

Template CoMFA Applied to 116 Biological Targets

Richard D. Cramer*

Certara Corporation, 240 N. Tucker Blvd., St. Louis, Missouri 63101, United States

Supporting Information

ABSTRACT: Statistically acceptable 3D-QSAR models were effortlessly obtained, by application of automatic template CoMFA, for the vast majority, arguably more than 95%, of 116 biological targets. Among these targets, 76 were structure-based, pooling multiple templates (as protein-bound conformation of ligands) and training sets into a single model, and the other 40 were ligand-based, with a low-energy conformation of an exemplar training set structure becoming the single template. Criteria proposed for statistical acceptability are a leave-one-out $q^2 > 0.4$, or a standard error of leave-one-out prediction <1.0 , or a ratio of best-fit r^2 to count of PLS components >0.2 . The structure-based 3D-QSAR models provide direct visual comparisons of SAR-derived 3D-QSAR contours with cavity surfaces.

Target	Tmp	n	q^2	SDEP	#Cp	r^2	s
3-oxoacyl-[acyl-carrier-protein]							
1 synthase	2	24	0.794	0.29	5	0.998	0.03
2 Aldose Reductase	8	47	0.822	0.76	3	0.921	0.51
:	:	:	:	:	:	:	:
73 Transferrin	2	22	0.147	1.57	4	0.983	0.44
74 VEGFR-2 (KDR)	4	51	0.398	0.88	2	0.703	0.62
75 WEE1	1	24	0.549	0.73	4	0.730	0.26
76 X-linked inhibitor of apoptosis	2	31	0.346	0.93	3	0.712	0.61
	505	6180	0.546	0.71	7	0.844	0.37
38 Topoisomerase IIa inhib	1 ^a	25	0.444	0.71	4	0.893	0.31
39 Vascular H ⁺ -type ATPase	1 ^a	23	0.366	1.42	3	0.875	0.63
40 Yersinia Tyr phos inhib	1 ^a	39	0.413	1.34	3	0.759	0.86
	2993	0.380	0.92	5	0.780	0.52	

To a prospective user, adoption of a new CADD methodology involves a risk/benefit decision. There are relatively certain costs, with the time needed to install and learn a new method often being an obstacle greater than any licensing expense. The benefits are more speculative—what will the method contribute to current and future discovery activities?

Validation studies are intended to provide objective information that allows assessment of these benefits. Typically a single set of input data is taken through a workflow with results that at minimum are consistent with expectations and therefore show that the method's implementation works as described. For instance, CoMFA was introduced¹ with a single validating application to corticosteroid globulin binding. At that time, the central issue was whether PLS analysis of the far-more-columns-than-rows descriptor array, representing the fields surrounding a set of ligands, could yield a robust and informative and predictive model of experimental potencies, an outcome that all previous QSAR experience using multiple regression contradicted.

However, the structures of ligands and their targets are so astronomically diverse that such a single validation example leaves almost completely open the question of generality. How likely is the new method to contribute to a specific project? More examples can help, if chosen in an unbiased or at least appropriately biased way. When topomer CoMFA was introduced² as an approach for simplifying 3D-QSAR alignments, the central issue was whether the topomer approximations of 3D structure would maintain statistical quality. This was addressed by application to a relatively large number of validation data sets, the first 15 3D-QSAR publications randomly encountered by paging through recent journals. All yielded models with topomer CoMFA, on average slightly though not statistically significantly inferior to the published ones, yet seemingly as satisfactory for use in prediction. This 15

for 15 success rate from an unbiased data set collection would seem to justify an expectation that, for structurally appropriate data sets, topomer alignments are as likely to be as effective as tortuous manual alignments in yielding useful 3D-QSAR models. However, because of the bias that 3D-QSAR had already succeeded with these data sets, this study did not address the more fundamental concern, whether application to a new data set is likely to produce a useful 3D-QSAR model.

Template CoMFA, a very new 3D-QSAR methodology,³ introduces two significantly distinctive⁴ capabilities: (1) Completely automatic alignments, with the only required inputs being one or more templates, typically ligands superimposed in receptor-bound conformations, and SAR tables with connectivities only for their structures. (2) Generation of a single 3D-QSAR multiseries model for a target, simply by pooling all the available template and SAR information.

Contrary to the expectation that structural locality will yield better QSAR models, these “pooled” template CoMFA models (automatically “anchored”) for factor Xa and p38 map kinase have overall statistical properties at least as good as their constituent single template-and-series template CoMFA models.³ A second study⁵ compared template CoMFA models (manually “anchored”) with X-ray-for-every-ligand models for urokinase, protein-tyrosine-phosphatase-1B (PTP-1B), and checkpoint kinase 1 (Chk-1). These template CoMFA models were at least statistically equivalent to those from X-ray alignments and more interpretable because of more focused coefficient contours.

The goal of the current studies is to determine the likelihood that a useful template CoMFA model can be obtained for an arbitrary biological target of interest. Breadth-first investigation,

Received: April 14, 2014



Table 1. Statistical Metrics for Automatic Template CoMFA of All bindingdb Targets Referencing Multiple Receptors for Two PLS Component Count Criteria^a

