Supporting Information

PHOENIX: A Scoring Function Derived Using High-Resolution Crystal Structures and Calorimetry Measurements To Predict Binding Affinities of Protein-Ligand Complexes

Yat T. Tang and Garland R. Marshall

1 Metrics for X-ray Crystal Structure Quality

- Free R Value (R_{free})

  Free R Value (R_{free}) is a measure of the degree to which an atomic model predicts a subset of observed diffraction data that has been omitted from the refinement. R_{free} is defined by the equation:

$$R_{free} = \frac{\sum \mid F_{obs} - F_{calc} \mid}{\sum \mid F_{obs} \mid} \tag{1}$$

  F_{obs} = the observed reflection amplitudes,
  F_{calc} = the reflection amplitudes calculated from the model.

- Diffraction-component Precision Index (DPI)

  DPI is measure of the quality of the structural model derived from the diffraction data. DPI is defined by the equation:

$$\sigma(x, B_{avg}) = 1.28 N_{atoms}^{1/2} V_a^{1/3} N_{obs}^{-5/6} R_{free}, \tag{2}$$

N_{atoms} = number of atoms in the unit cell,
V_a = volume of unit cell,
N_{obs} = number of crystallographic observations.

2 Metrics for Partial Least Squares Models

- Correlation Coefficient (r^2)

$$r^2 = 1 - \frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{\sum_i (X_{i,exp} - X_{i,mean})^2} \tag{3}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
X_{mean} = mean of experimentally measured thermodynamic parameter.

- Standard Error (s)

$$s = \sqrt{\frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{(N-1)}} \tag{4}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
N = number of thermodynamic parameters.

- F-value (F)

$$F = \frac{r^2(N-k-1)}{(1-r^2)k} \tag{5}$$

r = Pearson correlation coefficient,
N = training set size,
k = number of PLS components used.

- Cross-validation Correlation Coefficient (q^2)

$$q^2 = 1 - \frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{\sum_i (X_{i,exp} - X_{i,mean})^2} \tag{6}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
X_{mean} = mean of experimentally measured thermodynamic parameter.

- Cross-validation Standard Error (S_{PRESS})

$$S_{PRESS} = \sqrt{\frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{(N - k - 1)}} \tag{7}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
N = number of thermodynamic parameters,
k = number of PLS components used.

- Predictive Correlation Coefficient (r^2_{pred})

$$r^2_{pred} = 1 - \frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{\sum_i (X_{i,exp} - X_{i,mean})^2} \tag{8}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
X_{mean} = mean of experimentally measured thermodynamic parameter.

- Predictive Standard Error (SE_{pred})

$$SE_{pred} = \sqrt{\frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{(N - 1)}} \tag{9}$$

X_{pred} = predicted thermodynamic parameter,
X_{exp} = experimentally measured thermodynamic parameter,
N = number of thermodynamic parameters.

3 Metrics for Correlation Evaluations

- Pearson Correlation Coefficient (R_P)

$$R_P = 1 - \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \tag{10}$$

- Spearman Correlation Coefficient (R_S)

$$R_S = 1 - \frac{6 \times \sum_i (R_i - S_i)^2}{n(n^2 - 1)} \tag{11}$$

R_i = experimentally measured rank,
S_i = predicted rank,
n = number of values.

- Standard Deviaton (SD)

$$SD = \sqrt{\sum_i [y_i - (ax_i + b)]^2/(N-2)} \tag{12}$$

a = slope of regression line,
b = coefficient of regression line,
N = number of values.

- Mean Error (ME)

$$ME = \sum_i \mid y_i - (ax_i + b) \mid /N \tag{13}$$

a = slope of regression line,
b = coefficient of regression line,
N = number of values.

