Supporting Information:

Efficient Multi-task chemogenomics for drug specificity prediction

Benoit Playe,[†,¶] Chloé-Agathe Azencott,[†,¶] and Véronique Stoven[*,†,¶]

†*Mines ParisTech, PSL Research University, Centre for Computational Biology, 35 Rue Saint-Honoré, F-77305 Fontainebleau Cedex, France*

‡*Institut Curie F-75248 Paris, France*

E-mail: benoit.playe@mines-paristech.fr

Supplementary Figure F1

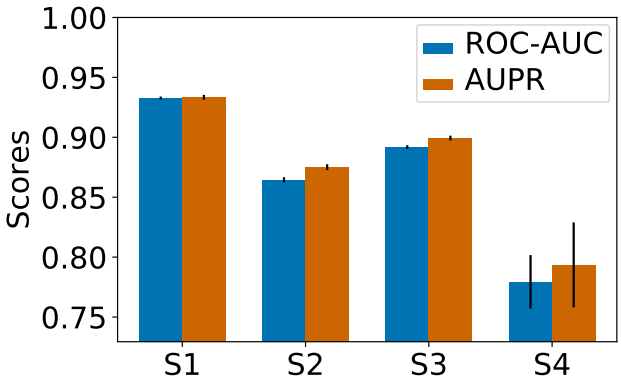

Figure 1: Scores of *MT* SVM chemogenomics on $S_1 - S_4$ settings obtained with the 5-fold CV scheme.

Supplementary Figure F2

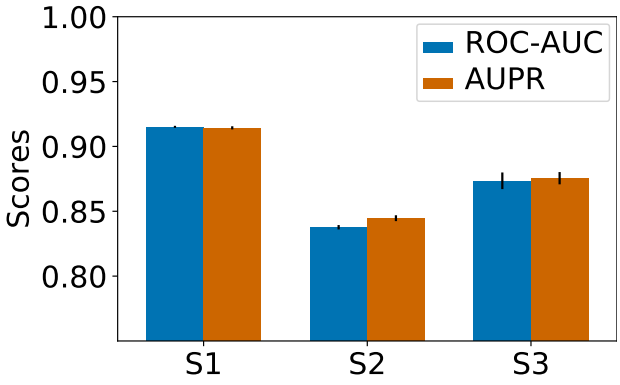

Figure 2: Scores of *MT* Kernel Ridge regression on S1/2/34 in nested-5foldCV

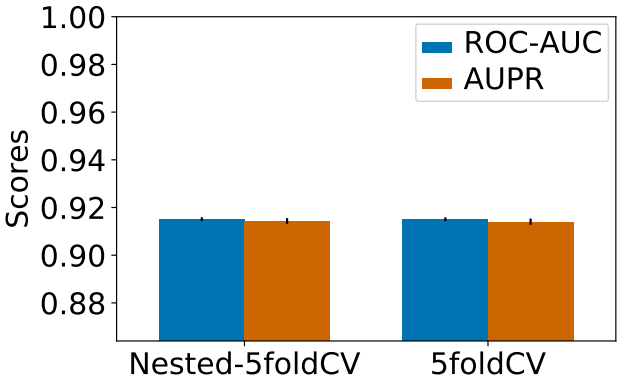

Figure 3: Scores of *MT* Kernel Ridge regression on S1 depending on CV scheme

.

Supplementary Figure F3

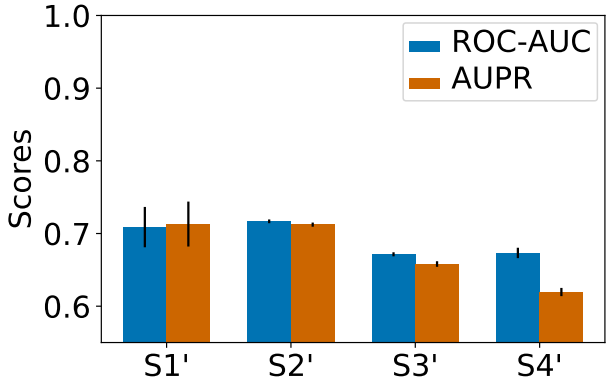

Figure 4: Scores of *MT* SVM method on S1', S2', S3' and S4' datasets with nested-5-fold CV

Supplementary Figure F4

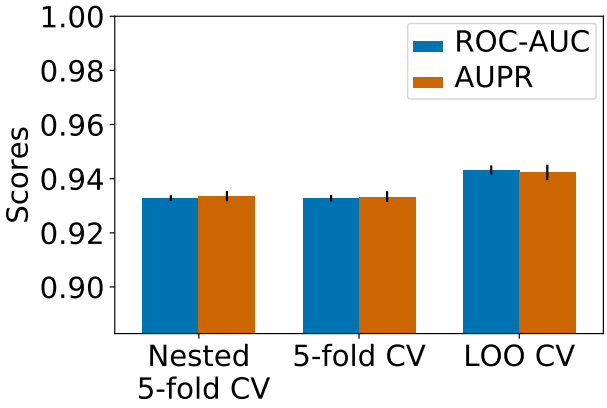

Figure 5: Scores of the *MT* method on S1 depending on CV scheme.

