J Comput Aided Mol Des (2007) 21:23–32
DOI 10.1007/s10822-006-9100-0

ORIGINAL PAPER

Pushing the boundaries of 3D-QSAR

Richard D. Cramer · Bernd Wendt

Received: 22 November 2006 / Accepted: 19 December 2006 / Published online: 26 January 2007

Abstract Based primarily on further studies of a collection of eleven publications reporting fifteen successful 3D-QSAR relations, several phenomena are preliminarily described. The RMS error of 133 ligand binding energy predictions based on these successful 3D-QSARs is 0.75 kcal/mole, which compares favorably to the prediction accuracies of approaches that include the receptor. A similar result is obtained when topomer alignments are substituted for those published, with seemingly profound implications for the future of 3D-QSAR. The ‘‘alignment-averaged’’ molecular properties, log *P* and molar refractivity, have very little correlative power for these data sets, either alone or in combination with the 3D-QSAR field descriptors. The q^2 metric for the number of PLS components necessarily tends to discard any unique or unconfirmed SAR information. Large drops in q^2 are thus to be expected whenever such unique information is first encountered. Predictive r^2 values from an exploratory new ‘‘series trajectory’’ analysis of these 3D-QSAR though highly variable do not differ much from their q^2 values, a phenomenon that seems to encourage prediction even when there are so few structures underlying a 3D-QSAR so that almost all information is unique.

Keywords 3D-QSAR · CoMFA · Topomer · q^2 · Log *P* · Prediction · Series trajectory

R. D. Cramer (✉) · B. Wendt
Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA
e-mail: cramer@tripos.com

Introduction

Whenever a methodology has been successful in gaining a wider following, there is usually a critical moment when that methodology is first evaluated by a respected practitioner who is unaffiliated with its development. For CoMFA (of course the first and probably the most widely used 3D-QSAR methodology), that ‘‘earliest adopter’’ and ‘‘initial champion’’ was Yvonne Martin [1]. Thus it seems appropriate on an occasion that commemorates her distinguished career to present some thoughts on 3D-QSAR and its possible future development.

This publication is intended as an essay or mini-review that considers five general and loosely linked topics related to 3D-QSAR methodologies.

1. *The ability of a good CoMFA model to predict ligand binding affinities often exceeds the abilities of methods that also consider the ligand’s binding partner*. That seems surprising.
2. *Performing CoMFA on standardized and context-ignorant topomer alignments has no effect on that predictive ability*. Though even more surprising, by providing a robust, reproducible, and rapid starting point for 3D-QSAR as well as an unprecedented structural searching capability, this ‘‘topomer CoMFA’’ result suggests the possibility of major improvements in the scope and effectiveness of 3D-QSAR approaches.
3. *Although the biological measurements that underlie all SAR models must often be influenced by differences in ligand transport, the relevant ‘‘alignment-averaged’’ ligand properties (e.g., log P) are not explicitly addressed by (alignment-dependent)*

3D-QSAR descriptors. Does this omission have an important effect on the performance of 3D-QSAR? If not, why not?

4. *The leave-one-out cross-validated r^2 or q^2 has become the major measure of merit for a 3D-QSAR.* What are the strengths and weaknesses of this q^2 centrality?
5. *Are there any regularities among the fluctuations in r^2 and q^2 values and predictivity that are encountered as new rows are added to an SAR table?*

These topics will be treated in a relatively speculative manner, for this deadline-constrained mini-review is intended more to raise and highlight questions than to provide answers beyond the preliminary. (Fortunately, however, in many of these cases, these preliminary answers do seem suggestive.) The emphasis will be perhaps disproportionately on the original CoMFA formalism [2], rather than other worthy 3D-QSAR approaches such as COMSIA [3], HINT [4], GOLPE [5], COMPASS [6], or HASL [7], simply because the original method is the one whose behavior is most familiar to us. We assume the reader has some familiarity with at least one 3D-QSAR approach.

Wherever our speculations are supported by preliminary work, all results are based on a particular collection of fifteen biological measurements taken from eleven publications and eleven distinct sets of structures [8]. (Two publications reported three and two different biological measurements on the same compounds, while a fifteenth data set resulted from a second treatment of one series. So, recognizing that the ‘‘fifteen data sets’’ actually represent only eleven sets of structures, several of the following investigations were limited to the eleven structurally distinct series, arbitrarily selecting a single biological response wherever there were several possibilities.) The fifteen sets were originally ‘‘selected’’ simply by being the first eleven publications encountered that reported a successful 3D-QSAR derivation, in a random perusal of journals at hand at that moment. Fifteen (or eleven when structural redundancy is excluded) seemed a large enough number of diverse samples to eliminate artifact or chance as a cause of any uniform trends that might be observed. Perhaps it should be repeated for emphasis that there was no filtering of any sort in assembling these data sets, other than the successful 3D-QSAR analysis. Of course, that 3D-QSAR requirement does significantly bias the data sets to have a ‘‘good behavior’’, that is, a high internal consistency, especially compared with the ‘‘unpolished’’ collections that CADD practitioners typically encounter first. On the other hand, data sets that lack internal consistency will be challenging and probably unproductive to handle using any approach. In particular, when using 3D-QSAR, the presence of ‘‘good behavior’’ would be unambiguously indicated (by a satisfactory q^2), in order for any conclusions to be drawn. So trends so far observed only within data sets that have a ‘‘good behavior’’ would seem to remain appropriate and relevant to other 3D-QSAR studies.

