Electronic Supplementary Material (ESI) for Toxicology Research.

Supporting information

Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data

Fredrik Svensson[1], Ulf Norinder[2,3], Andreas Bender[1]

[1] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

[2] Swedish Toxicology Sciences Research Center, SE-151 36 Södertälje, Sweden

[3] Dept. Computer and Systems Sciences, Stockholm Univ., Box 7003, SE-164 07 Kista, Sweden

Contents

List of RDKit descriptors used in this study

'Chi0' , 'Chi0n' , 'Chi0v' , 'Chi1' , 'Chi1n' , 'Chi1v' , 'Chi2n' , 'Chi2v' , 'Chi3n' , 'Chi3v' , 'Chi4n' , 'Chi4v' , 'EState_VSA1' , 'EState_VSA10' , 'EState_VSA11' , 'EState_VSA2' , 'EState_VSA3' , 'EState_VSA4' , 'EState_VSA5' , 'EState_VSA6' , 'EState_VSA7' , 'EState_VSA8' , 'EState_VSA9' , 'FractionCSP3' , 'HallKierAlpha' , 'HeavyAtomCount' , 'Ipc' , 'Kappa1' , 'Kappa2' , 'Kappa3' , 'LabuteASA' , 'MolLogP' , 'MolMR' , 'MolWt' , 'NHOHCount' , 'NOCount' , 'NumAliphaticCarbocycles' , 'NumAliphaticHeterocycles' , 'NumAliphaticRings' , 'NumAromaticCarbocycles' , 'NumAromaticHeterocycles' , 'NumAromaticRings' , 'NumHAcceptors' , 'NumHDonors' , 'NumHeteroatoms' , 'NumRotatableBonds' , 'NumSaturatedCarbocycles' , 'NumSaturatedHeterocycles' , 'NumSaturatedRings' , 'PEOE_VSA1' , 'PEOE_VSA10' , 'PEOE_VSA11' , 'PEOE_VSA12' , 'PEOE_VSA13' , 'PEOE_VSA14' , 'PEOE_VSA2' , 'PEOE_VSA3' , 'PEOE_VSA4' , 'PEOE_VSA5' , 'PEOE_VSA6' , 'PEOE_VSA7' , 'PEOE_VSA8' , 'PEOE_VSA9' , 'RingCount' , 'SMR_VSA1' , 'SMR_VSA10' , 'SMR_VSA2' , 'SMR_VSA3' , 'SMR_VSA4' , 'SMR_VSA5' , 'SMR_VSA6' , 'SMR_VSA7' , 'SMR_VSA8' , 'SMR_VSA9' , 'SlogP_VSA1' , 'SlogP_VSA10' , 'SlogP_VSA11' , 'SlogP_VSA12' , 'SlogP_VSA2' , 'SlogP_VSA3' , 'SlogP_VSA4' , 'SlogP_VSA5' , 'SlogP_VSA6' , 'SlogP_VSA7' , 'SlogP_VSA8' , 'SlogP_VSA9' , 'TPSA' , 'VSA_EState1' , 'VSA_EState10' , 'VSA_EState2' , 'VSA_EState3' , 'VSA_EState4' , 'VSA_EState5' , 'VSA_EState6' , 'VSA_EState7' , 'VSA_EState8' , 'VSA_EState9'

Modeling results using random forest and 100 trees.

Table S1. Validity for RDKit model using 100 trees at different confidence levels. The results display a strong tendency towards over conservative predictions.

Conf. lvl	70		75		80		85		90	
AID	Toxic	Non-toxic	Toxic	Non-toxic	Toxic	Non-toxic	Toxic	Non-toxic	Toxic	Non-toxic
463	0.761	0.751	0.827	0.805	0.865	0.850	0.918	0.896	0.972	0.937
1486	0.724	0.784	0.785	0.825	0.857	0.873	0.932	0.909	0.993	0.943
1825	0.743	0.778	0.803	0.828	0.866	0.866	0.933	0.909	0.988	0.942
598	0.724	0.714	0.778	0.765	0.829	0.814	0.880	0.864	0.929	0.911
648	0.741	0.755	0.795	0.798	0.844	0.841	0.892	0.884	0.957	0.923
719	0.760	0.745	0.800	0.800	0.858	0.848	0.912	0.887	0.966	0.927
847	0.856	0.790	0.948	0.844	0.995	0.887	1	0.920	1	0.950
903	0.775	0.725	0.760	0.827	0.828	0.870	0.902	0.900	0.988	0.934
504648	0.902	0.703	0.822	0.924	0.940	0.937	0.985	0.947	0.998	0.956
588856	0.733	0.777	0.790	0.827	0.847	0.865	0.924	0.899	0.990	0.935
624418	0.765	0.908	0.891	0.931	0.958	0.945	0.994	0.955	1.000	0.963
430	0.731	0.740	0.786	0.787	0.839	0.834	0.902	0.879	0.946	0.921
620	0.745	0.796	0.835	0.837	0.934	0.886	0.986	0.923	1	0.947
602141	0.739	0.825	0.793	0.848	0.869	0.870	0.958	0.915	0.995	0.945
2275	0.694	0.791	0.751	0.815	0.824	0.848	0.881	0.898	1	0.931
2717	0.736	0.756	0.787	0.800	0.842	0.838	0.893	0.882	0.944	0.922