Overall, all CV schemes provide high prediction performance on this dataset, in the range of 0.93-0.94 in AUC and AUPR. The *Nested 5-fold CV* leads to performance very close to those of *5-fold CV*, showing that on the S1 dataset, *5-fold CV* did not suffer from overestimation of the performance due to data over-fitting. *LOO CV* leads to slightly better results, although very close to those of the other CV schemes. In general, the *LOO CV* scheme is expected to provide better results because the model is trained on more data points than *5-fold CV*. Again, this problem seems to be limited here, since the performance of *LOO CV* does not differ much from that of *Nested 5-fold CV*.

Supplementary Tables

Table 1: scores corresponding to 5-fold cross-validated *MT* on S1–S4

Setting	AUC	AUPR
S1	93.4 ± 0.11	93.4 ± 0.19
S2	86.3 ± 0.32	87.4 ± 0.35
S3	89.2 ± 0.58	89.9 ± 0.46
S4	81.0 ± 2.08	82.2 ± 2.40

Table 2: scores corresponding to *MT* on S1 depending on the CV scheme

CV scheme	AUC	AUPR
Nested-5-fold CV	93.3 ± 0.11	93.4 ± 0.19
5-fold CV	93.3 ± 0.12	93.3 ± 0.20
LOO-CV	94.3 ± 0.17	94.2 ± 0.29

Table 3: scores corresponding to Nested 5-fold cross-validated *MT* on S1'–S4'

Dataset	AUC	AUPR
S1'	69.3 ± 5.16	79.8 ± 1.82
S2'	69.4 ± 0.68	66.9 ± 0.53
S3'	66.3 ± 0.66	62.4 ± 0.95
S4'	46.3 ± 2.33	34.1 ± 0.95

Table 4: values corresponding to LOO-CV *ligand-based ST* and *MT-intra* on S1

model/nb_neg in intra task	1	2	5	10	50	full
ligand-based ST	91.02 ± 0.26	91.39 ± 0.32	92.2 ± 0.23	92.84 ± 0.2	93.43 ± 0.14	91.62 ± 0.18
KronSVM	94.19 ± 0.13	94.72 ± 0.18	95.19 ± 0.18	95.49 ± 0.12	95.59 ± 0.09	95.29 ± 0.04

Table 5: values corresponding to LOO-CV *NN-MT* on S1

nb_pos/nb_neg in NN intra task	1	2	5	10
0	95.49 ± 0.12	95.49 ± 0.12	95.49 ± 0.12	95.49 ± 0.12
1	95.79 ± 0.1	95.82 ± 0.1	95.78 ± 0.1	95.69 ± 0.13
5	96.09 ± 0.15	96.05 ± 0.16	95.87 ± 0.13	95.7 ± 0.11
10	96.2 ± 0.13	96.11 ± 0.16	95.93 ± 0.12	95.7 ± 0.09
50	96.18 ± 0.09	96.18 ± 0.07	96.0 ± 0.08	95.75 ± 0.09

Table 6: values corresponding to LOO-CV *RN-MT* on S1

nb_pos/nb_neg in RN extra task	1	2	5	10
0	95.49 ± 0.12	95.49 ± 0.12	95.49 ± 0.12	95.49 ± 0.12
1	95.65 ± 0.14	95.63 ± 0.14	95.59 ± 0.14	95.54 ± 0.14
5	95.89 ± 0.15	95.83 ± 0.15	95.69 ± 0.14	95.52 ± 0.14
10	96.02 ± 0.16	95.93 ± 0.16	95.72 ± 0.15	95.48 ± 0.14
50	95.86 ± 0.17	95.76 ± 0.16	95.55 ± 0.14	95.33 ± 0.14

Table 7: values corresponding to LOO-CV *MT-intra* on S1 with similarity constraint on intra-task pairs

θ	1	2	10	50
20	66.37 ± 1.18	66.9 ± 0.57	67.19 ± 0.72	67.34 ± 0.35
30	70.41 ± 0.85	70.87 ± 0.61	71.65 ± 0.54	71.63 ± 0.51
50	72.3 ± 0.52	72.52 ± 0.71	73.79 ± 0.54	73.78 ± 0.59
80	75.55 ± 0.35	76.46 ± 0.43	77.83 ± 0.51	77.49 ± 0.26