1. The ability of a good CoMFA model to predict ligand binding affinities often exceeds the abilities of methods that also consider the ligand’s binding partner. Table 1 summarizes various 3D-QSAR analyses of these fifteen data sets. The ‘‘CoMFA Prediction’’ block, at the far right of Table 1, contains the RMS errors for predictions of potency for 133 compounds within eleven of the fifteen data sets. The contents of its Lit column were taken from the original publications. Here the RMS error over all 133 predictions is 0.553, in units of log(IC50), and the differences in that RMS error among the eight of eleven groups that reported such predictions are modest. Standardizing measurement units, at room temperature this RMS error of prediction equals 0.75 kcal/mol.

Although the error in potency predictions using receptor-based approaches of course depends on the method and the practitioner as well as the data set, here are a few recent and relatively reliable and relevant data points for comparison. From an analysis of errors, Jorgensen concludes [9] that uncertainties only in the changes in (rather flexible) ligand conformational free energies on binding, even when estimated using ab initio or DFT, cannot be less than 5 kcal/mol. More empirically, it is evident from an extensive evaluation by Warren et al. [10] for example their Figs. 9–11, that obtaining an RMS error of potency prediction as small as 2 kcal/mol for the application of a random docking engine to a random data set is a rare and unpredictable occurrence.

It may well be objected that a direct comparison of the 0.75 kcal/mole RMS error with, say, a 3.0 kcal/mole error from receptor-based methods is unreasonable. For example, the lack of any receptor structure would make such calculations, hence any direct comparison of results, impossible for all but a few of the Table 1 data sets. Also the underlying philosophies of the two approaches are very different, with the receptor-based methods hoping to use universalistic though approximate physics to establish a single model globally applicable to any non-covalent molecular assemblage, while 3D-QSAR seeks only a locally valid model, for some particular ligand class(es) binding to its receptor site(s). On the other hand, most drug discovery programs are primarily attempting to optimize

Table 1 Statistical parameters of model derivation and the external prediction errors, for the fifteen 3D QSAR literature studies and their repetitions with topomeric CoMFA, generating topomers by either the original or the current protocol

Dataset		CoMFA Model Construction													CoMFA Prediction			
		#	x-validated q^2			x-val SDEP			# Compnts			Final r^2			#	RMS pred error		
ID	Name	cpds	Lit	TopB[e]	TopC[f]	Lit[b]	TopB	TopC	Lit	TpB	TpC	Lit	TopB	TopC	cpds	Lit[a]	TopB	TopC
1	ICEc	36	0.630	0.362	0.429	0.816	1.002	0.964	6	5	6	0.970	0.883	0.961	9	0.568	0.740	1.580
2	ICEb	38	0.630	0.433	0.362	0.816	0.951	1.023	6	3	4	0.970	0.806	0.844	10	0.553	0.595	0.606
3	thrombin	72	0.687	0.533	0.410	0.594	0.726	0.736	4	4	2	0.881	0.838	0.743	16	0.673	0.619	0.646
4	trypsin	72	0.629	0.657	0.632	0.556	0.531	0.559	5	4	9	0.916	0.886	0.966	16	0.524	0.523	0.421
5	factorXa	72	0.374	0.186	0.305	0.515	0.591	0.592	3	4	8	0.680	0.747	0.897	16	0.278	0.340	0.363
6	MAOa	71	0.440	0.566	0.549	1.025	0.926	0.931	2	4	2	0.680	0.813	0.668				
7	MAOb	71	0.430	0.483	0.520	1.253	1.214	1.200	2	2	3	0.880	0.640	0.761				
8	hiv	25	0.680	0.389	0.609	0.571	0.845	0.730	3	3	6	0.950	0.878	0.985	7	0.823	0.449	0.094
9	a2a	78	0.541	0.226	0.341	0.563	0.742	0.684	4	3	3	0.817	0.555	0.623	23	0.668	0.761	0.737
10	d4	29	0.739	0.636	0.745	0.734	0.802	0.703	7	5	7	0.996	0.957	0.985				
11	flav	38	0.752	0.763	0.745	0.475	0.495	0.559	4	5	10	0.969	0.952	0.964	4	0.337	1.314	0.958
12	cannab	61	0.592	0.423	0.469	0.570	0.696	0.669	4	3	3	0.905	0.777	0.744	6	0.452	0.540	0.708
13	ACEest	41	0.937	0.746	0.811	0.346	0.726	0.626	4	3	3	0.990	0.916	0.962	7	0.413	0.478	0.580
14	5ht3	61	0.645	0.295	0.463	1.193	1.804	1.663	5	2	8	0.913	0.519	0.896				
15	rvtrans	82	0.837	0.830	0.810	0.567	0.587	0.619	4	4	4	0.936	0.916	0.899	19	0.791	0.608	0.671
	Total/Avg	847	0.636	0.502	0.547	0.581[c]	0.717[c]	0.706[c]	4.2	3.6	5.2	0.897	0.806	0.86	133	0.553	0.633	0.669
																0.574[d]	0.565[d]	0.641[d]

[a] For ICEc, ICEb, hiv and a2a, the individual prediction values were read from the graphs in Figure 3, Figure 3, Figure 6, and Figure 3, respectively, of the original publications. Others were taken directly from tables

[b] For MAOa, MAOb, hiv, a2a, flac, cannab, and ACEest the sdep was calculated from the original variance in biological activity and the reported q^2. Other values were taken directly from the tables

[c] Average excluding MAOa, MAOb, d4, and 5ht3 (to permit comparison with RMS CoMFA Prediction error)

[d] Average excluding flav (see text in original paper for discussion)

[e] The TopB protocol represents ‘‘standard topomeric CoMFA’’ settings and the original topomer generation protocol

[f] The TopC protocol represents ‘‘standard topomeric CoMFA’’ settings and the current topomer generation protocol. At the deadline for submission of this manuscript, the current protocol is not quite as effective as previous protocols for topomer CoMFA on this data set (probably a bug), but for comparability within Table 2 these inferior results appear here

such a locally valid model, while a difference in predictive error between 0.75 and 3.0 kcal/mole will equate to the difference between helpful and an ineffective methodologies.

