2 is the third biggest target, and has one of highest number of active compounds at 6.8% (28,230).

The confusion matrices, F1 score for precise predictions, and ranking precision for NFE2L2 are described below in Tables 5, 6 and 7. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Combination of p-values by taking minimum returned the highest F1 score when predicting inactive compounds. Whereas, Fisher and Fisher ECDF were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

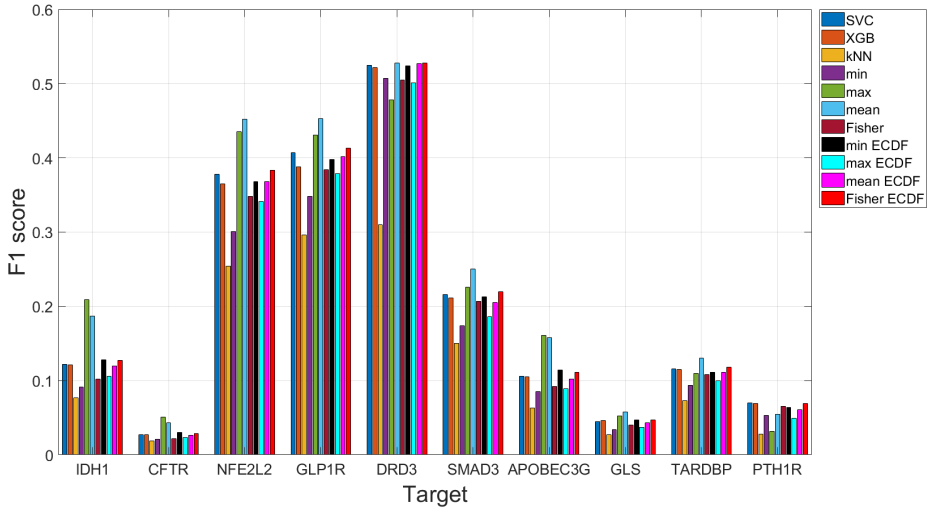
3.6. Detailed results for GLP1R target

GLP1R is the fourth biggest target, and has one of highest number of active compounds at 8.1% (33,311).

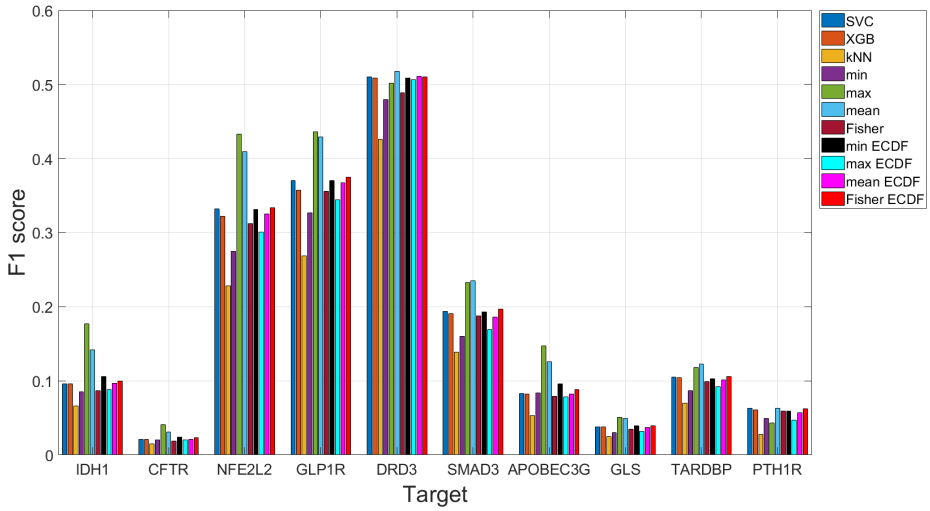
The confusion matrices, F1 score for precise predictions, and ranking precision for GLP1R are described below in Tables 8, 9 and 10. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher returned the highest F1 score when predicting inactive compounds. Whereas, Fisher ECDF and combination of p-values by taking the mean were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.



