Supporting Information: "Improvement of Multi-Task Learning by Data Enrichment: Application for Drug Discovery"

Ekaterina A. Sosnina,[*,†] Sergey Sosnin,[‡] and Maxim V. Fedorov[†,¶]

†*Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Bol'shoy Bul'var, 30, Moscow 143026, Russia*

‡*Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Josef-Holaubek-Platz 2, Vienna 1190, Austria*

¶*Sirius University of Science and Technology, Olympiisky prospect 1, Sochi 354340, Russia*

E-mail: ekaterina.sosnina@skoltech.ru

Phone: +79260256249

The number of targets applied in the metrics calculation.

For each dataset, the "i" and "c" subsets remained unchanged across all scenarios (Table 1 in the Supporting Materials). We performed predictions for all data from the subsets, but in some cases, the evaluation metrics did not use all the data due to the specifics of their calculation. Thus, for some metrics, we applied only part of the interactions and had to eliminate some targets from the calculation. For example, the targets were excluded for:

– the calculation of R^2 when all interaction values in training or test subsets were the same (possessed the same value). In this case, we can not calculate the R^2. There are two

solutions: either excluding these targets from analysis or setting the R^2 score for them to 0. Instead of artificially setting R^2 to zero, which would lead to incorrect calculation of R^2_{median} and further prediction performance analyses, we preferred to exclude these targets from the analysis. This resulted in the use of 157 (for the subset "*i*") and 159 (for the subset "*c*") targets in the case of the pQSAR(159) dataset, and 2,975 (for the subset "*i*") and 4,204 (for the subset "*c*") targets in the case of the pQSAR(4276) dataset.

– the calculation of the ROC AUC and PR AUC scores when they possess data of only one class. These metrics cannot be defined when only one class is present. The only right way is to exclude such targets from the calculation of the ROC AUC and PR AUC scores. This resulted in the use of 62 (for the subset "*i*") and 64 (for the "subset "*c*") targets in the case of the ViralChEMBL dataset.

Table S1: Data subsets

		Number (proportion of the total number) of *						Density of training data, %
		*targets		*compounds		*interactions		
pQSAR(159)								
whole dataset		159	(1.0)	13,190	(1.0)	114,317	(1.0)	
Sc1	subset "trn"	159	(1.0)	10,134	(0.77)	85,675	(0.75)	5.32
	subset "c"	159	(1.0)	3,056	(0.23)	13,074	(0.11)	
	subset "i"	158	(0.99)	619	(0.05)	15,568	(0.14)	
Sc2	subset "trn"	159	(1.0)	10,134	(0.77)	101,243	(0.89)	6.28
	subset "c"	159	(1.0)	3,056	(0.23)	13,074	(0.11)	
	subset "i"	-	-	-	-	-	-	
Sc3	subset "trn"	159	(1.0)	13,190	(1.0)	98,749	(0.86)	4.71
	subset "c"	-	-	-	-	-	-	
	subset "i"	158	(1.0)	619	(0.05)	15,568	(0.14)	
*Sc4***	subset "trn"	159±0	(1.0±0.0)	13,173±149	(0.99±0.01)	106,787±2,969	(0.93±0.03)	5.1±0.17
	subset "c"	1±0	(0.006±0.0)	82±147	(0.006±0.011)	82±147	(0.001±0.0)	
	subset "i"	1±0	(0.006±0.0)	98±36	(0.007±0.003)	98±36	(0.001±0.001)	
pQSAR(4276)								
whole dataset		4,276	(1.0)	496,946	(1.0)	1,368,500	(1.0)	
Sc1	subset "trn"	4,276	(1.0)	409,253	(0.82)	1,024,751	(0.75)	0.06
	subset "c"	4,247	(0.99)	87,693	(0.18)	146,405	(0.11)	
	subset "i"	3,265	(0.76)	89,991	(0.18)	197,344	(0.14)	
Sc2	subset "trn"	4,276	(1.0)	409,253	(0.82)	1,222,095	(0.89)	0.07
	subset "c"	4,247	(0.99)	87,693	(0.18)	146,405	(0.11)	
	subset "i"	-	-	-	-	-	-	
Sc3	subset "trn"	4,276	(1.0)	496,946	(1.0)	1,171,156	(0.86)	0.06
	subset "c"	-	-	-	-	-	-	
	subset "i"	3,265	(0.76)	89,991	(0.18)	197,344	(0.14)	
*Sc4***	subset "trn"	4,276±0	(1.0±0.0)	443,983±18,648	(0.89±0.04)	1,093,758±45,595	(0.8±0.03)	0.06±0.0
	subset "c"	1±0	(0.0±0.0)	34±238	(0.0±0.0)	34±238	(0.0±0.0)	
	subset "i"	1±0	(0.0±0.0)	46±368	(0.0±0.001)	46±368	(0.0±0.0)	
ViralChEMBLE								
whole dataset		158	(1.0)	258,951	(1.0)	412,238	(1.0)	
Sc1	subset "trn"	158	(1.0)	200,712	(0.78)	309,207	(0.75)	0.98
	subset "c"	86	(0.54)	58,239	(0.22)	85,207	(0.21)	
	subset "i"	86	(0.54)	9,835	(0.04)	17,824	(0.04)	
Sc2	subset "trn"	158	(1.0)	200,712	(0.78)	327,031	(0.79)	1.03
	subset "c"	86	(0.54)	58,239	(0.22)	85,207	(0.21)	
	subset "i"	-	-	-	-	-	-	
Sc3	subset "trn"	158	(1.0)	258,951	(1.0)	394,414	(0.86)	0.96
	subset "c"	-	-	-	-	-	-	
	subset "i"	86	(0.54)	9,835	(0.04)	17824	(0.04)	
*Sc4***	subset "trn"	158±0	(1.0±0.0)	258,684±1,371	(0.99±0.01)	411,338±3,647	(0.99±0.02)	1.01±0.01
	subset "c"	1±0	(0.003±0.003)	539±2,626	(0.002±0.01)	539±2,629	(0.001±0.006)	
	subset "i"	1±0	(0.003±0.003)	112±467	(0.0±0.001)	113±468	(0.0±0.001)	