Nevertheless, intuition insists that any analysis that ignores available receptor information must be suboptimal, that its proper use would improve the predictive accuracy of a 3D-QSAR approach. This intuition is expressed, for example, in a general belief that the most effective ligand alignment protocol for 3D-QSAR development should be the receptor-bound conformation [11]. Several groups have developed interesting approaches that include the receptor structure within a 3D-QSAR-like statistical modeling approach, for example VALIDATE [12], COMBINE [13], and AFMoC [14]. However, there is a significant cost of obtaining receptor structures for sufficient ligands of interest to an optimization program, which has perhaps slowed the dissemination of such intuitively appealing approaches.

Most aficionados of receptor-based CAMD will perhaps wish to attribute any apparently smaller error of potency prediction from 3D-QSAR to fortunate alignment procedures, which first generated good approximations of the bound ligand conformations, and perhaps were then augmented by practitioner bias. So let us turn to potency predictions resulting from a ligand alignment procedure that is clearly unbiased and is at most coincidentally related to any bound conformation.

2. Comparing potency prediction errors from 3D-QSAR using ‘‘topomer’’ alignments with those from receptor-based methods. As the basis for very rapidly comparing molecular shapes and thereby identifying ‘‘lead hops’’ (structurally dissimilar molecules having similar biological properties) [15], a ‘‘topomer’’ protocol had been developed for canonically generating a 3D alignment for any molecular fragment from its atom-bond connectivity. Topomer alignment [16] starts from a rule-based (CONCORD-generated) assignment of valence and ring geometries. Its subsequent torsion and chirality standardizations completely ignore conformational energy, for example any interpenetrating atomic radii. For this 3D-QSAR application, it is especially important to appreciate that the topomer alignment protocol is an entirely localized operation, not in any explicit way cognizant of how its product is to be compared with any other ligand fragment structure or any receptor cavity.

When an internal need arose for 3D-QSAR, with its frustrating input requirement for 3D alignments, the convenience of already having the topomer protocol in hand overcame obvious misgivings about the effectiveness of such a context-indifferent alignment approach [8]. Our fifteen data sets were thus originally assembled as a benchmark, to compare the published predictions that had resulted from a variety of arduous alignment heuristics with the predictions that the topomer alignments might yield from whatever 3D-QSAR emerged. The only variable component of "topomer CoMFA" is how structures are disconnected into commensurate fragments, and for all but one of these rather congeneric data sets there was no doubt about the optimal fragmentation "rule".

The most salient results appear in the two right-most columns of Table 1. Using either of two generations of the topomer protocol, it can be seen that the potency predictions from topomer-based CoMFA models are not significantly less accurate than those from the original authors' various 3D-QSAR alignments. Furthermore, from the preceding columns of data, it is further clear that the topomer alignments performed equivalently to the published alignments at every intermediate stage of the 3D-QSAR analysis, in every one of the fifteen independent trials. (For the current publication, ten "biodata scrambling" runs were also repeatedly performed for each of the eleven sets of structures, yielding a total of only three q^2 values (artifactually) greater than 0.05 among the 110 trials.) We find such an unequivocal outcome from an appropriately extensive benchmarking experiment to be startling and significant. At the least, doesn't any impression that any relative superiority of 3D-QSAR predictions is a result of 3D alignments that artifactually either duplicate the receptor bound conformation or somehow memorize the SAR data merit some critical reconsideration?

Anyway it seems that relatively accurate potency predictions are possible starting only from topomer ligand shapes that at best coincidentally resemble the receptor-bound conformations. What is going on? What if anything is special about these topomer shapes? The intent of the topomer protocol is that similar fragment connectivities will yield similar fragment shapes. Receptor-based alignments yield much more variable ligand shapes, because within any system as mechanically rich as a ligand–receptor complex, the overall outcome of docking relaxations is not a simple or a consistent function of any localized ligand structural change. Figure 1 illustrates this variety of outcomes, using the docking of a small combinatorial library into a receptor (the receptor then being removed). Although most of the atoms are identical among all structures (the constant library scaffold structure is inset into the upper right hand corner of Fig. 1), the docking process has responded to the

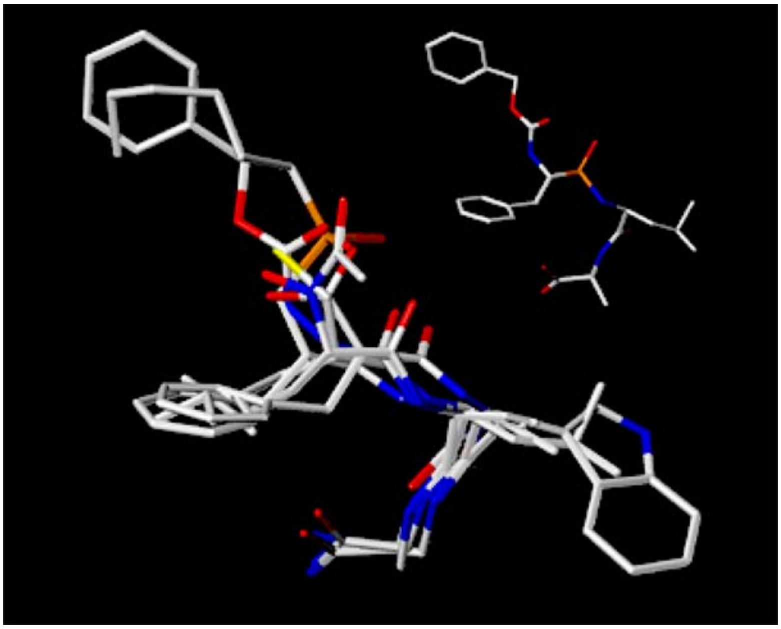


Fig. 1 Conformations of a small combinatorial library docked into a receptor (receptor not shown). The docking process has rearranged a structurally constant scaffold (shown for reference in the upper right corner) into a variety of configurations, caused by the small structural differences among the side chains