Correlation between RDKit descriptors and assay outcome

For each dataset we list the ten descriptors with highest absolute Person correlation to the assay outcome.

AID 463
MolLogP 0.053
NumAromaticRings 0.049
FractionCSP3 -0.047
NumAromaticCarbocycles 0.040
NumSaturatedHeterocycles -0.033
HallKierAlpha -0.030
NumAliphaticRings -0.027
EState_VSA2 -0.027
NumAliphaticHeterocycles -0.023
NumAromaticHeterocycles 0.023

AID 1486
EState_VSA8 0.067
NumHeteroatoms -0.045
NOCount -0.043
NumHAcceptors -0.041
EState_VSA2 -0.039
EState_VSA10 -0.038
MolLogP 0.037
PEOE_VSA2 -0.035
SMR_VSA3 -0.032
EState_VSA3 -0.028

AID 598
MolLogP 0.171
NumAromaticRings 0.151
RingCount 0.134
MolMR 0.128
SMR_VSA7 0.126
HallKierAlpha -0.122
NumAromaticCarbocycles 0.121
Chi1 0.120
LabuteASA 0.119
HeavyAtomCount 0.118

AID 648
MolLogP 0.092
MolMR 0.066
RingCount 0.064
NumAromaticRings 0.061
NumAromaticCarbocycles 0.061
LabuteASA 0.061
Chi1 0.060
HeavyAtomCount 0.059
MolWt 0.059
SMR_VSA7 0.058

AID 719
MolLogP 0.072
NumAromaticRings 0.056

SMR_VSA7 0.051
NumAromaticCarbocycles 0.050
RingCount 0.049
MolMR 0.048
MolWt 0.044
LabuteASA 0.044
Chi1 0.043
HeavyAtomCount 0.043

AID 847
MolWt 0.026
NumHeteroatoms 0.026
Chi0 0.025
HeavyAtomCount 0.024
Chi1 0.023
LabuteASA 0.023
Chi0v 0.023
Chi1v 0.023
MolMR 0.023
Kappa1 0.022

AID 903
Kappa1 -0.064
Kappa2 -0.061
Chi0 -0.058
Chi0v -0.058
MolWt -0.058
Chi0n -0.056
LabuteASA -0.056
HeavyAtomCount -0.054
Chi1 -0.053
Chi1n -0.052

AID 504648
EState_VSA8 0.028
SMR_VSA4 0.020
Chi4n 0.019
Chi3n 0.018
Chi2n 0.017
MolMR 0.016
Chi0 0.016
HallKierAlpha -0.016
HeavyAtomCount 0.016
EState_VSA1 0.015

AID 588856
MolLogP 0.054
Estate_VSA8 0.037
PEOE_VSA6 0.037
SMR_VSA7 0.031
NumAromaticCarbocycles 0.030
MolMR 0.024

NumAromaticRings 0.023
Estate_VSA4 0.021
FractionCSP3 0.020
LabuteASA 0.020

AID 624418
EState_VSA8 0.018
EState_VSA2 -0.015
EState_VSA10 -0.015
EState_VSA9 0.013
NOCount -0.013
EState_VSA3 -0.012
FractionCSP3 -0.012
NumRotatableBonds -0.012
MolLogP 0.011
EState_VSA6 -0.010

AID 430
MolLogP 0.108
NumAromaticRings 0.080
NumAromaticCarbocycles 0.061
MolMR 0.059
MolWt 0.056
LabuteASA 0.055
Ipc 0.055
HallKierAlpha -0.055
Chi1 0.054
FractionCSP3 -0.054

