Topomer CoMFA: A Design Methodology for Rapid Lead Optimization

Richard D. Cramer*

Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri 63144

Received May 6, 2002

To provide an objective QSAR methodology that might accelerate lead optimization, the CoMFA and topomer technologies have been merged, with surprisingly good results. A series of input structures are each broken into two or more fragments at central acyclic single bonds, while removing any core fragment structurally common to the entire series. Standard topomer 3D models are automatically constructed for each fragment, and a set of steric and electrostatic fields ("CoMFA column") is generated for each set of topomers. Application of "topomer CoMFA" to 15 3D-QSAR analyses taken from the literature (847 structures) were all successful, with an average q^2 of 0.520 (literature average $q^2 = 0.636$) and an average standard deviation of true prediction (SDEP) of 0.688 (literature average SDEP = 0.553) for 133 structures. Topomer CoMFA results are particularly promising as queries into virtual libraries already composed of topomer structures, to directly seek structures having increased potency. Accordingly, in 13 of the 15 such "topomer CoMFA searches" attempted, combinations of commercially offered fragments were retrieved that were predicted to be more potent than any structure described in the original publication (average predicted potency increase = 20×), showing in principle how optimization could occur.

When optimizing a lead structure, typically a major goal is to improve potency in the primary assay by several orders of magnitude. Searching for "similar" structures within large databases,[1] where similarity is defined by a property such as pharmacophores or fingerprint Tanimotos,[2] is of limited value during lead optimization. For a "similarity search" implies that *every* structural change is more or less undesirable, whereas potency improvement can be achieved only by discovering some *particular* structural change that is not bad but good. Therefore, although serendipity is always welcome, potency improvement can be systematically sought only by iterative analysis of accumulating SAR data. A sufficiently rapid, automatic, and general method of QSAR analysis and application, analogous to lead discovery by similarity searching of large databases, could become most valuable in lead optimization, by shortening and maximally utilizing its repetitive design−synthesis−test cycles.

One QSAR methodology with particularly widespread[3] successes in analyzing structure−activity data[4,5] is comparative molecular field analysis (CoMFA).[6] Yet CoMFA has weaknesses. The greatest of these is its input requirement that each ligand structure be represented as a 3D model, suitably "aligned" (by selecting an absolute orientation of a single conformation) with respect to all the other ligand structures being considered. The optimal CoMFA alignment of a ligand is widely assumed to be its experimental "receptor-bound conformation". Such alignments can yield excellent CoMFA models, sometimes superior in their predictive accuracy to direct calculations of binding energy.[7−9] In the absence of a reliable receptor site structure, the analyst must choose among a multitude of ligand alignment protocols. Such protocols can yield CoMFA statistics superior to those for the favored "receptor-bound conformations",[10−13] and indeed alignment has been advocated as another variable for optimization.[14] However, only one approach exists for comparing alignment protocols, maximization of various q^2 (cross-validated r^2) statistics, which average the internal predictive accuracy for every structure/activity observation. Considering only the measurement uncertainties attached to any structure/activity observation, it is doubtful that moderate changes in q^2 can provide dependable alignment guidance.

The other major difficulty with CoMFA is in applying its results. As with other QSAR methodologies, a CoMFA can only passively filter lists of candidate structures—generation of those candidate lists is an exercise left to other means. Furthermore and peculiar to CoMFA, the structural alignment issue must be addressed somehow for each individual candidate structure.[15]

These two CoMFA difficulties would be addressed by a completely objective and universal methodology for generating a structural alignment, both a conformation and its orientation in a Cartesian space. Of course, such a methodology would by definition ignore any specific receptor requirements, which in principle should degrade the resulting CoMFA model. Yet some of the subjective judgments needed to perform CoMFA alignments may introduce as much "noise" (irrelevant differences among structures) as signal into the input data.

This supposition could be conveniently evaluated, because there already exists an objective and universal "topomer" methodology for generating an alignment of a structural fragment.[16] Structural fragments by definition contain a common feature, the "open valence" or "attachment bond". The topomer methodology overlaps this common feature to provide an absolute orientation for any fragment. A single fragment conformation is

* To whom correspondence should be addressed. Tel: (505)995-4425 or (314)647-1099. FAX: (505)995-4439. E-mail: cramer@tripos.com.

10.1021/jm020194o CCC: $25.00
Published on Web 01/01/2003

then generated from a standardized 3D model by rule-based adjustments to acyclic single bond torsions and chiralities. Previous applications of topomers then proceed to characterize and compare aligned 3D fragments by steric fields (as in CoMFA though with subtle differences) and, recently, by the locations of pharmacophoric features. Here only the aligned 3D topomer structures themselves would be used.

There would remain the issue of how to convert structure/activity observations into sets of mutually comparable fragments. Four different cases may be considered, roughly in order of increasing generality:

Case 1. A purely congeneric series, for example, a combinatorial library sharing a common core. Differences in activity can originate only from differing portions of the structures. Hence, in this case the variable "side chain(s)" which are responsible for the differences in potency can be "clipped off" the common core, to become the set(s) of topomerically modeled fragments. In this case, ambiguity about fragment definition is unlikely.

Case 2. A roughly homologous series, with each individual structure consisting of more than one large group connected by one or more acyclic bonds, but with none of those large groups identical throughout the series. This case could be considered as similar to case 1, except that the largest "common core" comprises only one of those acyclic connecting bonds. Thus, two topomer fragments are produced simply by splitting each series member at a chosen acyclic bond. However, there may be ambiguity in choosing the acyclic bond within each of the input structures.

Case 3. A roughly homologous series containing only one large group, which is similar though not identical across the series. The steroid data set of the first CoMFA publication[6] would be an example. Such series are poorly suited for topomer alignment because the few acyclic bonds are not structurally central, although they are relatively easy to handle semiautomatically by other alignment methods, such as RMS superposition of their maximum heavy-atom subgraph.

Case 4. Series having negligible homology. In the absence of recognizable commonalities, it may be helpful to identify subseries, each having both a structural commonality such as those suggested above and a few active individual structures.

If topomer alignment[17] were to produce a satisfactory CoMFA, the resulting QSAR would also be immediately suitable for searching enormous databases of 3D fragments that already are aligned by the same topomeric procedure. Only a few minutes would then be needed to translate screening results for an initial combinatorial library, through topomer CoMFA, into predictions of which among several hundred billion other side chain synthon combinations are most likely to confer greater potency.[18] Alternatively, such a topomer CoMFA model could also be used to search among conventional databases of complete structures[19] for constituent fragments likely to increase potency.

But rapid analysis and searching are relevant only if the predictions are reliable. To establish any confidence in the reliability of predictions for phenomena as subtle as drug activities, a number of examples must be considered. Fortunately, the literature abounds in successful CoMFA applications. We therefore have set out to repeat many of those studies, starting with the same 2D structures and activity values, but using, instead of the carefully considered alignments of the original authors, the context-ignorant topomer alignment method. We compare our results with those previously published mainly on the basis of two criteria:

(i) Their internal self-predictivity (as measured by the q^2 statistic)

(ii) Their external predictive accuracy (RMS error of all potency predictions for compounds not included in the training set)

To get some impression of what structures and activity predictions might then result from "topomer CoMFA searching", we then used each of the topomer CoMFA models to search an in-house database of roughly a half-million topomerically aligned side chains.