4 Supporting Tables

PDB	Res.($\mathring{A}$)	R_{free}	DPI	ΔG	ΔH	TΔS
1a30	2	0.23	0.28	-6.34	-3.6	2.74
1adl	1.6	0.22	0.17	-7.38	-6.77	0.6
1anf	1.67	na	na	-7.08	6.24	13.31
1ax0	1.9	0.2	0.16	-4.28	-5.49	-1.22
1ax1	1.95	0.2	0.18	-4.49	-9.84	-5.37
1ax2	1.95	0.18	0.16	-5.42	-11.25	-5.93
1b05	2	0.24	0.25	-9.7	1.89	11.51
1b0h	1.9	0.22	0.21	-9.12	4.97	14.09
1b1h	1.8	0.22	0.18	-9.63	1.72	11.32
1b2h	1.9	0.22	0.21	-6.19	16.58	22.78
1b32	1.75	0.21	0.16	-9.67	3.49	13.11
1b3f	1.8	0.22	0.18	-9.39	4.92	14.28
1b3g	2	0.24	0.25	-9.15	4.9	14
1b3h	2	0.24	0.26	-8.46	4.87	13.52
1b3l	2	0.24	0.34	-8.03	3.37	11.35
1b46	1.8	0.22	0.16	-7.19	3.96	11.11
1b4h	1.9	0.23	0.21	-7.43	10.6	18.01
1b4z	1.75	0.21	0.16	-7.12	1.93	90.19
1b51	1.8	0.22	0.24	-10.03	2.13	12.09
1b58	1.8	0.22	0.26	-8.96	4.94	13.83

Table 1: The final training set (n = 112) (1 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

PDB	Res.($\mathring{A}$)	R_{free}	DPI	ΔG	ΔH	TΔS
1b5h	1.9	0.23	0.22	-8.19	10.68	18.86
1b5i	1.9	0.22	0.23	-9.6	1.84	11.44
1b5j	1.8	0.23	0.19	-10.13	2.72	12.75
1b6h	1.8	0.22	0.2	-10.68	1.89	12.53
1b7h	2	0.26	0.28	-11.01	4.68	15.66
1b9j	1.8	0.22	0.18	-8.1	5.88	13.92
1bbz	1.65	0.27	0.04	-7.7	-21.9	-14.2
1bkj	1.8	0.21	0.16	-9.3	-21.5	-12.2
1dzk	1.48	0.24	0.12	-9.2	-23.22	-14.02
1e4w	1.95	0.27	0.25	-10.3	-26.7	-16.4
1e4x	1.9	0.31	0.31	-10.3	-24.8	-14.5
1g6s	1.5	0.17	0.1	-9.37	-19	-9.57
1g6t	1.6	0.19	0.12	-6.86	-5.2	1.62
1gic	2	na	na	-5.5	-4.5	-0.11
1gz9	1.7	0.24	0.17	-4.78	-4.71	0.07
1gzc	1.58	0.23	0.14	-4.76	-7.15	-2.39
1hpv	1.9	na	na	-12.6	-7.3	5.3
1hpx	2	na	na	-13.3	-5.4	7.9
1hsg	2	na	na	-12.7	2.1	14.8
1hw5	1.82	0.3	0.07	-6.78	-0.98	5.82

Table 2: The final training set (n = 112) (2 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

PDB	Res.($\mathring{A}$)	R_{free}	DPI	ΔG	ΔH	TΔS
1hxw	1.8	na	na	-13.63	-2.5	11.1
1i06	1.9	0.25	0.27	-8.38	-11.2	-2.78
1i82	1.9	0.24	0.03	-8.01	-11.07	-3.06
1i8a	1.9	0.22	0.04	-4.83	-7.45	-2.72
1is0	1.9	0.27	0.04	-9.55	-5.91	5.07
1jet	1.2	0.26	0.08	-9.82	4.8	14.59
1jeu	1.25	0.25	0.08	-9.29	2.7	11.94
1jev	1.3	0.23	0.08	-9.36	7	16.29
1jyr	1.55	0.23	0.17	-9.12	-7.94	1.18
1k1n	2	na	na	-9.15	-5.9	2.53
1k21	1.86	0.26	0.23	-11.42	-13.81	-2.39
1k22	1.93	0.21	0.21	-11.44	-8.84	2.6
1kjl	1.4	0.22	0.11	-5.6	-9.6	-4
1lcj	1.8	na	na	-9.58	-9.24	0.33
1lid	1.6	na	na	-7.77	-6.05	1.72
1lzb	1.5	na	na	-7	-14.1	-7.1
1mk5	1.4	0.19	0.11	-18	-24.9	-6.6
1ndc	2	na	na	-4.68	-3.2	1.5
1nzl	1.9	0.27	0.25	-9.7	-7	2.7
1qka	1.8	0.22	0.19	-8.07	8.6	16.65