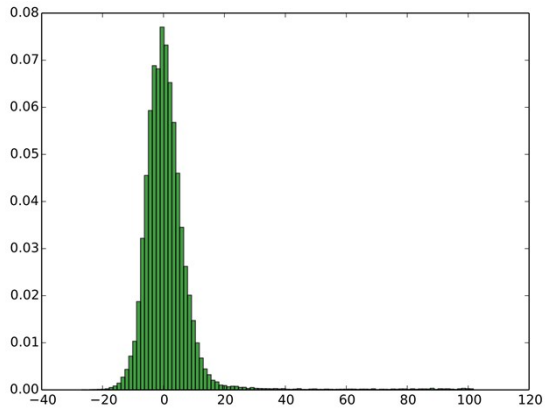
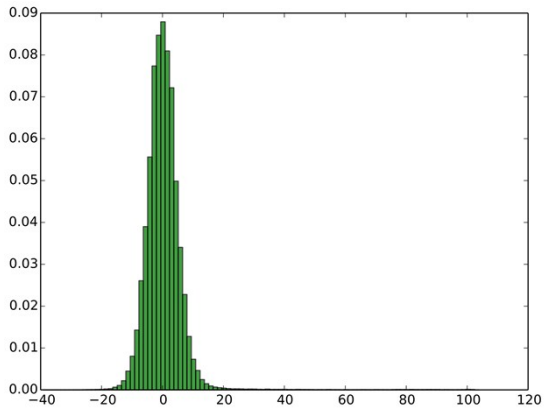
AID 620
MolLogP 0.038
NumAromaticCarbocycles 0.027
FractionCSP3 -0.027
NumAromaticRings 0.022
EState_VSA9 0.018
EState_VSA7 0.016
NumSaturatedHeterocycles -0.015
HallKierAlpha -0.014
EState_VSA2 -0.014
MolMR 0.013

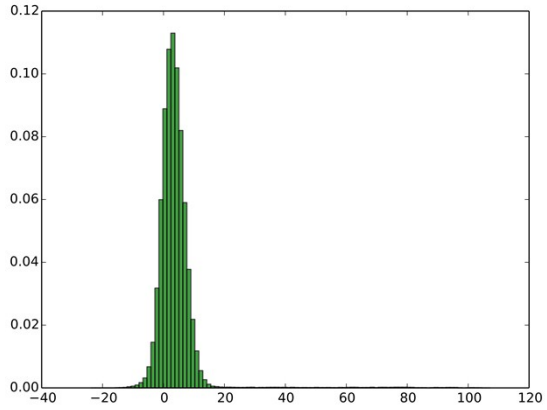
AID 602141
MolLogP 0.038
NumAromaticRings 0.026
SMR_VSA7 0.024
SMR_VSA10 0.023
NumAliphaticCarbocycles 0.022
Chi4n 0.022
FractionCSP3 -0.021
MolMR 0.021
LabuteASA 0.021
MolWt 0.021

AID 2275
NumSaturatedRings 0.155
NumAliphaticRings 0.131
NumAliphaticCarbocycles 0.128
SMR_VSA5 0.120
NumSaturatedHeterocycles 0.120
NumSaturatedCarbocycles 0.116
Chi4n 0.106
SMR_VSA4 0.105
EState_VSA1 0.104
Chi3n 0.081

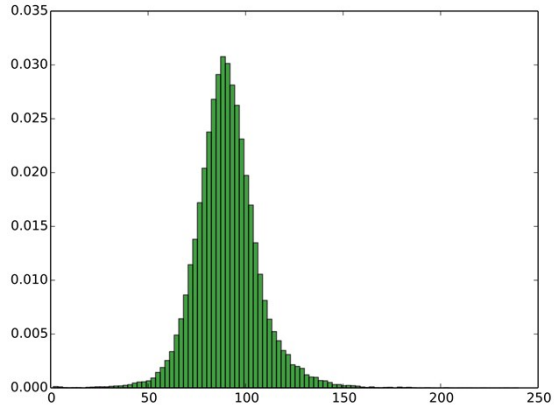
AID 2717
MolLogP 0.107
PEOE_VSA6 0.064
NumAromaticCarbocycles 0.060
SMR_VSA7 0.058
NumAromaticRings 0.058
EState_VSA8 0.055
MolMR 0.055
LabuteASA 0.048
MolWt 0.046
Chi0v 0.044