(a) $\varepsilon = 0.05$

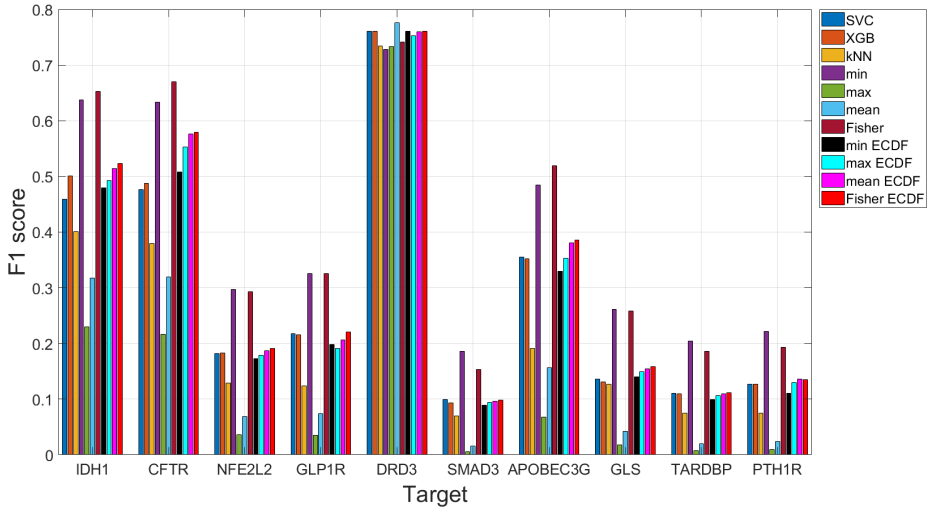


(b) $\varepsilon = 0.1$

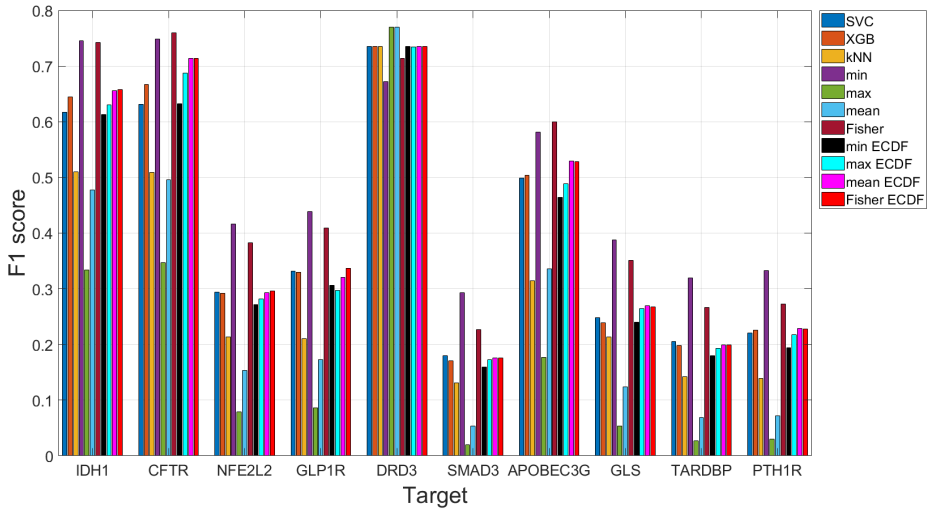


(c) $\varepsilon = 0.15$

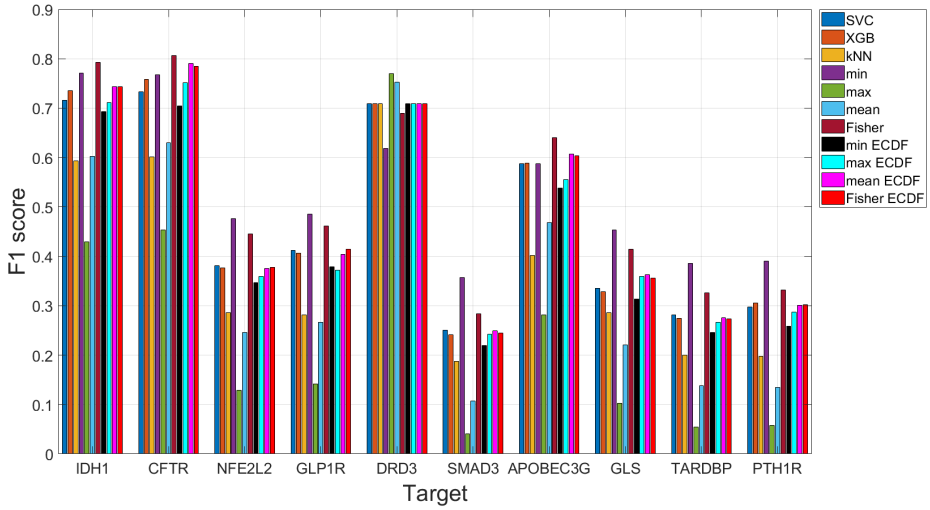
Figure 2: Algorithms ranking for Active class



(a) $\varepsilon = 0.05$



(b) $\varepsilon = 0.1$



(c) $\varepsilon = 0.15$

Figure 3: Algorithms ranking for Inactive class

Table 2: Prediction error rate for CFTR target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 159,914, in which only about 405 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	146.70	3.22	15744.74	1584.40	0.00	142434.94	9.928e-03
XGB	136.54	2.88	12981.80	1585.10	0.00	145207.68	9.930e-03
kNN	70.68	3.94	13378.94	1595.84	0.00	144864.60	1.000e-02
min	174.94	7.54	29109.68	3753.78	59.72	126808.34	2.389e-02
max	53.50	0.34	2547.60	86.92	0.00	157225.64	5.457e-04
mean	62.24	0.94	6423.06	148.04	0.00	153279.72	9.316e-04
Fisher	203.56	13.18	51320.10	5464.28	0.68	102912.20	3.426e-02
min ECDF	137.48	6.92	26099.18	1574.98	23.64	132071.80	1.004e-02
max ECDF	104.48	4.60	23411.06	1592.70	0.00	134801.16	9.988e-03
mean ECDF	114.76	4.60	25087.44	1592.64	0.00	133114.56	9.988e-03
Fisher ECDF	146.92	5.70	29114.72	1597.06	0.00	129049.60	1.002e-02

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	211.58	18.88	52449.84	7982.60	0.00	99251.10	0.050
XGB	206.50	19.08	54310.02	8002.42	0.00	97375.98	0.050
kNN	126.22	22.22	39227.12	7944.74	0.00	112593.70	0.050
min	257.06	32.42	77529.84	16292.38	2099.78	63702.52	0.115
max	80.80	3.66	20382.98	615.50	0.00	138831.06	0.004
mean	114.08	6.20	31705.28	1557.30	0.00	126531.14	0.010
Fisher	263.92	33.08	84069.18	15185.48	72.18	60290.16	0.096
min ECDF	208.44	19.32	57017.10	7616.28	395.84	94657.02	0.050
max ECDF	167.30	20.80	64052.06	7986.46	0.00	87687.38	0.050
mean ECDF	190.64	21.08	67878.28	7997.52	0.00	83826.48	0.050
Fisher ECDF	223.04	21.88	68260.18	7987.00	5.54	83416.36	0.050

Table 3: F1 score for CFTR target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	**0.075**	0.041	0.027	0.021	0.169	0.476	0.631	0.733
XGB	0.070	0.040	0.027	0.021	0.142	0.488	0.667	0.758
kNN	0.037	0.025	0.019	0.015	0.145	0.380	0.509	0.601
min	0.057	0.027	0.021	0.020	0.294	0.633	0.748	0.768
max	0.046	0.056	**0.051**	**0.041**	0.030	0.217	0.347	0.454
mean	0.052	**0.059**	0.043	0.031	0.074	0.319	0.496	0.630
Fisher	0.052	0.030	0.022	0.019	**0.469**	**0.670**	**0.760**	**0.807**
min ECDF	0.070	0.042	0.030	0.024	0.266	0.508	0.632	0.704
max ECDF	0.054	0.032	0.023	0.020	0.242	0.553	0.687	0.752
mean ECDF	0.059	0.037	0.026	0.021	0.257	0.576	0.714	0.791
Fisher ECDF	**0.075**	0.043	0.029	0.023	0.293	0.579	0.714	0.785

Table 4: CFTR target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.868	0.842	0.696	0.485	0.330	0.160
	Highest p_1	0.894	0.850	0.700	0.485	0.330	0.159
XGB	Lowest p_0	0.922	0.896	0.704	0.491	0.331	0.157
	Highest p_1	0.950	0.911	0.700	0.490	0.329	0.157
kNN	Lowest p_0	0.942	0.923	0.684	0.379	0.208	0.099
	Highest p_1	0.974	0.932	0.678	0.378	0.206	0.098
min	Lowest p_0	0.874	0.862	0.718	0.468	0.305	0.149
	Highest p_1	**0.978**	**0.942**	0.731	0.492	0.322	0.149
max	Lowest p_0	0.962	0.934	0.734	0.481	0.295	0.138
	Highest p_1	0.920	0.894	0.711	0.443	0.255	0.121
mean	Lowest p_0	0.964	0.940	0.737	0.485	0.307	0.147
	Highest p_1	0.976	0.939	0.740	0.497	0.329	0.166
Fisher	Lowest p_0	0.962	0.941	**0.743**	**0.498**	**0.336**	**0.174**
	Highest p_1	0.976	0.940	0.740	0.497	0.327	0.162

