Supplementary Information for

Knowledge-based H-bond prediction to aid experimental polymorph screening

Peter T. A. Galek, László Fábián, Frank H. Allen* & Neil Feeder*

Aromaticity Parameter

The aromaticity function, $a_c(i)$, measures the total number of covalent bonds in a molecular structure that have a degree of π character ($b_{arom,c}$), as defined by their internal CSD flag (*e.g.* double, aromatic, delocalised, π) as a proportion of the total number of covalent bonds that have this potential(b_c), that is all non-terminal covalent bonds. The function returns a value between 0 and 1 describing the aromatic nature of a molecule, and can relate to a reduced potential for hydrogen bond formation.

$$a_c(i) = \frac{\sum_c b_{arom,c}}{\sum_c b_c} \qquad \text{(S1)}$$

LHP model for ritonavir

The methodology is summarised in the text. The coefficient size and sign in Table S1 indicates a degree of influence over the model predictions (positive favouring true outcomes, and negative false).

Table S1. Ritonavir LHP model coefficients and fitting statistics.

Parameter, x_k	Coeff. Value, β_k	Standard Error	Pr > χ^2	+/– [a]
Intercept, α	2.005	0.157	< 0.0001	0.308
Competition	-0.064	0.008	< 0.0001	0.015
Aromaticity	-1.851	0.181	< 0.0001	0.355
Steric density (don.)	-0.066	0.054	0.220	0.105
Steric density (acc.)	-0.240	0.051	< 0.0001	0.099
Donor-hydroxyl	0.000	0.000	-	0.000
Donor-other	0.405	0.079	< 0.0001	0.154
Donor-amido	-0.008	0.152	0.959	0.298
Donor-carbamate	0.264	0.106	0.013	0.208
Donor-ureido	-0.891	0.083	< 0.0001	0.163
Acceptor-hydroxyl	0.000	0.000	-	0.000
Acceptor-amido	-0.257	0.075	0.001	0.147
Acc.-carbamate	-0.021	0.186	0.910	0.365
Acceptor-ether	-3.195	0.102	< 0.0001	0.200
Acceptor-aromatic N	-2.044	0.124	< 0.0001	0.243
Acceptor-other	-2.122	0.078	< 0.0001	0.153
Acceptor-ureido	-2.164	0.183	< 0.0001	0.359
Acceptor-thiazoyl	-0.428	0.119	0.000	0.233

[a] The error bars of the coefficient value: the value falls within this range at the 95% confidence level, based on a χ^2 distribution.

Ritonavir Model Regression Statistics

Relative category frequencies are presented in table S1. The correlation matrix, Table S2, represents how distinctly the quantitative parameters vary. Table S3 provides goodness-of-fit statistics. The fitted model equation is compared against an independent model (unoptimised coefficients). The included statistics may be referred to in Hosmer & Lemeshow (2000). The null hypothesis test, Table S4, determines any improvement in a fitted model from a null model (the 'null probability', P_0 = 0.29 is always returned). Type III analysis, Table S5, is used to observe the significance of each parameter by systematic removal from the model. Table S6 compares the classification of a validation sample (true and false observations in a subset of data set aside from model fitting: 3000 observations selected randomly) with classification of the training data (5731 observations).

Table S2. Observed frequencies of categorical variables

Variable	Categories	Frequencies	%
Donor group label	hydroxyl	5347	53.8
	amido	2014	20.3
	carbamate	366	3.68
	other	698	7.02
	ureido	1518	15.2
Acceptor group label	hydroxyl	1791	18.0
	amido	1594	16.0
	carbamate	159	1.60
	ether	2854	28.7
	nitrogen, aromatic	639	6.43
	other	2250	22.6
	thiazoyl	249	2.50
	ureido	407	4.09

Table S3. Correlation Matrix

Variables	competition	Aromaticity	Donor Steric Density	Acceptor Steric Density
competition	**1.000**	-0.213	0.007	0.126
Aromaticity	-0.213	**1.000**	0.280	0.031
Donor Steric Density	0.007	0.280	**1.000**	0.312
Acceptor Steric Density	0.126	0.031	0.312	**1.000**

Table S4. Goodness-of-fit statistics

Statistic	Independent	Full
Observations	9943	9943
Sum of weights	9943.000	9943.000
DF	9942	9927
-2 Log(Likelihood)	12067.114	8924.709
R^2(McFadden)	0.000	0.260
R^2(Cox and Snell)	0.000	0.271
R^2(Nagelkerke)	0.000	0.386
AIC	12069.114	8956.709
SBC	12076.319	9071.983
Iterations	0	6

Table S5. Test of the null hypothesis H0: Y=0.356 (Variable bond exists)