** For the Sc4 scenario, mean ± standard deviation were calculated across all targets.

Table S2: Hyperparameters of the selected models

Algorithm validation

Dataset	No. of neurones in layers	Dropout value	Epoch
pQSAR(159)	1024, 768, 512, 384	0.2, 0.3, 0.2, 0.2	250
ViralChEMBL	1024, 512, 256, 128	0.1, 0.4, 0.3, 0.2	140

Scenarios Sc1, Sc2, Sc3, Sc4

Dataset	No. of neurones in layers	Dropout value	Epoch
pQSAR(159)	1024, 768, 512, 384	0.2, 0.3, 0.2, 0.2	250
pQSAR(4276)	768, 512, 256, 128	0.2, 0.3, 0.2, 0.2	440
ViralChEMBL	1024, 512, 256, 128, 64	0.1, 0.4, 0.3, 0.2, 0.1	90

* Learning rate - 0.0001

Optimizer - RAdam

Activation function - ELU

For the y-scrambling tests, we applied the same hyperparameters as for the ordinary ones.

Table S3: Assessment of DNN-based algorithm on the *pQSAR(159)* dataset.

Evaluation metrics	*pQSAR(159)*					
	Test subset "i"			*Test subset "c"*		
	Sc1	*Sc3*	*Sc4*	*Sc1*	*Sc2*	*Sc4*
RMSE	0.74	0.69	0.61	1.02	1.01	1.02
RMSE 5%	0.66	0.58	0.48	0.75	0.69	0.78
RMSE 10%	0.62	0.56	0.50	0.70	0.67	0.68
$RMSE_{median}$	0.71	0.67	0.60	1.04	1.01	1.03
$RMSE_{mean}$±STD	0.71±0.19	0.68±0.17	0.60±0.16	1.00±0.23	0.98±0.24	0.99±0.23
R^2	0.40	0.47	0.59	0.03	0.07	0.04
R^2_{median}	0.30	0.37	0.53	-0.11	-0.04	-0.08
targets with R^2 >0.0	139*	146*	149*	32**	62**	46**
targets with R^2 >0.7	2*	2*	17*	0**	0**	0**
targets with R^2 >0.9	0*	0*	0*	0**	0**	0**

* - out of 157 targets, ** - out of 159 targets

Table S4: The absolute and relative (%) performance gain for the *pQSAR(159)* dataset.