remaining structural differences by scrambling the initial superpositions of the structurally constant atoms. That scrambling creates variations within the ligand fields that are the explanatory variables in 3D-QSAR. Those field variations within a region of space where the ligand structures are actually constant are ‘‘noise’’ (unless they happen to be related to the biological response) and will therefore weaken the ability of partial least squares (PLS) to detect possible ‘‘signal’’ within other regions of space [17]. Or, putting the same ideas a bit differently, the differential biological responses that a drug discovery project seeks to improve are affected only by ligand variation, since everything else that might alter a response is held constant. The mechanism by which a ligand variation produces a response change may be mechanically obscure and complicated, so for purposes of selecting the next ligand variation, interpretation may well be easier if the input data are organized to focus on ligand variations, as topomers by design do. Or putting matters differently yet again, regardless of interpretation, as a practical matter, Table 1 shows that any gains in 3D-QSAR ‘‘signal strength’’ that resulted from the physicochemically more realistic ligand alignments used by the original researchers tended to be offset by increases in ‘‘noise’’, from sources such as Fig. 1 depicts.

Though still mostly unexplored, the potential implications of continued success in using topomers for 3D-QSAR input seem profound. First, the cost and time of trying 3D-QSAR on a data set almost disappear (for example, replicating Table 1, starting with only the simple ‘‘2D’’ structure activity tables in computer readable form, takes only a few minutes, surely many orders of magnitude less time and effort than the original publications’ 3D-QSAR investigations required). Second, topomer-based 3D-QSAR models immediately become search queries into databases composed of other topomer fragments, obtaining results in a few hours with engines such as dbtop [18] (10^6 candidate structures, whose topomer fragments are generated on the fly), ChemSpace [19] (10^{13} candidates, each a combination of available topomerized building blocks), and AllChem [20] (10^{20} candidates, whose topomerized building blocks are produced by applying short sequences of reactions to available building blocks). Finally, this standardized protocol for generating 3D-QSAR alignments, by eliminating the highly variable and subjective current alignment process along with its time and cost, provides a consistent starting point for addressing other 3D-QSAR issues in greater depth. Much of the remainder of this mini-review tries to make further use of this consistency.

Would some canonical alignment approach other than the topomer protocol also have worked as well? This reasonable possibility has not been directly investigated, though we suspect that the emphasis of topomers on fragments rather than complete structures may be important for their effectiveness. The topomer protocol also seems critical pragmatically, for enabling 3D-QSAR-based searches of candidate databases, especially the large ones.

3. Although the biological measurements that underlie all SAR models must often be influenced by differences in ligand transport, ‘‘alignment-averaged’’ ligand properties (e.g., log *P*) are not explicitly addressed by (alignment-dependent) 3D-QSAR descriptors. There are recurring reports that docking scores are not much more accurate predictors of binding than is molecular weight [21], or that raw atom counts are almost as good as ‘‘2D fingerprints’’ [22]. Such reports suggest that these 3D-QSAR performances also should be compared to QSAR based on such simple descriptors. It is also worth recalling that the earliest QSAR descriptors, octanol/water log *P* and molecular refractivity, are together more effective summaries than molecular weights or atom counts of all such ‘‘alignment averaged’’ physicochemical properties [23]. We have therefore added log *P* (CLOGP) and molecular refractivity (CMR) descriptors to the topomer CoMFA analyses reported in Table 1, using ‘‘standard CoMFA’’ weighting which gives each of these two scalar descriptors the same weight as an entire CoMFA field. (It is not always appreciated that results from PLS, unlike multiple regression, are strongly dependent on the relative scalings of the various descriptors.) Also we derived QSAR using only CLOGP and CMR as descriptors.

The results of these studies appear in Table 2, including q^2, r^2, and the coefficients observed for the CLOGP and CMR terms when these were included. The most important result, surprising to us, is that for these fifteen data sets the CLOGP and CMR parameters seldom had any correlative power whatever, either alone or in combination with the 3D-QSAR fields. Certainly the few and poor correlations with CLOGP and CMR alone show that these alignment-averaged parameters cannot be surrogates for the 3D-QSAR fields. A modest improvement in average q^2 when CLOGP/CMR columns are added is rather offset by the increase in the number of PLS components.

The single weak exception to this trend was the ‘‘factor X’’ data set, where the 3D-QSAR correlation was the weakest among any of the fifteen data sets, while the CLOGP/CMR-only correlation is the second strongest. As a result, the Factor X correlation using