Methods

Selection and Preparation of Datasets. Recent issues of various journals at hand were scanned, to accumulate 11 publications that included 3D-QSAR models (either CoMFA or the GRID-GOLPE variant) for a total of 14 sets of biological data. These models are summarized in the left-hand side of Table 1 (including a 15th topomer CoMFA example resulting from an alternative structure fragmentation tried for the ICE structures). From left to right, Table 1 provides for each model a short "dataset name" as subsequent identifier, the literature reference footnote, the biological endpoint, the general kinds of structures, the general methods originally used to generate and then orient their 3D models, and the PLS variation used to derive the 3D-QSAR. Whenever an article described multiple 3D-QSAR models, the model chosen for comparison was that whose derivation conditions seemed most similar to those of "standard CoMFA".

In repeating these studies, the general approach was to simplify the setup and analysis procedures to an extent that there should be no doubt about whether they could be automated. To begin with:

(i) All compounds reported were included in the study, with exceptions as noted in the Omit IDs column of Table 1 for the reasons footnoted, and with best estimates of any biological potency values not directly reported.

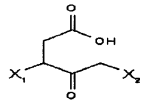
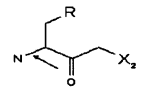
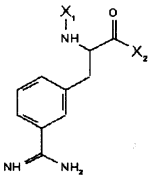
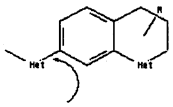
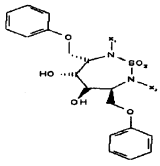
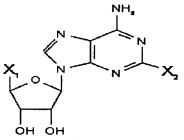
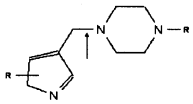
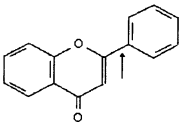
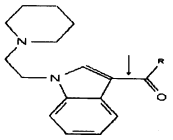
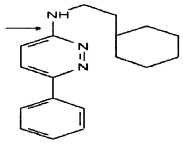
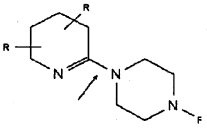
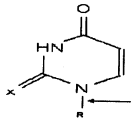
(ii) All ionizable groups were entered in their uncharged forms.

(iii) Stereoisomerism was ignored.[31] The averaged potency value was used wherever potencies were reported for multiple resolved stereoforms.

All of the compounds in each publication were inspected manually for a multi-heavy atom "common core" (corresponding to Case 1 in the introduction). If a common core was not found, a "commonly located" acyclic single bond was identified (corresponding to Case 2 in the introduction), as the bond that was closest to the largest group most common to all structures. These designations appear in the Frag Case column of Table 1. For the Case 1 series, the "Common Core or Fragmentation Bond" column shows the entire common core structure, with the variable "side chains" denoted by X_1 and X_2. For Case 2 series, a partial generic structure is instead shown with an arrow pointing to the bond where the input structures were split. Note that both fragmentation methods were tried for ICE, yielding the ICEc and ICEb examples, with two compounds necessarily omitted in ICEc as lacking the common core.

Entry of structures was by typing SLN representations of each topomeric fragment (as a 2D structure only) into an ASCII file, the three items per compound-line of each file being an R1 fragment SLN, an R2 fragment SLN, and the biological potency (in the familiar −log[concentration] formulation, the logit transformation being used where necessary to convert %binding observations to crude logIC50 values). Safeguards

Table 1. Data and Methodologies for the Fifteen 3D QSAR Literature Studies Repeated with Topomer CoMFA

				Literature CoMFA methods			Topomer CoMFA		
Dataset Name	**Ref #**	**Biological activity**	**Structural class**	**Conformer**	**Orient**	**PLS[a]**	**Omit IDs**	**Frag Case[g]**	**Common Core or Fragmentation Bond**
ICEc	20	Interleukin 1-b converting enzyme inhibitor	doubly blocked mono- to tripeptides	docking to enzyme 1ICE; mutation; minimization	RMS fit to backbone atoms	A	17,30 T1[b]	1	
ICEb							none	2	
thrombin	21	inhibition of thrombin	N-sulfonylated-C-amino derivatives of 3-amidino-phenylalanine	docking into 1ETS, 1PPH, 1HCG; systematic search; minimization		A	none	1	
trypsin		inhibition of trypsin							
factorXa		inhibition of Factor Xa							
MAOa	22	inhibition of MAO-A	mostly 7-benzyloxy-coumarins	coumarins only; minimization of "flat" conformer	RMS fit to coumarin ring	B	all 71 used (lit. used < 45)[f]	2	
MAOb		inhibition of MAO-B							
hiv	23	inhibition of HIV-1 Protease	derivatives of cyclic sulfamides	crystal structure; dock with fixed core; Monte Carlo minimization of side chains (all 2048 asymmetry combinations examined with standard CoMFA including parabolic fields)			none	1	
a2a	24	A_{2A} adenosine receptor agonists	adenosine derivatives	substituents added to X-ray structure of adenosine	RMS fit to N3, C6, N7, N9 of purine ring	A	113[c]	1	
d4	25	D4 receptor antagonism	heterocyclic-CH2-piperazines	substituents added to a clozapine-inspired core, some minimized	"field-fit" using ASP	A	none	2	
flav	26	binding to benzodiazepine site in $GABA_A$ receptors	flavonoids	lowest energy conformer (systematic search, minimization)	RMS fit to aryl ring centroids and carbonyl O	A	none	2	
cannab	27	displacing WIN-55212-2 from CB_1 receptors	aminoalkyl-indoles	lowest energy conformer around C=O (systematic search, minimization)	RMS fit to indole N, carbonyl O, another C	A	none	2	
ACest	28	inhibition of acetylcholin esterase from Torpedo californica	4-phenylamino-pyridazines	docking into 1ACL, 2ACK, 1ACJ, 1VOT		B	6k[d]	2	
5ht3	29	displacing 5-HT_3 from NG 108-15 cells	heterocyclic-piperazines	CATALYST hypothesis		A	19-23, 68-74[e]	2	
rvtrans	30	protection of MT-4 cell from HIV-1	thymines	minimization of ALCHEMY models with GAUSSIAN94	RMS fit to thymine heavy atoms	A	none	2	

[a] A denotes the standard CoMFA method within SYBYL. B indicates use of GOLPE. [b] These structures did not contain the common structure (Asp). [c] Structure not given for this compound. [d] Structure not understood for this compound. [e] These structures would have been fragmented at a double bond, which topomerically is not yet defined. [f] By including non-coumarin compounds, compounds which for solubility reasons did not reach 50% inhibition (logIC50 estimated by logit transform) and compounds showing no activity at the solubility limit (assigned an arbitrary low logIC50 value of 3.0). [g] See description of Cases in the Introduction.

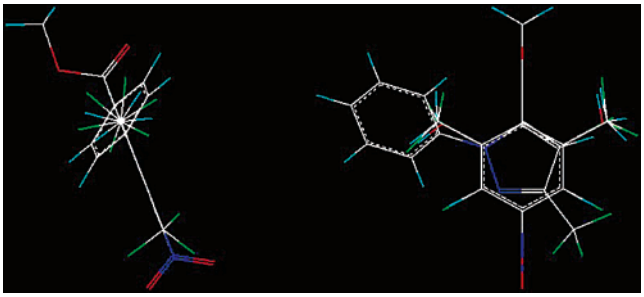


Figure 1. Orthogonal views of the overlaid topomer alignments for the 36 R2 fragments from the ICEc data set.

against structure transcription errors occurred mostly in the subsequent analysis as follows: there should be only one open valence; the valence-filled structure should be modeled by Concord; structures yielding the largest residuals after PLS were rechecked against the original publication.

