Similarity Boosted QSAR – Supporting information

Tobias Girschick,[*,†] Pedro R. Almeida,[‡] Stefan Kramer,[¶] and Jonna Stålring[§]

Technische Universität München, Institut für Informatik/I12, Boltzmannstr. 3, 85748 Garching b. München, Germany,

EngInMotion Ltd, Avenida Infante D. Henrique, n. 145, 3510-070 Viseu, Portugal,

Institut für Informatik, Johannes Gutenberg-Universität Mainz, Staudingerweg 9, 55128 Mainz, Germany, and

Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Pepparedsleden 1, 431 53 Mölndal, Sweden

E-mail: tobias.girschick@in.tum.de

$\mathscr{R}^{lit}$ reference compounds

Table S1: SMILES representation of hERG reference compounds.

1	CC(C)(C)C1=CC=C(C=C1)C(CCCN2CCC(CC2)C(C3=CC=CC=C3)(C4=CC=CC=C4)O)O
2	COC1CN(CCC1NC(=O)C2=CC(=C(C=C2OC)N)Cl)CCCOC3=CC=C(C=C3)F
3	COC1=CC=C(C=C1)CCN2CCC(CC2)NC3=NC4=CC=CC=C4N3CC5=CC=C(C=C5)F
4	CN(CCC1=CC=C(C=C1)NS(=O)(=O)C)CCOC2=CC=C(C=C2)NS(=O)(=O)C
5	C1CN(CCC1C2=CN(C3=C2C=C(C=C3)Cl)C4=CC=C(C=C4)F)CCN5CCNC5=O
6	CC1CN(CCN1)C2=C(C(=C3C(=C2)N(C=C(C3=O)C(=O)O)C4CC4)C)F

[*]To whom correspondence should be addressed
[†]Technische Universität München
[‡]EngInMotion
[¶]Universität Mainz
[§]AstraZeneca

Table S2: SMILES representation of AhR reference compounds.

1	C1=C2C(=CC(=C1Cl)Cl)OC3=CC(=C(C=C3O2)Cl)Cl
2	C1=CC(=C(C(=C1)Cl)C=NN=C(N)N)Cl
3	C1=CC=C2C3=C4C(=CC2=C1)C=CC5=C4C(=CC=C5)C=C3
4	C1=CC(=C(C=C1C2=CC(=C(C(=C2)Cl)Cl)Cl)Cl)Cl
5	CC1=C2CCC3=C2C(=CC4=C3C=CC5=CC=CC=C54)C=C1
6	C1=CC=C(C=C1)C2=CC(=O)C3=C(O2)C=CC4=CC=CC=C43
7	C1=CC=C2C(=C1)NC(=N2)C3=CSC=N3
8	CC1=CN=C(C(=C1OC)C)CS(=O)C2=NC3=C(N2)C=C(C=C3)OC
9	C1C(C(C2=CC=CC=C21)N)O
10	SC1=CC=CC=C1N
11	CN(C1=CC=CC=C1)N
12	C1=CC2=C(C=CC=C2N)C(=C1)N

Table S3: SMILES representation of ER reference compounds.

1	C[C@]12CCC3c4ccc(cc4CCC3C1CC[C@@H]2O)O
2	c1cc(ccc1c2c(c3ccc(cc3s2)O)C(=O)c4ccc(cc4)OCCN5CCCCC5)O
3	c1cc(ccc1c2cc3cc(cc(c3o2)Br)O)O
4	c1cc(ccc1C2C3=C(CCOc4c3ccc(c4)O)c5ccc(cc5O2)F)OCCN6CCCCCC6
5	CN(C)CCOc1ccc(cc1)[C@H]2[C@H](Sc3cc(ccc3O2)O)c4cccc(c4)O
6	c1cc(ccc1C2c3cc(ccc3Cc4c2c5ccc(cc5cc4)O)O)OCCN6CCCCC6
7	CC/C(=C(/CC)\c1ccc(cc1)O)/c2ccc(cc2)O

Table S4: SMILES representation of SRC-1 reference compounds.

1	CN(CCOc1ccc(cc1)C[C@@H](C(=O)O)Nc2ccccc2C(=O)c3ccccc3)c4ccccn4
2	c1ccc(cc1)NC(=O)c2cc(ccc2Cl)N(=O)=O
3	Cc1c(nc(o1)c2ccccc2)CCOc3ccc(cc3)CC4C(=O)NC(=O)S4
4	Cc1cccc(c1)C(=O)c2ccccc2NC(Cc3ccc(cc3)OCCN(C)c4nc5ccccc5o4)C(=O)O
5	COC(=O)C(Cc1ccc(cc1)OCCn2c3ccc(cc3sc2=O)C(=O)c4ccccc4)C(=O)O
6	Cc1c(nc(o1)c2cccs2)CCOc3ccc(cc3)CC(C)(C(=O)O)Oc4ccccc4
7	CN(CCOc1ccc(cc1)CC2C(=O)NC(=O)S2)c3ccccn3

Table S5: SMILES representation of THR reference compounds.