Table 2 Statistical parameters of model derivation and coefficients for log *P* and MR, for the fifteen topomer CoMFA models (Cfa), for log *P* and MR models only (PM), and for topomer CoMFA and log *P*/MR together (Cfa/PM)

	q^2			# Component			r^2			Log *P* Coeff		MR Coeff	
	Cfa	Cfa/PM	PM	Cfa	Cf/PM	PM	Cfa	Cfa/PM	PM	Cfa/PM	PM	Cfa/PM	PM
ICEb	0.429	0.435	0.182	6	6	2	0.961	0.929	0.311	0.05	–0.28	0.55	0.28
ICEc	0.362	0.424	0.234	4	2	2	0.844	0.670	0.373	0.01	–0.15	0.65	0.26
thrombin	0.410	0.510	–0.026	2	6	1	0.743	0.893	0.046	0.14	0.07	0.10	0.07
trypsin	0.632	0.683	0.233	9	6	2	0.966	0.915	0.289	0.05	–0.13	0.43	0.28
factor X	0.305	0.324	0.328	8	1	1	0.897	0.380	0.369	0.32	0.12	0.34	0.12
MAOa	0.549	0.536	–0.089	2	5	1	0.668	0.787	0.028	0.02	–0.01	0.09	0.22
MAOb	0.520	0.534	0.079	3	4	2	0.761	0.717	0.172	0.16	0.64	0.27	–0.34
hiv	0.609	0.682	–0.277	6	9	1	0.985	0.990	0.023	0.07	0.55	0.14	0.47
a2a	0.341	0.314	–0.050	3	4	1	0.623	0.643	0.023	0.25	0.10	0.05	0.02
d4	0.745	0.779	–0.141	7	9	1	0.985	0.989	0.074	0.52	0.14	0.43	0.16
flav	0.746	0.776	0.277	10	4	1	0.964	0.911	0.372	0.33	0.52	0.23	0.74
cannab	0.469	0.611	0.422	3	3	1	0.744	0.762	0.474	0.18	–0.65	0.27	–0.52
acest	0.811	0.756	0.325	3	3	1	0.962	0.902	0.393	0.02	0.41	0.31	0.34
5ht3	0.463	0.487	0.040	8	13	2	0.896	0.917	0.123	0.75	–0.81	0.23	0.43
rvtrans	0.810	0.830	0.260	4	9	2	0.899	0.964	0.306	0.10	0.61	0.22	–0.06
Average	0.547	0.579	0.120	5.2	5.6	1.4	0.860	0.825	0.225	0.20	0.07	0.29	0.17

both parameter sets is indistinguishable from the CLOGP/CMR correlation only. Although the ''cannab'' data set has the highest CLOGP/CMR-only correlation, the negative signs for both its CLOGP and CMR terms are strongly opposite the usual tendency for binding to increase with ligand lipophilicity and size. Therefore these CLOGP/CMR terms are probably summarizing receptor-specific interactions, though roughly, and indeed with the addition of the 3D-QSAR terms more appropriate for characterizing receptor interactions, these CLOGP and CMR coefficients revert to their expected positive values.

Potency values within these exclusively in vitro assays simply were almost completely unaffected by transport factors, in striking contrast to the behavior of the in vivo assay systems that dominated discovery projects during the infancy of QSAR approaches. As many other observers are today emphasizing, slighting transport factors until after commitment has been made to a series or candidate based mainly on in vitro data has probably been unfortunate. These data further support the desirability of overall changes in discovery decision-making protocols and priorities to emphasize earlier consideration of transport behavior, now generally underway.

4. What are the weaknesses of leave-one-out cross-validated r^2 or q^2 as the measure of merit for a 3D-QSAR? Because the purpose of deriving a QSAR is useful prediction, cross-validation (repeated compound omission, model rederivation and prediction of omitted values) has become a procedural standard. Several authorities have pointed out that actually q^2 is not a very reliable indicator of accuracy in truly external predictions [24, 25], and further model validation refinements such as ''progressive scrambling'' [26] have been introduced. However here we would like to call attention to a different concern. *It is certainly possible, perhaps probable, that the cross-validation criterion **discards** useful SAR information.* In essence, cross-validation says ''don't believe anything until you've heard it more than once''!

Suppose a QSAR table contains a single structure that has desirable biological activity only because it contains features (column property values) unlike those of the others. During cross-validation, whenever this structure is omitted from QSAR derivation, its unique features aren't included so its activity must be predicted to be the average of all the others. Not only is this useful but unique QSAR information self-evidently being discarded, but also the q^2 value is being ''artificially'' depressed, potentially by enough to reach the wrong conclusion that no useful QSAR exists. In the context of an ongoing discovery project, actually such a promising singleton should (and probably would) be treated as an opportunity, to be validated and exploited by further synthesis and testing, rather than an artifact to be discarded. Indeed, the most consistently successful use of QSAR that either of us has experienced began with a graphically-derived hypothesis that barely passed statistical significance tests when a computer became available [27].

On the other hand, we are certainly not suggesting the abandonment of q^2. Indeed, with thousands of candidate molecular descriptors conveniently available today for correlation with biological activity, the long-established risks of futile activity [28] by taking seriously every result of ‘‘data mining’’ have become still greater. The risk of such ‘‘chance correlations’’ does depend on the statistical methodology, being much higher with multiple regression (MR) that manipulates individual columns than with partial least squares (PLS) that manipulates blocks of columns. (Indeed, it has been shown empirically [17] that PLS responds to added columns of random data values with a decreasing ability to recognize true correlations rather than a tendency to generate more artifactual ones.)

For us, a consideration even more important than statistical methodology in assessing the artifactuality of an interesting QSAR result is its consistency with some established physicochemical model [29]. When 3D-QSAR was first introduced, its necessary reliance on the unproven PLS algorithm for extracting useful information from ‘‘tables with more columns than rows’’ was only one very good reason for recommending a conservative and absolute dependence on q^2. Today with several decades of experience and thousands of published applications, it may be worthwhile to push the original q^2 constraints on 3D-QSAR, for example to seek extrapolable relationships by extracting more PLS components than can be justified by cross-validation. Especially within active discovery projects, for it must be remembered that the real value of any CADD technique in drug discovery may come more from highlighting a promising but uncertain trend for experimental pursuit, than from providing accurate predictions of mediocre potency values, however numerous. On the other hand, QSAR hypotheses that emerge only from ‘‘data mining’’ approaches, essentially by correlating more-or-less random combinations of random molecular descriptors with a biological response, seem much less promising to us almost regardless of any associated q^2 value.

