Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability.

Brandall L. Ingle,[†] *Brandon C. Veber*[‡,§] *John W. Nichols,*[‡] *Rogelio Tornero-Velez*[†*]

[†] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC 27709

[‡] U.S. Environmental Protection Agency, Office of Research and Development, National Health Exposure Effects Research Laboratory, Duluth, MN 55804

[§] Oak Ridge Institutes for Science and Education, Oak Ridge, TN 37830

List of Contents

Figure S1. The (a) MAE and (b) RMSE with between the error predicted by LCV and the actual error as impacted by the number of neighbors used in the LCV.

Table S1. The relevant properties and ranges of the descriptors used to construct the 3 F_{ub} models.

Table S2. Chemicals outside of the AD defined by the bounded box of descriptor ranges of the training set in each F_{ub} model.

Table S3. Chemicals outside of the AD defined by the range of all principal components.

Table S4. Performance of Fub models relative to experimental F_{ub}.

Table S5. Consensus model 3D reliability estimates, based on the average distance to 5 nearest neighbors, standard deviation across kNN, SVM and RF models, and the F_{ub} prediction.

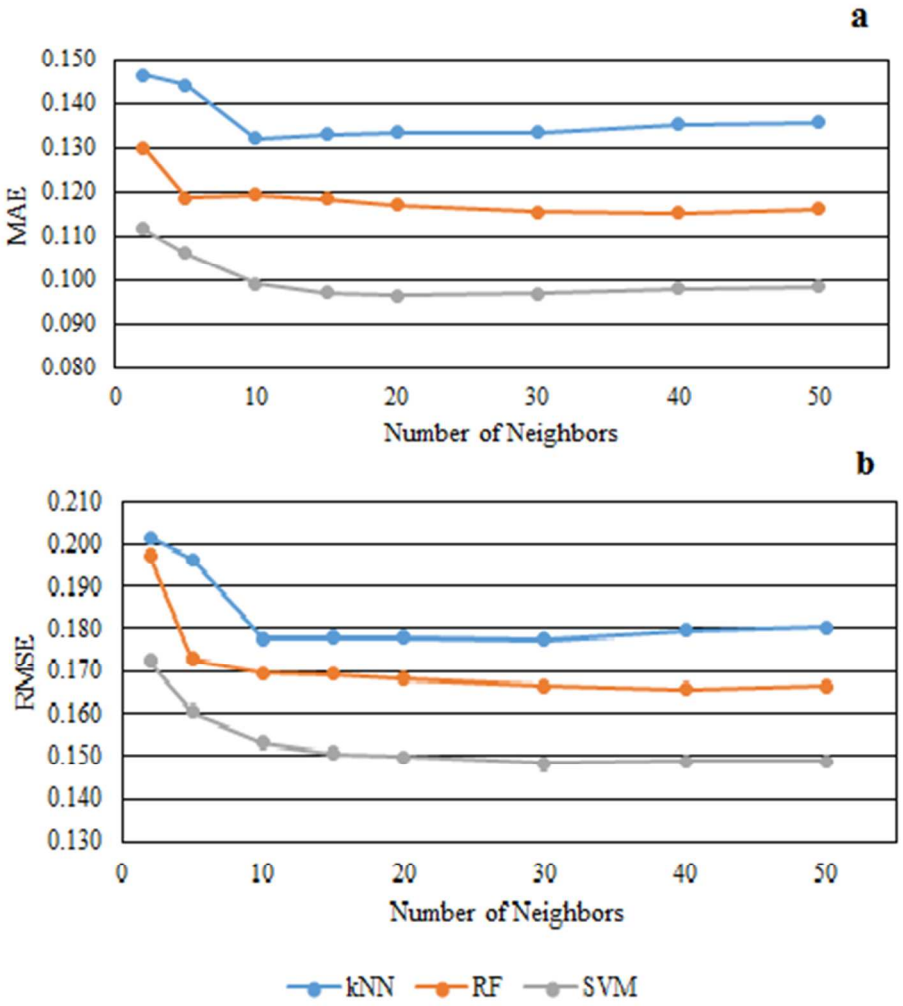

Figure S1. The (a) MAE and (b) RMSE with between the error predicted by LCV and the actual error as impacted by the number of neighbors used in the LCV.