Topomer CoMFA. There are two main phases in topomer CoMFA, the first being generation of the topomer 3D models for each of the "side chains", and the second the CoMFA analysis itself. In this work, each analysis was performed in two different ways on the same sets of 3D models, one as "standard CoMFA" with SYBYL's usual defaults, and the other as "standard topomer CoMFA" using several adaptations mostly needed to support the subsequent searches.

Procedures for generating the topomer conformation have been detailed elsewhere.[15] In brief summary:

(i) A structurally distinctive "cap" is attached to the open valence, and a Concord model is generated for the resulting complete structure.

(ii) This model is oriented to superimpose the "cap" attachment bond onto a vector fixed in Cartesian space.

(iii) Proceeding away from this "root" attachment bond, only as required to place the "most important" (typically the largest) unprocessed group farthest from the root and the next most important counterclockwise relative to the largest looking along a vector pointing back to the root, stereocenters are inverted[31] and torsional angles adjusted.

(iv) Removal of the cap completes the topomer conformation.

Orthogonal views of the topomer alignments for three of the 24 sets of structures (two R-groups times twelve series) appear in Figures 1−3. These three sets are also the input 3D models for the four topomer CoMFA search results shown in Figures 5−8, with Figure 2 showing the topomer models underlying both Figures 6 and 7. In the left-hand view of Figures 1 and 2, the aligning attachment or root bond is the left uppermost bond shown (perpendicular to the left margin of the figure) and the viewing direction is perpendicular to the XY plane. In the left-hand view of Figure 3, the root bond is the leftmost

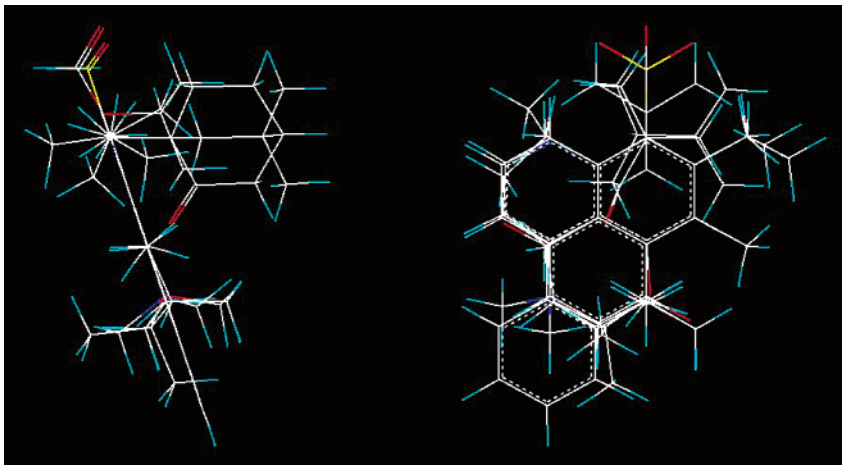


Figure 2. Orthogonal views of the overlaid topomer alignments for the 72 R1 fragments from the thrombin/trypsin/factor-Xa data set.

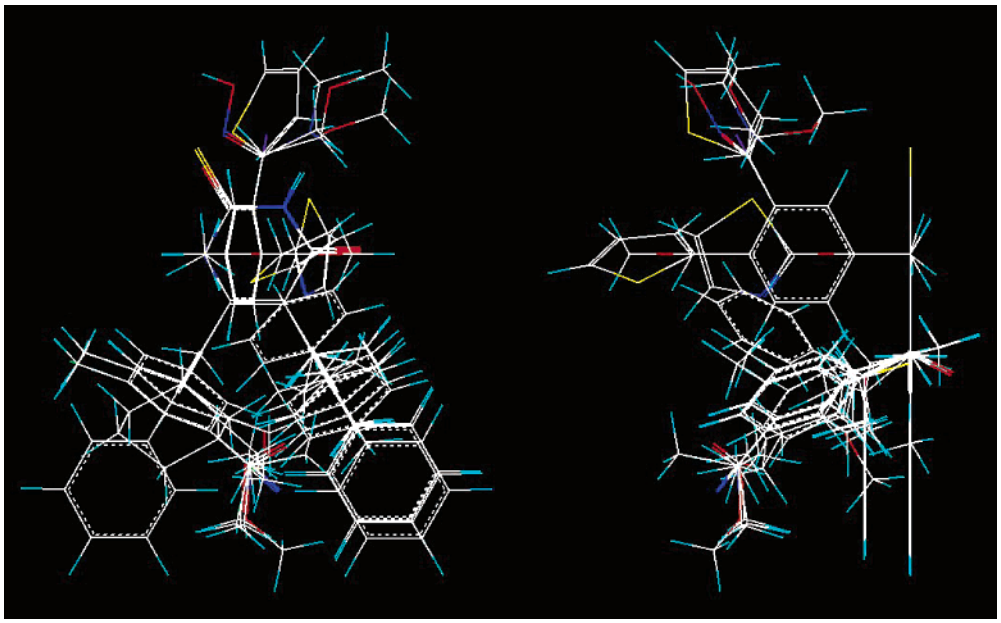


Figure 3. Orthogonal views of the overlaid topomer alignments for the 82 R1 fragments from the rvtrans data set.

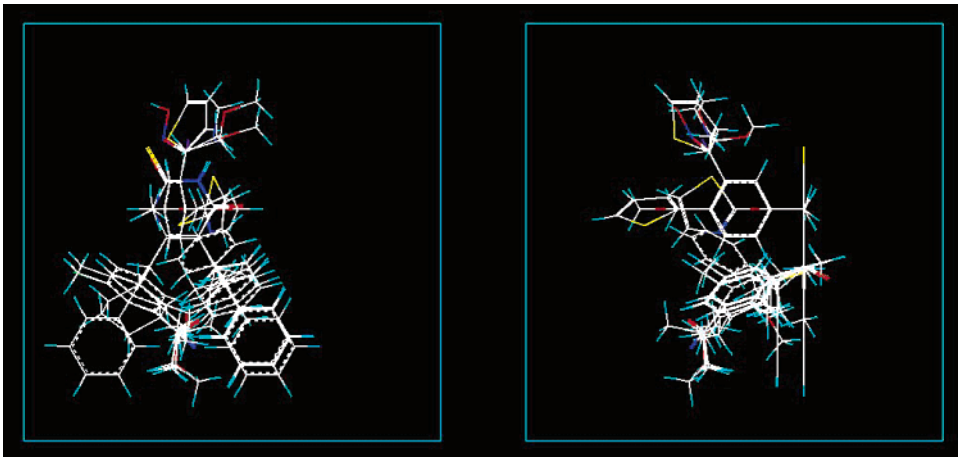


Figure 4. Orthogonal views of the standard topomer CoMFA lattice boundary, surrounding the overlaid topomer alignments for the 82 R1 fragments from the rvtrans data set.