5. Are there any generalities among the variations in r^2 and q^2 values and predictivity that are encountered as new results are added to a data set? Given a collection of structures with measured activities, such as the outcome of an HTS campaign, most would agree on the likelihood therein of sets of ‘‘related structures’’, presenting the same shape in the same relevant way, perhaps (or not) by virtue of a common scaffold structure. (Within HTS results there are also many ‘‘singletons’’ and false negatives and positives.) Such sets are usually (though not necessarily) identified only after determining the biological activities, and of course their memberships depend strongly on how structural similarity is defined. We have begun investigating how effectively topomer similarity methods can automatically identify such sets of related structures, evaluating the sets by concurrently generating topomer-based 3D-QSAR.

In general, our most effective and relevant protocol so far is to begin with small ‘‘nucleating clusters’’ of the three most topomerically similar structures and then to iteratively add to a cluster, one by one, the topomerically nearest structure (where nearness is the distance to the cluster centroid). After every addition, we derive a topomer CoMFA from the cluster and predict the activities of the structures still outside the cluster, tracking q^2, r^2, predictive r^2, components, and potency spread.

It may well be noted that beginning with the most similar structures does not accord with the general precepts of experimental design, which recommend widely separated values of potential explanatory values with the intent of rapidly establishing any causative relationship. On the other hand, the SAR in discovery-targeted investigations are always more or less local, such that larger structural changes may well exceed the boundaries in applicability of any useful regularity in SAR. Certainly the structural variety within typical HTS data sets is far too great to expect the emergence of any but localized SAR trends.

Such a ‘‘series trajectory’’ protocol seems to be generating another dimension for a 3D-QSAR study to consider, in particular possible insights into how sensitively the 3D-QSAR responds to a systematic expansion of the (topomer field) descriptor space. Note that every such prediction necessarily represents extrapolation, not interpolation. Also, toward the end of a trajectory, the 3D-QSAR will be more complete and hence presumably more reliable, but the few remaining structures will be the most dissimilar to those underlying the model and so perhaps the most difficult to predict. We have been applying this protocol to the same eleven 3D-QSAR successful structure sets, so there must exist a satisfactory q^2 for at least the last iteration (and presumably many of the others). However, at the start of a trajectory, with clusters containing few and similar structures, we wondered whether the spread in biological properties (which of course drive any QSAR) would be large enough to afford acceptable 3D-QSAR. The performance of small models is particularly relevant to lead optimization applications of 3D-QSAR, because for the secondary assays that are increasingly important in lead optimization there will be far fewer structural data points.

It is generally believed that the more similar a structure is to a training set structure, the more accurate the prediction of its activity is likely to be. To investigate this expectation, we also restricted predicted potencies into groups based on the ratio of their topomer distance to the current cluster centroid to the current cluster radius, excluding the less topomer-similar structures, at cutoffs of 1.2×, 1.5×, 1.7×, and 2.0× that radius. These restrictions are most meaningful earlier in the trajectory, with the larger numbers of predictions possible and the smaller cluster radii. We are also investigating the effect on predictions of using more PLS components than those prescribed by the q^2 criterion (see theme 4 above), but results were not available at the submission deadline for this manuscript.

In general, satisfactory q^2 values (greater than 0.2) were obtained in 76% of these 574 trials. Concurrent biodata scrambling experiments reinforce a general belief that 0.2 is an appropriate q^2 significance threshold in 3D-QSAR, with only 4% of 5024 such trials yielding a q^2 greater than 0.2. With the smaller and more self-similar clusters the likelihood of successful correlation decreases greatly and the likelihood of chance correlation increases somewhat. For clusters containing from three to ten structures, $q^2 > 0.2$ was obtained in 29% of 77 trials, with biodata scrambling of these clusters producing a $q^2 > 0.2$ in 6% of 770 trials.

Some initial results of this investigation appear in Fig. 2. Each of these eleven plots tracks the effects of iterative addition of structures to one of the data sets in Table 1, by the procedure just described. The plot lines are consistently color coded, with deep blue indicating q^2 and the other colors the predictive r^2 when predictions are restricted to structures no farther from the cluster centroid than 20%, 50%, 75%, and 100% of the current cluster radius. (All negative values are represented in the plot as 0.0.) Predictions were based on the number of PLS components that minimized the cross-validated SE of prediction, but always at least one component. From examination of these plots, we now tentatively suggest some general phenomena, all needing further investigation.

Obviously the plot lines are much more variable than consistent, both within a particular series and when comparing plots. The q^2 lines usually trend upward or remain constant as structures are added but can suddenly drop. This falls are usually accompanied by a sharp drop in the q^2-acceptable number of components (not shown). Surely these q^2 drops are caused by the first inclusion of some novel type of structural change that profoundly affects the biological response (see the discussion above on the downside of q^2 dependence). Even though these discontinuities are statistically troubling, they should also be particularly