Table 5: Prediction error rate for NFE2L2 target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 114,510, in which only about 7,798 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	758.78	80.14	4631.90	1054.64	0.00	107984.54	0.010
XGB	768.54	77.80	5356.82	1063.10	0.00	107243.74	0.010
kNN	725.26	79.64	3256.54	1062.66	0.00	109385.90	0.010
min	1308.04	201.02	9585.02	2272.56	4.56	101138.80	0.022
max	282.54	3.18	644.18	198.44	0.00	113381.66	0.002
mean	420.70	6.88	1180.98	336.10	0.00	112565.34	0.003
Fisher	1805.88	282.76	13884.40	3602.58	0.00	94934.38	0.034
min ECDF	819.56	81.54	5060.74	1064.90	0.08	107483.18	0.010
max ECDF	843.12	79.14	5266.24	1063.00	0.00	107258.50	0.010
mean ECDF	867.90	79.64	5553.64	1062.44	0.00	106946.38	0.010
Fisher ECDF	899.96	81.02	5905.90	1067.26	0.00	106555.86	0.010

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	2101.96	393.78	16679.94	5346.64	0.00	89987.68	0.050
XGB	2000.18	388.82	16813.16	5330.28	0.00	89977.56	0.050
kNN	1355.22	400.02	11552.74	5349.26	0.00	95852.76	0.050
min	2903.96	865.06	29165.26	10960.70	682.60	69932.42	0.109
max	735.16	33.92	3091.62	832.14	0.00	109817.16	0.008
mean	1045.16	87.44	5912.44	1473.80	0.00	105991.16	0.014
Fisher	2835.76	772.72	28710.98	9064.24	1.10	73125.20	0.086
min ECDF	2005.68	388.36	15778.18	5229.30	98.34	91010.14	0.050
max ECDF	1893.98	393.70	16392.30	5333.64	0.00	90496.38	0.050
mean ECDF	2068.08	396.80	17250.64	5330.06	0.00	89464.42	0.050
Fisher ECDF	2202.18	394.18	17587.46	5323.06	0.08	89003.04	0.050

Table 6: F1 score for NFE2L2 target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.379	0.436	0.378	0.332	0.054	0.182	0.294	0.381
XGB	0.382	0.420	0.365	0.322	0.062	0.183	0.292	0.377
kNN	0.364	0.305	0.254	0.228	0.038	0.129	0.214	0.286
min	0.453	0.362	0.301	0.275	0.109	**0.297**	**0.416**	**0.476**
max	0.211	0.391	0.435	**0.433**	0.008	0.036	0.079	0.129
mean	0.285	0.443	**0.452**	0.409	0.014	0.069	0.153	0.246
Fisher	**0.475**	0.402	0.348	0.312	**0.154**	0.293	0.383	0.445
min ECDF	0.402	0.425	0.368	0.331	0.059	0.173	0.272	0.347
max ECDF	0.411	0.402	0.341	0.301	0.061	0.179	0.282	0.360
mean ECDF	0.421	0.431	0.368	0.325	0.064	0.187	0.293	0.376
Fisher ECDF	0.432	**0.453**	0.383	0.334	0.068	0.191	0.296	0.378

Table 7: NFE2L2 target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.542	0.525	0.523	0.522	0.513	0.212
	Highest p_1	0.542	0.522	0.525	0.522	0.512	0.212
XGB	Lowest p_0	0.682	0.673	0.676	0.653	0.602	0.210
	Highest p_1	0.692	0.670	0.672	0.654	0.602	0.209
kNN	Lowest p_0	0.774	0.740	0.711	0.671	0.637	0.157
	Highest p_1	0.792	0.750	0.711	0.673	0.636	0.157
min	Lowest p_0	0.706	0.675	0.662	0.643	0.613	0.212
	Highest p_1	0.772	0.768	0.730	0.694	0.651	0.214
max	Lowest p_0	0.748	0.756	0.721	0.691	0.650	0.205
	Highest p_1	0.680	0.642	0.628	0.614	0.596	0.201
mean	Lowest p_0	0.782	0.773	0.730	0.696	0.657	0.223
	Highest p_1	**0.800**	0.775	0.735	**0.697**	**0.658**	0.230
Fisher	Lowest p_0	0.786	0.761	0.726	0.695	0.658	**0.231**
	Highest p_1	0.800	**0.776**	**0.736**	0.697	0.658	0.227

Table 8: Prediction error rate for GLP1R target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 112,076, in which only about 9,056 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	727.20	90.52	5681.44	1028.20	0.00	104548.64	0.010
XGB	731.38	92.30	5485.66	1026.84	0.00	104739.82	0.010
kNN	729.28	94.42	2966.32	1022.58	0.00	107263.40	0.010
min	1306.74	229.04	10512.34	2303.00	10.94	97713.94	0.023
max	274.02	4.06	409.48	151.38	0.00	111237.06	0.001
mean	391.44	9.22	813.40	275.02	0.00	110586.92	0.003
Fisher	1806.34	333.22	16495.66	3649.02	0.02	89791.74	0.036
min ECDF	820.38	97.90	5109.16	1026.36	0.46	105021.74	0.010
max ECDF	804.06	87.40	4914.62	1021.82	0.00	105248.10	0.010
mean ECDF	830.72	88.78	5201.18	1021.34	0.00	104933.98	0.010
Fisher ECDF	871.84	94.62	5927.60	1023.62	0.00	104158.32	0.010

(a) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	2094.70	451.58	20397.72	5137.26	0.00	83994.74	0.050
XGB	1982.46	460.38	20252.54	5128.82	0.00	84251.80	0.050
kNN	1503.00	452.60	11028.60	5139.24	0.00	93952.56	0.050
min	3098.46	967.00	32613.72	10463.80	694.02	64239.00	0.108
max	709.78	48.96	2996.82	812.50	0.00	107507.94	0.008
mean	1032.08	115.38	6446.82	1500.06	0.00	102981.66	0.014
Fisher	3027.02	862.18	32613.46	9039.30	2.10	66531.94	0.088
min ECDF	2028.18	453.16	18345.90	5061.74	100.92	86086.10	0.050
max ECDF	1972.68	450.56	17661.00	5144.56	0.00	86847.20	0.050
mean ECDF	2091.10	454.02	19165.16	5127.06	0.00	85238.66	0.050
Fisher ECDF	2194.10	456.48	20799.98	5143.88	0.10	83481.46	0.050