Table S1. The relevant properties and ranges of the descriptors used to construct the 3 F_{ub} models.

Model	Descriptor	Property*	Drug Training	Drug Test		ToxCast 1		ToxCast 2	
			Range	Range	Out	Range	Out	Range	Out
kNN	PEOE_VSA_FPPOS	Charge – Fractional positive polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +0.83	0 to +0.50	0	0 to +0.71	0	0 to +0.45	0
	SlogP	Hydrophobicity - Log octonal/water partition coefficent from atomic contribution model	-8.90 to +10.86	-7.29 to +9.35	0	-0.96 to +6.93	0	-0.93 to +7.51	0
	logS	Solubility - Log of aqueous solubility from linear atomic type model	-18.36 to +2.38	-15.01 to +2.08	0	-7.98 to +0.89	0	-12.16 to -0.14	0
	logP(o/w)	Hydrophobicity - Log octonal/water partition coefficent from linear atom type model	-11.56 to +10.54	-7.20 to +9.15	0	-1.47 to +7.95	0	-2.08 to +9.72	0
	GCUT_SMR_0	Molecular shape – Distance matrix considering both interatomic distance and atomic contribution to molar refractivity	-0.65 to -0.40	-0.61 to -0.43	0	-0.60 to -0.39	5	-0.60 to -0.37	11
	BCUT_SLOGP_0	Molecular shape – Adjacency matrix considering both bond order and atomic contribution to logP	-3.10 to -1.98	-3.03 to -1.86	1	-2.92 to -1.92	4	-3.04 to -1.85	11
	GCUT_PEOE_0	Molecular shape – Distance matrix considering both interatomic distance and partial charges	-1.11 to -0.68	-0.95 to -0.75	0	-0.89 to -0.67	1	-1.01 to -0.68	0
	GCUT_SLOGP_0	Molecular shape – Distance matrix considering both interatomic distance and atomic contribution to logP	-1.49 to -0.75	-1.38 to -0.76	0	-1.46 to -0.56	3	-1.37 to -0.49	21
	GCUT_PEOE_1	Molecular shape – Distance matrix considering both interatomic distance and partial charges	-0.54 to -0.13	-0.54 to -0.24	0	-0.55 to -0.20	1	-0.53 to -0.14	0
	BCUT_SLOGP_2	Molecular shape – Adjacency matrix considering both bond order and atomic contribution to logP	+0.12 to +0.97	+0.12 to +1.21	2	0.12 to + 1.20	5	+0.12 to +1.21	6
	PEOE_VSA+1	Charge – Low positive charge on van der Waals surface area from partial equalization of orbital electronegativities	0 to +221.3	0 to +150.3	0	0 to +116.6	0	0 to +168.2	0
	PEOE_VSA+4	Charge – Moderate positive charge on van der Waals surface area from partial equalization of orbital electronegativities	0 to +158.9	0 to +112.6	0	0 to +130.4	0	0 to +108.0	0
	PEOE_VSA_PPOS	Charge – Total positive polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +321.3	0 to +330.9	1	0 to +135.9	0	0 to +170.3	0
	PEOE_VSA_POL	Charge – Total polar van der Waals surface	0 to +647.1	0 to +697.9	1	0 to +194.3	0	0 to +333.5	0