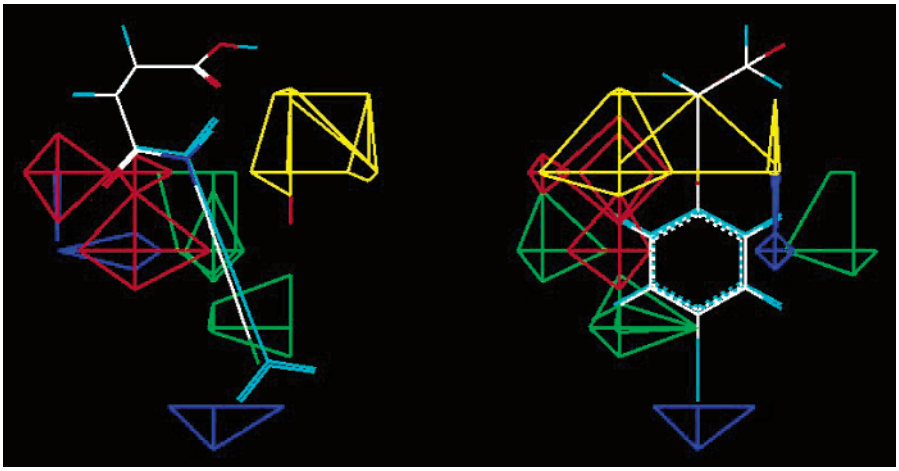


Figure 5. Orthogonal views of some topomer CoMFA search results for the R2 fragment of the ICEc data set. Overlaid are the topomer CoMFA contours, the R2 group in the most active structure originally reported (colored cyan), and the predicted "best" R2 group (colored "by atom type").

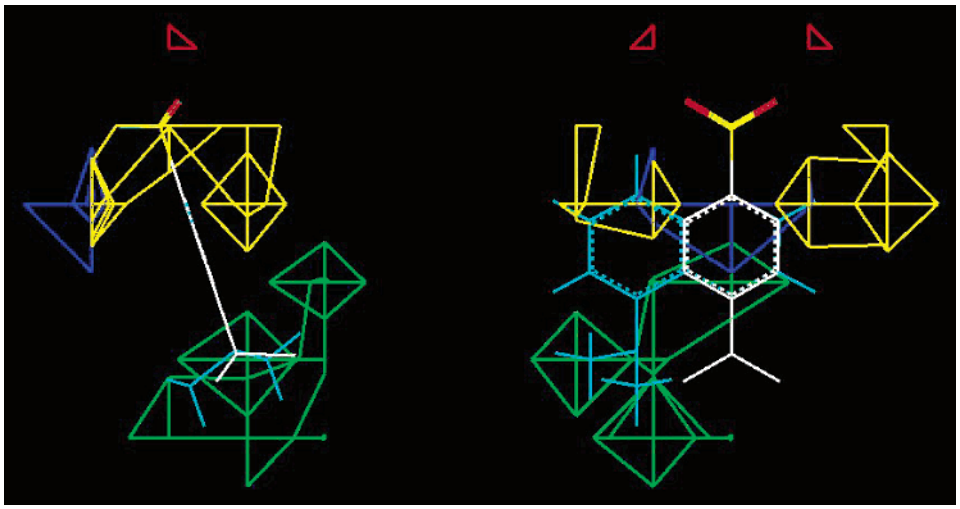


Figure 6. Orthogonal views of some topomer CoMFA search results for the R1 fragment of the thrombin data set. Overlaid are the topomer CoMFA contours, the R1 group in the most active structure originally reported (colored cyan), and the predicted "best" R1 group (colored "by atom type").

attached to the center group of rings (again perpendicular to the left margin of the figure) and the viewing direction is perpendicular to the XZ plane (to better show the spread of the structures).

Since there are two sets of 3D models, one for each of the R groups, two "CoMFA columns" are needed rather than the familiar single "CoMFA column". Fortunately, the internal architecture of standard releases of SYBYL already supports this scheme, so that a rather simple SPL script suffices to perform the remainder of the topomer CoMFA analysis. A pair of columns is created for each varying "side chain" or R-group position, one of CONFORMER type into which a "full" SLN of

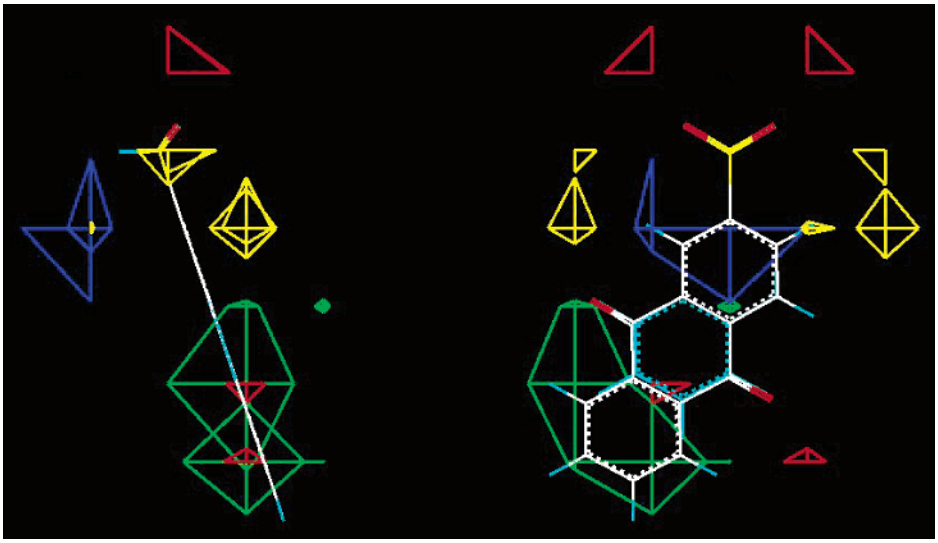


Figure 7. Orthogonal views of some topomer CoMFA search results for the R1 fragment of the trypsin data set. Overlaid are the topomer CoMFA contours, the R1 group in the most active structure originally reported (colored cyan), and the predicted "best" R1 group (colored "by atom type").

the 3D topomer model for that side chain from the compound-row is written, and the other of COMFA type. The TABLE CONFORMER command links the field generation with the appropriate 3D model.

Two other general procedural points about the automatic CoMFA setup and analyses are worth noting.

(i) All atomic charges were calculated by the Gasteiger-Marsilli method for the topomer structure (its open valence being filled with hydrogen), and without any assignment of formal charge.

(ii) The lattice is a 2 Å grid with its lowest valued corner at (−4, −12, −8) and its highest valued corner at (+14, +6, +10). (This "standard topomer" grid is intended as the 1000 point cube that is best positioned to contain a topomer, with its root vector endpoint coordinates of [0,0,0], [1.5,0,0].) Figure 4 shows the superposition of these lattice boundaries on the topomerically aligned side chains of Figure 3. Atoms that happen to extend beyond these boundaries (encountered in roughly 4% of a random sample of 100 000 topomerically modeled fragments) will have negligible influence on the fields and the CoMFA.

"Standard Topomer CoMFA". Two sets of analyses were performed on the same 3D model sets. To facilitate comparison with the literature results, one set used SYBYL's "standard CoMFA" default settings for field calculations and PLS with leave-one-out cross-validation, the only differences from usual practice being the topomer models and the two CoMFA columns. The other set of "standard topomer CoMFA" analyses involved the three following methodological changes, the first two for compatibility with the already-stored properties of the topomer structures to be searched:

(i) An "attenuation factor" reduces the field contributions of fragment atoms more distant from the attachment bond. The steric or electrostatic contribution of an atom to the field value at a lattice point is multiplied by 0.85^n, where n is the number of acyclic single bonds along any path between that atom and the open attachment bond. (Attenuation reduces the influence on dissimilarity of atoms whose link to the fragment root is more flexible, as being more able to adjust themselves to optimize receptor interactions.)