Statistic	DF	Chi-square	Pr > Chi²
-2 Log(Likelihood)	15	3142.406	< 0.0001
Score	15	2878.983	< 0.0001
Wald	15	2131.928	< 0.0001

Table S6. Type III analysis

Source	DF	Chi-square (Wald)	Pr > Wald	Chi-square (LR)	Pr > LR
competition	1	64.713	< 0.0001	76.561	< 0.0001
Aromaticity	1	104.672	< 0.0001	107.218	< 0.0001
Donor steric density	1	1.504	0.220	1.507	0.220
Acceptor steric density	1	22.450	< 0.0001	22.772	< 0.0001
Donor group label	4	178.244	< 0.0001	190.561	< 0.0001
Acceptor group label	7	1676.641	< 0.0001	2128.375	< 0.0001

Figure S1. ROC curve for the ritonavir model showing degree of correct positive (sensitivity) to false negative predictions (specificity) for the training observations, with changing threshold for whether a propensity score is considered to predict a *true* or *false* outcome. AUC = area under curve and measures predictive ability. The dashed line represent a purely random outcome and a curve with AUC > 0.5 has predictive power.

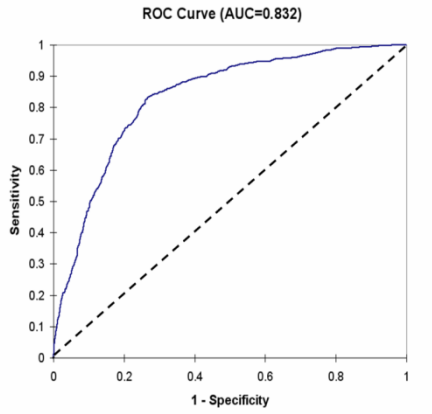

Table S7. Validation tests.

S7.1 Training data classification (with true/false cutoff = 0.3)

observed\ predicted	*False*	*True*	Total	% correct
False	2509	1178	3687	68.05%
True	245	1799	2044	88.01%
Total	2754	2977	5731	75.17%

S7.2 Validation data classification (with true/false cutoff = 0.3)

Observed\ Predicted	*False*	*True*	Total	% correct
False	1345	589	1934	69.54%
True	139	927	1066	86.96%
Total	1484	1516	3000	75.73%

Table S8 Propensity predictions and associated parameter values for the ritonavir molecule.

Donor group	Acceptor group	Competition	Aromaticity	Donor Steric density	Acceptor Steric Density	propensity	+/– (χ^2)	Form I	Form II
amido	carbamate	5.33	0.359	4.27	2.61	0.618	0.094	×	×
amido	hydroxyl	8.00	0.359	4.27	3.13	0.551	0.090	×	✓
carbamate	carbamate	5.33	0.359	2.94	2.61	0.538	0.052	✓	×
hydroxyl	carbamate	5.33	0.359	3.13	2.61	0.537	0.090	×	×
amido	amido	5.33	0.359	4.27	3.62	0.501	0.055	✓	×
carbamate	hydroxyl	8.00	0.359	2.94	3.13	0.470	0.078	×	×
amido	ureido	5.33	0.359	4.27	2.94	0.499	0.072	×	×
hydroxyl	hydroxyl	8.00	0.359	3.13	3.13	0.469	0.037	×	×
carbamate	amido	5.33	0.359	2.94	3.62	0.420	0.083	×	✓
hydroxyl	amido	5.33	0.359	3.13	3.62	0.419	0.045	×	×
carbamate	ureido	5.33	0.359	2.94	2.94	0.418	0.088	×	×
hydroxyl	ureido	5.33	0.359	3.13	2.94	0.417	0.058	×	✓
ureido	carbamate	5.33	0.359	3.36	2.61	0.319	0.086	×	✓
ureido	hydroxyl	8.00	0.359	3.36	3.13	0.263	0.041	×	×
ureido	amido	5.33	0.359	3.36	3.62	0.225	0.040	×	×
ureido	ureido	5.33	0.359	3.36	2.94	0.224	0.044	✓	×
amido	thiazoyl a	8.00	0.359	4.27	2.14	0.152	0.054	×	×
amido	thiazoyl b	8.00	0.359	4.27	2.47	0.142	0.050	×	×
carbamate	thiazoyl a	8.00	0.359	2.94	2.14	0.115	0.044	×	×
hydroxyl	thiazoyl a	8.00	0.359	3.13	2.14	0.114	0.039	✓	×
carbamate	thiazoyl b	8.00	0.359	2.94	2.47	0.107	0.041	×	×
hydroxyl	thiazoyl b	8.00	0.359	3.13	2.47	0.106	0.036	×	×
ureido	thiazoyl a	8.00	0.359	3.36	2.14	0.049	0.020	×	×
ureido	thiazoyl b	8.00	0.359	3.36	2.47	0.046	0.018	×	×