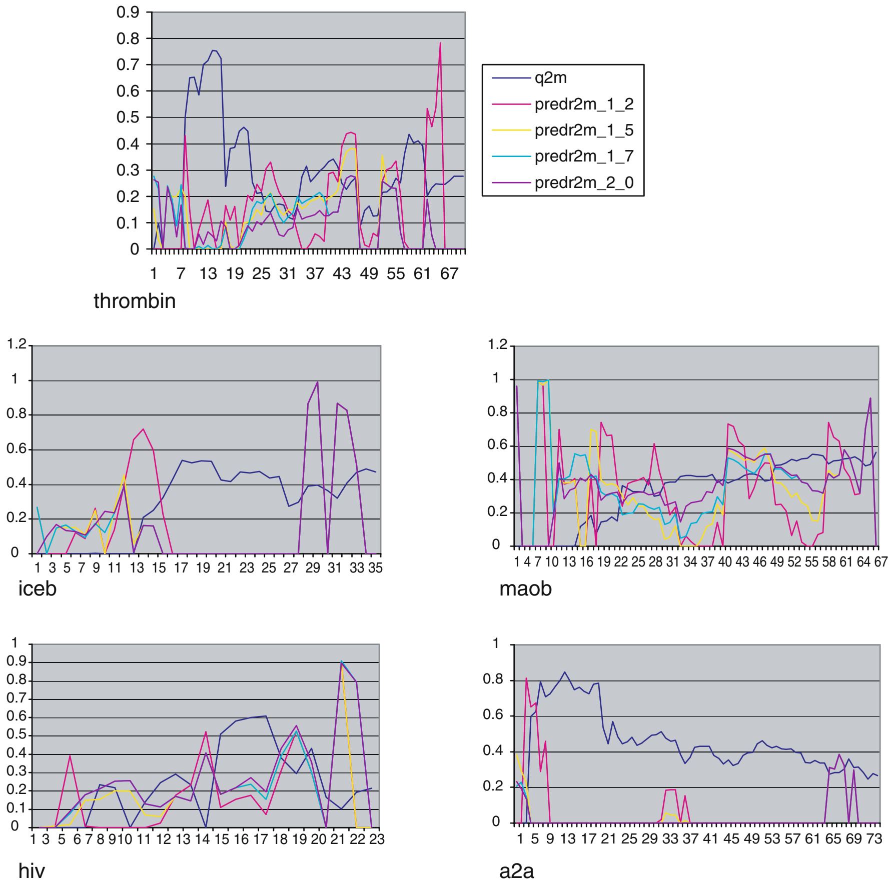
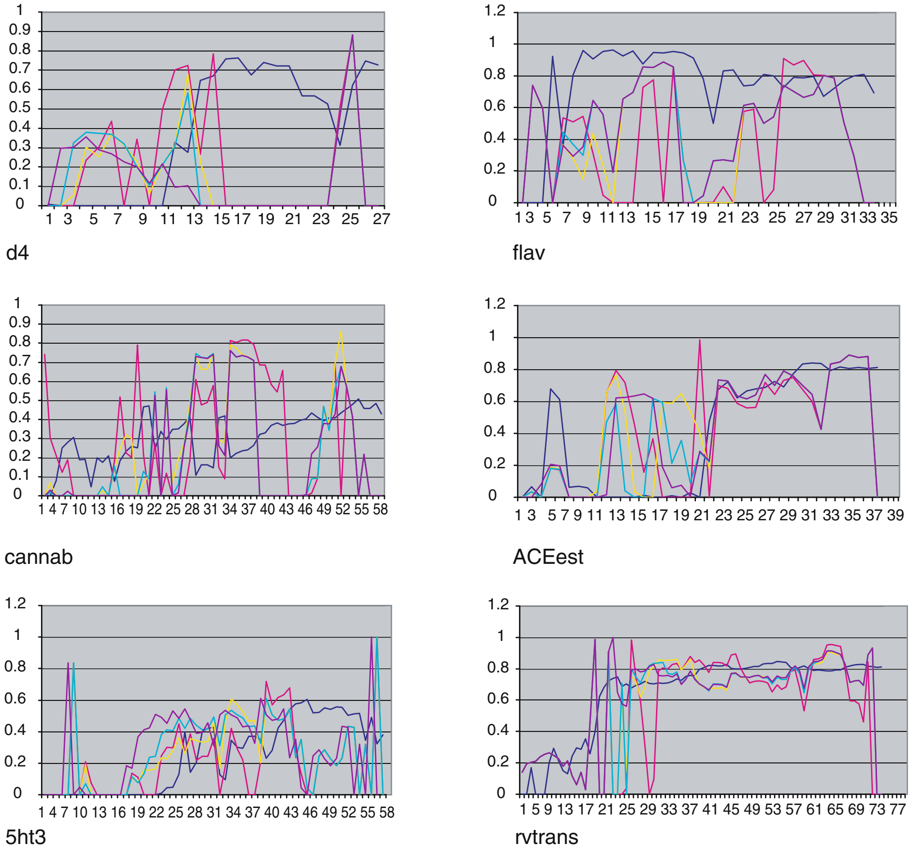


Fig. 2 The progress of q^2 and of various predictive r^2 metrics, as structures are added to a data set, in order of increasing topomer distance from the topomer center, and a 3D-QSAR rederived, for each of eleven data sets for which a 3D-QSAR of the whole existed. See text for details

Fig. 2 continued



informative and so perhaps quite beneficial to the discovery goal of a project. How rapidly the q^2 recovers would then depend on whether the next few structures to be added, those next most topomerically similar, reinforce the important new finding, or instead either revisit a known trend or explore some other inconsequential structural change.