target	code	# tmpl	n	no 3D model/ fields				# cmp minimizing SDEP				# cmp exceeding # Tmp!				$r^2/\#$ cmp
				q^2	SDEP	# cmp	r^2	s	$r^2/\#$ cmp	q^2	SDEP	# cmp	r^2	s	$r^2/\#$ cmp	
3'-oxoacyl[acyl-carrier-protein]	3-oxoacyl	2	24	0	0.794	0.29	5	0.998	0.03	0.20	0.666	0.34	2	0.804	0.26	0.40
aldred	aldred	8	47	0	0.822	0.76	3	0.921	0.51	0.31	0.738	0.90	2	0.889	0.59	0.44
androgen receptor	androgen	11	106	0	0.413	0.71	3	0.576	0.61	0.19	0.413	0.71	3	0.576	0.61	0.19
aurora kinase A	aurora	7	76	0	0.441	0.63	4	0.758	0.41	0.19	0.441	0.63	4	0.758	0.41	0.19
β -secretase	bace	14	194	2	0.909	0.49	17	0.989	0.15	0.06	0.858	0.53	6	0.913	0.41	0.15
carbonic anhydrase 2	carbanhyd2	20	158	0	0.827	0.48	14	0.988	0.12	0.07	0.800	0.50	9	0.964	0.21	0.11
carbonic anhydrase IV	carbanhyd4	3	38	0	0.316	0.71	5	0.912	0.25	0.18	0.228	0.72	2	0.599	0.52	0.30
cathepsin B	cathB	3	47	0	0.822	0.76	3	0.921	0.51	0.31	0.822	0.76	3	0.921	0.51	0.31
cathepsin K	cathK	8	109	5	0.649	0.94	10	0.860	0.59	0.09	0.584	0.99	4	0.728	0.80	0.18
cdk2	cdk2	20	234	1	0.794	0.49	17	0.973	0.18	0.06	0.771	0.51	10	0.921	0.30	0.09
chk1	chk1	20	263	65	0.480	1.00	13	0.925	0.38	0.07	0.461	1.01	9	0.868	0.50	0.10
chk2	chk2	3	25	0	0.177	0.41	5	0.946	0.11	0.19	-0.138	0.45	2	0.723	0.22	0.36
cmet	cmet	5	69	0	0.749	0.47	17	0.999	0.02	0.06	0.090	0.78	2	0.279	0.69	0.14
cycA2	cycA2	20	194	0	0.602	0.76	8	0.839	0.48	0.10	0.602	0.76	8	0.839	0.48	0.10
dihydrofolate reductase (DHFR)	dfr	10	137	0	0.779	0.75	10	0.900	0.50	0.09	0.739	0.80	4	0.824	0.65	0.21
dihydroorotate dehydrogenase (DHODH)	dhood	11	147	10	0.110	0.95	2	0.260	0.87	0.13	0.110	0.95	2	0.260	0.87	0.13
dUTP	dUTP	2	25	0	0.844	0.44	2	0.933	0.28	0.47	0.844	0.44	2	0.933	0.28	0.47
epoxyhydase	epoxyhy	10	119	0	0.749	0.84	8	0.950	0.37	0.12	0.699	0.91	5	0.887	0.56	0.18
estrogen	estrogen	6	19	0	0.571	0.56	3	0.872	0.31	0.29	0.571	0.56	3	0.872	0.31	0.29
eukar	eukar	2	28	0	0.818	0.59	6	0.994	0.11	0.17	0.459	0.93	2	0.684	0.71	0.34
Factor Xa (FXa)	facXa	12	270	8	0.579	0.86	6	0.766	0.64	0.13	0.579	0.86	6	0.766	0.64	0.13
farn	farn	4	36	0	0.517	0.70	6	0.979	0.15	0.16	0.182	0.86	2	0.626	0.58	0.31
glucarib	glucarib	3	25	0	0.213	1.37	5	0.898	0.49	0.18	-0.347	1.66	2	0.504	1.01	0.25
glutamate receptor ionotropic kainate 2	glutkain	5	24	0	0.248	1.31	2	0.742	0.77	0.37	0.248	1.31	2	0.742	0.77	0.37
glutathione S-transferase Pi	glutath	5	47	4	1.000	0.00	9	1.000	0.00	0.11	0.752	0.25	2	0.808	0.22	0.40
HCV NSSB polymerase	hcvNSSB	8	121	30	0.603	0.72	17	0.999	0.03	0.06	0.475	0.76	4	0.769	0.50	0.19
hepatocyte growth factor receptor	hepat	14	115	9	0.748	0.74	14	0.990	0.15	0.07	0.552	0.94	3	0.684	0.79	0.23
HIV-1 protease	hivprot	10	69	25	0.330	1.02	2	0.535	0.85	0.27	0.330	1.02	2	0.535	0.85	0.27
HIV-1 reverse transcriptase RNase H	hivRNase	4	43	0	0.403	0.78	11	0.986	0.12	0.09	0.128	0.83	2	0.425	0.67	0.21
17- β -hydroxysteroid dehydrogenase 5	hsdh5	3	36	0	0.491	0.72	8	0.637	0.55	0.08	0.411	0.64	2	0.467	0.59	0.23
heat shock protein 90 (Hsp90)	hsp90	4	39	27	-0.254	0.61	2	0.663	0.32	0.33	-0.254	0.61	2	0.663	0.32	0.33
HSP90AA1 protein	hspAA1	9	108	0	1.000	0.01	11	1.000	0.01	0.09	0.947	0.12	4	0.960	0.11	0.24
enoyl-ACP reductase (InhA)	InhA	6	65	1	0.180	0.72	2	0.316	0.66	0.16	0.180	0.72	2	0.316	0.66	0.16
integrin α -IIb	integrin	2	29	0	0.551	0.99	4	0.910	0.45	0.23	0.476	1.03	2	0.803	0.63	0.40
tyrosine-protein kinase ITK	itk	2	18	0	0.589	1.12	3	0.988	0.12	0.33	0.538	1.15	2	0.942	0.41	0.47
tyrosine-protein kinase Jak2	jak2	6	83	2	0.201	1.20	2	0.340	1.09	0.17	0.201	1.20	2	0.340	1.09	0.17
kinesin	kinesin	13	149	0	0.857	0.58	13	0.970	0.27	0.07	0.814	0.65	6	0.909	0.45	0.15
LHX	LHX	4	43	0	-0.126	1.43	2	0.107	1.28	0.05	-0.126	1.43	2	0.107	1.28	0.05
mapk	mapk	7	131	0	0.558	0.96	6	0.840	0.58	0.14	0.534	0.98	3	0.671	0.82	0.22
matrigip	matrigip	2	25	3	0.116	0.42	7	0.999	0.05	0.14	-0.065	0.42	2	0.603	0.25	0.30
metap2	metap2	3	75	0	0.393	1.14	14	0.948	0.33	0.07	0.078	1.28	2	0.382	1.05	0.19