(ii) The field values at each lattice point are discretized by rounding up as follows. Steric field values can be 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, or 30 kcal/mol. Electrostatic field values can be −13, −11, −9, −7, −5, −3, −1, +1, +3, +5, +7, +9, +11, +13, or +15 kcal/mol. Any values greater than the highest value shown are rounded down. Such rounding also substantially reduces the number of terms in the CoMFA QSAR equations, as many of the lattice points then experience identical field values from all structures. (The steric fields to be searched were already discretized in this fashion, so as to conserve storage space while preserving over 90% of the intrafield variance.)

(iii) The number of PLS components is chosen more conservatively, as that yielding the (first local) minimum standard error of prediction rather than the (first local) maximum q^2. (Now that there is 20 years' collective experience with PLS in CoMFA, this change in standard practice would be favored by most knowledgeable opinion, a view for which these results will provide a bit of incremental support.)

Potency Predictions. For 11 of the 15 datasets (from eight of the 11 publications), the CoMFA model had been validated by predicting the potencies of compounds completely omitted from its development. Because of this exemplary practice, predictions could also be made from the topomer CoMFA models and their accuracies could be compared with the accuracies of the published predictions. Each of the "test compounds" to be predicted was fragmented and modeled in exactly the same way as the "training compounds" used to develop the topomer CoMFA. The QSAR PREDICT operation of SYBYL then automates the calculation of fields and the plugging of those field values into the appropriate CoMFA model, yielding a potency prediction.

Implementation of Topomer CoMFA Searching. A topomer CoMFA search predicts the activity contribution or "partial activity", using a topomer CoMFA QSAR, for each side chain that also might partially satisfy a topomer similarity search criterion. Thus, topomer CoMFA searching evaluates the topomer similarity of each candidate, as well as its predicted partial potency. Of course, to maintain the speed of topomeric searching, the CoMFA QSAR is evaluated by the topomer searching program rather than within SYBYL. Potency predictions also required electrostatic fields (again with Gasteiger-Marsilli atomic charges based on the H-capped fragment and no formal charge assignments) to be calculated for all the topomer structures to be searched.

Less obviously, usually the "topomer similarity" aspect of a query is to be defined in terms of a set of structures (the topomer CoMFA inputs) rather than a single exemplary structure. The two descriptors currently used in determining topomer similarity are the steric fields (evaluated over the standard topomer CoMFA lattice) and the list of pharmacophoric features (where each feature is characterized by a class and the Cartesian coordinates of a key atom).[19] To formulate these similarity descriptors from a CoMFA model input set, *for similarity searching only*, each query steric field value is taken to be the average of the steric field values over the CoMFA input set, and the query pharmacophoric feature list includes only those features that are found among at least 2/3

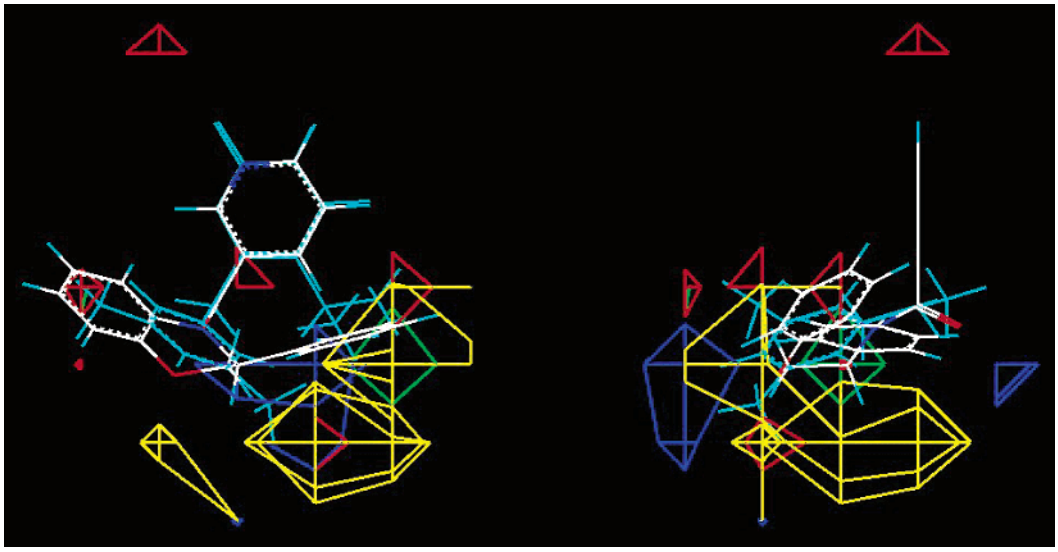


Figure 8. Orthogonal views of some topomer CoMFA search results for the R1 fragment of the rvtrans data set. Overlaid are the topomer CoMFA contours, the R1 group in the most active structure originally reported (colored cyan), and the predicted "best" R1 group (colored "by atom type").

of the CoMFA input set at the same topomer location (within an 0.2 Å tolerance). (Features that are present in less than 2/3 of the CoMFA input set are thus expected to have their influence expressed in the CoMFA QSAR, if they have any importance for activity.) Also the feature similarity calculation during topomer CoMFA excludes "missing candidate features", which otherwise would penalize features that are present in the candidate match but not in the query.[19]

Although the units of topomer dissimilarity are inherited directly from standard CoMFA steric fields, there are two complications. First, lattice-point-by-lattice-point topomer steric field differences ("distances") are combined in a Euclidean but not very intuitive root-mean-square manner. Thus, if the only change from one structure to another is a methyl group, the topomer difference will be 90 units, but if there is already a difference of 200 units, the same additional methyl will increase that difference by only 19 units (calculated as sqrt(200 * 200 + 90 * 90) − 200). Or, put differently, the shape change represented by a topomer difference of 180 is more than three times as great as the structural change represented by a topomer difference of 90. Second, the introduction of pharmacophoric features, by providing new ways that two fragments may differ, increases the average topomer difference between groups. The amount of this increase depends on the relative scaling of pharmacophore feature differences relative to steric differences. In the searches to be described, all topomer differences (similarities) included feature differences with default relative scaling.[19]

Topomer CoMFA Searches. Topomer CoMFA searching was performed on a current version of the "two-piece" sublibrary within the ChemSpace virtual library, for which electrostatic fields were additionally calculated as described above. This sublibrary combines 69 751 nucleophilic synthons representing 19 families with 89 509 electrophilic synthons representing 26 families, or roughly 6.2×10^9 product structures (possible Class 2 data set members) or side chain pairings (possible Class 1 data set members). All but a few of the synthon structures originated from reagent catalogs, and thus most of the product structures are presumed readily available.

The contributions of each side chain to a topomer CoMFA model potency prediction are completely additive and independent "partial potencies". Therefore, the output of topomer CoMFA searching emphasizes "hitlists of R1's and R2's", suitable for inspection within a SYBYL molecular spreadsheet containing among its columns "partial potency", "partial topomer similarity", "R1 or R2", and "nucleo-/electrophile".