Evaluation metrics (M)	*pQSAR(159)*							
	Test subset "i"				*Test subset "c"*			
	ΔM(*Sc1 – Sc3*)		ΔM(*Sc1 – Sc4*)		ΔM(*Sc1 – Sc2*)		ΔM(*Sc1 – Sc4*)	
RMSE	0.05	(6.8%)	0.13	(17.6%)	0.01	(1.0%)	0.00	(0.0%)
RMSE 5%	0.08	(12.1%)	0.18	(27.3%)	0.06	(8.0%)	-0.03	(-4.0%)
RMSE 10%	0.06	(9.7%)	0.12	(19.4%)	0.03	(4.3%)	0.02	(2.9%)
$RMSE_{median}$	0.04	(5.6%)	0.11	(15.5%)	0.03	(2.9%)	0.01	(1.0%)
$RMSE_{mean}$	0.03	(4.2%)	0.11	(15.5%)	0.02	(2.0%)	0.01	(1.0%)
R^2	0.07	(17.5%)	0.19	(47.5%)	0.04	(133.3%)	0.01	(33.3%)
R^2_{median}	0.07	(23.3%)	0.23	(76.7%)	0.07	(-63.6%)	0.03	(-27.3%)
targets with R^2 >0.0	7	(5.0%)	10	(7.2%)	30	(93.8%)	14	(43.8%)
targets with R^2 >0.7	0	(0.0%)	15	(750.0%)	0	(0.0%)	0	(0.0%)
targets with R^2 >0.9	0	(0.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)

Table S5: Assessment of DNN-based algorithm on the *pQSAR(4276)* dataset.

Evaluation metrics	*pQSAR(4276)*					
	Test subset "i"			*Test subset "c"*		
	Sc1	*Sc3*	*Sc4*	*Sc1*	*Sc2*	*Sc4*
RMSE	0.78	0.78	0.77	0.95	0.94	0.94
RMSE 5%	1.08	1.06	1.08	1.26	1.28	1.28
RMSE 10%	0.98	0.97	0.98	1.29	1.28	1.28
$RMSE_{median}$	0.63	0.62	0.62	0.90	0.89	0.89
$RMSE_{mean}$±STD	0.70±0.47	0.67±0.43	0.69±0.45	0.97±0.40	0.95±0.39	0.96±0.40
R^2	0.50	0.50	0.51	0.62	0.62	0.62
R^2_{median}	0.27	0.29	0.26	-0.04	-0.02	-0.05
targets with R^2 >0.0	2,081*	2,118*	2,081	1,879**	1,971**	1,880**
targets with R^2 >0.7	563*	584*	618*	27**	43**	45**
targets with R^2 >0.9	153*	153*	136*	0**	1**	2**

* - out of 2,975 targets, ** - out of 4,204 targets

Table S6: The absolute and relative (%) performance gain for the *pQSAR(4276)* dataset.

Evaluation metrics (M)	*pQSAR(159)*							
	Test subset "i"				*Test subset "c"*			
	ΔM(*Sc1 – Sc3*)		ΔM(*Sc1 – Sc4*)		ΔM(*Sc1 – Sc2*)		ΔM(*Sc1 – Sc4*)	
RMSE	0.0	(0.0%)	0.01	(1.3%)	0.01	(1.1%)	0.01	(1.1%)
RMSE 5%	0.02	(1.9%)	0.0	(0.0%)	-0.02	(-1.6%)	-0.02	(-1.6%)
RMSE 10%	0.01	(1.0%)	0.0	(0.0%)	0.01	(0.8%)	0.01	(0.8%)
$RMSE_{median}$	0.01	(1.6%)	0.01	(1.6%)	0.01	(1.1%)	0.01	(1.1%)
$RMSE_{mean}$	0.03	(4.3%)	0.01	(1.4%)	0.02	(2.1%)	0.01	(1.0%)
R^2	0.0	(0.0%)	0.01	(2.0%)	0.0	(0.0%)	0.0	(0.0%)
R^2_{median}	0.02	(7.4%)	-0.01	(-3.7%)	0.02	(-50.0%)	-0.01	(25.0%)
targets with R^2 >0.0	37	(1.8%)	0	(0.0%)	92	(4.9%)	1	(0.1%)
targets with R^2 >0.7	21	(3.7%)	55	(9.8%)	16	(59.3%)	18	(66.7%)
targets with R^2 >0.9	0	(0.0%)	-17	(-11.1%)	1	(-%)	2	(-%)

Table S7: Assessment of DNN-based algorithm on the *ViralChEMBL* dataset.