1	c1cc(c(cc1Oc2c(cc(cc2I)CC(=O)O)I)I)O
2	CC1(C2CCC(C1C2)NC(=O)c3cc(ccc3O)Oc4c(cc(cc4Cl)n5c(=O)[nH]c(=O)cn5)Cl)C
3	COc1ccccc1CCNC(=O)c2cc(ccc2O)Oc3c(cc(cc3Br)CC(=O)O)Br
4	c1ccc(cc1)C(CNC(=O)c2cc(ccc2O)Oc3c(cc(cc3Br)CC(=O)O)Br)c4ccccc4
5	c1cc(c(cc1Oc2c(cc(cc2Cl)n3c(=O)[nH]c(=O)cn3)Cl)C(=O)N4CCOCC4)O

Table S6: SMILES representation of KCNQ2 reference compounds.

1	c1ccc2c(c1)C(=O)c3ccccc3C2(Cc4ccncc4)Cc5ccncc5
2	CCN1CCOc2c1cc(cc2)C(C)NC(=O)/C=C/c3ccccc3F
3	CC(c1ccc(c(c1)N2CCOCC2)F)NC(=O)/C=C/c3ccc(cc3)F
4	CC(c1ccc2c(c1)OCO2)NC(=O)/C=C/c3ccccc3Cl
5	CCN1CCCc2c1cc(cc2)C(C)NC(=O)/C=C/c3cc(ccc3F)F

Table S7: SMILES representation of M1 reference compounds.

1	c1ccc(cc1)C(c2ccccc2)([C@@H]3CCN(C3)CCc4ccc5c(c4)CCO5)C(=O)N
2	c1cc2c(cc1)Sc3c(cccc3)N2C[C@H]4[C@H]5CCN(C4)CC5
3	CC(C)N(CCC(c1ccccc1)c2cc(ccc2O)CO)C(C)C
4	c1ccc(cc1)c2ccoc2C3=CN4CCC3CC4
5	c1ccc2c(c1)C(=O)Nc3cccnc3N2C(=O)CN4CCNCC4
6	C[N+]1(C2CCC1CC(C2)CC(CO)(c3ccccc3)c4ccccc4)C
7	CCCCCCSc1c(nccn1)O[C@@H]2CN3CCC2C3
8	CN1CCN(CC1)C2=Nc3cc(ccc3Nc4c2cccc4)Cl
9	Cc1ccc(c(c1)[C@@H](CCN(C(C)C)C(C)C)c2ccccc2)O
10	c1ccc(cc1)C(c2ccccc2)(C(=O)OC3CC4CCC(C3)[N+]45CCCC5)O
11	CCN(CC)CC#CCOC(=O)C(c1ccccc1)(C2CCCCC2)O
12	CN1CCC=C(C1)c2c(nsn2)OCCCCCCCOc3c(nsn3)C4=CCCN(C4)C
13	CN1[C@@H]2CC[C@H]1CC(C2)OC(=O)C(CO)c3ccccc3
14	CC(CC(=O)O)N1CCC(=C2c3ccccc3OCc4c2ccc(c4)F)CC1

Additional Result Tables

Table S8: Example of CPU runtimes in seconds. Shown are ten-fold cross-validation running times for the descriptor set $SD(\mathscr{S}_{ALL}, \mathscr{R}^{act})$

Dataset	RF	SVM
hERG	34737	30713
AhR	143038	558594
ER	9263	7528
SRC-1	6921	5362
THR	5456	5637
KCNQ2	26322	57498
M1	6103	5167

Table S9: RandomForest (RF) ten-fold cross-validation results for using single similarity measures and their combination. The reference set is always $\mathscr{R}^{lit}$.

	Random Forests					
Dataset	$\{s_{MK}\}$	$\{s_{topo}\}$	$\{s_{ECFP}\}$	$\{s_{FCFP}\}$	$\{s_{AP}\}$	$\mathscr{S}_{ALL}$
hERG	0.592	0.551	0.566	0.566	0.566	0.613
AhR	0.738	0.632	0.739	0.711	0.717	0.772
ER	0.627	0.645	0.672	0.662	0.661	0.711
SRC-1	0.625	0.599	0.625	0.627	0.596	0.696
THR	0.537	0.572	0.583	0.569	0.571	0.650
KCNQ2	0.589	0.594	0.625	0.650	0.632	0.686
M1	0.653	0.592	0.653	0.625	0.608	0.653

Table S10: Support Vector Machine (SVM) ten-fold cross-validation results for using single similarity measures and their combination. The reference set is always $\mathscr{R}^{lit}$.