Distribution of assay data

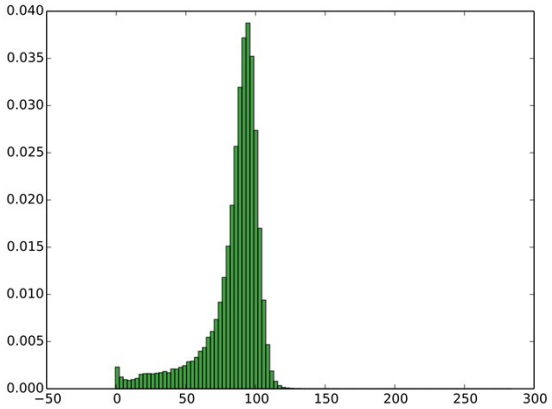
AID 463	**AID 1825**
	
AID 1486	**AID 847**



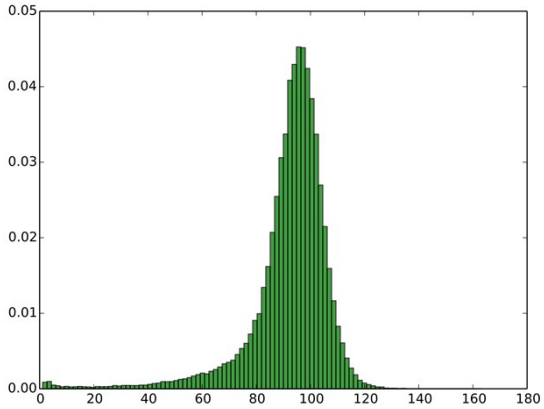
AID 598



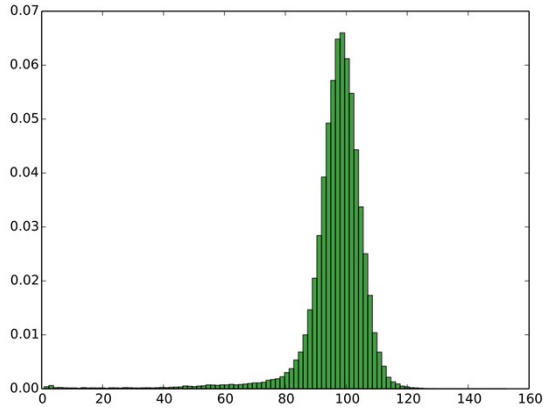
AID 648



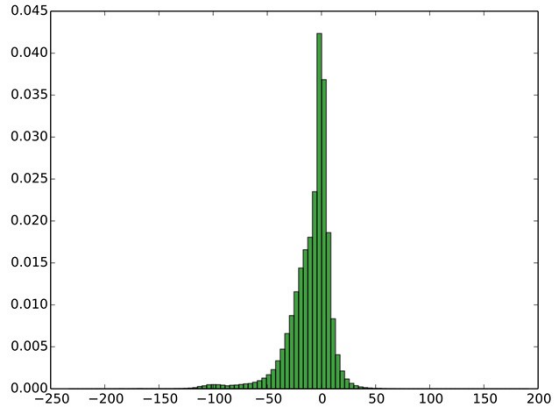
AID 719



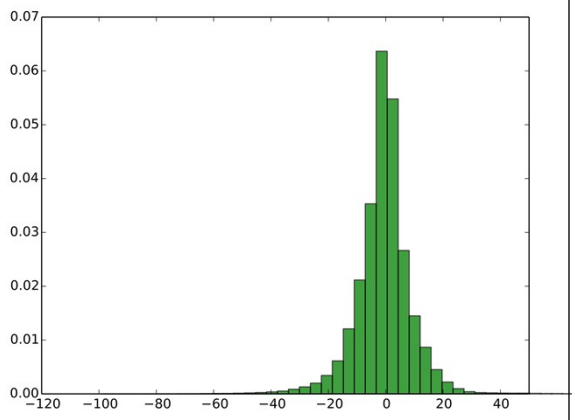
AID 588856



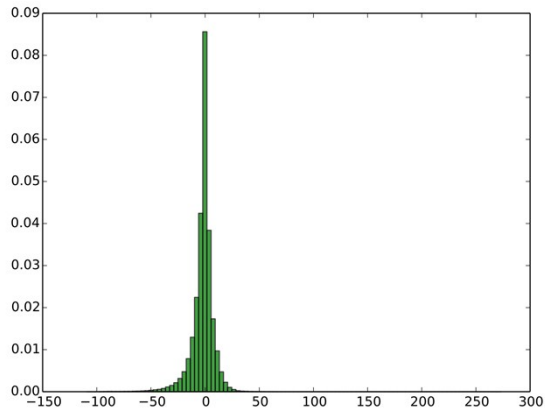
AID 624418



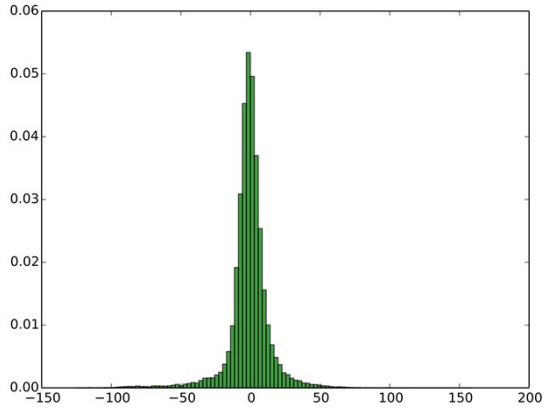
AID 504648



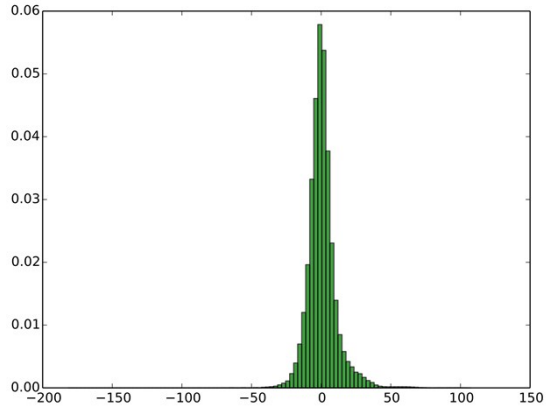
AID 903



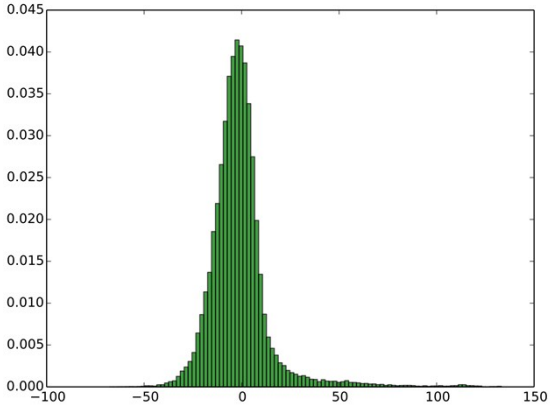
AID 620



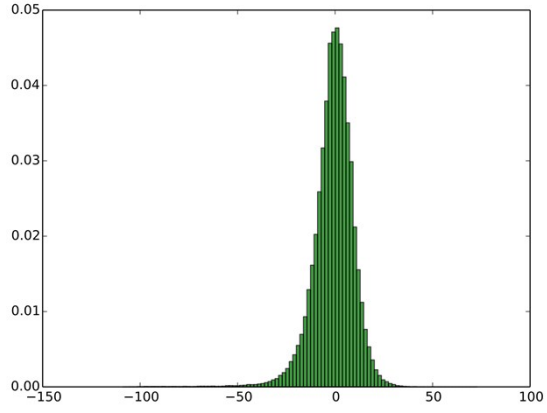
AID 430



AID 602141



AID 2717



AID 2275

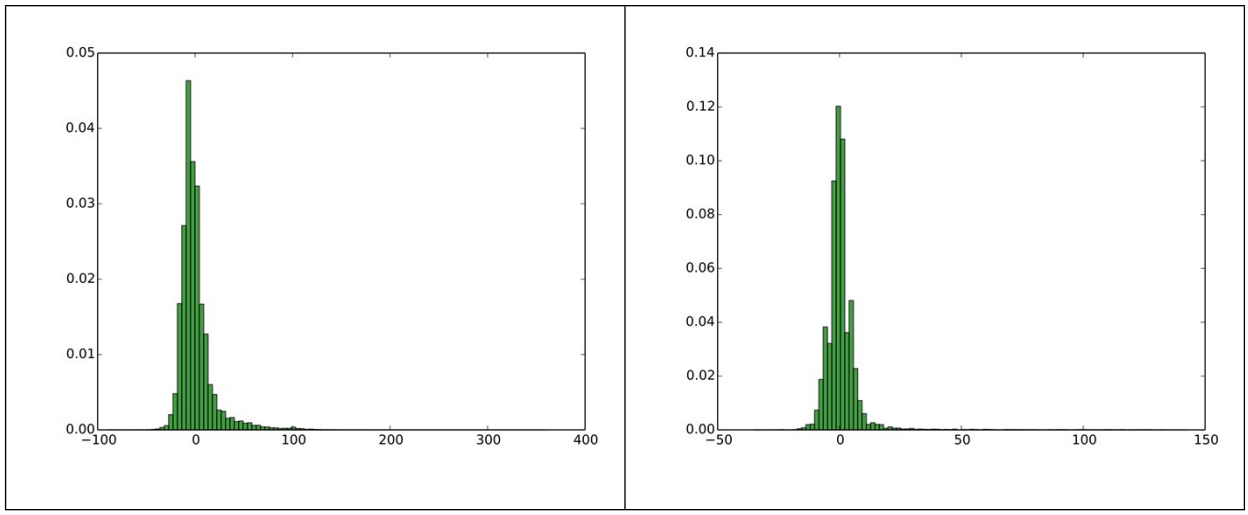


Figure S1. Distribution of raw assay data. Either measured as percent inhibition or percent cell viability depending on assay

Molprint2D Results

Table S5. Accuracy on the single class predictions and coverage for the models built using Molprint2D descriptors at 70 and 80 % confidence levels.

	70 %				**80 %**			