		area from partial equalization of orbital electronegativities							
	SMR_VSA6	Surface area – Accessible van der Waals surface area and molar refractivity	0 to +278.4	0 to +127.6	0	0 to +194.3	0	0 to +146.5	0
SVM	logS	Solubility - Log of aqueous solubility from linear atomic type model	-18.36 to +2.38	-15.01 to +2.08	0	-7.98 to +0.89	0	-12.16 to -0.14	0
	VAdjEq	Atomic bonds – Vertex adjacency information (equality) for neighboring heavy atoms and bonds	+0.12 to +1.00	+0.13 to +0.99	0	0.21 to +0.90	0	+0.13 to +0.86	0
	PEOE_VSA_FPPOS	Charge – Fractional positive polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +0.83	0 to +0.50	0	0 to +0.71	0	0 to +0.45	0
	SlogP	Hydrophobicity - Log octonal/water partition coefficent from atomic contribution model	-8.90 to +10.86	-7.29 to +9.35	0	-0.96 to +6.93	0	-0.93 to +7.51	0
	a_nS	Atom count – Number of sulfur atoms	0 to +6	0 to +3	0	0 to +3	0	0 to +4	0
	a_base	Basicity – Number of basic atoms	0 to +16	0 to +2	0	0 to +2	0	0 to +1	0
	a_nN	Atom count – Number of nitrogen atoms	0 to +18	0 to +17	0	0 to +6	0	0 to +6	0
	SlogP_VSA6	Surface area – Accessible van der Waals surface area and log octanol/water partition coefficient	0 to +15.52	0 to +17.50	1	0 to +22.52	2	0 to +18.44	2
	logP(o/w)	Hydrophobicity - Log octonal/water partition coefficent from linear atom type model	-11.56 to +10.54	-7.20 to +9.15	0	-1.47 to +7.95	0	-2.08 to +9.72	0
	PEOE_VSA_FPOL	Charge – Fractional polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +1	0 to +0.77	0	0 to +1	0	0 to +0.69	0
RF	logS	Solubility - Log of aqueous solubility from linear atomic type model	-18.36 to +2.38	-15.01 to +2.08	0	-7.98 to +0.89	0	-12.16 to -0.14	0
	logP(o/w)	Hydrophobicity - Log octonal/water partition coefficent from linear atom type model	-11.56 to +10.54	-7.20 to +9.15	0	-1.47 to +7.95	0	-2.08 to +9.72	0
	SlogP	Hydrophobicity - Log octonal/water partition coefficent from atomic contribution model	-8.90 to +10.86	-7.29 to +9.35	0	-0.96 to +6.93	0	-0.93 to +7.51	0
	PEOE_VSA_FPPOS	Charge – Fractional positive polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +0.83	0 to +0.50	0	0 to +0.71	0	0 to +0.45	0
	GCUT_SMR_0	Molecular shape – Distance matrix considering both interatomic distance and atomic contribution to molar refractivity	-0.65 to -0.40	-0.61 to -0.43	0	-0.60 to -0.39	5	-0.60 to -0.37	11
	BCUT_SLOGP_0	Molecular shape – Adjacency matrix	-3.10 to -1.98	-3.03 to -1.86	1	-2.92 to -1.92	4	-3.04 to -1.85	11

		considering both bond order and atomic contribution to logP							
	GCUT_PEOE_0	Molecular shape – Distance matrix considering both interatomic distance and partial charges	-1.11 to -0.68	-0.95 to -0.75	0	-0.89 to -0.67	1	-1.01 to -0.68	0
	GCUT_PEOE_1	Molecular shape – Distance matrix considering both interatomic distance and partial charges	-0.54 to -0.13	-0.54 to -0.24	0	-0.55 to -0.20	1	-0.53 to -0.14	0
	GCUT_SLOGP_0	Molecular shape – Distance matrix considering both interatomic distance and atomic contribution to logP	-1.49 to -0.75	-1.38 to -0.76	0	-1.46 to -0.56	3	-1.37 to -0.49	20
	PEOE_VSA_PPOS	Charge – Total positive polar van der Waals surface area from partial equalization of orbital electronegativities	0 to +321.3	0 to +330.9	1	0 to +135.9	0	0 to +170.3	0

* Property descriptions from MOE. Descriptors listed in order of importance towards model based on reduction in cross validation mean square error.