	SVM					
Dataset	$\{s_{MK}\}$	$\{s_{topo}\}$	$\{s_{ECFP}\}$	$\{s_{FCFP}\}$	$\{s_{AP}\}$	$\mathscr{S}_{ALL}$
hERG	0.596	0.548	0.577	0.568	0.580	0.602
AhR	0.701	0.618	0.725	0.698	0.725	0.772
ER	0.640	0.663	0.698	0.676	0.676	0.725
SRC-1	0.640	0.624	0.614	0.629	0.638	0.694
THR	0.527	0.601	0.564	0.580	0.607	0.633
KCNQ2	0.597	0.597	0.664	0.653	0 640	0.697
M1	0.646	0.588	0.638	0.649	0.653	0.680

Table S11: Statistical significance analysis of single similarities and their combination. Null hypothesis is that the combination is not better than a single similarity. Shown are mean accuracy values ± standard deviations over 100 hold-out runs[a]. The reference set is always $\mathscr{R}^{lit}$.

	Random Forests					
Dataset	$\{s_{MK}\}$	$\{s_{AP}\}$	$\{s_{FCFP}\}$	$\{s_{ECFP}\}$	$\{s_{topo}\}$	$\mathscr{S}_{ALL}$
hERG	0.587±0.011	0.565±0.013	0.559±0.012	0.559±0.013	0.551±0.012	0.608±0.013 •
AhR	0.735±0.004	0.736±0.005	0.705±0.005	0.714±0.005	0.629±0.005	0.766±0.004 •
ER	0.627±0.015	0.668±0.013	0.659±0.015	0.663±0.015	0.646±0.014	0.708±0.013 •
SRC-1	0.614±0.019	0.615±0.018	0.613±0.018	0.583±0.017	0.600±0.015	0.681±0.018 •
THR	0.538±0.019	0.573±0.017	0.571±0.017	0.564±0.017	0.563±0.020	0.631±0.017 •
KCNQ2	0.587±0.009	0.627±0.009	0.650±0.007	0.623±0.009	0.589±0.008	0.686±0.009 •
M1	0.646±0.018	0.645±0.018	0.624±0.018	0.611±0.018	0.581±0.020	0.651±0.017 •

•/○ statistically significant improvement in all five cases/some of the five cases.

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means. [1]

Table S12: Statistical significance analysis of BBRC and the variants for finding reference molecules. Null hypothesis is, that there is no improvement compared to column BBRC. Shown are mean accuracy values ± standard deviations over 100 hold-out runs[a]. $\mathscr{S} = \mathscr{S}_{ALL}$.

	Random Forests			
Dataset	BBRC	$\mathscr{R}^{lit}$	$\mathscr{R}^{act}$	$\mathscr{R}^{db}$
hERG	0.633±0.014	0.608±0.013 ○	0.634±0.015	0.633±0.013
AhR	0.782±0.005	0.766±0.004 ○	0.779±0.008 ○	**± **
ER	0.710±0.014	0.708±0.013	0.731±0.014 •	0.722±0.013 •
SRC-1	0.720±0.017	0.681±0.018 ○	0.728±0.018 •	0.725±0.020 •
THR	0.650±0.017	0.631±0.017 ○	0.669±0.016 •	0.636±0.020 ○
KCNQ2	0.677±0.009	0.686±0.009 •	0.738±0.010 •	0.727±0.009 •
M1	0.645±0.022	0.651±0.017 •	0.669±0.017 •	0.663±0.020 •

•/○ statistically significant improvement/degradation wrt. column BBRC

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means. [1]

Table S13: Statistical significance analysis of $ECFP_{r1}$ and the variants for finding reference molecules. Null hypothesis is, that there is no improvement compared to column $ECFP_{r1}$. Shown are mean accuracy values ± standard deviations over 100 hold-out runs[a]. $\mathscr{S} = \mathscr{S}_{ALL}$.

