Supporting Information

RELIABILITY OF CoMFA MODELS: EFFECTS OF DATA SCALING AND VARIABLE SELECTION USING A SET OF HUMAN SYNOVIAL FLUID PHOSPHOLIPASE A_2 INHIBITORS.

Angel R. Ortiz[*†#], Manuel Pastor[§#], Albert Palomer[‡], Gabriele Cruciani[§], Federico Gago[#], Rebecca C. Wade[*†]

Table 1. Details of GRID (version 11) calculations.

Alignment	Cα atoms of modeled ligand-receptor complexes
Charges	ESP/MNDO
Probes	CH_3 and H^+
Electrostatic cutoff (kcal/mol)	30.0
Van der Waals cutoff (kcal/mol)	30.0
Relative dielectric constant	4
Grid spacing (Å)	2
Box size	26 x 28 x 22
Number of variables	6720

CH_3 = Methyl probe (r=1.95Å; q=0.0e); H^+ = Proton probe (r=0.00Å; q=1.0e)

Table 2. Block unscaled weights for X-matrices with different dielectric models

X-Matrix dielectric model[a]	**Block unscaled weights** steric field	electrostatic field
C	1.554	0.796
R	1.548	0.795
W	1.029	0.972
H	1.106	0.919

[a]Dielectric model: C: constant (e = 4); R: distance dependent; W: Warshel model; H: Hingerty model.

Table 3. Predictive performance for the initial X-matrices without application of variable selection[a].

Method[b]	LV[c]	SDEC(%)	R^2	SDEP(%)	Q^2	Xsel[d]
C-N	2	13.24	0.64	21.19	0.077	4775
C-A	1	20.41	0.14	21.96	0.009	3803
C-B	2	12.42	0.68	20.68	0.121	4775
R-N	1	15.97	0.47	20.64	0.124	4599
R-A	4	7.46	0.88	21.60	0.041	4599
R-B	2	12.03	0.70	21.12	0.083	4599
W-N	2	10.66	0.76	21.16	0.080	4600
W-A	3	11.12	0.74	21.37	0.062	4600
W-B	2	11.06	0.74	21.33	0.065	4600
H-N	2	10.30	0.78	21.03	0.091	4600
H-A	4	7.29	0.89	21.52	0.048	4600
H-B	2	11.44	0.73	21.21	0.072	4600

[a]A minimum standard deviation sigma cutoff of 0.05 was used to filter out variables of very low variance before statistical analysis.

[b]Dielectric model: C: constant (e = 4); R: distance dependent; W: Warshel; H: Hingerty. Variable scaling procedure: N: No scaling; A: Autoscaling; B: Block scaling.

[c]Number of latent variables

[d]Number of selected X variables in the model

Table 4. Predictive performance of regression models obtained with a cutoff value of 5 kcal/mol for the computed interaction fields[a].

Method	LV	SDEC(%)	R^2	SDEP(%)	Q^2	Xsel
N-T-C	1	20.17	0.16	21.79	0.020	92
	2	16.49	0.44	19.89	0.187	
	3	13.95	0.60	19.42	0.224	
	4	12.22	0.69	18.95	0.261	
	5	10.94	0.79	19.62	0.208	
B-T-C	1	19.23	0.24	21.37	0.061	171
	2	16.82	0.41	20.16	0.164	
	3	15.19	0.52	23.20	-0.105	
	4	12.70	0.66	25.46	-0.331	
	5	9.82	0.80	27.72	-0.579	

[a]Key as for Table 3 with T denoting use of the Q^2-GRS variable selection method.

Table 5. Effect of grid spacing on predictive performance for the Q^2-GRS variable selection method[a].

Spacing[b]	LV	SDEC(%)	R^2	SDEP(%)	Q^2	Xsel
1	0	22.06	0.00	23.10	-0.09	199
	1	16.73	0.42	19.97	0.18	
	2	12.11	0.70	17.69	0.36	
	3	11.08	0.75	16.52	0.44	
	4	10.22	0.78	17.25	0.39	
	5	9.15	0.83	18.59	0.29	
2	0	22.06	0.00	23.10	-0.09	133
	1	14.87	0.54	18.91	0.27	
	2	10.92	0.75	15.43	0.51	
	3	8.78	0.84	14.26	0.58	
	4	7.37	0.89	13.54	0.62	
	5	6.77	0.90	13.44	0.63	

[a] Key as for Tables 3 and 4. The N-T-C matrix was used and the Q^2-cutoff was set to 0.1 for selecting small boxes for the final model.

[b]Grid spacing (in Å) used in the small sub box calculations (see methods for description). In both cases, final PLS models were obtained with a 2 Å grid spacing.