Table 1. continued

target	code	# tmp	n	no 3D model/ fields	# cmp minimizing SDEP				# cmp exceeding # Tmp				$r^2/\#$ cmp		
					q^2	SDEP	# cmp	r^2	SDEP	# cmp	q^2	SDEP			
MAPK-activated protein Kinase 2 (MK2)	MK2	4	78	0	0.884	0.40	13	0.993	0.10	0.08	0.844	0.43	3	0.899	0.35
ML-IAP-BIR	milapbir	2	27	0	0.536	0.55	4	0.907	0.24	0.23	0.278	0.65	2	0.626	0.47
matrix metalloproteinase-1 (MMP-1)	mamp1	3	40	2	-0.116	0.96	2	0.318	0.75	0.16	-0.116	0.96	2	0.318	0.75
matrix metalloproteinase-3	mamp3	4	69	0	0.463	0.69	9	0.960	0.19	0.11	0.289	0.75	2	0.460	0.66
MurD	murD	4	67	2	0.954	0.09	9	0.964	0.08	0.11	0.633	0.23	2	0.700	0.21
serine/threonine protein kinase NEK2	nek2	9	158	0	0.749	0.39	17	0.900	0.25	0.05	0.551	0.50	4	0.674	0.43
Poly [ADP-ribose] polymerase-1	PADPr	2	49	0	0.548	0.58	2	0.670	0.49	0.34	0.548	0.58	2	0.670	0.49
Phosphodiesterase 10A1	pde10A1	2	27	11	0.191	0.82	2	0.807	0.40	0.40	0.191	0.82	2	0.807	0.40
Phosphodiesterase Type 5 (PDE5A)	pde5	4	42	2	0.565	0.55	3	0.773	0.40	0.26	0.281	0.70	2	0.511	0.58
enoyl-ACP reductase (PENR)	PENR	4	53	0	0.273	0.48	10	0.996	0.04	0.10	-0.153	0.56	2	0.307	0.43
PLM1	pim1	14	191	2	0.557	0.76	8	0.732	0.59	0.09	0.704	0.63	6	0.685	0.64
protein kinase B (Akt 1)	pBbAkt1	2	23	0	0.525	0.86	3	0.935	0.32	0.31	0.365	0.97	2	0.850	0.47
protein kinase B (Akt 2)	pBbAkt2	2	24	0	0.494	0.64	5	0.901	0.28	0.18	0.411	0.64	2	0.736	0.43
plasmeprin 2	plasmeprin	2	26	0	0.342	1.00	5	0.962	0.24	0.19	0.203	1.03	2	0.785	0.53
phenylethanolamine- N-methyltransferase (PNMT)	pnmt	5	60	11	0.513	0.80	13	0.950	0.25	0.07	0.233	0.87	2	0.564	0.65
PPTase Pin1	PPTase	2	17	0	0.836	0.27	2	0.988	0.17	0.49	0.836	0.27	2	0.988	0.17
prolyl endopeptidase	prolyl	2	24	0	0.028	1.12	5	0.923	0.32	0.18	-0.195	1.15	2	0.661	0.61
protein-tyrosine phosphatase 1B (PTP1B)	ptp1b	20	243	39	0.914	0.44	17	0.977	0.23	0.06	0.877	0.52	9	0.948	0.34
protein tyrosine kinase 2 β	ptyk2b	2	29	0	0.563	0.49	3	0.828	0.31	0.28	0.474	0.53	2	0.685	0.41
glycogen phosphorylase (PYGM)	PYGM	14	90	2	0.626	0.77	17	0.964	0.22	0.06	0.588	0.71	7	0.886	0.37
quinone reductase 1	qred1	2	23	0	0.475	0.94	2	0.775	0.61	0.39	0.475	0.94	2	0.775	0.61
quinone reductase 2	qred2	12	147	0	0.823	0.42	13	0.982	0.13	0.08	0.777	0.46	5	0.899	0.31
renin	renin	20	366	55	0.694	0.71	17	0.931	0.34	0.05	0.643	0.72	9	0.855	0.46
RNase A	Rnase	14	148	7	0.998	0.05	6	0.999	0.03	0.17	0.998	0.05	6	0.999	0.03
receptor-type tyrosine-protein phosphatase γ	rtypP	5	44	0	0.829	0.35	8	0.978	0.13	0.12	0.565	0.52	2	0.792	0.36
S-adenosylmethionine decarboxylase (AdoMetDC)	sadmet	6	75	0	0.998	0.23	17	1.000	0.01	0.06	0.907	0.52	3	0.937	0.43
S-enolpyruvylshikimate-3-phosphate synthase	shik	2	51	22	0.890	0.86	4	0.944	0.62	0.24	0.823	1.05	2	0.879	0.87
tyrosine-protein kinase SRC	src	7	85	3	0.835	0.70	7	0.952	0.38	0.14	0.764	0.81	3	0.840	0.67
thymidylate synthase	thymys	7	85	0	0.221	1.26	4	0.268	1.22	0.07	0.221	1.26	4	0.268	1.22
thymidylate synthase (TS)	thys	4	41	1	0.382	0.82	5	0.770	0.50	0.15	0.142	0.93	2	0.338	0.82
Tie-2	tie	3	40	0	0.589	0.65	5	0.901	0.32	0.18	0.351	0.79	2	0.713	0.52
transthyretin	transthy	2	22	0	0.147	1.57	4	0.933	0.44	0.23	0.046	1.57	2	0.519	1.11
vegf	vegf	4	51	0	0.398	0.88	2	0.703	0.62	0.35	0.398	0.88	2	0.703	0.62
VEGFR-2 (KDR)	WEE1	2	24	0	0.549	0.73	4	0.750	0.26	0.18	0.389	0.81	2	0.837	0.42
WEE1	Xlink	2	31	13	0.346	0.93	3	0.712	0.61	0.24	0.130	1.03	2	0.651	0.65
X-linked inhibitor of apoptosis		506	6180	5	0.546	0.71	7	0.844	0.37	0.12	0.432	0.77	3	0.698	0.55

^aSee text for details.

application of template CoMFA to as many data sets as possible, seems an approach more relevant than careful depth-first investigations of a much smaller number of targets. Furthermore, such a somewhat “quick-and-dirty” approach is expected to yield poorer results for each individual target and thus should yield a relatively conservative estimate of template CoMFA’s scope of applicability.

Most fortunately for this approach, at the bindingdb URL⁶ there recently appeared validation sets for around 100 different targets. For each of these targets, bindingdb supplies one or more 3D templates and 2D training sets, exactly matching template CoMFA’s input requirements in a format very convenient for use. This major expansion of bindingdb contents, coupled with the simplicity and rapidity of automatically anchored template CoMFA, enabled an unprecedentedly large validation study, focused entirely on template CoMFA’s distinctive abilities to automatically generate a single 3D-QSAR model for a target from multiple templates and training sets. As of early 2014 there were available 76 targets referencing more than one .pdb entry. An assessment so extensive should yield an exceptionally robust estimate of the likelihood of obtaining a statistically acceptable 3D-QSAR model for an arbitrary new target by applying automatic template CoMFA.

Of course, statistical acceptability does not guarantee a QSAR model’s relevance to project goals. A 3D-QSAR model allows a comparison of its trends to such other experimental data as an X-ray structure, which can improve understanding of and confidence in its predictions. Therefore, such visual comparisons are presented here for six randomly chosen targets.

Finally, for many biological targets there are no X-ray structures available to provide templates. However, in principle any reasonable 3D structure might serve as a template. To assess template CoMFA’s general applicability in this situation, it was also applied to another collection of 40 downloadable “QSAR validation” training sets,⁵ using as a single template for each the Concord model (or where necessary an energy minimized conformer) of a rule-selected structure from that training set.

METHODS

Data Preparation. For each of the 76 multitemplate targets, all of the .pdb structures referenced at bindingdb, up to a maximum of the first 20 such entries per target, were downloaded from the Protein Data Bank. One ligand was extracted from each of these .pdb files, as follows. To generate the requisite common coordinate reference frame for the ligands binding to a particular target, these .pdb entries were read into SYBYL and all but the first superimposed onto that first, by applying SYBYL’s Biopolymer → Compare Structures → Align Structures by Homology ... menu item to all C-alphas. (The algorithm generating these homologies is the Orchestrar⁷ technology with all defaults.) Qualitatively, over 75% of the resulting alpha-carbon RMSD’s were <1.0 Å, and over 95% were <2.0 Å, the larger RMSD values being associated with extensive sequence insertions or deletions. Hydrogens were then added to each structure by applying the Biopolymer → Prepare Structures → Add Hydrogens ... operation to each.

To select the template ligand to be extracted from a .pdb structure, all atoms that were not .pdb-designated ligand atoms were hidden from the display. Whenever the remaining atoms defined more than one chemical structure, selection of the single ligand first considered any distinctive structure features

suggested by inspection of the associated training set structures, without referring to the candidate’s pocket’s location. Any remaining ambiguity was resolved by selecting the structure that seemed most similarly located to the others (as seen from the graphical projection originally established by the coordinates of the .pdb for that target’s first bindingdb entry). The selected structure was extracted and appended to a file containing the other such templates, which thus in its final version contained one template from each of the .pdb files.

Training sets were simply the “2D” .sdf files downloaded from the bindingdb URL, pooled for each target, and used without any inspection of the structures unless some unexpected downstream behavior required diagnosis. The biological potencies, given in concentration units, were transformed logarithmically to their free energy difference equivalents, as is usual for QSAR studies.