Table 8: values corresponding to LOO-CV *ligand-based ST* on S1 with similarity constraint on intra-task pairs

centile/ratio	1	2	10	50
20	66.0 ± 1.08	65.84 ± 1.11	69.62 ± 0.63	72.24 ± 0.36
30	64.86 ± 0.64	65.87 ± 0.42	70.04 ± 0.9	72.3 ± 0.74
50	67.11 ± 0.81	67.35 ± 0.79	71.28 ± 0.81	73.09 ± 0.6
80	69.13 ± 1.02	70.18 ± 0.33	74.09 ± 0.85	75.77 ± 1.24

Table 9: values corresponding to LOO-CV *NN-MT* on S1 with similarity constraint on intra-task pairs ($\theta = 20$)

nb_pos/nb_neg ratio in NN extra task	1	2	5
0	66.37 ± 1.18	66.37 ± 1.18	66.37 ± 1.18
1	85.64 ± 0.83	84.61 ± 0.73	82.89 ± 0.55
5	85.99 ± 0.76	84.81 ± 0.37	83.15 ± 0.5
10	84.65 ± 0.86	83.09 ± 0.36	81.69 ± 0.52
50	77.83 ± 0.58	77.11 ± 0.39	77.49 ± 0.3

Table 10: values corresponding to LOO-CV *NN-MT* on S1 with similarity constraint on intra-task pairs ($\theta = 80$)

nb_pos/nb_neg ratio in NN extra task	1	2	5
0	75.55 ± 0.35	75.55 ± 0.35	75.55 ± 0.35
1	86.12 ± 0.34	85.76 ± 0.36	84.87 ± 0.19
5	86.98 ± 0.37	86.52 ± 0.42	87.51 ± 0.22
10	86.82 ± 0.41	86.02 ± 0.44	87.24 ± 0.44
50	82.72 ± 0.49	82.73 ± 0.42	81.66 ± 0.27

Table 11: values corresponding to LOO-CV *RN-MT* on S1 with similarity constraint on intra-task pairs ($\theta = 20$)

nb_pos/nb_neg ratio in RN extra task	1	2	5
0	66.37 ± 1.18	66.37 ± 1.18	66.37 ± 1.18
1	63.83 ± 1.19	63.83 ± 0.81	66.06 ± 0.97
5	65.77 ± 0.89	65.98 ± 0.47	67.31 ± 0.35
10	67.55 ± 1.07	67.06 ± 1.05	65.65 ± 0.4
50	69.64 ± 0.98	70.19 ± 0.88	70.37 ± 0.56

Table 12: values corresponding to LOO-CV *RN-MT* on S1 with similarity constraint on intra-task pairs ($\theta = 80$)

nb_pos/nb_neg ratio in RN extra task	1	2	5
0	75.55 ± 0.35	75.55 ± 0.35	75.55 ± 0.35
1	72.15 ± 0.25	72.35 ± 0.19	75.51 ± 0.51
5	74.57 ± 0.38	75.73 ± 0.32	78.46 ± 0.36
10	75.78 ± 0.24	75.1 ± 0.63	77.43 ± 0.35
50	76.94 ± 0.44	75.91 ± 0.22	75.31 ± 0.27

Table 13: values corresponding to LOO-CV *MT-intra* , *NN-MT*, *RN-MT* on S1 with similarity constraint on intra-task pairs

model/centile	20	30	50	80
MT-intra	66.37 ± 1.18	70.41 ± 0.85	72.3 ± 0.52	75.55 ± 0.35
NN-MT	84.65 ± 0.86	85.68 ± 0.52	86.24 ± 0.45	86.82 ± 0.41
RN-MT	67.55 ± 1.07	70.55 ± 0.66	72.36 ± 0.36	75.78 ± 0.24

Table 14: values corresponding to LOO-CV *NN-MT* on S1 with similarity constraint on intra- and extra-task pairs ($\theta = 20$)

nb_pos/nb_neg ratio in NN extra task	1	2	5
1	63.39 ± 0.82	64.15 ± 1.3	65.87 ± 0.5
5	64.63 ± 0.84	64.98 ± 0.56	66.02 ± 0.24
10	65.64 ± 0.87	65.61 ± 0.7	64.53 ± 0.85
50	64.88 ± 0.67	64.24 ± 0.82	63.04 ± 0.27

Table 15: values corresponding to LOO-CV *NN-MT* on S1 with similarity constraint on intra- and extra-task pairs ($\theta = 80$)

nb_pos/nb_neg ratio in NN extra task	1	2	5
1	72.14 ± 0.33	71.91 ± 0.37	75.65 ± 0.34
5	73.35 ± 0.09	75.5 ± 0.27	77.58 ± 0.31
10	73.88 ± 0.35	73.83 ± 0.62	75.9 ± 0.59
50	72.49 ± 0.49	72.15 ± 0.54	70.72 ± 0.67