Table 9: F1 score for GLP1R target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.368	0.444	0.407	0.370	0.066	0.218	0.332	0.412
XGB	0.370	0.426	0.388	0.357	0.064	0.216	0.330	0.407
kNN	0.370	0.340	0.296	0.269	0.035	0.124	0.210	0.282
min	0.450	0.393	0.348	0.327	0.119	**0.326**	**0.439**	**0.486**
max	0.209	0.382	0.431	**0.436**	0.005	0.035	0.086	0.142
mean	0.274	0.437	**0.453**	0.429	0.010	0.074	0.173	0.267
Fisher	**0.472**	0.425	0.384	0.356	**0.180**	**0.326**	0.409	0.462
min ECDF	0.406	0.437	0.398	0.370	0.059	0.198	0.306	0.379
max ECDF	0.400	0.424	0.379	0.344	0.057	0.191	0.297	0.372
mean ECDF	0.411	0.444	0.402	0.367	0.061	0.206	0.321	0.404
Fisher ECDF	0.426	**0.460**	0.413	0.375	0.069	0.221	0.337	0.415

Table 10: GLP1R target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.916	0.894	0.836	0.737	0.643	0.226
	Highest p_1	0.902	0.891	0.838	0.736	0.643	0.226
XGB	Lowest p_0	0.980	0.986	0.982	0.903	0.745	0.222
	Highest p_1	1.000	0.999	0.992	0.905	0.745	0.223
kNN	Lowest p_0	0.974	0.980	0.984	0.914	0.774	0.177
	Highest p_1	1.000	0.999	0.994	0.916	0.774	0.177
min	Lowest p_0	0.972	0.962	0.959	0.905	0.763	0.227
	Highest p_1	1.000	1.000	0.985	0.903	0.771	0.228
max	Lowest p_0	0.998	0.992	0.979	0.894	0.771	0.221
	Highest p_1	0.978	0.954	0.932	0.842	0.726	0.219
mean	Lowest p_0	0.998	0.994	0.982	0.910	0.775	0.237
	Highest p_1	1.000	1.000	0.992	0.919	0.777	0.242
Fisher	Lowest p_0	0.998	0.998	0.991	0.931	0.783	**0.244**
	Highest p_1	1.000	1.000	0.992	0.919	0.777	0.240

3.7. Detailed results for DRD3 target

DRD3 is the fifth biggest target, with 0.83% of active compounds (3,381).

The confusion matrices, F1 score for precise predictions, and ranking precision for DRD3 are described below in Tables 11, 12 and 13. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher and combination of p-values by taking the mean returned the highest F1 score when predicting inactive compounds. Whereas, Fisher and combination of p-values by taking the mean were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.8. Detailed results for SMAD3 target

SMAD3 is the sixth biggest target, and has one of highest number of active compounds at 5.4% (21,995).

The confusion matrices, F1 score for precise predictions, and ranking precision for DRD3 are described below in Tables 14, 15 and 16. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher and combination of p-values by taking the min returned the highest F1 score when predicting inactive compounds. Whereas, Fisher and combination of p-values by taking the mean were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.9. Detailed results for APOBEC3G target

APOBEC3G is the seventh biggest target, and has 1% of active compounds (4,040).

The confusion matrices, F1 score for precise predictions, and ranking precision for DRD3 are described below in Tables 17, 18 and 19. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher returned the highest F1 score when predicting inactive compounds. Whereas, Fisher ECDF and combination of p-values by taking the mean were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.10. Detailed results for GLS target

GLS is the eighth biggest target, and has 0.71% of active compounds (2,870).

The confusion matrices, F1 score for precise predictions, and ranking precision for GLS are described below in Tables 20, 21 and 22. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher and combination of p-values by taking the minimum returned the highest F1 score when predicting inactive compounds. Whereas, Fisher ECDF was the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.11. Detailed results for TARDBP target

TARDBP is the ninth biggest target, and has 2.67% of active compounds (10,765).

Table 11: Prediction error rate for DRD3 target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 108,536, in which only about 898 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	870.58	8.06	89750.30	1080.56	0.00	16826.50	1.003e-02
XGB	868.18	7.88	86141.92	1078.40	0.00	20439.62	1.001e-02
kNN	785.02	8.60	61365.38	1071.58	0.00	45305.42	9.952e-03
min	871.88	9.20	99133.88	1283.62	1244.76	5992.66	2.338e-02
max	777.06	0.62	49073.54	81.80	0.00	58602.98	7.594e-04
mean	808.80	1.56	70375.58	136.62	0.00	37213.44	1.273e-03
Fisher	846.60	12.56	104185.42	200.78	3290.64	0.00	3.228e-02
min ECDF	866.38	6.26	87688.08	769.12	312.88	18893.28	1.003e-02
max ECDF	843.98	8.40	89944.64	1082.16	0.92	16655.90	1.006e-02
mean ECDF	855.76	8.22	97041.18	1083.52	0.44	9546.88	1.006e-02
Fisher ECDF	875.68	9.36	99697.92	1063.92	17.06	6872.06	1.005e-02

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	853.78	14.64	102260.80	375.84	5030.94	0.00	0.050
XGB	853.46	14.46	102262.18	480.00	4925.90	0.00	0.050
kNN	837.88	42.86	96690.24	5358.82	18.44	5587.76	0.050
min	821.82	5.66	95348.34	122.70	12237.48	0.00	0.114
max	829.62	15.16	96336.00	462.26	1.06	10891.90	0.004
mean	849.52	24.94	105752.56	536.46	522.90	849.62	0.010
Fisher	795.00	6.24	98116.92	47.26	9570.58	0.00	0.089
min ECDF	851.96	10.78	102237.34	507.88	4898.68	29.36	0.050
max ECDF	842.86	18.44	100611.98	812.30	4596.30	1654.12	0.050
mean ECDF	848.22	14.66	102126.74	361.84	5044.38	140.16	0.050
Fisher ECDF	852.16	9.72	102259.68	283.08	5131.36	0.00	0.050