Template CoMFA Alignments. In the “automatic” mode used for all of these studies, this step is literally a single mouse click. The underlying algorithms⁵ may be summarized as two steps: (1) Placement of the candidate training (or test) set structure, by overlaying its “best-matching” bond onto that template bond which it best “matches”, where a major determinant of “match” is the count of connected atoms that “match” between candidate and template. (2) Placement of the individual atoms within the candidate training (or test) set structure, by “copying” coordinates to candidate atoms from those template atoms that “match” them and applying the topomer protocol⁹ to candidate atoms that do not “match” template atoms.

Template CoMFA Model Generation. The usual method of partial least-squares (PLS) with leave-one-out (LOO) cross-validation was applied. One continuing issue with PLS is the choice of the number of PLS components to combine when producing a final “best-fit” model. The CoMFA default is the number that minimizes the standard deviation of the error of prediction (SDEP). However, this criterion can yield a final model whose “*s*” (standard error of best-fit residuals) is suspiciously less than the expected experimental uncertainties of the biological potency values being fit to. Therefore, for the bindingdb targets having the multiple templates, an objective if arbitrary additional constraint was also tried, such that component extraction ended whenever the count of those components exceeded one-half of the number of templates.

RESULTS

The statistical measures of quality for each of these 116 validation biological targets appear in Table 1, for the bindingdb multitemplate targets, and in Table 2, for the single-largest-template QSAR validation sets. From left to right, the columns appearing in both tables are the target description; a code providing a convenient reference to the target; the count of pooled templates; the count of the pooled training set structures; the count of training set structures that could not be included in the model because of failure in either structure preparation, structure alignment, or electrostatic field calculation; the maximum LOO q^2 and minimum SDEP values during cross-validation; the number of PLS components that yielded the minimum SDEP value and was therefore used in the final “best-fit” PLS model; and the r^2 and *s* values for that final model.

Table 1 then repeats the q^2 , SDEP, number of components, r^2 , and *s* for the models applying the above-mentioned additional constraint that the number of components must

Table 2. Statistical Metrics for Automatic Template CoMFA of QSAR Benchmark Data Sets for Two Template Selection Rules^a

target	code	# tmp1	n	no 3D model/ fields	q^2	SDEP	# cmp	$r^2/\#$ cmp	s	q^2	SDEP	# cmp	r^2	s	$r^2/\#$ cmp	
angiotensin converting enzyme	ACE	1	114	7	0.536	1.60	6	0.777	1.11	0.13	0.536	1.60	6	0.777	1.11	0.13
AChE inhibition	ACHE	1	111	35	0.351	0.99	2	0.528	0.84	0.26	0.527	0.87	8	0.930	0.34	0.12
irreversible AChE inhib	ACHE_TOX	1	30	0	-0.501	1.13	2	0.660	0.54	0.33	-0.375	1.07	2	0.698	0.50	0.35
steroid aromatase	AI	1	78	2	0.621	0.68	3	0.808	0.48	0.27	0.621	0.68	3	0.808	0.48	0.27
amphiphysin I	AMPHI	1	130	72	0.510	0.54	6	0.951	0.17	0.16	0.510	0.54	6	0.951	0.17	0.16
anticancer	APZ	1	17	1	0.433	0.56	2	0.867	0.27	0.43	0.344	0.60	2	0.761	0.36	0.38
angiotensin receptor	ARB	1	28	3	0.033	0.98	2	0.855	0.38	0.43	0.332	0.91	2	0.756	0.55	0.38
antituberculosis	ATA	1	94	1	0.117	1.18	3	0.514	0.88	0.17	0.108	1.20	3	0.558	0.84	0.19
benzodiazepine receptor	BZR	1	163	0	0.229	0.98	3	0.416	0.85	0.14	0.391	0.88	7	0.819	0.48	0.12
cannabinoid I	CBRA	1	31	0	0.400	0.85	3	0.834	0.45	0.28	0.400	0.85	3	0.834	0.45	0.28
catechol OMe transferrase	COMT	1	92	1	0.498	0.63	12	0.880	0.13	0.07	0.647	0.51	6	0.910	0.26	0.15
cyclooxygenase II	COX2	1	322	0	0.338	1.17	5	0.552	0.96	0.11	0.357	1.16	6	0.636	0.87	0.11
dopamine 2 receptor	D2A	1	26	0	-0.445	1.70	2	0.614	0.88	0.31	0.169	1.32	3	0.783	0.67	0.26
dopamine transporter	DAT	1	42	2	0.266	0.69	4	0.777	0.38	0.19	0.366	0.64	4	0.866	0.30	0.22
dihydrofolate reductase	DHFR	1	397	3	0.615	0.93	8	0.809	0.66	0.10	0.565	0.99	6	0.825	0.63	0.14
diazepam recep. isoform	DIAZEPAM	1	42	0	0.756	0.60	8	0.975	0.19	0.12	0.756	0.60	8	0.975	0.19	0.12
dopamine receptor	DR	1	38	10	0.482	0.95	2	0.797	0.59	0.40	0.476	0.98	2	0.859	0.51	0.43
ecdysone	ECR	1	50	0	0.216	0.73	9	0.901	0.26	0.10	0.216	0.73	9	0.901	0.26	0.10
estrogen receptor	EDC	1	123	6	0.451	1.33	3	0.699	0.98	0.23	0.451	1.33	3	0.699	0.98	0.23
enterovirus 71 inhib	EZ71	1	41	0	0.442	0.76	5	0.921	0.29	0.18	0.442	0.76	5	0.921	0.29	0.18
growth hormone secretagogue	GHS	1	31	0	0.230	0.96	7	0.939	0.27	0.13	0.455	0.75	3	0.832	0.42	0.28
glycogen phosphorylase B inhib	GPB	1	66	0	0.378	0.92	4	0.737	0.60	0.18	0.378	0.92	4	0.737	0.60	0.18
GSK3B	GSK3B	1	42	0	0.640	0.52	7	0.861	0.32	0.12	0.643	0.50	5	0.838	0.24	0.17
HIVPR	HIVPR	1	113	0	0.011	1.36	2	0.247	1.19	0.12	-0.080	1.42	2	0.246	1.19	0.12
HVRT	HVRT	1	101	6	0.777	0.64	5	0.910	0.41	0.18	0.777	0.64	5	0.910	0.41	0.18
human protein tyrosine phosphotase inhib	HPTP	1	135	10	0.636	0.38	5	0.841	0.25	0.17	0.636	0.38	5	0.841	0.25	0.17
κ opioid inhib	KOA	1	39	16	0.666	0.54	3	0.865	0.34	0.29	0.666	0.54	3	0.865	0.34	0.29
melatonin receptor	MT	1	56	0	0.722	0.39	11	0.986	0.09	0.09	0.930	0.20	12	0.984	0.10	0.08
nutagenic potential	MX	1	29	0	0.689	0.94	4	0.909	0.51	0.23	0.706	0.93	5	0.920	0.46	0.18
PDE IV inhib	PDE	1	29	0	-0.159	0.68	2	0.444	0.47	0.22	-0.079	0.65	2	0.446	0.47	0.22
phospholipase A2 inhib	PLA2	1	11	0	0.396	1.21	2	0.790	0.71	0.40	0.396	1.21	2	0.790	0.71	0.40
enantiomeric selectivity of phase-transfer catalysis	PTC	1	40	11	0.674	0.60	3	0.907	0.24	0.30	0.674	0.60	3	0.907	0.24	0.30
ryanodine	RYR	1	18	1	0.278	0.96	2	0.814	0.49	0.41	0.278	0.96	2	0.814	0.49	0.41
steroids	STEROIDS	1	21	0	0.826	0.52	2	0.966	0.23	0.48	0.878	0.43	2	0.953	0.27	0.48
TCHK	TCHK	1	42	16	0.158	0.98	2	0.480	0.77	0.24	-0.036	1.06	2	0.456	0.77	0.23
TERM	TERM	1	76	1	0.310	1.89	15	0.911	0.68	0.06	0.425	1.59	5	0.842	0.83	0.17
THR	THR	1	88	15	0.396	0.86	6	0.912	0.33	0.15	0.158	0.97	5	0.819	0.45	0.16
TP2A	TP2A	1	25	0	0.444	0.71	4	0.893	0.31	0.22	0.404	0.72	3	0.836	0.38	0.28
VATP	VATP	1	23	4	0.366	1.42	3	0.875	0.63	0.29	0.366	1.42	3	0.875	0.63	0.29
YOPH	YOPH	1	39	4	0.413	1.34	3	0.759	0.86	0.25	0.082	1.67	4	0.862	0.65	0.22
		2993	6	0.380	0.92	5	0.780	0.52	0.17	0.412	0.89	4	0.801	0.50	0.19	