Table S2. Chemicals outside of the AD defined by the bounded box of descriptor ranges of the training set in each F_{ub} model.

Model	Test Set	Chemical	Descriptor	AD Range	Value	Exp. F_{ub}*	Prd. F_{ub}*
kNN	Drugs	Pentachlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.86	0.010	0.050
			BCUT_SLOGP_2	0.12 to 0.97	1.21		
			PEOE_VSA_PPOS	0 to +321.3	17.50		
		Thyroxine	BCUT_SLOGP_2	0.12 to 0.97	1.05	0.010	0.042
		Daptomycin	PEOE_VSA_POL	0 to + 647.1	697.86	0.080	0.649
			PEOE_VSA_PPOS	0 to +321.3	330.89		
	TCP1	Anilazine	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.005	0.043
		Boscalid	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.032	0.036
		Bromoxynil	BCUT_SLOGP_0	-3.10 to -1.98	-1.94	0.500	0.087
			BCUT_SLOGP_2	0.12 to 0.97	0.98		
		Clofentezine	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.005	0.017
		Cyromazine	GCUT_PEOE_1	-0.54 to -0.13	-0.55	0.935	0.779
		Dichlobenil	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.062	0.059
			BCUT_SLOGP_2	0.12 to 0.97	1.01		
			GCUT_SLOGP_0	-1.49 to -0.75	-0.56		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		Dichloran	BCUT_SLOGP_2	0.12 to 0.97	1.01	0.005	0.083
		Fluazinam	PEOE_VSA_PPOS	0 to +321.3	22.52	0.005	0.023
		Nitrapyrin	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.005	0.045
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Quinoxyfen	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.005	0.017
		Quintozene	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.005	0.046
			BCUT_SLOGP_2	0.12 to 0.97	1.20		
			GCUT_PEOE_0	-1.11 to -0.68	-0.67		
			PEOE_VSA_PPOS	0 to +321.3	17.50		
		Thiabendazole	BCUT_SLOGP_0	-3.10 to -1.98	-1.95	0.138	0.182
		Triclopyr	BCUT_SLOGP_2	0.12 to 0.97	0.97	0.022	0.058
	TCP2	1,2,3-trichlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.89	0.016	0.059
			BCUT_SLOGP_2	0.12 to 0.97	1.06		
			GCUT_SLOGP_0	-1.49 to -0.75	-0.53		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		1,2,4,5-tetrachlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.85	0.001	0.059
			BCUT_SLOGP_2	0.12 to 0.97	1.06		
			GCUT_SLOGP_0	-1.49 to -0.75	-0.49		