Table 12: F1 score for DRD3 target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.420	0.499	0.525	0.510	0.699	0.761	0.735	0.709
XGB	0.419	0.485	0.522	0.509	0.678	0.761	0.735	0.709
kNN	0.388	0.200	0.310	0.426	0.537	0.734	0.735	0.709
min	0.402	**0.524**	0.507	0.480	0.746	0.728	0.672	0.618
max	0.509	0.476	0.478	0.502	0.452	0.733	**0.770**	**0.770**
mean	0.515	0.475	**0.528**	**0.518**	0.593	**0.777**	**0.770**	0.753
Fisher	**0.522**	0.523	0.505	0.489	**0.769**	0.741	0.714	0.690
min ECDF	0.453	0.481	0.524	0.509	0.687	0.761	0.735	0.709
max ECDF	0.410	0.440	0.501	0.506	0.701	0.753	0.734	0.709
mean ECDF	0.414	0.499	0.527	0.511	0.736	0.760	0.735	0.709
Fisher ECDF	0.424	0.513	**0.528**	0.510	0.749	0.761	0.735	0.709

Table 13: DRD3 target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.994	0.996	0.996	0.995	0.995	0.945
	Highest p_1	1.000	1.000	1.000	0.999	0.998	0.946
XGB	Lowest p_0	0.998	0.998	0.998	0.997	0.996	0.944
	Highest p_1	1.000	1.000	1.000	1.000	1.000	0.945
kNN	Lowest p_0	0.998	0.998	0.997	0.997	0.996	0.853
	Highest p_1	1.000	1.000	1.000	1.000	0.997	0.853
min	Lowest p_0	0.994	0.994	0.991	0.990	0.992	0.944
	Highest p_1	1.000	1.000	1.000	1.000	0.999	0.941
max	Lowest p_0	1.000	0.999	0.999	0.999	0.998	0.921
	Highest p_1	1.000	1.000	1.000	1.000	0.999	0.912
mean	Lowest p_0	1.000	1.000	1.000	0.999	0.999	0.933
	Highest p_1	1.000	1.000	1.000	1.000	1.000	0.947
Fisher	Lowest p_0	0.998	0.998	0.999	0.999	0.999	**0.955**
	Highest p_1	1.000	1.000	1.000	1.000	1.000	0.954

Table 14: Prediction error rate for SMAD3 target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 107,094, in which only about 5,774 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	425.68	57.88	1824.16	1001.52	0.00	103784.76	0.010
XGB	429.80	56.44	1714.98	1011.66	0.00	103881.12	0.010
kNN	335.74	59.60	1344.00	1014.68	0.00	104339.98	0.010
min	675.50	151.88	4192.60	2320.80	5.74	99747.48	0.023
max	163.92	0.80	22.66	123.50	0.00	106783.12	0.001
mean	203.34	2.38	70.36	185.54	0.00	106632.38	0.002
Fisher	800.36	171.38	5236.80	2861.42	0.00	98024.04	0.028
min ECDF	449.56	60.96	1716.08	1009.76	0.26	103857.38	0.010
max ECDF	411.38	56.14	1768.38	1015.52	0.00	103842.58	0.010
mean ECDF	434.84	57.12	1833.48	1015.76	0.00	103752.80	0.010
Fisher ECDF	481.42	59.12	1898.26	1011.58	0.00	103643.62	0.010

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	1018.68	289.54	8665.00	5042.96	0.00	92077.82	0.050
XGB	986.78	286.54	8135.44	5051.44	0.00	92633.80	0.050
kNN	670.90	290.42	6016.94	5054.06	0.00	95061.68	0.050
min	1531.22	661.26	17130.90	11107.30	506.32	76157.00	0.115
max	286.60	15.10	496.74	432.08	0.00	105863.48	0.004
mean	398.12	41.58	1330.42	843.26	0.00	104480.62	0.008
Fisher	1330.90	490.34	13878.64	7825.58	3.20	83565.34	0.078
min ECDF	996.48	289.10	7743.38	5007.00	48.64	93009.40	0.050
max ECDF	864.84	288.78	8254.50	5034.32	0.00	92651.56	0.050
mean ECDF	956.20	290.74	8395.62	5035.92	0.00	92415.52	0.050
Fisher ECDF	1061.16	291.02	8603.28	5047.44	0.26	92090.84	0.050

Table 15: F1 score for SMAD3 target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.235	0.247	0.216	0.194	2.165e-02	0.099	0.180	0.251
XGB	0.236	0.240	0.211	0.191	2.037e-02	0.093	0.171	0.241
kNN	0.189	0.169	0.150	0.139	1.600e-02	0.070	0.131	0.188
min	0.260	0.206	0.174	0.160	4.904e-02	**0.186**	**0.293**	**0.358**
max	0.132	0.197	0.226	0.233	2.720e-04	0.006	0.020	0.041
mean	0.157	0.232	**0.250**	**0.235**	8.442e-04	0.016	0.053	0.107
Fisher	**0.273**	0.235	0.207	0.188	**6.088e-02**	0.153	0.227	0.284
min ECDF	0.246	0.243	0.213	0.193	2.038e-02	0.089	0.159	0.219
max ECDF	0.227	0.214	0.186	0.169	2.099e-02	0.094	0.173	0.243
mean ECDF	0.239	0.234	0.205	0.186	2.176e-02	0.096	0.176	0.249
Fisher ECDF	0.261	**0.256**	0.220	0.197	2.252e-02	0.098	0.176	0.245

Table 16: SMAD3 target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.648	0.634	0.599	0.578	0.526	0.124
	Highest p_1	0.650	0.630	0.599	0.580	0.526	0.124
XGB	Lowest p_0	0.712	0.684	0.658	0.617	0.553	0.124
	Highest p_1	0.698	0.678	0.659	0.618	0.553	0.124
kNN	Lowest p_0	0.708	0.700	0.685	0.662	0.593	0.095
	Highest p_1	0.700	0.694	0.690	0.661	0.594	0.095
min	Lowest p_0	0.708	0.682	0.662	0.635	0.580	0.127
	Highest p_1	0.716	0.721	0.713	0.691	0.626	0.118
max	Lowest p_0	**0.748**	0.714	0.710	0.691	0.626	0.115
	Highest p_1	0.702	0.674	0.654	0.625	0.576	0.123
mean	Lowest p_0	0.730	0.718	0.710	**0.697**	0.630	0.126
	Highest p_1	0.724	0.720	0.719	**0.697**	0.632	0.130
Fisher	Lowest p_0	0.730	**0.732**	**0.725**	0.695	**0.635**	**0.135**
	Highest p_1	0.724	0.720	0.720	**0.697**	0.631	0.127