^aSee text for details.

not exceed the ceiling of one-half of the number of templates. Since there is only one template for the targets in Table 2, this additional constraint is not applicable. Instead, the second block contains the same five statistics with the ligand having the highest ligand efficiency¹⁰ as the template, as an alternative to the largest molecular weight structure used as the template for the first block.

Which among these measures of statistical quality are the best indicators of whether a particular model will be useful to a project? There are several considerations. Most analysts focus on q^2 . However, a model's guidance value to a project primarily depends on the magnitude of its errors in predicting individual biological potencies, which an SDEP value in LOO cross-validation directly represents, while q^2 depends as much on the variation among the training set's experimental potencies as on SDEP. A further complication is that LOO cross-validation can reject a useful model whenever its training set includes structures that uniquely combine extreme potency observations with unusual structural features. These points are discussed below in more detail.

From these considerations, the following series of guidelines is proposed as a single statistical criterion for the statistical acceptability of a PLS-derived QSAR model: (1) The SDEP should be 1.00 or less; or (2) the q^2 should be >0.4 ; or (3) the ratio of r^2 to the number of components should be >0.2 .

Application of this criterion to Table 1 indicates that only four of its 76 models (5%) are statistically unacceptable: (jak2) tyrosine protein kinase; (LHX) Liver X receptor β ; (metap2) methionine aminopeptidase type 2; (thyms) thymidylate synthase. In Table 2, three of the 40 models (7.5%) are unacceptable: (ATA) antituberculosis; (COX2) cyclooxygenase II; (HIVPR) HIV-protease inhibition.

Relaxation of the q^2 threshold from 0.4 to 0.2 (the floor value in SYBYL for rejecting a topomer CoMFA model) additionally accepts all except the LHX model in Table 1 (99%) and all except ATA and HIVPR in Table 2 (95%). At an intermediate q^2 threshold of 0.3 the rejected model counts are three in Table 1 (4%) and two in Table 2 (5%).

To provide visual impressions of these models' contents, Figures 1–5 were created for a few targets randomly chosen from Table 1. The viewing directions for all 3D images are inherited from the first .pdb reference for that target at the bindingdb URL. Their scalings may slightly differ. The top pair of images in each Figure show the CoMFA steric stdev*coef "fields", contoured at standardized levels, surrounding the aligned training set structures, with cyan rather than the familiar green enclosing regions where bulk favors potency and yellow enclosing the disfavoring regions. Its right-hand image also includes a 10 Å Z-clipped green-colored target cavity surface centered near the $Z = 1.0$ coordinate, again as determined by that target's first .pdb reference at the bindingdb URL. The middle pair of images provides the same sort of information for the CoMFA electrostatic contours, with the target cavity surface colored by its electrostatic potential on a scale that crosses the entire visible spectrum from red (most positive) to purple (most negative). The bottom image pair shows the templates for that target, individually as 2D structures and collectively overlaid, including all of the 3D structural information actually used for the training set alignments.

Presumably the right-hand images that also include the cavity surfaces are of greatest interest. Although these views are objectively if arbitrarily standardized, rather than manually adjusted to optimize the portrayal of some structural feature, in

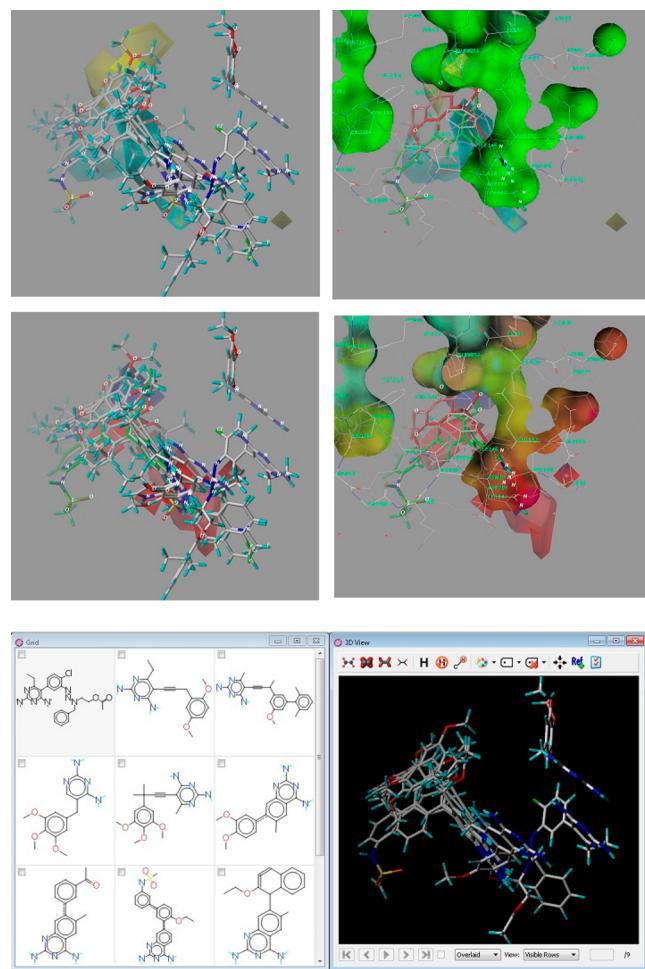


Figure 1. Steric and electrostatic stdev* coefficient template CoMFA contours, without and with 1VJ3.pdb added, and templates, for dihydrofolate reductase. See text for details.

general these images do seem in accord with expectations. Wherever there has been enough ligand variation to generate a QSAR trend, in the steric maps the boundaries of the bulk-desirable cyan CoMFA contours do tend to just touch the green contours that delineate the surface of the otherwise empty pocket. The yellow bulk-undesirable contours produce yellow-green hues wherever they overlap pocket walls, although no good examples appear in any of these images. Turning to the electrostatic images, their (nonadjustable) use of the entire visible spectrum to represent variation in electrostatic potential at the cavity surface is awkward, especially where superposition of these surfaces yields a brown projected image. However, there are several occurrences of the expected red field contours adjacent to red cavity surfaces and blue contours adjacent to violet surfaces.