Table 3: The final training set (n = 112) (3 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

PDB	Res.($\AA$)	R_{free}	DPI	ΔG	ΔH	TΔS
1qkb	1.8	0.22	0.17	-10.01	5.35	15.36
1qy1	1.7	0.21	0.19	-8.1	-10.64	-2.54
1qy2	1.75	0.21	0.19	-9.2	-11.44	-2.24
1rst	1.7	0.23	0.26	-6.071	-12.559	-6.49
1s0r	1.02	0.14	0.03	-6.35	-4.51	1.84
1s3z	2	0.21	0.24	-8.3	-19.9	-11.6
1slt	1.9	na	na	-5.47	-7.76	-2.29
1sre	1.78	na	na	-5.27	1.7	6.97
1srg	1.8	na	na	-7.23	1.28	8.51
1sri	1.65	na	na	-8.29	2.15	10.44
1t3r	1.2	0.18	0.06	-15.2	-12.1	3.1
1t7i	1.35	0.2	0.09	-13.7	-10	3.7
1uae	1.8	na	na	-6	6.88	12.87
1w19	2	0.19	0.22	-7.82	-12.72	-4.9
1w3j	2	0.26	0.28	-8.61	-11.15	-2.52
1y93	1.03	0.17	0.06	-3.01	-3.18	-0.17
1yon	1.95	0.21	0.19	-5.8	-3.01	2.79
1yp6	1.8	0.22	0.22	-8.43	-7.51	0.92
1yv5	2	0.25	0.3	-8.8	1.8	10.5
1znk	1.6	0.26	0.19	-9.27	-15.19	-5.92

Table 4: The final training set (n = 112) (4 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

PDB	Res.($\mathring{A}$)	R_{free}	DPI	ΔG	ΔH	TΔS
2a2r	1.4	0.21	0.1	-11.6	-29.9	-18.3
2aqu	2	0.24	0.34	-15	-4.6	10.4
2bba	1.65	0.19	0.15	-9.8	-15.5	-5.7
2c92	1.6	0.22	0.14	-7.88	-15.14	-7.27
2c94	1.9	0.22	0.22	-9.45	-9.83	-0.38
2c97	2	0.22	0.25	-8.52	-10.52	-2
2dm6	2	0.22	0.2	-5.4	-20.2	-14.8
2dqt	1.8	0.25	0.22	-13.46	-9.31	4.15
2dqu	1.7	0.25	0.17	-13.46	-9.31	4.15
2flh	1.2	0.19	0.08	-5.323	-3.27	2.04
2fqw	1.71	0.21	0.17	-8.95	-10.3	-1.38
2fqx	1.7	0.22	0.18	-9.58	-12.6	-3.05
2fqy	1.9	0.23	0.24	-8.81	-12.9	-4.07
2gud	0.94	0.16	0.02	-5.53	-0.081	5.45
2h2d	1.7	0.24	0.16	-7.36	-5.02	2.3
2h2g	1.63	0.23	0.14	-6.87	-5.91	0.961
2hyq	2	0.24	0.3	-5.65	-0.072	5.58
2hyr	1.51	0.19	0.11	-4.72	-0.048	4.67
2ikg	1.43	0.2	0.09	-8.46	-6.11	2.34
2ikh	1.55	0.21	0.11	-7.48	-2.08	5.4

Table 5: The final training set (n = 112) (5 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