			GCUT_SMR_0	-0.65 to -0.40	-0.37		
		1,2-benzisothiazolin-3-one	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.000	0.629
		1,3-diisopropylbenzene	GCUT_SLOGP_0	-1.49 to -0.75	-0.63	0.012	0.078
		1,4-Dichlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.91	0.023	0.096
			GCUT_SLOGP_0	-1.49 to -0.75	-0.52		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		1,5,9-cyclododecatriene	GCUT_SLOGP_0	-1.49 to -0.75	-0.60	0.041	0.044
		2,4,5-trichlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.005	0.067
		2,4,6-trichlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.015	0.067
			BCUT_SLOGP_2	0.12 to 0.97	1.01		
		7,12-dimethylbenz(a)anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.60	0.000	0.058
		Acenaphthylene	GCUT_SLOGP_0	-1.49 to -0.75	-0.55	0.011	0.053
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.025	0.041
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Benz[a]anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.011	0.027
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Benzo[b]fluoranthene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.011	0.025
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Biphenyl	GCUT_SLOGP_0	-1.49 to -0.75	-0.55	0.016	0.041
		Coumarin	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.192	0.126
		Dibenzofuran	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.000	0.041
			GCUT_SLOGP_0	-1.49 to -0.75	-0.54		
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Didecyl dimethyl ammonium chloride	PEOE_VSA_PPOS	0 to +321.3	18.44	0.000	0.520
		Diphenylenemethane	GCUT_SLOGP_0	-1.49 to -0.75	-0.69	0.016	0.063
		Fluoranthene	BCUT_SLOGP_0	-3.10 to -1.98	-1.96	0.014	0.037
			GCUT_SLOGP_0	-1.49 to -0.75	-0.54		
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Heptachlor	BCUT_SLOGP_2	0.12 to 0.97	1.04	0.043	0.021
		Mirex	BCUT_SLOGP_2	0.12 to 0.97	1.12	0.015	0.044
		Naphthalene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.016	0.064
		O,p'-ddt	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.000	0.029
		P,p'-ddd	GCUT_SLOGP_0	-1.49 to -0.75	-0.74	0.031	0.023
		P,p'-DDT	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.000	0.029
		Pentachlorophenol (=2,4-d)	BCUT_SLOGP_0	-3.10 to -1.98	-1.86	0.054	0.050
			BCUT_SLOGP_2	0.12 to 0.97	1.21		

			PEOE_VSA_PPOS	0 to +321.3	17.50		
		Phenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.96	0.620	0.256
		Pyrene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.024	0.027
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		Quinoline	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.204	0.096
		Resorcinol	BCUT_SLOGP_0	-3.10 to -1.98	-1.97	0.745	0.269
SVM	Drug	Pentachlorophenol	SlogP_VSA6	0 to +15.52	17.50	0.010	0.010
	TCP1	Fluazinam	SlogP_VSA6	0 to +15.52	22.52	0.005	0.012
		Quintozene	SlogP_VSA6	0 to +15.52	17.50	0.005	0.006
	TCP2	Didecyl dimethyl ammonium chloride	SlogP_VSA6	0 to +15.52	18.44	0.000	0.002
		Pentachlorophenol (=2_4-d)	SlogP_VSA6	0 to +15.52	17.50	0.054	0.010
RF	Drug	Pentachlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.86	0.01	0.015
		Daptomycin	PEOPE_VSA_PPOS	0 to +321.3	330.9	0.08	0.842
	TCP1	Anilazine	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.005	0.013
		Boscalid	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.032	0.036
		Bromoxynil	BCUT_SLOGP_0	-3.10 to -1.98	-1.94	0.5	0.04
		Clofentezine	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.005	0.009
		Cyromazine	GCUT_PEOE_1	-0.54 to -0.13	-0.55	0.935	0.154
		Dichlobenil	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.062	0.036
			GCUT_SLOG_0	-1.49 to -0.75	-0.56		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		Nitrapyrin	GCUT_SLOG_0	-1.49 to -0.75	-0.71	0.005	0.034
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Quinoxyfen	GCUT_SLOG_0	-1.49 to -0.75	-0.71	0.005	0.023
		Quintozene	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.005	0.012
			GCUT_PEOE_0	-1.11 to -0.67	-0.67		
		Thiabendazole	BCUT_SLOGP_0	-3.10 to -1.98	-1.95	0.138	0.085
	TCP2	1,2,3-trichlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.89	0.016	0.009
			GCUT_SLOGP_0	-1.49 to -0.75	-0.53		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		1,2,4,5-tetrachlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.85	0.001	0.012
			GCUT_SLOGP_0	-1.49 to -0.75	-0.49		
			GCUT_SMR_0	-0.65 to -0.40	-0.37		
		1,2-benzisothiazolin-3-one	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.000	0.160