Figures 6 and 7 provide visual impressions from Table 2, where each template is the Concord model of a particular training set structure. To provide some sense of how the selection of template structure can change the resulting model, the two figures are from the 2 of the 40 targets for which the template selection rule change (greatest molecular weight on the left vs greatest ligand efficiency on the right) produced the largest change in model statistics. The three overlaid structures in each image, which despite appearances are 2D identical within a figure pair, are the most (green) and the least (red)

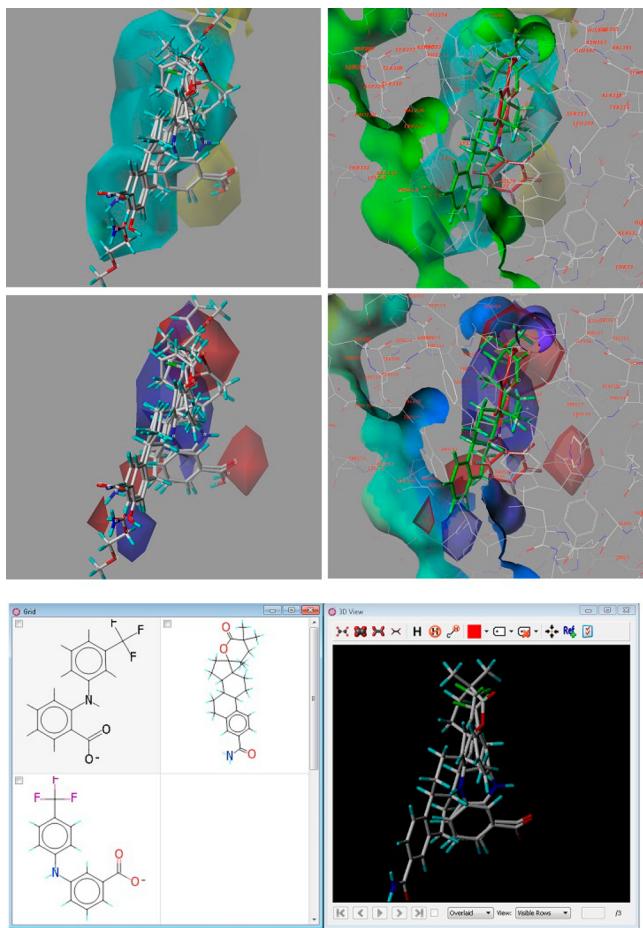


Figure 2. Steric and Electrostatic stdev* coefficient template CoMFA contours, without and with 1S2C.pdb added, and templates, for 17- β -hydroxysteroid dehydrogenase 5. See text for details.

active and the template itself (white). Clearly evident are the very large effects on alignments => models => predictions that a change in a single template structure can produce. As an extreme example, in Figure 7 the (rather ligand-inefficient!) structure having greatest potency generates a very extended training set, with the most and least active structures being optimally aligned at its opposite ends.

■ DISCUSSION

First some caveats should be acknowledged: any study that includes 116 targets can only superficially consider the individual targets; any guidelines for statistical acceptability are somewhat arguable; statistical metrics are only one criterion of a model's quality and usefulness; and even this independently chosen collection of 116 targets must include some selection bias and thus may be imperfectly representative of "any biological target".

Nevertheless, application of template CoMFA arguably succeeded at least 95% of the time, by yielding statistically acceptable models in 109–113 of 116 attempts (count depending on the exact statistical criterion). This very high success rate would seem to provide extraordinarily strong justification—most probably unprecedentedly so among all CADD methodologies—for trying template CoMFA on almost any SAR table of interest. Surely a further encouragement must be the fact that these 116 trials occupied <40 h of the author's

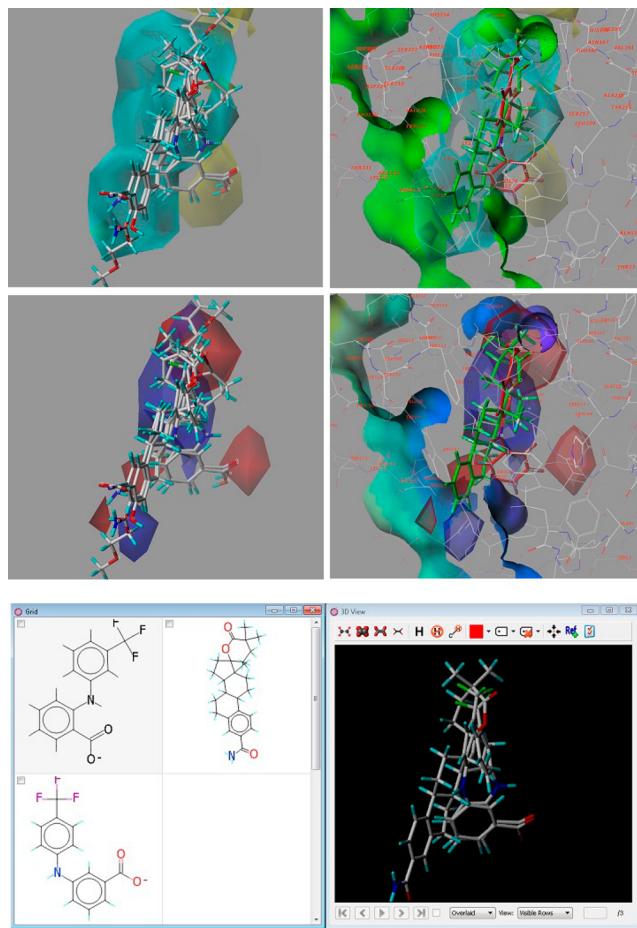


Figure 3. Steric and electrostatic stdev* coefficient template CoMFA contours, without and with 1HFS.pdb added, and templates for matrix metalloproteinase-3. See text for details.

time (excluding fixes of the bugs inevitably encountered within such a diversity of structures).

More specifically, these results also validate as general the two significantly distinctive capabilities that template CoMFA offers: (1) Completely automatic alignments, with the only required inputs being one or more templates, typically ligands superimposed in receptor-bound conformations, and SAR tables with only connectivities for their structures. (2) Generation of a single 3D-QSAR multiseries model for a target, by pooling all the available template and SAR information.