Table 17: Prediction error rate for APOBEC3G target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 106,341, in which only about 1,066 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	376.66	10.20	10400.70	1048.70	0.00	94504.74	0.010
XGB	363.94	10.06	10851.88	1046.56	0.00	94068.56	0.010
kNN	181.38	9.62	5158.94	1053.90	0.00	99937.16	0.010
min	464.60	24.60	20139.90	2313.44	48.40	83350.06	0.022
max	137.10	0.38	492.40	146.72	0.00	105564.40	0.001
mean	169.56	0.80	1200.88	216.44	0.00	104753.32	0.002
Fisher	556.90	38.34	32997.74	3293.56	0.18	69454.28	0.031
min ECDF	339.30	11.56	10963.50	1041.70	10.06	93974.88	0.010
max ECDF	286.12	9.36	10692.10	1047.00	0.00	94306.42	0.010
mean ECDF	316.82	9.34	11841.54	1042.86	0.00	93130.44	0.010
Fisher ECDF	384.52	11.92	14955.48	1042.56	0.00	89946.52	0.010

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	611.46	52.02	36043.96	5266.00	0.00	64367.56	0.050
XGB	595.46	50.12	35664.38	5249.22	0.00	64781.82	0.050
kNN	311.86	50.82	17657.90	5253.18	0.00	83067.24	0.050
min	713.32	93.04	53334.62	10296.08	1653.84	40250.10	0.113
max	228.12	4.50	5896.68	520.64	0.00	99691.06	0.005
mean	319.50	11.86	14122.20	1071.30	0.00	90816.14	0.010
Fisher	718.48	98.32	58422.58	9110.76	55.94	37934.92	0.087
min ECDF	595.86	47.90	32961.00	5003.98	254.36	67477.90	0.050
max ECDF	477.06	49.72	35805.92	5237.52	0.00	64770.78	0.050
mean ECDF	548.72	50.48	39219.94	5236.04	0.00	61285.82	0.050
Fisher ECDF	630.64	51.88	39884.14	5261.72	2.68	60509.94	0.050

Table 18: F1 score for APOBEC3G target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.208	0.152	0.106	0.083	0.117	0.355	0.499	0.588
XGB	0.202	0.148	0.105	0.082	0.122	0.352	0.504	0.589
kNN	0.106	0.080	0.063	0.053	0.060	0.191	0.314	0.402
min	0.187	0.108	0.085	0.084	0.215	0.485	0.581	0.587
max	0.111	0.155	**0.161**	**0.147**	0.006	0.068	0.177	0.281
mean	0.131	**0.178**	0.158	0.126	0.014	0.156	0.336	0.469
Fisher	0.184	0.120	0.092	0.079	**0.330**	**0.519**	**0.600**	**0.640**
min ECDF	0.190	0.153	0.114	0.096	0.123	0.330	0.464	0.538
max ECDF	0.162	0.121	0.089	0.078	0.120	0.353	0.489	0.556
mean ECDF	0.178	0.138	0.102	0.082	0.132	0.381	0.529	0.607
Fisher ECDF	**0.212**	0.156	0.111	0.088	0.164	0.386	0.528	0.604

Table 19: APOBEC3G target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.638	0.629	0.613	0.581	0.531	0.216
	Highest p_1	0.672	0.627	0.612	0.581	0.530	0.215
XGB	Lowest p_0	0.712	0.644	0.592	0.542	0.492	0.202
	Highest p_1	0.708	0.653	0.592	0.543	0.493	0.202
kNN	Lowest p_0	0.718	0.692	0.646	0.594	0.455	0.103
	Highest p_1	0.716	0.690	0.647	0.592	0.455	0.103
min	Lowest p_0	0.704	0.674	0.615	0.570	0.517	0.198
	Highest p_1	0.812	0.748	0.706	0.643	0.543	0.203
max	Lowest p_0	0.826	0.752	0.706	0.642	0.533	0.174
	Highest p_1	0.684	0.654	0.620	0.569	0.517	0.157
mean	Lowest p_0	0.818	0.758	0.710	0.648	0.539	0.197
	Highest p_1	0.816	0.758	0.707	**0.650**	**0.552**	0.229
Fisher	Lowest p_0	**0.834**	**0.766**	**0.710**	0.636	0.547	**0.233**
	Highest p_1	0.816	0.758	0.707	**0.650**	**0.552**	0.224

Table 20: Prediction error rate for GLS target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 104,997, in which only about 746 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	121.38	6.62	2485.08	1049.00	0.00	101334.92	1.005e-02
XGB	122.28	7.14	2495.30	1049.10	0.00	101323.18	1.006e-02
kNN	45.82	7.24	2698.06	1039.38	0.00	101206.50	9.968e-03
min	160.30	18.02	6476.04	2457.14	18.88	95866.62	2.375e-02
max	26.02	0.10	124.60	64.58	0.00	104781.70	6.160e-04
mean	34.68	0.28	296.28	109.98	0.00	104555.78	1.050e-03
Fisher	200.02	27.68	10165.32	3270.96	0.00	91333.02	3.142e-02
min ECDF	116.26	9.28	3486.48	1047.62	3.84	100333.52	1.010e-02
max ECDF	85.72	7.24	2956.10	1037.86	0.00	100910.08	9.954e-03
mean ECDF	97.42	8.08	3176.60	1047.08	0.00	100667.82	1.005e-02
Fisher ECDF	130.24	8.46	3572.80	1044.62	0.00	100240.88	1.003e-02

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	228.62	36.50	12207.56	5199.68	0.00	87324.64	0.050
XGB	226.94	36.90	11710.04	5232.64	0.00	87790.48	0.050
kNN	112.50	37.96	11348.92	5203.82	0.00	88293.80	0.050
min	303.76	78.72	25078.72	11259.30	651.54	67624.96	0.114
max	54.88	3.28	1504.78	362.82	0.00	103071.24	0.003
mean	89.60	9.00	3575.42	874.42	0.00	100448.56	0.008
Fisher	296.60	72.60	24651.88	9462.80	11.72	70501.40	0.091
min ECDF	216.38	37.80	12565.52	5124.76	102.48	86950.06	0.050
max ECDF	168.66	37.04	13462.10	5224.70	0.00	86104.50	0.050
mean ECDF	200.48	36.96	13946.74	5219.92	0.00	85592.90	0.050
Fisher ECDF	239.02	39.52	14356.06	5212.78	0.40	85149.22	0.050