Table 16: values corresponding to LOO-CV *RN-MT* on S1 with similarity constraint on intra- and extra-task pairs ($\theta = 20$)

nb_pos/nb_neg ratio in RN extra task	1	2	5
1	63.0 ± 0.55	64.24 ± 0.34	65.99 ± 0.64
5	65.68 ± 0.15	64.53 ± 0.92	67.27 ± 1.1
10	66.01 ± 0.74	64.39 ± 1.24	62.99 ± 0.8
50	66.45 ± 0.66	65.75 ± 0.47	62.23 ± 0.66

Table 17: values corresponding to LOO-CV *RN-MT* on S1 with similarity constraint on intra- and extra-task pairs ($\theta = 80$)

nb_pos/nb_neg ratio in RN extra task	1	2	5
1	72.43 ± 0.11	71.99 ± 0.44	75.48 ± 0.36
5	73.92 ± 0.15	75.48 ± 0.33	78.1 ± 0.28
10	74.5 ± 0.18	73.77 ± 0.57	76.7 ± 0.34
50	73.17 ± 0.27	71.74 ± 0.54	69.14 ± 0.44

Table 18: GPCR dataset: values corresponding to LOO-CV *NN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42
1	96.43 ± 0.34	96.55 ± 0.3	96.77 ± 0.28	96.9 ± 0.28	96.98 ± 0.26
10	96.46 ± 0.2	96.65 ± 0.22	96.84 ± 0.18	97.01 ± 0.18	97.05 ± 0.19
50	96.32 ± 0.35	96.5 ± 0.25	96.66 ± 0.21	96.63 ± 0.17	96.65 ± 0.22
100	95.31 ± 0.27	95.88 ± 0.26	96.57 ± 0.22	96.74 ± 0.19	96.56 ± 0.21

Table 19: GPCR dataset: values corresponding to LOO-CV *NN-MT* with sequence based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23
1	92.92 ± 0.19	93.22 ± 0.22	93.66 ± 0.21	94.06 ± 0.18	94.32 ± 0.19
10	93.55 ± 0.24	93.97 ± 0.24	94.6 ± 0.27	95.14 ± 0.26	95.47 ± 0.37
50	93.04 ± 0.36	93.69 ± 0.27	95.05 ± 0.29	95.67 ± 0.34	95.88 ± 0.41
100	92.25 ± 0.26	93.65 ± 0.29	94.91 ± 0.29	95.54 ± 0.34	95.83 ± 0.39

Table 20: GPCR dataset: values corresponding to LOO-CV *RN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42
1	96.31 ± 0.4	96.32 ± 0.39	96.33 ± 0.39	96.34 ± 0.38	96.37 ± 0.4
10	96.21 ± 0.34	96.23 ± 0.35	96.26 ± 0.35	96.31 ± 0.36	96.52 ± 0.34
50	96.04 ± 0.3	96.19 ± 0.33	96.47 ± 0.39	96.65 ± 0.4	96.79 ± 0.39
100	95.72 ± 0.31	96.08 ± 0.36	96.57 ± 0.45	96.8 ± 0.51	96.99 ± 0.34

Table 21: GPCR dataset: values corresponding to LOO-CV *RN-MT* with sequence based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23	92.68 ± 0.23
1	93.23 ± 0.15	93.19 ± 0.13	93.17 ± 0.13	93.14 ± 0.14	93.08 ± 0.16
10	93.8 ± 0.23	93.86 ± 0.19	94.03 ± 0.28	94.13 ± 0.22	94.24 ± 0.22
50	94.08 ± 0.1	94.43 ± 0.11	94.73 ± 0.21	94.76 ± 0.25	95.18 ± 0.27
100	92.76 ± 0.12	93.84 ± 0.13	94.85 ± 0.2	94.9 ± 0.3	95.47 ± 0.27

Table 22: Ion Channel dataset: values corresponding to LOO-CV *NN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25
1	97.04 ± 0.27	97.05 ± 0.27	97.1 ± 0.27	97.14 ± 0.26	97.18 ± 0.27
10	97.22 ± 0.22	97.29 ± 0.2	97.38 ± 0.19	97.45 ± 0.16	97.45 ± 0.18
50	96.82 ± 0.15	97.0 ± 0.16	97.28 ± 0.15	97.42 ± 0.15	97.41 ± 0.12
100	96.54 ± 0.19	96.79 ± 0.16	97.24 ± 0.13	97.35 ± 0.1	97.36 ± 0.13

Table 23: Ion Channel dataset: values corresponding to LOO-CV *NN-MT* with sequence based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42
0	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17
1	96.56 ± 0.19	96.64 ± 0.16	96.69 ± 0.13	96.67 ± 0.13	96.66 ± 0.13
10	96.72 ± 0.16	96.81 ± 0.15	96.89 ± 0.14	96.81 ± 0.13	96.84 ± 0.16
50	96.28 ± 0.16	96.42 ± 0.15	96.64 ± 0.15	96.77 ± 0.13	96.97 ± 0.14
100	95.39 ± 0.14	95.73 ± 0.12	96.38 ± 0.16	96.74 ± 0.21	96.98 ± 0.18