Referring to Fig. 2, here are several specific examples of these phenomena. At datapoint 17 within the thrombin plot q^2 dropped from 0.724 down to 0.237, with the first inclusion of a donor-group (amide) attached to the 4-position of a piperidine sidechain. Based on previous SAR an activity around 7.5 was expected, but this compound showed an activity of 6.8. Datapoint 6 within the hiv plot corresponds to the first inclusion of a large p-substituent (a thiophenyl) to a phenyl sidechain that formerly showed only none or small substituents in the meta-position, producing a large drop in activity from 7.2 to 5.22 and thus in q^2 from 0.2 to –0.27. More welcome, datapoint 11 represents the first inclusion of a donor group (hydroxymethyl) to 4-position of phenyl sidechain, which led to a large jump in activity to 8.14 (highest activity within the training set), even though q^2 therefore dropped from 0.23 to 0. With datapoint 5 of flav, replacement of Cl/Br by F at the 6-position of coumarin depressed the pI50 from 7.64/7.7 to 6.7 with a corresponding q^2-drop of 0.93 to 0.37. At datapoint 6 in ACEest, the first inclusion of 4-and 6-substitutions (both in one compound) on a phenyl sidechain (other compounds having 2-substitutions only) caused a large drop in activity 5.53 (lowest activity before was 6.49) producing a drop in q^2 from 0.61 to 0.06. Finally the q^2 dip at datapoint 26 in 5ht3 results from a behavior irregularity in the topomeric alignment protocol itself. In structures Mol_29, Mol_30 and Mol_33 the pyrrolidine of a 3-ring-system composed of a central pyrazine with fused pyrrolidine and pyridine or benzene is oriented to the left hand side of the plane, but in structures Mol_21, Mol_23 and Mol_32 this pyrrolidine is oriented to the other side of the plane.

There is no evidence of a relationship between the dissimilarity of a group of structures being predicted and the predictive-r^2 that results. (Ignoring the dark blue q^2 line, the magenta line for the most similar predictive group fluctuates rather uniformly around any other lines.) It seems that much larger samples would be needed to confirm the expectation that predictions will be most accurate for the structures most similar to the training set. These sorts of investigations will be particularly useful for better calibrating topCoMFA recommendations, which consider topomer similarity as well as topCoMFA model potency predictions.

The most interesting phenomenon is that the external predictivities of these models (predictive r^2) can be better than their leave-one-out internal predictivities (q^2). In every one of the eleven plots, there is some segment of the magenta ''predrm_1_2'' line that exceeds the corresponding segment of dark blue q^2 line. Such segments are most common at the left sides of the plots, associated with the smaller and more self-similar training sets and larger numbers of predictions. At the same time, the regions where the dark blue q^2 line consistently exceeds the other predictive lines are confined to the a2a, iceB, and thrombin data sets. These results do tend to encourage predictions even when the q^2 is not encouraging, especially when the training set is small.

Another tendency worth noting is the occasional absence of any q^2 values (actually q^2 was negative) at the left of most plots. In isolation this result is usually interpreted as ''no significant 3D-QSAR exists'', so in some hypothetical project where these structures were made and tested in exactly this order, work might have ended prematurely. On the other hand, it is not usually expected that a 3D-QSAR would emerge from fewer than ten or so biological measurements, so if these too-close-to-the-left-margin regions of the plots are excluded, only the 5ht3 and ACEest sets would have generated misleadingly low q^2 early in the development of these eleven series.

In conclusion, it must be emphasized that all of these preliminary results are biased to an unknown degree by the data sets being ''well-behaved'', as all yield a satisfactory q^2 from 3D-QSAR analysis of their entirety. At the start of a series development, no one knows whether the data set that emerges will actually be ''well-behaved''. It might be argued that the best strategy at the start is to assume good behavior, until that assumption is contradicted by the SAR data. One wonders how often a clinical candidate emerges from a generally poorly behaved series (though localized discontinuities in q^2 such as those in Fig. 2 can signal dramatic turning points within a successful project). However, all of our interpretations must be considered tentative until similar studies can be done for comparison, for example on data sets that do not yield satisfactory (topomer-based) 3D-QSAR models.

References

1. Martin YC (1998) Persp Drug Disc Design 12:3
2. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5939
3. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130
4. Kellogg GE, Semus SF, Abraham DJ (1991) J Comp-Aided Mol Des 5:545
5. Baroni M, Costantino G, Cruciani G, Riganelli D, Valigi R, Clementi S (1993) Quant Struct-Act Relat 12:9
6. Jain AN, Koile K, Chapman D (1994) J Med Chem 37:2315
7. Doweyko AM, Mattes WB (1992) Biochem 31:9388
8. Cramer RD (2003) J Med Chem 46:374
9. Tirado-Rives J, Jorgensen W (2006) J Med Chem 49:5880
10. Warren GL, Andrews CW, Capelli A-M, Clarke B, Lalonde B, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912
11. Lemmen C, Lengauer T (2000) J Computer-Aided Mol Des 14:215
12. Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, Marshall GR (1996) J Am Chem Soc 118:3959
13. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) J Med Chem 38:2681
14. Gohlke H, Klebe G (2002) J Med Chem 45:4153
15. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RD (2004) J Med Chem 47:6777
16. Jilek RJ, Cramer RD (2004) J Chem Inf Comp Sci 44:1221
17. Clark M, Cramer RD (1993) Quant Struct-Act Relat 12:137
18. Cramer RD, Jilek RJ, Andrews KM (2002) J Mol Graphics Model 20:447
19. Andrews KM, Cramer RD (2000) J Med Chem 43:1723
20. Cramer RD, Soltanshahi F, Jilek R, Campbell B (in press) J Comp-Aided Mol Des
21. Pan Y, Huang N, Cho S, MacKerell AD (2003) J Chem Inf Comput Sci 43:267
22. Bender A, Mussa HY, Glen RC (2004) J Chem Inf Comput Sci 44:1708
23. Cramer RD (1980) J Am Chem Soc 102:1837
24. Clark RD (2003) J Comput-Aided Mol Des 17:265
25. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269
26. Clark RD, Fox PC (2004) J Comput-Aided Mol Des 18:563
27. Cramer RD, Snader KM, Willis CR, Chakrin LW, Thomas J, Sutton BM (1979) J Med Chem 22:714
28. Topliss JG, Edwards RP (1979) J Med Chem 22:1238
29. Unger SH, Hansch C (1973) J Med Chem 16:745