Of course this extraordinarily high success rate partially depends on the accompanying newly proposed criterion for statistical acceptability. As an alternative, the almost universal QSAR practice (which ironically the author may have initiated)¹¹ is to consider only q^2 as a yes–no model acceptability criterion, applying such threshold values as 0.2, 0.4, or 0.6. With these q^2 values as the only acceptability thresholds, the rates and counts for acceptable models in Tables 1 and 2 combined are for $q^2 = 0.2$, 86% (100), for $q^2 = 0.4$, 66% (77), and for $q^2 = 0.6$, 38% (44). This drop in the number of acceptable models for a q^2 -only threshold as demanding as 0.6 may be interpreted as an indication that this newly proposed three-component criterion is too aggressive. In practice, different q^2 thresholds are usually used for the differing needs as a project evolves, of course also considering other criteria

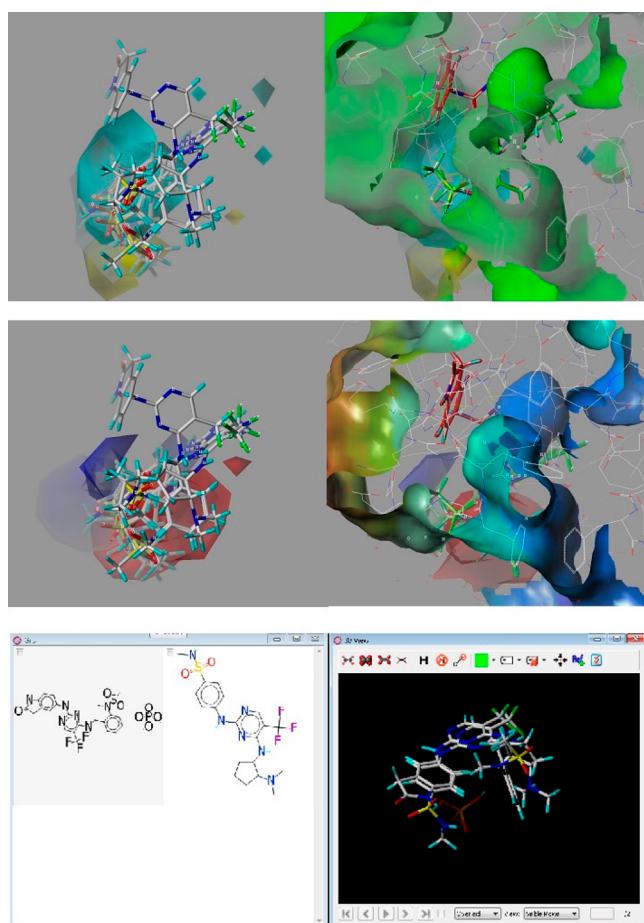


Figure 4. Steric and electrostatic $stdev^*$ coefficient template CoMFA contours, without and with 3FZR.pdb added, and templates for protein tyrosine kinase 2 β . See text for details.

such as diversity, druglikeness, and in particular costs and human factors.

However, conventional practice implicitly ignores two weaknesses of q^2 . The first is that q^2 depends as much on the variation of biological potency within the training set as on the accumulated errors in leave-out predictions. Tables 1 and 2 very well illustrate this concern. Their SDEP columns provide the most direct estimates of the likely error in forecasts of biological potency, with q^2 being computed therefrom as $(SD - SDEP)/SD$, where SD is the standard deviation among the reported training set potencies. (Note that if the SDEP error of prediction is greater than the SD of the training set potencies, implying that the forecasting error from the model is worse than the error from simply assuming that a new potency value is best estimated as the average of potencies already measured, then q^2 indeed becomes negative.) This often overlooked impact of SD on q^2 values is amply illustrated in Table 1, where for a maximal SDEP acceptability, say, between 0.7 and 0.8, the associated q^2 values range from 0.835 down to 0.180. While the models with lower q^2 are indeed less reliable guides than those with higher q^2 , the cause of that diminished reliability is the lower SD of its training set potencies, not inadequacies in the modeling approach.

The second weakness of LOO q^2 is its great sensitivity to outliers. Any unique structure, one whose potency greatly diverges from existing trends and so is very badly predicted upon its LOO omission, can easily double the overall SDEP.

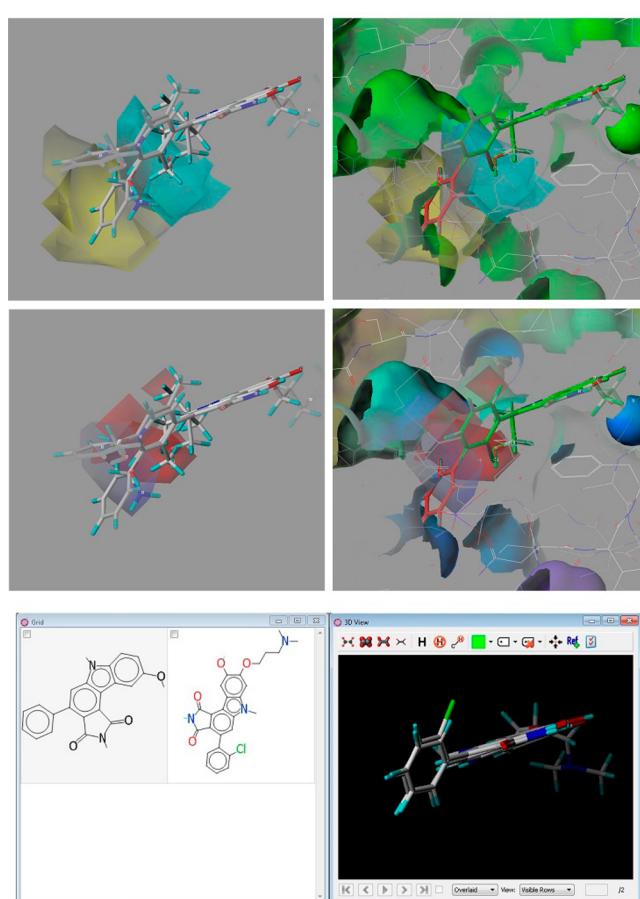


Figure 5. Steric and electrostatic $stdev^*$ coefficient template CoMFA contours, without and with 1X8B.pdb added, and templates for WEE1. See text for details.

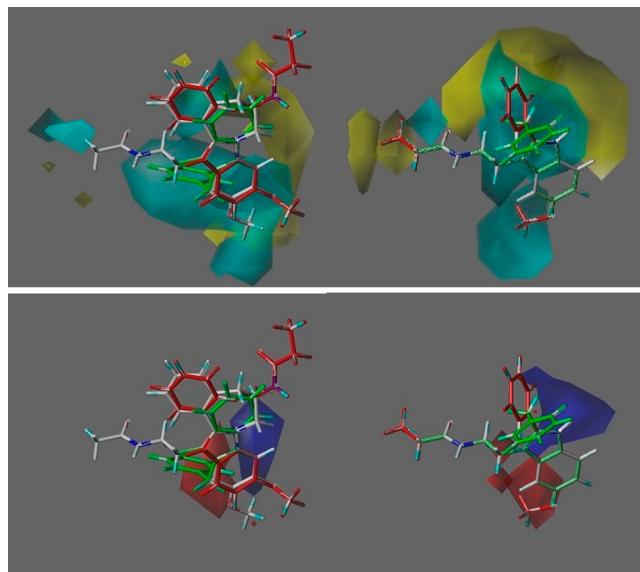


Figure 6. Steric and electrostatic $stdev^*$ coefficient template CoMFA contours, with highest molecular weight and highest ligand efficiency structures as templates, for melatonin receptor. See text for details.