Table 24: Ion Channel dataset: values corresponding to LOO-CV *RN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42	96.3 ± 0.42
0	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25	96.96 ± 0.25
1	97.0 ± 0.25	97.0 ± 0.25	97.0 ± 0.25	96.99 ± 0.25	96.99 ± 0.25
10	97.22 ± 0.23	97.21 ± 0.23	97.2 ± 0.23	97.19 ± 0.23	97.19 ± 0.08
50	97.28 ± 0.22	97.29 ± 0.22	97.31 ± 0.21	97.36 ± 0.2	97.36 ± 0.12
100	97.03 ± 0.29	97.11 ± 0.26	97.26 ± 0.23	97.34 ± 0.19	97.45 ± 0.14

Table 25: Ion Channel dataset: values corresponding to LOO-CV *RN-MT* with sequence based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17	96.34 ± 0.17
1	96.43 ± 0.15	96.42 ± 0.15	96.42 ± 0.15	96.41 ± 0.15	96.38 ± 0.15
10	96.76 ± 0.12	96.76 ± 0.12	96.74 ± 0.12	96.67 ± 0.11	96.56 ± 0.11
50	96.81 ± 0.03	97.01 ± 0.03	97.06 ± 0.08	96.92 ± 0.1	96.7 ± 0.11
100	96.3 ± 0.02	96.74 ± 0.04	97.05 ± 0.08	96.91 ± 0.11	96.78 ± 0.12

Table 26: Kinase dataset: values corresponding to LOO-CV *NN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49
1	89.36 ± 0.54	89.58 ± 0.54	89.87 ± 0.52	90.14 ± 0.5	90.41 ± 0.48
10	89.68 ± 0.51	89.93 ± 0.51	90.36 ± 0.53	90.75 ± 0.52	91.16 ± 0.46
50	87.78 ± 0.57	88.48 ± 0.51	89.57 ± 0.44	90.54 ± 0.44	91.28 ± 0.43
100	86.16 ± 0.37	87.02 ± 0.38	88.14 ± 0.48	89.72 ± 0.45	90.9 ± 0.43

Table 27: Kinase dataset: values corresponding to LOO-CV *NN-MT* with sequence based kernel

nb_pos/nb_neg ratio in NN extra task	1	2	5	10	20
0	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24
1	90.78 ± 0.44	90.76 ± 0.46	90.74 ± 0.46	90.47 ± 0.5	90.3 ± 0.49
10	93.23 ± 0.22	92.94 ± 0.32	92.53 ± 0.34	92.18 ± 0.28	91.51 ± 0.24
50	91.86 ± 0.36	91.55 ± 0.34	91.34 ± 0.28	91.57 ± 0.27	91.55 ± 0.18
100	89.83 ± 0.32	89.67 ± 0.28	90.12 ± 0.33	90.77 ± 0.19	90.84 ± 0.19

Table 28: Kinase dataset: values corresponding to LOO-CV *NN-MT* with family's hierarchy based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49	90.99 ± 0.49
1	90.6 ± 0.49	90.6 ± 0.49	90.6 ± 0.49	90.6 ± 0.49	90.6 ± 0.49
10	89.59 ± 0.37	89.62 ± 0.37	89.71 ± 0.37	89.84 ± 0.39	90.11 ± 0.53
50	87.61 ± 0.79	88.18 ± 0.75	89.31 ± 0.59	90.18 ± 0.51	90.86 ± 0.47
100	85.28 ± 1.13	86.89 ± 1.04	89.15 ± 0.76	90.35 ± 0.56	91.06 ± 0.48

Table 29: Kinase dataset: values corresponding to LOO-CV *NN-MT* with sequence based kernel

nb_pos/nb_neg ratio in RN extra task	1	2	5	10	20
0	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24	89.18 ± 0.24
1	89.7 ± 0.2	89.69 ± 0.21	89.66 ± 0.21	89.6 ± 0.21	89.49 ± 0.21
10	92.13 ± 0.21	92.01 ± 0.22	91.6 ± 0.22	90.9 ± 0.18	89.97 ± 0.33
50	92.74 ± 0.26	93.21 ± 0.23	93.39 ± 0.19	92.34 ± 0.19	90.45 ± 0.3
100	91.73 ± 0.1	92.69 ± 0.18	93.3 ± 0.2	92.06 ± 0.18	90.23 ± 0.26