All topomer CoMFA searches used the same maximal allowable dissimilarity (to the average input structure as described above) of 150 topomer units per R group, and the same minimal acceptable total predicted potency of 4 log units (equivalent to a 100 μm IC50). To simplify their inspection, the resulting R hitlists were merged, with filtering to exclude duplicates (same synthon arising from multiple synthetic routes); certain undesirable substructures that had persisted through virtual library construction (notably two nitrogens connected by any acyclic bond); and charged moieties whose activity predictions would represent unreasonable extrapolations of the topomer CoMFA models (which as mentioned were derived strictly from uncharged input structures). A further partial activity cutoff was selected manually to limit the size of each R-group molecular spreadsheet to a few hundred "best" synthon structures.

Selection of R-Group Structure Examples. The two search criteria, similarity to the average CoMFA input structure (s) and CoMFA predicted potency (p), were combined to yield a single R-group "score" according to the following formula: $p + 0.01 * (150 - s)$. Stated differently, since 150 was the similarity cutoff when searching for an R-group, the resulting "score" is the predicted potency incremented by an amount between 0.0 and 1.5 log units. The side chain structure having the highest such "score" was examined further, in particular by overlaying its 3D topomer model over the contoured topomer CoMFA model (see Figures 5−8 for examples), mainly to detect possible procedural error.

Results

Table 2 shows the results for the 30 topomer alignment-based CoMFA models, i.e., the two procedural variations applied to each of 15 studies. The key results from Table 2 are also depicted in bar graphs, Figures 9−11, to facilitate individual comparisons. In both presentations, each topomer result, in columns or bars labeled TopA (standard CoMFA) or TopB (standard topomer CoMFA), accompanies the corresponding result from the literature, in columns or bars labeled Lit.

The left-hand block of Table 2 compares the parameters involved in CoMFA model construction. From left to right within this block, there appear subblocks for each of the 15 studies containing:

(i) the count of compounds used to derive the topomer-based model;

(ii) three q^2 values (leave-one-out cross-validated r^2), shown also in Figure 9;

Table 2. Statistical Parameters of Model Derivation and the External Prediction Errors, for the 15 3D QSAR Literature Studies and their Repetitions with Topomeric CoMFA

		CoMFA Model Construction															CoMFA Prediction			
dataset		#	x-validated q^2			x-val SDEP			# compnts			final r^2			% steric		#	RMS pred error		
ID	name	cpds	lit	TopA[e]	TopB[f]	lit[b]	TopA	TopB	lit	TpA	TpB	lit	TopA	TopB	lit	TopA	cpds	lit[a]	TopA	TopB
1	ICEc	36	0.630	0.346	0.362	0.816	1.032	1.002	6	6	5	0.970	0.879	0.883	59	53	9	0.568	0.728	0.740
2	ICEb	38	0.630	0.435	0.433	0.816	0.949	0.951	6	5	3	0.970	0.885	0.806	59	56	10	0.553	0.686	0.595
3	throm-bin	72	0.687	0.512	0.533	0.594	0.741	0.726	4	4	4	0.881	0.822	0.838	62	52	16	0.673	0.596	0.619
4	trypsin	72	0.629	0.678	0.657	0.556	0.522	0.531	5	6	4	0.916	0.939	0.886	66	51	16	0.524	0.472	0.523
5	factorXa	72	0.374	0.221	0.186	0.515	0.578	0.591	3	4	4	0.680	0.761	0.747	70	49	16	0.278	0.329	0.340
6	MAOa	71	0.440	0.528	0.566	1.025	0.974	0.926	2	5	4	0.680	0.822	0.813		26				
7	MAOb	71	0.430	0.460	0.483	1.253	1.250	1.214	2	3	2	0.880	0.652	0.640		25				
8	hiv	25	0.680	0.479	0.389	0.571	0.894	0.845	3	8	3	0.950	0.986	0.878	66	46	7	0.823	1.133	0.449
9	a2a	78	0.541	0.296	0.226	0.563	0.723	0.742	4	6	3	0.817	0.802	0.555	52	58	23	0.668	0.594	0.761
10	d4	29	0.739	0.730	0.636	0.734	0.723	0.802	7	7	5	0.996	0.983	0.957	77	56				
11	flav	38	0.752	0.733	0.763	0.475	0.553	0.495	4	8	5	0.969	0.946	0.952	54	50	4	0.337	1.344	1.314
12	cannab	61	0.592	0.402	0.423	0.570	0.716	0.696	4	4	3	0.905	0.816	0.777	80	78	6	0.452	0.602	0.540
13	ACEest	41	0.937	0.784	0.746	0.346	0.668	0.726	4	3	3	0.990	0.934	0.916		53	7	0.413	0.513	0.478
14	5ht3	61	0.645	0.393	0.295	1.193	1.752	1.804	5	7	2	0.913	0.858	0.519		53				
15	rvtrans	82	0.837	0.806	0.830	0.567	0.634	0.587	4	6	4	0.936	0.919	0.916	64	71	19	0.791	0.574	0.608
	total/avg	847	0.636	0.520	0.502	.581[c]	.728[c]	.717[c]	4	5.5	3.6	0.897	0.867	0.806	64	56	133	0.553 0.574[d]	0.688 0.623[d]	0.633 0.565[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 CoMFA Prediction RMS error). [d] Average excluding flav (see text for discussion). [e] TopA was derived using "standard CoMFA" settings. [f] TopB was derived using "standard topomeric CoMFA" settings, and then used for topomer CoMFA searching.