		1,3-diisopropylbenzene	GCUT_SLOGP_0	-1.49 to -0.75	-0.63	0.012	0.031
		1,4-Dichlorobenzene	BCUT_SLOGP_0	-3.10 to -1.98	-1.91	0.023	0.018
			GCUT_SLOGP_0	-1.49 to -0.75	-0.52		
			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		1,5,9-cyclododecatriene	GCUT_SLOGP_0	-1.49 to -0.75	-0.60	0.041	0.097
		2,4,5-trichlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.005	0.026
		2,4,6-trichlorophenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.90	0.015	0.025
		7,12-dimethylbenz(a)anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.60	0.000	0.020
		Acenaphthylene	GCUT_SLOGP_0	-1.49 to -0.75	-0.55	0.011	0.027
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.025	0.029
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Benz[a]anthracene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.011	0.010
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Benzo[b]fluoranthene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.011	0.016
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Biphenyl	GCUT_SLOGP_0	-1.49 to -0.75	-0.55	0.016	0.025
		Coumarin	GCUT_SMR_0	-0.65 to -0.40	-0.40	0.192	0.075
		Dibenzofuran	BCUT_SLOGP_0	-3.10 to -1.98	-1.92	0.000	0.056
			GCUT_SLOGP_0	-1.49 to -0.75	-0.54		
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Diphenylenemethane	GCUT_SLOGP_0	-1.49 to -0.75	-0.69	0.016	0.038
		Fluoranthene	BCUT_SLOGP_0	-3.10 to -1.98	-1.96	0.014	0.014
			GCUT_SLOGP_0	-1.49 to -0.75	-0.54		
			GCUT_SMR_0	-0.65 to -0.40	-0.40		
		Naphthalene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.016	0.038
		O,p'-ddt	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.000	0.012
		P,p'-ddd	GCUT_SLOGP_0	-1.49 to -0.75	-0.74	0.031	0.030
		P,p'-DDT	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.000	0.012
		Pentachlorophenol (=2,4-d)	BCUT_SLOGP_0	-3.10 to -1.98	-1.86	0.054	0.015
		Phenol	BCUT_SLOGP_0	-3.10 to -1.98	-1.96	0.620	0.773
		Pyrene	GCUT_SLOGP_0	-1.49 to -0.75	-0.54	0.024	0.015

			GCUT_SMR_0	-0.65 to -0.40	-0.39		
		Quinoline	GCUT_SLOGP_0	-1.49 to -0.75	-0.71	0.204	0.213
		Resorcinol	BCUT_SLOGP_0	-3.10 to -1.98	-1.97	0.745	0.660

***** Experimental (Exp.) and predicted (Prd.) F_{ub} values as predicted by the specified model listed for each chemical outside the range of the training set (AD Range). Calculated value for the out of range descriptors are listed per chemical.

Table S3. Chemicals outside of the AD defined by the range of all principal components.

Model	Test Set	Name	Exp. F_{ub}*	Pred. F_{ub}*
kNN	Drugs	Pentachlorophenol	0.010	0.050
		Pipecuronium	0.980	0.694
	ToxCast I	Cacodylic acid	0.906	0.950
		Dimethoate	0.965	0.863
		Fentin hydroxide	0.005	0.032
		Fluoxastrobin	0.036	0.034
		Methidathion	0.268	0.494
		Pentadecafluorooctanoic acid	0.005	0.030
		Pirimiphos-methyl	0.005	0.609
		Quinoxyfen	0.005	0.017
		Quintozene	0.005	0.046
		Thidiazuron	0.028	0.061
	ToxCast II	1,2,3-trichlorobenzene	0.016	0.059
		1,2,4,5-tetrachlorobenzene	0.001	0.059
		1,2-dinitrobenzene	0.069	0.220
		1,3-diphenylguanidine	0.784	0.078
		2,4-dinitrophenol	0.027	0.225
		7,12-dimethylbenz(a)anthracene	0.000	0.058
		Benz[a]anthracene	0.011	0.027
		Benzo[b]fluoranthene	0.011	0.025
		Ethion	0.000	0.039
		Fluoranthene	0.014	0.037
		Mirex	0.015	0.044
		Nitrobenzene	0.384	0.147
		O,p'-ddt	0.000	0.029
		P,p'-DDT	0.000	0.029
		Pentachlorophenol (=2,4-d)	0.054	0.050
		Pentadecafluorooctanoic acid ammonium salt	0.004	0.030
		Perfluorodecanoic acid	0.000	0.028
		Perfluoroheptanoic acid	0.002	0.017
		Perfluorohexanoic acid	0.017	0.030