Supporting Information: list of withdrawn drugs

Based on,[1,2] we consider a list of 174 withdrawn drugs:

name	year	country	DrugBank id
Adderall	2005	Canada	DB00182
Alatrofloxacin	2006	Worldwide	
Alclofenac	1979	UK	DB13167
Alpidem	1995	Worldwide	
Alosetron	2000	US	DB00969
Althesin	1984	France, Germany, UK	
Amineptine	1999	France, US	DB04836

Aminopyrine	1999	France, Thailand	DB01424
Amobarbital	1980	Norway	
Amoproxan	1970	France	
Anagestone	1969	Germany	
Antrafenine	1984	France	
Aprotinin	2008	US	
Ardeparin	2001	US	
Astemizole	1999	US, Malaysia, Multiple Nonspecified Markets	DB00637
Azaribine	1976	US	
Bendazac	1993	Spain	
Benoxaprofen	1982	Germany, Spain, UK, US	DB04812
Benzarone	1992	Germany	
Benziodarone	1964	France, UK	
Beta-ethoxy-lacetanilanide	1986	Germany	
Bezitramide	2004	Netherlands	
Bithionol	1967	US	DB04813
Broazolam	1989	UK	
Bromfenac	1998	US	
Bucetin	1986	Germany	
Buformin	1978	Germany	DB04830
Bunamiodyl	1963	Canada, UK, US	
Butamben	1964	US	
Canrenone	1986	Germany	
Cerivastatin	2001	US	DB00439
Chlormadinone	1970	UK, US	

Chlormezanone	1996	European Union, US, South Africa, Japan	DB01178
Chlorphentermine	1969	Germany	DB01556
Cianidanol	1985	France, Germany, Spain, Sweden	
Cinepazide	1988	Spain	
Cisapride	2000	US	
Clioquinol	1973	France, Germany, UK, US	DB04815
Clobutinol	2007	Germany	
Cloforex	1969	Germany	DB00631
Clomacron	1982	UK	
Clometacin	1987	France	
Co-proxamol	2004	UK	
Cyclobarbital	1980	Norway	
Cyclofenil	1987	France	
Dantron	1963	Canada, UK, US	
Dexfenfluramine	1997	European Union, UK, US	DB01191
Propoxyphene	2010	Worldwide	DB00647
Diacetoxydiphenolisatin	1971	Australia	
Diethylstilbestrol	1970s		DB00255
Difemerine	1986	Germany	
Dihydrostreptomycin	1970	US	
Dilevalol	1990	UK	
Dimazole	1972	France, US	
Dimethylamylamine	1983	US	
Dinoprostone	1990	UK	

Dipyrone	1975	UK, US, Others	
Dithiazanine	1964	France, US	
Dofetilide	2004	Germany	DB00204
Drotrecogin	2011	Worldwide	
Ebrotidine	1998	Spain	
Efalizumab	2009	Germany	
Encainide	1991	UK, US	DB01288
Ethyl	1963	Canada, UK, US,	
Etretinate	1989	France	DB00926
Exifone	1989	France	
Fen-phen	1997		
Fenclofenac	1984	UK	
Fenclozic	1970	UK, US	
Fenfluramine	1997	European Union, UK, US, India, South Africa, others	DB00574
Fenoterol	1990	New Zealand	DB01288
Feprazone	1984	Germany, UK	
Fipexide	1991	France	
Flosequinan	1993	UK, US	
Flunitrazepam	1991	France	
Gatifloxacin	2006	US	DB01044
Gemtuzumab	2010	US	
Glafenine	1984	France, Germany	
Grepafloxacin	1999	Withdrawn Germany, UK, US others	DB00365
Hydromorphone	2005		

Ibufenac	1968	UK	
Indalpine	1985	France	DB08953
Indoprofen	1983	Germany, Spain, UK	DB08951
Iodinated	1964	US	
Iproniazid	1964	Canada	DB04818
Isaxonine	1984	France	
Isoxicam	1983	France, Germany, Spain, others	DB08942
Kava	2002	Germany	
Ketorolac	1993	France, Germany, others	DB00465
L-tryptophan	1989	Germany, UK	
Levamisole	1999	US	
Levomethadyl	2003	US	DB01227
Lumiracoxib	2007ÃŘ2008	Worldwide	
Lysergic	1950sÃŘ1960s		
Mebanazine	1975	UK	
Methandrostenolone	1982	France, Germany, UK, US, others	
Methapyrilene	1979	Germany, UK, US	
Methaqualone	1984	South Africa (1971), India (1984), United Nations (1971-1988)	
Metipranolol	1990	UK, others	DB01214
Metofoline	1965	US	
Mibefradil	1998	European Union, Malaysia, US, others	