AID	Accuracy non-toxic	Coverage non-toxic	Accuracy toxic	Coverage toxic	Accuracy non-toxic	Coverage non-toxic	Accuracy toxic	Coverage toxic
463	77.2	92.4	79.3	95.2	77.4	81.7	83.5	87.4
1486	72.8	98.2	73.1	98.7	74.3	68.1	79.2	82.4
1825	84.8	86.0	82.7	89.8	80.7	90.6	83.2	93.6
598	75.7	92.8	78.8	94.1	77.3	84.5	81.2	86.8
648	80.1	89.5	80.3	91.6	79.1	87.3	81.6	90.2
719	75.1	95.9	76.5	96.4	76.8	76.7	82.1	84.0
847	47.0	49.3	67.9	56.2	14.4	19.2	94.3	36.1
903	91.9	80.5	86.7	84.6	83.8	97.7	84.9	98.2
504648	90.5	94.6	77.8	94.7	87.3	85.0	81.6	92.5
588856	79.0	92.7	77.8	94.5	78.6	81.1	81.9	87.5

624418	87.8	95.4	72.6	95.4	81.0	53.4	85.9	77.1
430	77.0	92.2	79.5	91.8	77.8	84.2	81.3	87.4
620	74.3	96.0	74.4	95.6	73.8	60.5	86.5	73.4
602141	85.6	89.1	80.1	92.1	82.8	88.9	81.3	93.2
2275	86.2	83.2	84.7	81.3	80.4	96.3	80.9	97.4
2717	84.3	85.0	84.9	87.5	80.4	94.9	83.1	96.3

Number of trees and model validity

Previous studies applying conformal prediction have used RFs consisting of 100 trees with good results.[S1,S2] However, when applied to the data sets in this study the resulting validities were over-conservative, i.e. the validity is above the set confidence level (Table S1). This has previously been observed for aggregated conformal predictors as well as being reported as an effect when the number of calibration examples is small, which may sometimes be the case for the minority class in this study.[S3-S5] Too conservative models may pose a problem since the increase in validity can generate a decrease in the efficiency of the classifier. Using the data from AID 2275 we investigated how the number of trees affected the validity of the predictions, the results are shown in Figure S1. When using only 20 trees the validity was 95.1 % but as the number of trees increases the model validity approaches the set confidence level of 80 % with a validity of 80.6 % for 1000 trees. This was largely due to fewer and fewer compounds being assigned to the *both* class as the number of trees increased. Based on the results we decided on using 500 trees as the model had a validity of 81.3 % and this was deemed an appropriate balance between validity and computational time.