		Perfluorononanoic acid	0.001	0.028
		Perfluoroundecanoic acid	0.000	0.026
		Pyrene	0.024	0.027
		Tannic acid	1.000	0.103
SVM	Drugs	Eflornithine	1.000	0.909
		Ethanol	1.000	0.964
		Pentachlorophenol	0.010	0.010
		Daptomycin	0.080	0.721
	ToxCast I	Cacodylic acid	0.906	0.976
		Fluazinam	0.005	0.012
		Quintozene	0.005	0.006
	ToxCast II	Didecyl dimethyl ammonium chloride	0.000	0.002
		Ethion	0.000	0.030
		Pentachlorophenol (=2,4-d)	0.054	0.010
RF	Drugs	Daptomycin	0.080	0.842
	ToxCast I	Cacodylic acid	0.906	0.966
		Difenzoquat methyl sulfate	0.851	0.071
		Fentin hydroxide	0.005	0.124
		Lindane	0.005	0.032
		Pentadecafluorooctanoic acid	0.005	0.020
		Quintozene	0.005	0.012
		Thidiazuron	0.028	0.285
	ToxCast II	1,2,3-trichlorobenzene	0.016	0.009
		1,2,4,5-tetrachlorobenzene	0.001	0.012
		1,3-diphenylguanidine	0.784	0.070
		1,4-Dichlorobenzene	0.023	0.018
		Heptachlor epoxide isomer B	0.011	0.010
		Mirex	0.015	0.022
		Nitrobenzene	0.384	0.119
		Pentadecafluorooctanoic acid ammonium salt	0.004	0.027
		Perfluorodecanoic acid	0.000	0.013
		Perfluoroheptanoic acid	0.002	0.038
		Perfluorohexanoic acid	0.017	0.054
		Perfluorononanoic acid	0.001	0.017

		Perfluoroundecanoic acid	0.000	0.014
		Tannic acid	1.000	0.040

* Experimental (Exp.) and predicted (Prd.) F_{ub} values as predicted by the specified model listed for each chemical outside applicability domain (AD) as defined by the range of the training set principal components for the k nearest neighbor (kNN), support vector machine (SVM) and random forest (RF) models.

Table S4. Performance of Fub models relative to experimental F_{ub}.

		Training*		Drug Test Set		ToxCast I		ToxCast II		Universal	
	Method	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
$F_{ub} < 0.15$	kNN	0.093	0.159	0.095	0.159	0.101	0.161	0.068	0.129	0.089	0.151
	SVM	0.106	0.164	0.107	0.175	0.115	0.166	0.133	0.197	0.119	0.178
	RF	0.080	0.144	0.086	0.147	0.072	0.127	0.060	0.115	0.071	0.128
	Consensus	0.090	0.144	0.095	0.149	0.093	0.140	0.081	0.128	0.090	0.138
$0.15 < F_{ub} < 0.85$	kNN	0.198	0.247	0.187	0.250	0.221	0.287	0.239	0.291	0.208	0.270
	SVM	0.209	0.264	0.216	0.270	0.171	0.234	0.183	0.220	0.195	0.249
	RF	0.190	0.239	0.188	0.244	0.220	0.294	0.199	0.247	0.200	0.261
	Consensus	0.181	0.223	0.174	0.233	0.181	0.253	0.152	0.201	0.172	0.233
$F_{ub} > 0.85$	kNN	0.275	0.344	0.297	0.380	0.432	0.535	0.349	0.480	0.340	0.443
	SVM	0.244	0.316	0.258	0.360	0.352	0.476	0.323	0.440	0.294	0.407
	RF	0.193	0.290	0.270	0.358	0.541	0.621	0.280	0.449	0.337	0.452
	Consensus	0.232	0.298	0.270	0.349	0.439	0.518	0.308	0.449	0.318	0.416