Put differently, to be assured of a high q^2 with LOO cross-validation, the training set structures must include a duplicate for any association of an outstanding potency with an unusual feature. Such caution is no doubt appropriate for QSAR

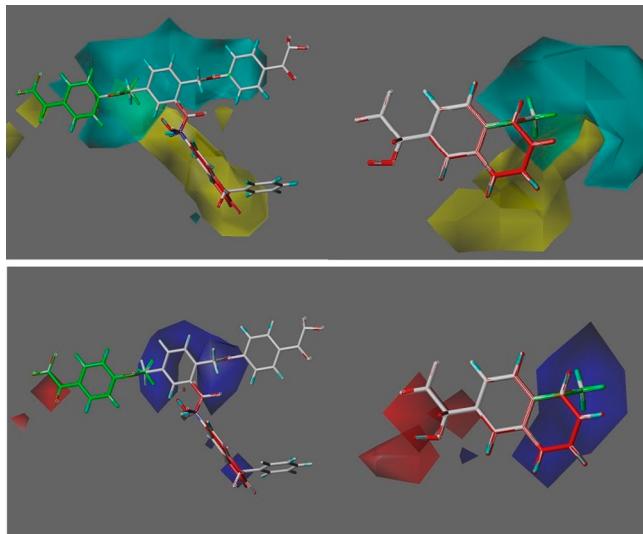


Figure 7. Steric and electrostatic stdev^* coefficient template CoMFA contours, with highest molecular weight and highest ligand efficiency structures as templates, for *Yersinia* tyrosine phosphate inhibitors. See text for details.

whenever uncertainty exists about the relevance of the descriptors or the soundness of the modeling approach, but with tens of thousands of publications using 3D-QSAR's field descriptors and PLS statistics, such uncertainties about 3D-QSAR seem misplaced. In particular the bindingdb training sets, because they have been collected primarily for lengthy structure-based calculation, are trimmed of structures with redundant features, and so, as previously reported, the q^2 for template CoMFA models for its individual bindingdb entries for factor Xa and P38 map kinase are often even <0.0 . Yet combining these entries apparently restores enough redundancy to generate the quite satisfactory q^2 that are reported here in Table 1. And the best-fit r^2 values for these individual bindingdb entry models with the unacceptable q^2 are very good (see Tables 1 and 2 in ref 3). Since these individual entry models thus seem constructive despite their many horrific q^2 values, a third guideline was sought to allow this acceptability to be recognized. It is well-known that a high best-fit r^2 can also be produced by overfitting. So even though extreme overfit is rare with PLS, because its algorithm manipulates descriptors collectively rather than individually, the third proposed guideline further requires the best-fit r^2 to be that obtained using the LOO determined count of PLS components. Note that this third criterion's further division of the best-fit r^2 by that PLS component count, with the 0.2 threshold for the resulting ratio, cannot confer acceptability onto a model with more than five PLS components (with the ratio thresholds for this acceptability guideline ranging from 0.2 for a one component model to 1.0 for a five component model).

However, a novel criterion success criterion is undeniably being used here to help justify claims of unexpectedly strong performance by a novel methodology. Some skepticism is understandable. So a few additional points seem worth consideration: (1) Empirically, despite tens of thousands of publications, there is little confidence among QSAR practitioners that any general correlation exists between the value of any statistical metric of a QSAR and its reliability for prediction. (2) Theoretically, the potential dynamic complexities revealed, for example, by inspecting any target-bound ligand structure¹²

suggest that a QSAR prediction's reliability will instead depend mostly on its training set composition. (3) Practically, the more critical decisions that a QSAR prediction might be expected to guide are mostly binary, "yes–no", more appropriate for Bayesian than for frequentist reasoning. For example, a decision might be better supported by consistent predictions from multiple 3D-QSAR models with q^2 values around 0.4 than by a single prediction from a 3D-QSAR model with a q^2 of 0.9 (assuming the multiple models to have resulted from divergent alignment hypotheses).

Considering these points, perhaps the goal for the "statistical acceptability" of an individual 3D-QSAR model should be whether or not its inclusion in such a set of multiple models will be constructive. Indeed it is this viewpoint that motivated development of this criterion for statistically acceptable models, a model being "unconstructive" for inclusion in such a model ensemble only if its q^2 , SDEP and $r^2/\# \text{ comp}$ are all unsatisfactory.

The potential value of seeking multiple satisfactory models rather than single optimal models is further illustrated in Figures 6 and 7. These images provide reminders that, especially with templates whose alignments are not experimentally constrained, different selections of template configurations can yield divergent models, all statistically acceptable, that may yield conflicting potency predictions. How to proceed? First it must be recognized that any 3D-QSAR (and, by extension, any QSAR whatsoever!) is a very underdetermined model, with several thousands of variables available at least in principle to fit the many fewer training set observables. So occurrences of such divergent models should be expected. The current practice in these situations is then to prefer, and often exclusively, the model with the "best statistics". An alternative response would be to accept the inherent uncertainty of any single model and prediction and therefore to seek a multiplicity of statistically acceptable models. Then the better the agreement among their multiple predictions, the more confidence that could be placed in any resulting decision. However, perhaps only now with template CoMFA is it practical to generate such multiple 3D-QSAR models, by varying and extending the templates, including their relative alignments, and also the training set compositions. A first example of such a topomer-enabled approach is QSEA.¹³ As another example, one set of studies underway is calibrating how adding a few templates will automatically generate models more statistically comparable to the 3D-QSAR models published, inherently "multitemplate", for such thoroughly studied targets as ACE,⁴ D2A,¹⁴ and thermolysin.¹⁵ If this approach shows promise, then targets for future exploration, having broadest therapeutic relevance, include ion channels and cytochromes.

In summary, whenever improved noncovalently determined molecular properties are sought, such an extremely high frequency of acceptable and effortless model generation over such a broad scope of biological targets strongly supports the potential relevance of template CoMFA.

■ ASSOCIATED CONTENT

S Supporting Information

Tables 1 and 2 combined, in .xls format to enable further manipulation. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Richard.Cramer@certara.com; Tel: (505) 982-1055.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

For the assemblage of so many targets and training sets at bindingdb that enabled this work, we greatly thank Michael Gilson and his colleagues.

■ REFERENCES

- (1) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5939–5967.
- (2) Cramer, R. D. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003**, *46*, 374–389.
- (3) Cramer, R. D.; Wendt, B. Template CoMFA: The 3D-QSAR Grail? *J. Chem. Inf. Model.* **2014**, *54*, 660–671.
- (4) Depriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin Converting Enzyme and Thermolysin Inhibitors – a Comparison of Models Based on Deduced and Experimentally Determined Active-Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372–5384.
- (5) Wendt, B.; Cramer, R. D. Challenging the Gold Standard for 3D QSAR: Template CoMFA versus X-ray Alignment. Submitted.
- (6) Mittal, R. R.; McKinnon, R. A.; Sorich, M. J. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. *J. Chem. Inf. Model.* **2009**, *54*, 1810–1820.
- (7) www.bindingdb.org/validation_sets/. (Accessed January 15, 2014.).
- (8) Dolan, M. A.; Keil, M.; Baker, D. S. Comparison of composer and ORCHESTRAR. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 1243–1258.
- (9) Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for their Self-Consistent Generation. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1221–1227.
- (10) Hopkins, A. L.; Keseru, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discovery* **2014**, *13*, 105–121.
- (11) Cramer, R. D.; Bunce, J. D.; Patterson, D. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
- (12) Motlaw, H. N.; Wrabi, J. G.; Li, J.; Hilser, V. J. The ensemble nature of allostery. *Nature* **2014**, *508*, 331–338.
- (13) Wendt, B.; Cramer, R. D. Quantitative Series Enrichment Analysis (QSEA): a novel procedure for 3D-QSAR analysis. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 541–551.
- (14) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, L.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comp.-Aided Mol. Des.* **1993**, *7*, 83–102.
- (15) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis to Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42*, 458–477.