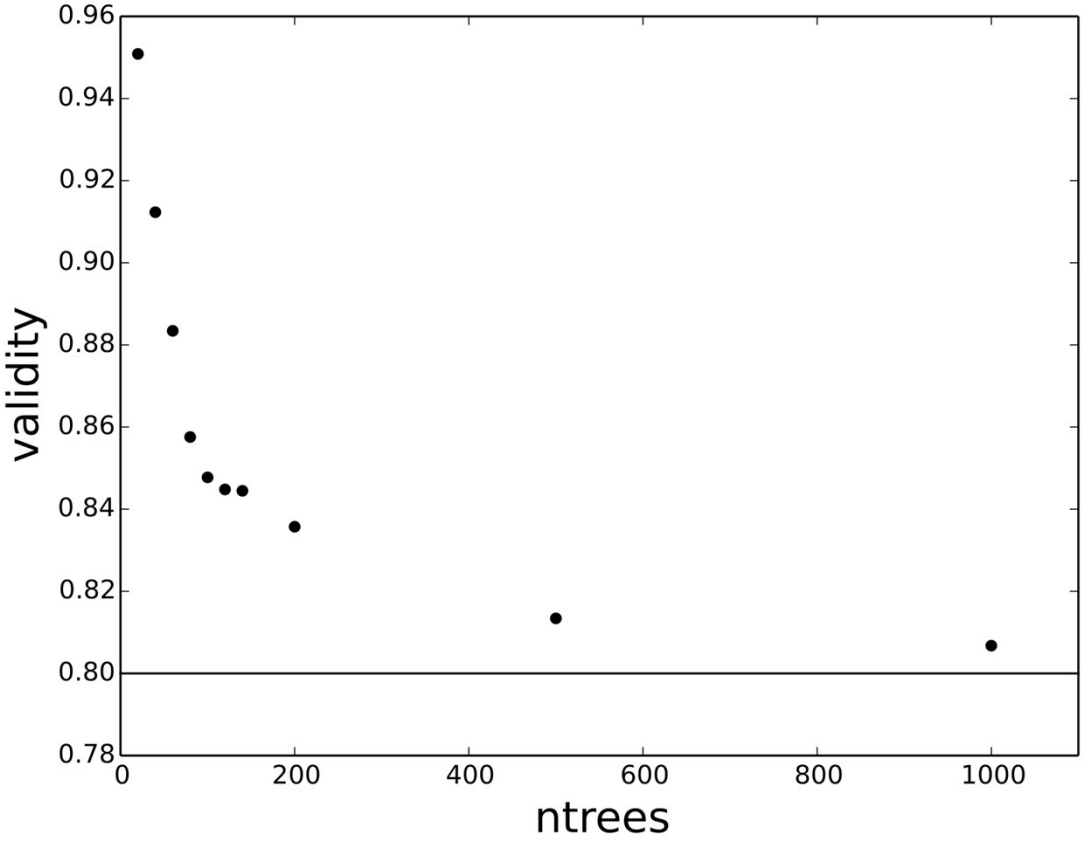


Figure S1. Validity of the predictions for AID 2275 at 80 % confidence level using different numbers of trees in the RF. The solid line indicates target performance for the 80 % confidence level. As the number of trees increases the obtained validity approaches the set confidence level.

S1. U. Norinder, L. Carlsson, S. Boyer and M. Eklund, Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination, J. Chem. Inf. Model., 2014, 54, 1596–1603.

S2. U. Norinder, L. Carlsson, S. Boyer and M. Eklund, Introducing conformal prediction in predictive modeling for regulatory purposes. A transparent and flexible alternative to applicability domain determination, Regul. Toxicol. Pharmacol., 2015, 71, 279–284.

S3. L. Carlsson, M. Eklund and U. Norinder, Aggregated Conformal Prediction, in Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings, ed. L. Iliadis, I. Maglogiannis, H. Papadopoulos, S. Sioutas and C. Makris, Springer International Publishing, Berlin, Heidelberg, 2014, pp. 231–240.

S4. L. Carlsson, E. Ahlberg, H. Boström, U. Johansson and H. Linusson, Modifications to p-Values of Conformal Predictors, in Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 2023, 2015, Proceedings, ed. A. Gammerman, V. Vovk and H. Papadopoulos, Springer International Publishing, Cham, 2015, pp. 251–259

S5. U. Johansson, E. Ahlberg, H. Boström, L. Carlsson, H. Linusson and C. Sönströd, Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors, in Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings, ed. A. Gammerman, V. Vovk and H. Papadopoulos, Springer International Publishing, Cham, 2015, pp. 271–280.

.

Overlapping compounds between the different data sets

Table S2. Overlap of tested compounds between the different data sets.