Evaluation metrics	*ViralChEMBL*					
	Test subset "i"			*Test subset "c"*		
	Sc1	*Sc3*	*Sc4*	*Sc1*	*Sc2*	*Sc4*
ROC AUC	0.94	0.95	0.96	0.86	0.87	0.86
ROC AUC_{median}	0.84	0.88	0.89	0.69	0.70	0.71
ROC AUC_{mean}±STD	0.77±0.20	0.81±0.19	0.79±0.24	0.65±0.20	0.67±0.22	0.66±0.22
targets with ROC AUC>0.8	36*	42*	43*	13**	17**	13**
BA	0.89	0.90	0.91	0.79	0.80	0.79
BA_{median}	0.67	0.80	0.80	0.62	0.67	0.66
BA_{mean}±STD	0.66±0.25	0.69±0.27	0.72±0.26	0.63±0.26	0.66±0.25	0.65±0.25
targets with BA>0.8	25***	39***	43***	21***	23***	23***

* - out of 62 targets, ** - out of 64 targets, *** - out of 86 targets

Table S8: The absolute and relative (%) performance gain for the *ViralChEMBL* dataset.

Evaluation metrics (M)	*ViralChEMBL*							
	Test subset "i"				*Test subset "c"*			
	ΔM(*Sc1 – Sc3*)		ΔM(*Sc1 – Sc4*)		ΔM(*Sc1 – Sc2*)		ΔM(*Sc1 – Sc4*)	
ROC AUC	0.01	(1.1%)	0.02	(2.1%)	0.01	(1.2%)	0.0	(0.0%)
ROC AUC$_{median}$	0.04	(4.8%)	0.05	(6.0%)	0.01	(1.4%)	0.02	(2.9%)
ROC AUC$_{mean}$	0.04	(5.2%)	0.02	(2.6%)	0.02	(3.1%)	0.01	(1.5%)
targets with ROC AUC>0.8	6	(16.7%)	7	(19.4%)	4	(30.8%)	0	(0.0%)
BA	0.01	(1.1%)	0.02	(2.2%)	0.01	(1.3%)	0.0	(0.0%)
BA$_{median}$	0.13	(19.4%)	0.13	(19.4%)	0.05	(8.1%)	0.04	(6.5%)
BA$_{mean}$	0.03	(4.5%)	0.06	(9.1%)	0.03	(4.8%)	0.02	(3.2%)
targets with BA>0.8	14	(56.0%)	18	(72.0%)	2	(9.5%)	2	(9.5%)

Table S9: Assessment of pQSAR and DNN-based algorithms on the *pQSAR(159)* dataset.

Evaluation metrics	*pQSAR(159)*		
	pQSAR	*DNN*	*Y-scrambling*
RMSE	0.64	0.69	0.99
RMSE 5%	0.60	0.58	0.93
RMSE 10%	0.56	0.54	0.93
RMSE$_{median}$	0.61	0.64	0.95
RMSE$_{mean}$±STD	0.61±0.15	0.63±0.16	0.95±0.24
R^2	0.51	0.53	0.01
R^2_{median}	0.46	0.53	-0.05
targets with R^2 >0.0	149*	152*	60*
targets with R^2 >0.7	10*	16*	0*
targets with R^2 >0.9	0*	0*	0*

* - out of 159 targets

Table S10: Assessment of SGIMC and DNN-based algorithms on the *ViralChEMBL* dataset.

Evaluation metrics	*ViralChEMBL*		
	SGIMC	*DNN*	*Y-scrambling*
ROC AUC	0.64	0.67	0.47
ROC AUC$_{median}$	0.64	0.64	0.47
ROC AUC$_{mean}$±STD	0.64±0.21	0.61±0.20	0.44±0.15
targets with ROC AUC>0.8	9*	3*	0*
BA	0.62	0.61	0.50
BA$_{median}$	0.65	0.62	0.50
BA$_{mean}$±STD	0.65±0.26	0.69±0.25	0.62±0.26
targets with BA>0.8	12**	17**	15**

* - out of 38 targets, ** - out of 56 targets