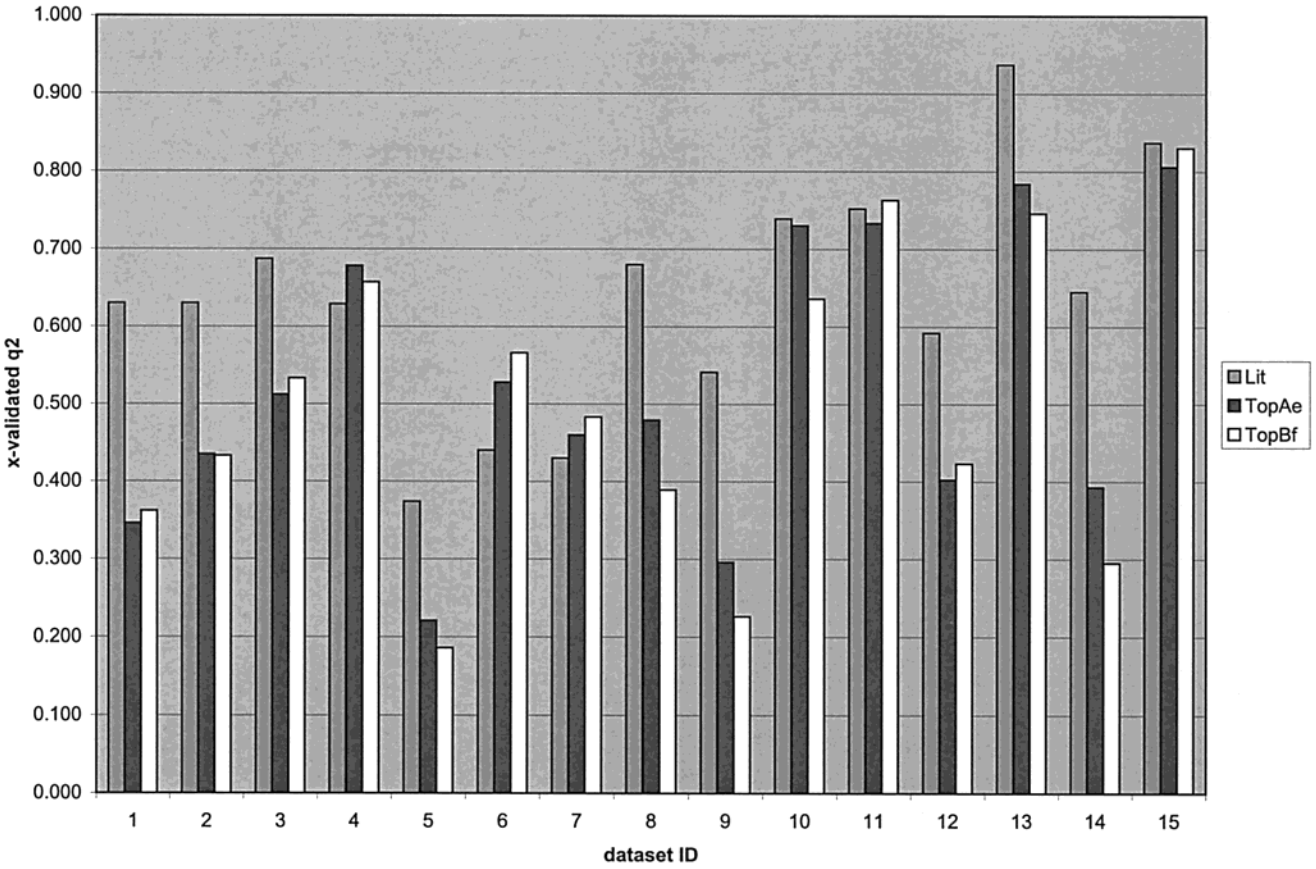


Figure 9. Cross-validated q^2 obtained from the three methods of model construction for each of the 15 different datasets. Data are taken from the "x-validated q^2 block" of Table 2.

(iii) three SDEP values (RMS of all prediction errors during cross-validation, weighted by degrees of freedom), shown also in Figure 10;

(iv) three values for the number of PLS components that yielded the q^2 shown; three r^2 values for the final model (using the corresponding numbers of PLS components);

(v) two contributions to the model from the steric fields, as a percentage of the total variation explained (the remaining contributions are of course from the electrostatic field).

An average or total value appears at the bottom of each column. Note that all averages are over data sets, unaffected by the number of compounds associated with

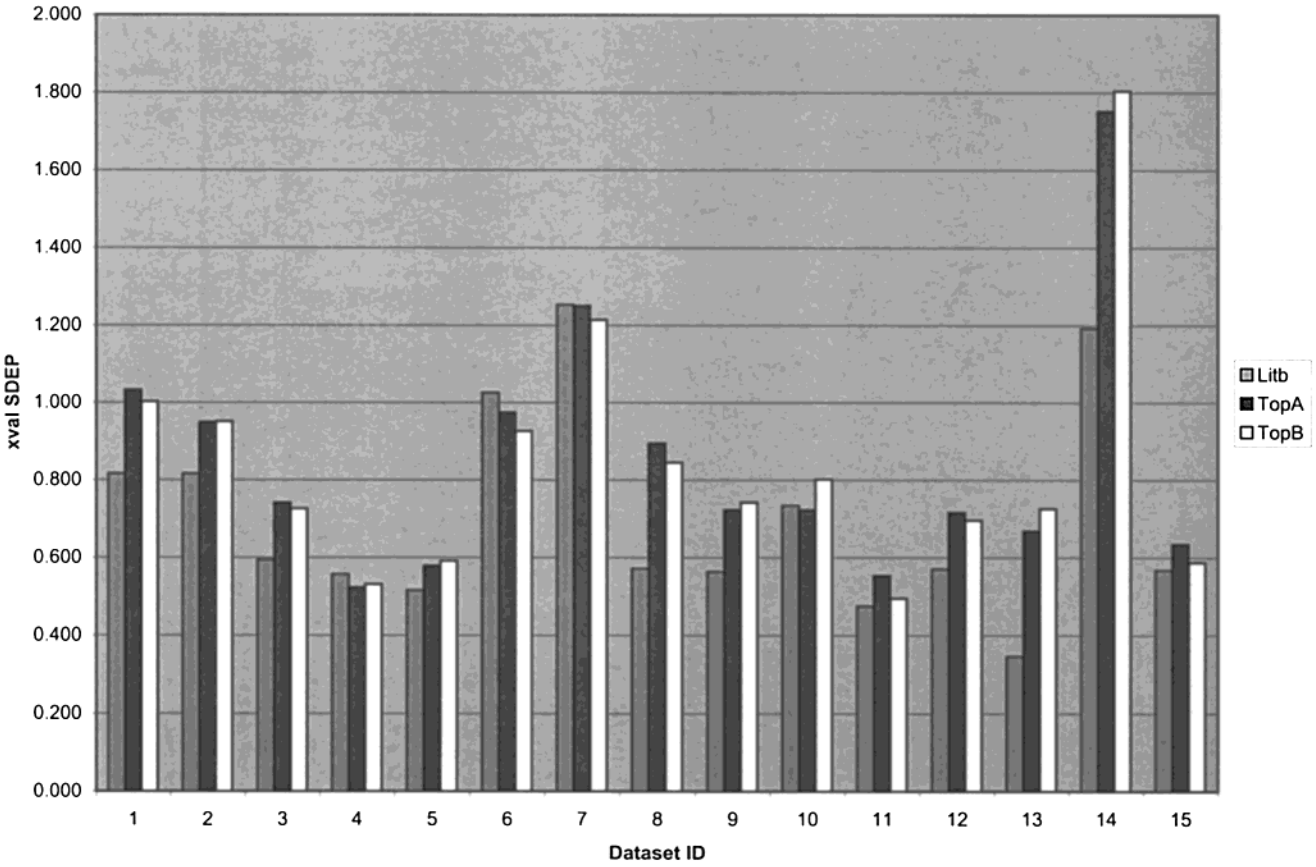


Figure 10. Standard errors of "prediction" during cross-validation, obtained from the three methods of model construction for each of the 15 different datasets. Data are taken from the "xval SDEP" block of Table 2.

a data set. Footnotes indicate where certain data sets were omitted from an average, to provide a more appropriate comparison with some other column average.

To summarize the results from model construction, the topomer alignment CoMFA models are generally weaker than the literature models in fitting the experimental data. However, the average decrease in precision is only a little over 0.1 in q^2 value. Much more surprising and encouraging is the consistency of the positive results obtained using the context-ignorant topomer alignment procedure. Every one of the 30 topomer CoMFA derivations provided a 3D-QSAR correlation that was both statistically significant and (see below) predictively useful when testable.

The standard CoMFA runs (TopA columns) provide the most direct comparison with the literature results. Here the largest individual declines in q^2, for the ICEc and 5ht3 data sets, are about 0.25, while there are even three studies, trypsin, MAOa, and MAOb, in which the q^2 values are a bit higher than the literature value. On the other hand, the q^2 value for factorXa does fall just below a traditional rule-of-thumb q^2 cutoff of 0.25,[32] and the average number of PLS components in the topomer CoMFA correlations (5.5) exceeds that for the literature models (4.2) by more than a single component.