PDB	Res.($\mathring{A}$)	R_{free}	DPI	ΔG	ΔH	TΔS
2iki	1.47	0.18	0.09	-10.2	-14.83	-4.63
2ikj	1.55	0.21	0.12	-10.08	-18.89	-8.81
2iko	1.9	0.23	0.28	-7.5	-9.5	-2
2o4k	1.6	0.21	0.13	-14.3	-4.2	10.1
2o4l	1.33	0.22	0.1	-12.5	-4.3	8.2
2o4n	2	0.23	0.24	-10.6	-3.6	6.9
2o4p	1.8	0.23	0.19	-14.6	-0.7	13.9
2o4s	1.54	0.22	0.13	-14.3	-2.4	11.9
2olb	1.4	na	na	-7.55	9.41	16.91
3mbp	1.7	na	na	-7.23	4.87	12.09
6gss	1.9	0.22	0.22	-5.53	-11.21	-5.65
9gss	1.97	0.23	0.14	-8.04	-16.13	-8.04

Table 6: The final training set (n = 112) (6 of 6) used in the PHOENIX scoring function. Protein Data Bank code (PDB), crystal structure quality metrics (nominal resolution (Res.), free R value (Rfree), diffraction-component precision index (DPI)), and ITC determined thermodynamic parameters (change in binding free energy (ΔG), change in enthalpy (ΔH), and change in entropy (T ΔS)) (kcal/mol) for each complex are listed.

Abbreviation	Description
EIE	Electrostatic Interaction Energy
SIE	Steric Interaction Energy
SF	Steric Fit
RB	Rotatable Bonds
LSE	Ligand Strain Energy
LLCSA1	Hydrophobic/Hydrophobic Contact Surface Area 1
HHCSA1OC	Hydrophilic/Hydrophilic Contact Surface Area 1 (Opposite Charge)
LHCSA1	Hydrophobic/Hydrophilic Contact Surface Area 1
LHCSA1SC	Hydrophilic/Hydrophilic Contact Surface Area 1 (Same Charge)
LLCSA2	Hydrophobic/Hydrophobic Contact Surface Area 2
HHCSA2OC	Hydrophilic/Hydrophilic Contact Surface Area 2 (Opposite Charge)
LHCSA2	Hydrophobic/Hydrophilic Contact Surface Area 2
HHCSA2SC	Hydrophilic/Hydrophilic Contact Surface Area 2 (Same Charge)
LTLSA	Ligand Total Hydrophobic Surface Area
LTHSA	Ligand Total Hydrophilic Surface Area
FI	Flexibility Index(Rot Bonds/ Non Term Bonds)

Table 7: List of the 42 descriptors (1 of 2) used in the final PHOENIX scoring function to derive equations to estimate the thermodynamic parameters.

Abbreviation	Description
LBLSA	Ligand Buried Hydrophobic Surface Area
LBHSA	Ligand Buried Hydrophilic Surface Area
LELSA	Ligand Exposed Hydrophobic Surface Area
LEHSA	Ligand Exposed Hydrophilic Surface Area
RBLSA	Receptor Buried Hydrophobic Surface Area
RBHSA	Receptor Buried Hydrophilic Surface Area
RELSA	Receptor Exposed Hydrophobic Surface Area
REHSA	Receptor Exposed Hydrophilic Surface Area
NLBLSA	Normalized Ligand Buried Hydrophobic Surface Area
NLBHSA	Normalized Ligand Buried Hydrophilic Surface Area
NLELSA	Normalized Ligand Exposed Hydrophobic Surface Area
NLEHSA	Normalized Ligand Exposed Hydrophilic Surface Area
TLRHB	Total Ligand/Receptor Hydrogen Bonds
LTDAC	Ligand Total Donor/Acceptor Count
LTHBA	Ligand Total Hydrogen Bond Atoms
LTBDAC	Ligand Total Buried Donor/Acceptor Count
RTDAC	Receptor Total Donor/Acceptor Count
RTBDAC	Receptor Total Buried Donor/Acceptor Count
XLOGP	Partition Coefficient
LIG_VOL	Ligand Volume
POCK_VOL	Pocket Volume
NB_AS	Number of Alpha Spheres
APOL_AS_PROP	Proportion of Apolar Alpha Spheres
MEAN_LOC_HYD_DENS	Mean Local Hydrophobic Density
POL_SCORE	Polarity Score
AS_DENSITY	Alpha Sphere Density

Table 8: List of the 42 descriptors (2 of 2) used in the final PHOENIX scoring function to derive equations to estimate the thermodynamic parameters.