*Training set predictions from 5-fold cross validation. Mean absolute error (MAE) and root mean square error (RMSE) values for k nearest neighbors (kNN), support vector machine (SVM), random forest (RF), and consensus models. The universal test set encompasses all chemicals in the pharmaceutical and environmental (ToxCast) test sets.

Table S5. Consensus model 3D reliability estimates, based on the average distance to 5 nearest neighbors, standard deviation across kNN, SVM and RF models, and the F_{ub} prediction.

Bin	Avg. Dist. 5NN	St. Dev. 3 model	F_{ub} Prediction	Train N	Train RMSE	Test N	Test RMSE	Error < binRMSE
1	< 0.14	< 0.040	< 0.13	89	0.080	0	N/A	N/A
2	0.14 – 2.00	< 0.040	< 0.13	93	0.102	13	0.117	100%
3	> 2.00	< 0.040	< 0.13	81	0.099	198	0.103	93%
4	< 0.14	0.040-0.104	< 0.13	26	0.074	0	N/A	N/A
5	0.14 – 2.00	0.040-0.104	< 0.13	36	0.094	2	0.540	50%
6	> 2.00	0.040-0.104	< 0.13	18	0.259	65	0.191	94%
7	< 0.14	> 0.104	< 0.13	2	0.064	0	N/A	N/A
8	0.14 – 2.00	> 0.104	< 0.13	1	0.059	0	N/A	N/A
9	> 2.00	> 0.104	< 0.13	2	0.213	8	0.314	88%
10	< 0.14	< 0.040	0.13-0.42	28	0.236	0	N/A	N/A
11	0.14 – 2.00	< 0.040	0.13-0.42	15	0.187	5	0.175	80%
12	> 2.00	< 0.040	0.13-0.42	7	0.164	12	0.277	58%
13	< 0.14	0.040-0.104	0.13-0.42	69	0.254	0	N/A	N/A
14	0.14 – 2.00	0.040-0.104	0.13-0.42	48	0.177	10	0.168	80%
15	> 2.00	0.040-0.104	0.13-0.42	42	0.263	68	0.252	82%
16	< 0.14	> 0.104	0.13-0.42	33	0.173	0	N/A	N/A
17	0.14 – 2.00	> 0.104	0.13-0.42	51	0.253	10	0.249	80%
18	> 2.00	> 0.104	0.13-0.42	55	0.282	91	0.209	88%
19	< 0.14	< 0.040	>0.42	4	0.277	0	N/A	N/A
20	0.14 – 2.00	< 0.040	>0.42	8	0.086	0	N/A	N/A
21	> 2.00	< 0.040	>0.42	21	0.228	7	0.268	86%
22	< 0.14	0.040-0.104	>0.42	25	0.221	0	N/A	N/A
23	0.14 – 2.00	0.040-0.104	>0.42	47	0.263	10	0.239	70%
24	> 2.00	0.040-0.104	>0.42	42	0.248	38	0.251	66%
25	< 0.14	> 0.104	>0.42	69	0.205	1	0.146	100%
26	0.14 – 2.00	> 0.104	>0.42	55	0.277	8	0.299	50%
27	> 2.00	> 0.104	>0.42	78	0.271	60	0.292	62%

*N is number of chemicals in each bin; Error < binRMSE is the percentage of test set chemicals in that bin with a prediction error lower than the RSME assigned to that bin (training RMSE).