Table 21: F1 score for GLS target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.072	0.060	0.045	0.038	0.029	0.136	0.248	0.335
XGB	0.073	0.059	0.046	0.038	0.029	0.131	0.239	0.328
kNN	0.028	0.030	0.027	0.025	0.032	0.127	0.213	0.286
min	0.067	0.044	0.034	0.030	0.075	**0.261**	**0.387**	**0.453**
max	0.023	0.042	0.052	**0.051**	0.001	0.018	0.054	0.103
mean	0.030	0.057	**0.058**	0.049	0.004	0.042	0.124	0.221
Fisher	0.071	0.050	0.040	0.035	**0.115**	0.258	0.351	0.415
min ECDF	0.069	0.057	0.047	0.039	0.041	0.140	0.240	0.314
max ECDF	0.052	0.044	0.037	0.032	0.035	0.149	0.264	0.359
mean ECDF	0.058	0.053	0.043	0.037	0.037	0.154	0.270	0.363
Fisher ECDF	**0.077**	**0.063**	0.047	0.039	0.042	0.158	0.268	0.356

Table 22: GLS target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.254	0.278	0.288	0.262	0.225	0.054
	Highest p_1	0.242	0.277	0.286	0.260	0.225	0.053
XGB	Lowest p_0	0.422	0.398	0.372	0.325	0.262	0.062
	Highest p_1	0.418	0.393	0.374	0.324	0.262	0.061
kNN	Lowest p_0	0.488	0.377	0.276	0.189	0.120	0.026
	Highest p_1	0.500	0.373	0.277	0.186	0.120	0.025
min	Lowest p_0	0.392	0.390	0.360	0.299	0.241	0.057
	Highest p_1	0.510	0.438	0.375	0.304	0.218	0.050
max	Lowest p_0	0.494	0.435	0.364	0.272	0.193	0.044
	Highest p_1	0.370	0.371	0.349	0.290	0.216	0.044
mean	Lowest p_0	0.500	**0.443**	0.361	0.281	0.207	0.051
	Highest p_1	0.490	0.434	0.379	0.322	0.245	0.061
Fisher	Lowest p_0	**0.524**	0.431	0.379	**0.325**	**0.268**	**0.068**
	Highest p_1	0.492	0.434	**0.380**	0.321	0.242	0.058

The confusion matrices, F1 score for precise predictions, and ranking precision for GLS are described below in Tables 23, 24 and 25. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher and combination of p-values by taking the minimum returned the highest F1 score when predicting inactive compounds. Whereas, Fisher and Fisher ECDF were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision.

3.12. Detailed results for PTH1R target

PTH1R is the tenth biggest target, and has 1.44% of active compounds (5,738).

The confusion matrices, F1 score for precise predictions, and ranking precision for PTH1R are described below in Tables 26, 27. Validity-wise, MCIP maintains good error rate below the significance level for all algorithms, except combination of p-values by taking the minimum and Fisher. Fisher and combination of p-values by taking the minimum returned the highest F1 score when predicting inactive compounds. Whereas, Fisher and Fisher ECDF were the best option when predicting active compounds. Fisher, again, returned the highest average ranking precision. and 28.

Table 23: Prediction error rate for TARDBP target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 102,570, in which only about 2,748 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	200.82	26.26	1923.18	998.36	0.00	99421.38	9.989e-03
XGB	218.52	28.44	2065.90	1000.72	0.00	99256.42	1.003e-02
kNN	124.74	26.90	1296.66	998.38	0.00	100123.32	9.996e-03
min	325.14	70.40	4503.82	2396.88	17.72	95256.04	2.423e-02
max	56.98	0.42	23.04	61.70	0.00	102427.86	6.056e-04
mean	72.52	1.28	73.18	99.28	0.00	102323.74	9.804e-04
Fisher	397.38	88.34	6480.16	2839.98	0.00	92764.14	2.855e-02
min ECDF	213.98	29.90	1881.74	999.24	4.42	99440.72	1.008e-02
max ECDF	189.18	27.56	2178.88	1012.00	0.00	99162.38	1.014e-02
mean ECDF	207.00	28.30	2211.94	1005.42	0.00	99117.34	1.008e-02
Fisher ECDF	238.08	29.22	2251.08	1006.20	0.00	99045.42	1.009e-02

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	501.08	136.98	9805.72	4994.52	0.00	87131.70	0.050
XGB	501.44	139.32	9706.38	5021.92	0.00	87200.94	0.050
kNN	282.34	139.28	6493.70	5007.94	0.00	90646.74	0.050
min	763.82	312.96	19006.92	11203.24	520.60	70762.46	0.117
max	108.42	7.90	604.26	302.18	0.00	101547.24	0.003
mean	170.98	21.18	1681.58	691.46	0.00	100004.80	0.007
Fisher	696.30	254.04	17114.18	8441.90	8.46	76055.12	0.085
min ECDF	474.56	138.00	8709.32	4910.40	74.82	88262.90	0.050
max ECDF	430.52	136.54	9453.36	5007.02	0.00	87542.56	0.050
mean ECDF	482.12	136.26	9705.66	5025.88	0.00	87220.08	0.050
Fisher ECDF	529.20	140.26	9886.76	5013.60	0.44	86999.74	0.050

Table 24: F1 score for TARDBP target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.118	0.130	0.116	0.105	2.281e-02	0.111	0.205	0.282
XGB	0.128	0.130	0.115	0.104	2.448e-02	0.110	0.198	0.275
kNN	0.075	0.075	0.073	0.070	1.544e-02	0.075	0.142	0.200
min	0.132	0.108	0.094	0.087	5.261e-02	**0.204**	**0.320**	**0.385**
max	0.049	0.083	0.110	0.118	2.765e-04	0.007	0.027	0.055
mean	0.061	0.112	**0.130**	**0.123**	8.780e-04	0.020	0.069	0.139
Fisher	**0.146**	0.123	0.108	0.099	**7.482e-02**	0.186	0.267	0.326
min ECDF	0.126	0.125	0.111	0.103	2.232e-02	0.099	0.180	0.246
max ECDF	0.111	0.113	0.100	0.092	2.581e-02	0.107	0.193	0.267
mean ECDF	0.121	0.125	0.111	0.101	2.619e-02	0.110	0.199	0.276
Fisher ECDF	0.138	**0.137**	0.118	0.106	2.665e-02	0.112	0.199	0.274