Minaprine	1996	France	DB00805
Moxisylyte	1993	France	
Muzolimine	1987	France, Germany, European Union	
Natalizumab	2005ÃŘ2006	US	
Nefazodone	2007	US, Canada, others	
Nialamide	1974	UK, US	
Nikethamide	1988	multiple markets	DB08989
Nitrefazole	1984	Germany	
Nomifensine	1981-1986	France, Germany, Spain, UK, US, others	DB4821
Oxeladin	1976	Canada, UK, US (1976)	
Oxyphenbutazone	1984-1985	UK, US, Germany, France, Canada	DB03585
Oxyphenisatin		Australia, France, Germany, UK, US	
Ozogamicin	2010	US	
Pemoline	1982		DB01230
Pentobarbital	1980	Norway	
Pemoline	1982		DB01230
Pergolide	2007	US	DB01186
Perhexilene	1985	UK, Spain	DB01074
Phenacetin	1975	Canada	DB03783
Phenformin	1977	France, Germany US	DB00914
Phenolphthalein	1997	US	DB04824
Phenoxypropazine	1966	UK	

Phenylbutazone	1985	Germany	
Phenylpropanolamine	2000	Canada, US	DB00397
Pifoxime	1976	France	
Pirprofen	1990	France, Germany, Spain	
Prenylamine	1988	Canada, France, Germany, UK, US, others	DB04825
Proglumide	1989	Germany	
Pronethalol	1965	UK	
Propanidid	1983	UK	
Proxibarbal	1998	Spain, France, Italy, Portugal, Turkey	
Pyrovalerone	1979	France	
Rapacuronium	2001	US, multiple markets	DB04834
Remoxipride	1993	UK, others	DB00409
Rimonabant	2008	Worldwide	DB06155
Rofecoxib	2004	Worldwide	DB00533
Rosiglitazone	2010	Europe	DB00412
Secobarbital		France, Norway, others.	
Sertindole	1998	European Union	DB06144
Sibutramine	2010	Australia, Canada, China, the European Union (EU), Hong Kong, India, Mexico, New Zealand, the Philippines, Thailand, the United Kingdom, and the United States	DB01105

Sitaxentan	2010	Germany	
Sorivudine	1993	Japan	
Sparfloxacin	2001	US	DB01208
Sulfacarbamide	1988	Germany	
Sulfamethoxydiazine	1988	Germany	
Sulfamethoxypyridazine	1986	UK	
Suloctidyl	1985	Germany, France, Spain	
Suprofen	1986-1987	UK, Spain, US	DB00870
Tegaserod	2007	US	DB01079
Temafloxacin	1992	US	DB01405
Temafloxacin	1992	US	DB01405
Temazepam	1999	Sweden, Norway	
Terfenadine	1997-1998	France, South Africa, Oman, others, US	DB00342
Terodiline	1991	Germany, UK, Spain, others	
Tetrazepam	2013	European Union	
Thalidomide	1961	Germany	DB01041
Thenalidine	1960	Canada, UK, US	DB04826
Thiobutabarbitone	1993	Germany	
Thioridazine	2005	Germany, UK	DB00679
Ticrynafen	1980	Germany, France, UK, US others	
Tolcapone	1998	European Union, Canada, Australia	

Tolrestat	1996	Argentina, Canada, Italy, others	
Triacetyldiphenolisatin	1971	Australia	
Triazolam	1991	France, Netherlands, Finland, Argentina, UK others	DB00897
Triparanol	1962	France, US	
Troglitazone	2000	US, Germany	
Trovafloxacin	1999-2001	European Union, US	
Valdecoxib	2004	US	DB00580
Vincamine	1987	Germany	
Xenazoic	1965	France	
Ximelagatran	2006	Germany	
Zimelidine	1983	Worldwide	DB04832
Zomepirac	1983	UK, Germany, Spain, US	DB04828

Table 30: List of widthdrawn drugs

We only considered withdrawn drugs for which we identified a DrugBank ID, we predicted targets and for which the side effects is specific and secondary target responsible for the side-effect has not been identified yet. This is the list of considered withdrawn drugs:

name	year	country	DrugBank id
Alosetron	2000	US	DB00969
Adderall	2005	Canada	DB00182
Cerivastatin	2001	US	DB00439
Cloforex	1969	Germany	DB00631
Dexfenfluramine	1997	European Union, UK, US	DB01191
Propoxyphene	2010	Worldwide	DB00647