AID	463	1825	1486	847	598	648	719	588856	624418	504648	903	620	430	602141	2717	2275
463	**56485**	54869	54833	34230	52080	52938	51795	54809	53153	41378	2541	55173	54006	54582	49761	961
1825	54869	**290731**	217964	40133	83541	84474	83224	286272	271744	217456	13365	84204	60798	285490	265906	3736
1486	54833	217964	**217964**	40133	83524	84457	83207	213872	202361	161460	11512	84145	60762	212981	198860	3167
847	34230	40133	40133	**41169**	37690	38448	37788	39769	38546	29777	1939	39944	38872	39564	36047	801
598	52080	83541	83524	37690	**85201**	85201	83921	82996	80074	62647	7530	80733	57615	82658	75657	1578
648	52938	84474	84457	38448	85201	**86160**	84880	83924	80959	63337	7559	81687	58523	83583	76498	1592
719	51795	83224	83207	37788	83921	84880	**84880**	82687	79757	62385	7469	80407	57360	82353	75368	1569
588856	54809	286272	213872	39769	82996	83924	82687	**404288**	382224	291391	52622	83851	60682	358597	296840	4779
624418	53153	271744	202361	38546	80074	80959	79757	382224	**386666**	278938	51529	80903	58845	340552	282111	4528
504648	41378	217456	161460	29777	62647	63337	62385	291391	278938	**292454**	40138	63231	45801	258183	225873	3626
903	2541	13365	11512	1939	7530	7559	7469	52622	51529	40138	**52789**	7913	2790	13218	12030	120
620	55173	84204	84145	39944	80733	81687	80407	83851	80903	63231	7913	**86742**	60270	83448	76336	1582
430	54006	60798	60762	38872	57615	58523	57360	60682	58845	45801	2790	60270	**62655**	60411	55093	1234
602141	54582	285490	212981	39564	82658	83583	82353	358597	340552	258183	13218	83448	60411	**359231**	295880	4733
2717	49761	265906	198860	36047	75657	76498	75368	296840	282111	225873	12030	76336	55093	295880	**300111**	4155
2275	961	3736	3167	801	1578	1592	1569	4779	4528	3626	120	1582	1234	4733	4155	**29940**

Table S3. Overlap of toxic compounds between the different data sets.

AID	463	1825	1486	847	598	648	719	588856	624418	504648	903	620	430	602141	2717	2275
463	**706**	201	246	20	383	60	110	28	6	12	0	147	234	24	42	3
1825	201	**2259**	827	11	271	72	120	225	51	84	5	176	171	195	300	10
1486	246	827	**2409**	11	310	80	138	158	51	72	5	187	171	128	207	10
847	20	11	11	**194**	56	31	37	15	3	7	1	12	40	13	20	0
598	383	271	310	56	**5140**	676	747	212	46	41	10	260	562	122	266	6
648	60	72	80	31	676	**925**	396	118	18	37	5	76	165	75	184	0
719	110	120	138	37	747	396	**937**	125	29	36	5	125	165	90	155	2
588856	28	225	158	15	212	118	125	**3021**	90	122	35	38	61	344	429	4
624418	6	51	51	3	46	18	29	90	**526**	57	4	11	9	66	55	0
504648	12	84	72	7	41	37	36	122	57	**606**	5	17	20	102	88	3
903	0	5	5	1	10	5	5	35	4	5	**338**	4	1	13	2	0
620	147	176	187	12	260	76	125	38	11	17	4	**364**	157	42	43	4
430	234	171	171	40	562	165	165	61	9	20	1	157	**1122**	36	112	4
602141	24	195	128	13	122	75	90	344	66	102	13	42	36	**1302**	282	7
2717	42	300	207	20	266	184	155	429	55	88	2	43	112	282	**3187**	14
2275	3	10	10	0	6	0	2	4	0	3	0	4	4	7	14	**193**

Table S4. Number of times each toxic compound has been tested vs. the number of times it has displayed toxicity.

Ntoxic \ Ntested	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	279	317	708	801	1278	2060	1811	172	274	706	998	791	1176	865	25	0
2		71	85	175	338	441	212	32	95	185	184	194	286	154	8	0
3			21	43	88	104	50	19	56	88	96	83	101	26	2	0
4				23	33	32	22	12	43	57	36	46	39	16	0	0
5					11	15	14	5	17	32	20	39	37	13	1	0
6						5	1	3	11	13	21	24	26	10	0	0
7							1	2	6	8	7	12	9	4	1	0
8								1	2	3	1	7	3	1	0	0
9									1	1	2	2	1	1	0	0
10										1	0	3	2	0	0	0
11											0	1	0	0	0	0
12												0	0	0	0	0
13													0	0	0	0
14														0	0	0
15															0	0