Table 25: TARDBP target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.446	0.429	0.403	0.360	0.308	0.071
	Highest p_1	0.432	0.430	0.405	0.360	0.308	0.071
XGB	Lowest p_0	0.582	0.566	0.512	0.447	0.373	0.077
	Highest p_1	0.596	0.566	0.514	0.446	0.373	0.077
kNN	Lowest p_0	0.744	0.709	0.621	0.479	0.343	0.052
	Highest p_1	**0.752**	**0.713**	**0.624**	0.480	0.343	0.052
min	Lowest p_0	0.672	0.621	0.574	0.494	0.389	0.076
	Highest p_1	0.730	0.670	0.596	0.504	0.384	0.072
max	Lowest p_0	0.686	0.651	0.592	0.511	0.381	0.069
	Highest p_1	0.592	0.543	0.525	0.469	0.379	0.070
mean	Lowest p_0	0.704	0.674	0.603	0.521	0.391	0.077
	Highest p_1	0.738	0.683	0.610	0.524	0.406	0.081
Fisher	Lowest p_0	0.726	0.690	0.630	**0.529**	**0.423**	**0.085**
	Highest p_1	0.738	0.686	0.611	0.524	0.406	0.079

Table 26: Prediction error rate for PTH1R target. ApA - Active predicted Active, ApI - Active predicted Inactive, IpI - Inactive predicted Inactive, IpA - Inactive predicted Active, refer to cases in which the region prediction contained only one label. Empty refers to cases in which the prediction set was empty. Uncertain are cases in which the region prediction contained more than one label. The values are averages over 50 splits. The number of test samples was 98,103, in which only about 1,409 on average across 50 splits are active compounds.

(a) $\varepsilon = 0.01$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	99.36	13.08	2317.84	962.04	0.00	94710.68	9.940e-03
XGB	99.58	14.18	2489.62	966.92	0.00	94532.70	1.000e-02
kNN	34.60	13.16	1613.52	972.86	0.00	95468.86	1.005e-02
min	158.30	35.46	5363.36	2363.30	26.52	90156.06	2.472e-02
max	8.52	0.24	87.68	29.28	0.00	97977.28	3.009e-04
mean	12.04	0.56	196.40	55.48	0.00	97838.52	5.712e-04
Fisher	189.96	40.64	7054.26	2366.88	0.32	88450.94	2.454e-02
min ECDF	85.16	15.52	2555.64	965.12	4.46	94477.10	1.004e-02
max ECDF	65.80	15.80	2992.16	967.98	0.00	94061.26	1.003e-02
mean ECDF	76.82	15.52	3050.08	965.52	0.00	93995.06	1.000e-02
Fisher ECDF	107.50	15.64	3122.64	972.64	0.00	93884.58	1.007e-02

(b) $\varepsilon = 0.05$

	ApA	ApI	IpI	IpA	Empty	Uncertain	Error rate
SVC	290.72	75.30	11297.58	4831.72	0.00	81607.68	0.050
XGB	279.62	71.02	11320.90	4834.30	0.00	81597.16	0.050
kNN	99.48	69.94	6491.04	4853.00	0.00	86589.54	0.050
min	412.04	156.98	20794.54	10640.44	833.50	65265.50	0.119
max	22.76	3.76	830.26	172.62	0.00	97073.60	0.002
mean	42.68	9.40	2032.92	395.98	0.00	95622.02	0.004
Fisher	363.84	116.56	17772.12	7277.16	14.34	72558.98	0.076
min ECDF	249.74	70.64	9795.90	4735.18	123.66	83127.88	0.050
max ECDF	185.74	71.36	11609.38	4846.02	0.00	81390.50	0.050
mean ECDF	231.64	71.84	12138.66	4834.72	0.00	80826.14	0.050
Fisher ECDF	288.96	74.58	12088.44	4832.46	3.04	80815.52	0.050

Table 27: F1 score for PTH1R target

	F1 for the active class				F1 for the inactive class			
epsilon	0.01	0.05	0.10	0.15	0.01	0.05	0.10	0.15
SVC	0.061	**0.079**	**0.070**	**0.063**	0.027	0.127	0.221	0.298
XGB	0.061	0.077	0.069	0.061	0.029	0.127	0.226	0.305
kNN	0.022	0.028	0.028	0.028	0.019	0.075	0.139	0.198
min	0.067	0.062	0.053	0.049	0.062	**0.222**	**0.332**	**0.390**
max	0.008	0.019	0.032	0.043	0.001	0.010	0.030	0.058
mean	0.011	0.032	0.055	**0.063**	0.002	0.024	0.072	0.134
Fisher	**0.080**	0.074	0.065	0.059	**0.081**	0.193	0.273	0.332
min ECDF	0.052	0.070	0.064	0.059	0.030	0.111	0.194	0.259
max ECDF	0.041	0.051	0.049	0.047	0.035	0.130	0.218	0.287
mean ECDF	0.047	0.064	0.061	0.057	0.036	0.136	0.229	0.301
Fisher ECDF	0.066	**0.079**	0.069	0.062	0.037	0.135	0.228	0.302

Table 28: PTH1R target's ranking precision for active compounds expressed in terms of precision-at-k, for k = 10, 25, 50, 100, 200 and in terms of Average Precision. Best values in each column are highlighted in bold.

		k=10	k=25	k=50	k=100	k=200	Avg prec
SVC	Lowest p_0	0.264	0.203	0.167	0.153	0.131	0.041
	Highest p_1	0.246	0.193	0.163	0.153	0.132	0.041
XGB	Lowest p_0	0.386	0.261	0.205	0.171	0.144	0.042
	Highest p_1	0.352	0.260	0.207	0.172	0.144	0.042
kNN	Lowest p_0	0.424	0.230	0.163	0.107	0.076	0.020
	Highest p_1	0.380	0.230	0.160	0.108	0.075	0.020
min	Lowest p_0	0.440	0.265	0.204	0.154	0.128	0.037
	Highest p_1	0.396	0.270	0.212	0.162	0.128	0.035
max	Lowest p_0	0.424	0.277	0.204	0.150	0.116	0.031
	Highest p_1	0.332	0.254	0.192	0.152	0.120	0.029
mean	Lowest p_0	0.428	0.272	0.206	0.153	0.125	0.037
	Highest p_1	0.422	0.279	0.218	0.171	0.137	0.041
Fisher	Lowest p_0	**0.456**	**0.297**	0.215	**0.178**	**0.151**	**0.045**
	Highest p_1	0.422	0.278	**0.218**	0.171	0.136	0.039