Zimelidine	1983	Worldwide	DB04832
Dofetilide	2004	Germany	DB00204
Encainide	1991	UK, US	DB01288
Etretinate	1989	France	DB00926
Fenfluramine	1997	European Union, UK, US, India, South Africa, others	DB00574
Fenoterol	1990	New Zealand	DB01288
Levomethadyl	2003	US	DB01227
Metipranolol	1990	UK, others	DB01214
Minaprine	1996	France	DB00805
Astemizole	1999	US, Malaysia, Multiple Nonspecified Markets	DB00637
Oxyphenbutazone	1984-1985	UK, US, Germany, France, Canada	DB03585
Bithionol	1967	US	DB04813
Pergolide	2007	US	DB01186
Phenacetin	1975	Canada	DB03783
Phenformin	1977	France, Germany US	DB00914
Phenolphthalein	1997	US	DB04824
Diethylstilbestrol	1970s		DB00255
Phenylpropanolamine	2000	Canada, US	DB00397
Prenylamine	1988	Canada, France, Germany, UK, US, others	DB04825
Remoxipride	1993	UK, others	DB00409
Rimonabant	2008	Worldwide	DB06155
Rofecoxib	2004	Worldwide	DB00533

Rosiglitazone	2010	Europe	DB00412
Sertindole	1998	European Union	DB06144
Sibutramine	2010	Australia, Canada, China, the European Union (EU), Hong Kong, India, Mexico, New Zealand, the Philippines, Thailand, the United Kingdom, and the United States	DB01105
Sparfloxacin	2001	US	DB01208
Tegaserod	2007	US	DB01079
Terfenadine	1997-1998	France, South Africa, Oman, others, US	DB00342
Thalidomide	1961	Germany	DB01041
Valdecoxib	2004	US	DB00580

Table 31: List of considered withdrawn drugs

Supporting Information: basic principles of SVM and Kernel Ridge Regression

In this section, we briefly recall the basic principles of SVM and Kernel Ridge Regression.

- Support Vector Machines[3] seeks to find the optimal hyperplane separating two classes of data points, for example, proteins that bind to a ligand and proteins that don't. Among the infinity of possible separating hyperplanes, the optimal hyperplane is the one which on the first hand, maximizes the margin defined as the closest distance from

any point to the separating hyperplane, and on the second hand, allows a small number of training errors. The trade-off between the optimization of the margin and the number of training errors is controlled via a parameter C.

In practice, given a set of labeled samples S = $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where $(x_i, y_i) \in X \times \{-1, +1\}$ for $i = 1, \ldots, N$, a constant b, and the normal vector parameterizing the separating hyperplane w, the SVM algorithm solves the following optimization problem:

$$\min \left[\frac{1}{n} \sum_{i=1}^{n} max(0, 1 - y_i < w, x_i > +b) \right] + \frac{1}{C}|w|^2 \tag{1}$$

It can be rewritten via the so-called slack variables $v_i = max(0, 1 - y_i(w.x_i + b))$ as

$$\underset{w,b,\epsilon}{\text{argmin}} \ \frac{1}{2}|w|^2 + C \sum_{i=1}^{l} \epsilon_i \tag{2a}$$

$$\text{subject to } y_i < w, x_i > +b \geq 1 - \epsilon_i, \forall i = 1, \ldots, l, \tag{2b}$$

$$\epsilon_i \geq 0, i = 1, \ldots, l. \tag{2c}$$

The constant C in the objective function 2a is meant to introduce a trade-off between the maximization of the margin, expressed by the term $\frac{1}{2}||w||^2$, and the classification error on the training set, expressed by the slack variables. In the present study, the optimal parameter C was searched between 10^{-5} and 10^5.

Once the SVM algorithm has learned the hyperplane separating the two classes, it predicts labels for new points according to their position with respect to this hyperplane. It provides a prediction score based on the distance of the new points to the learned hyperplane.

- Kernel Ridge Regression[4] is the kernelized version of a linear regression with L_2 norm

regularization which aims at solving the following optimization problem:

$$\min_{\mathbf{w}} ||\mathbf{X}^T.\mathbf{w} - \mathbf{y}|| + \lambda ||w||^2 \quad (3)$$

where $\mathbf{X}$ is the n_samples*n_features feature matrix, $\mathbf{y}$ the output vector and $\mathbf{w}$ the weight vector. As a kernel method, kernel RLS fully benefits from the kernel trick. The output of Kernel RLS is the score of the predicted protein-ligand interaction. The regularizing parameter of the kernel RLS was searched between 10^{-5} and 10^5.

In the case of an unbalanced number of positive and negative samples in the training set, a weight, inversely proportional to the proportion of sample of the corresponding class, was assigned to samples during the training phase.

References

(1) Qureshi, Z. P.; Seoane-Vazquez, E.; Rodriguez-Monguio, R.; Stevenson, K. B.; Szeinbach, S. L. *Pharmacoepidemiology and drug safety* **2011**, *20*, 772–777.

(2) Fung, M.; Thornton, A.; Mybeck, K.; Wu, J. H.-h.; Hornbuckle, K.; Muniz, E. *Drug Information Journal* **2001**, *35*, 293–317.

(3) Cortes, C.; Vapnik, V. *Machine learning* **1995**, *20*, 273–297.

(4) Saunders, C.; Gammerman, A.; Vovk, V. **1998**,