	Random Forests			
Dataset	$ECFP_{r1}$	$\mathscr{R}^{lit}$	$\mathscr{R}^{act}$	$\mathscr{R}^{db}$
hERG	0.661±0.012	0.608±0.013 ∘	0.634±0.015 ∘	0.633±0.013 ∘
AhR	0.792±0.015	0.766±0.004	0.779±0.008	**± **
ER	0.749±0.014	0.708±0.013 ∘	0.731±0.014 ∘	0.722±0.013 ∘
SRC-1	0.770±0.016	0.681±0.018 ∘	0.728±0.018 ∘	0.725±0.020 ∘
THR	0.680±0.018	0.631±0.017 ∘	0.669±0.016 ∘	0.636±0.020 ∘
KCNQ2	0.763±0.009	0.686±0.009 ∘	0.738±0.010 ∘	0.727±0.009 ∘
M1	0.679±0.021	0.651±0.017 ∘	0.669±0.017 ∘	0.663±0.020 ∘

•/∘ statistically significant improvement/degradation wrt. column ECFP

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means. [1]

Table S14: Mean sensitivity (*SN*) and specificity (*SP*) values including standard deviations over 100 hold-out runs[a]. Δ_{SN-SP} is the difference of sensitivity and specificity.

	Random Forests					
Dataset	BBRC			$ECFP_{r1}$		
	SN	*SP*	Δ_{SN-SP}	*SN*	*SP*	Δ_{SN-SP}
hERG	0.626±0.020	0.642±0.020	-0.016	0.619±0.012	0.699±0.041	-0.080
AhR	0.779±0.006	0.787±0.011	-0.008	0.788±0.010	0.797±0.020	-0.009
ER	0.696±0.018	0.726±0.020	-0.030	0.739±0.020	0.760±0.023	-0.021
SRC-1	0.713±0.023	0.730±0.026	-0.017	0.755±0.020	0.789±0.027	-0.034
THR	0.645±0.028	0.657±0.024	-0.012	0.677±0.027	0.685±0.030	-0.008
KCNQ2	0.679±0.012	0.674±0.010	0.005	0.764±0.011	0.760±0.013	0.004
M1	0.641±0.029	0.649±0.028	-0.008	0.676±0.032	0.684±0.031	-0.008

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means. [1]

Table S15: Mean sensitivity and specificity values including standard deviations over 100 hold-out runs[a]. Δ_{SN-SP} is the difference of sensitivity and specificity. $\mathscr{S} = \mathscr{S}_{ALL}$.

	Random Forests					
Dataset	$\mathscr{R}^{lit}$			$\mathscr{R}^{act}$		
	SN	*SP*	Δ_{SN-SP}	*SN*	*SP*	Δ_{SN-SP}
hERG	0.609±0.013	0.608±0.020	0.001	0.621±0.021	0.651±0.024	-0.030
AhR	0.787±0.008	0.747±0.007	0.040	0.777±0.009	0.782±0.006	-0.005
ER	0.698±0.020	0.719±0.019	-0.021	0.727±0.020	0.736±0.024	-0.009
SRC-1	0.682±0.028	0.681±0.025	0.001	0.729±0.025	0.728±0.026	0.001
THR	0.636±0.031	0.627±0.025	0.009	0.665±0.011	0.673±0.013	-0.008
KCNQ2	0.686±0.012	0.686±0.011	0.000	0.727±0.021	0.744±0.013	-0.017
M1	0.653±0.028	0.650±0.028	0.003	0.666±0.028	0.674±0.028	-0.008

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means. [1]

Table S16: Mean sensitivity and specificity values including standard deviations over 100 hold-out runs[a]. Δ_{SN-SP} is the difference of sensitivity and specificity. $\mathscr{S} = \mathscr{S}_{ALL}$.

	Random Forests		
Dataset	$\mathscr{R}^{db}$		
	SN	*SP*	Δ_{SN-SP}
hERG	0.624±0.019	0.643±0.024	-0.019
AhR	**± **	**± **	**
ER	0.715±0.018	0.731±0.023	-0.016
SRC-1	0.721±0.031	0.731±0.027	-0.010
THR	0.641±0.029	0.633±0.029	0.008
KCNQ2	0.738±0.015	0.717±0.012	0.021
M1	0.660±0.032	0.669±0.030	-0.009

[a]Please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means.[1]

Diversity measures

The used measures of classifier diversity are based on the cross-classification table (see Table S17) and defined as listed in Kuncheva and Whitaker:[2]

Table S17: Cross-classification table

		BBRC		
		correct	incorrect	
SD	correct	a	b	$(a+b)$
	incorrect	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	$n=a+b+c+d$

Yule's Q:

$$Q = \frac{ad-bc}{ad+bc} \tag{1}$$

Correlation Coefficient ρ:

$$\rho = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \tag{2}$$

The double-fault measure DF:

$$DF = \frac{d}{n} \tag{3}$$

References

(1) Cumming, G.; Fidler, F.; Vaux, D. Error bars in experimental biology. *Journal of Cell Biology* **2007**, *177*, 7–11.

(2) Kuncheva, L.; Whitaker, C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* **2003**, *51*, 181–207.