The standard topomer CoMFA results (TopB columns) are of much greater practical importance, as the only model formulation that enables topomer CoMFA searching. At first glance, a further decline in average q^2 (moving a second study below the traditional 0.25 q^2 cutoff) and a drop in average Final r^2 might seem disappointing. However, the much lower average number of components in the standard topomer CoMFA models (3.6), a result of the more conservative stopping criterion of minimizing SDEP rather than maximizing q^2, complicates this direct comparison. In fact, most experienced workers would greatly prefer the overall standard topomer CoMFA (TopB) results to the standard CoMFA (TopA) results, as the latter models require almost two more components to produce an average improvement of only 0.018 in q^2, and so the average xval-SDEP for TopA actually becomes slightly worse (larger) than that for TopB (because of the larger number of components, hence reduced degrees of freedom for TopA). Some might even regard the standard topomer CoMFA models as statistically equivalent in quality to the literature models, since the average improvement in q^2 of 0.134 did require an average increase of 0.6 in the number of PLS components.

Because a q^2 value is dependent on the spread of the biological potencies as well as their accuracy of prediction, the cross-validated SDEP can provide useful supplementary information about a QSAR. For example, the SDEP values for the factorXa CoMFA look a lot better than do their q^2 values, because the range of the factorXa potencies is so unusually small. On the other hand, the 5ht3 SDEP values are rather troublesome. It turns out that a major feature within the original SAR data for 5ht3 is an extraordinarily nonlinear behavior of two among the 61 compounds. A particular pairing of nonunique structural changes within a series whose −log(IC50) otherwise averages only 6.56 is reported to produce two picomolar ligands. Such an enormous

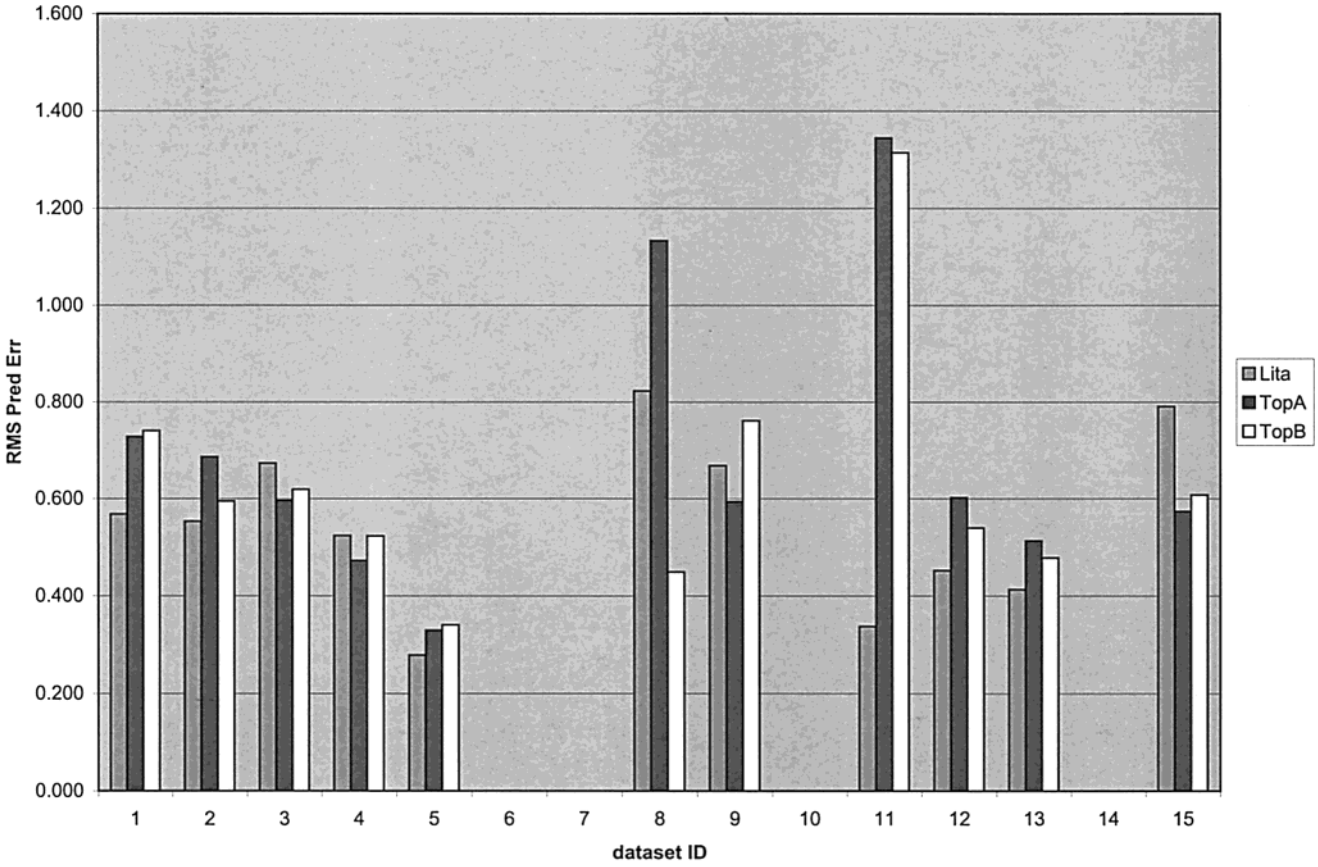


Figure 11. RMS error of potency prediction for compounds not included in model derivation, obtained from the three methods of model construction for each of the fifteen different datasets. Data are taken from the "RMS pred error" block of Table 2.

discontinuity in the underlying SAR data would severely challenge any QSAR methodology.

The right-hand CoMFA Prediction block of Table 2 compares the errors of the various CoMFA models in predicting the potencies of 133 structures not considered in their derivations, as the RMS of the individual errors. These results are probably the most important for evaluating overall performance and so are presented as bar graphs also in Figure 11. Although the literature alignment results are slightly better than those provided by topomers, the differences would have little practical significance, and it seems reasonable to characterize the average of the RMS prediction errors of the topomer CoMFA models as completely comparable to the literature models. Each of the three averaged errors from "true prediction" are actually somewhat smaller than the corresponding average SDEPs during crossvalidation, a result that supports the overall robustness of all the CoMFA results but which also suggests that many of the structures originally selected for potency prediction were not the most challenging.

To help in better visualizing how the results obtained from the three methods of model construction vary among the 15 different datasets, Figures 9 and 10 present as bar graphs the two most important data blocks in Table 2, the cross-validated q^2 for model derivations and the RMS error of true predictions.

Among the 15 individual prediction comparisons, the flav data set stands out for the very poor predictions yielded by its topomer CoMFA models. It happens that 75% of the topomer CoMFA RMS prediction error for flav is associated with a single compound, the flavonoid baicalin, which combines potency a log unit lower than any CoMFA input structure with a sugar moiety at a position where the only other substituents were -OH and -H. Given little structural precedent for such a large change, the topomer CoMFA models forecast a series-average potency for baicalin, a value that is in fact too high by two log units. However, the CoMFA alignment used in the original publication apparently did support the major extrapolation necessary to obtain an accurate prediction. Because of these peculiarities, the average of the CoMFA RMS prediction errors with the flav data set excluded is also presented, as the bottom line of the CoMFA Prediction block. Over the 129 predictions that then remain, it is evident, from the last line of data in the Pred RMS Error block of Table 2, that "standard topomer CoMFA" performed external predictions at least as accurately as did the collective literature CoMFA models.

Conversely, it is the hiv data set that most improves the average prediction performance for "standard topomer CoMFA" (TopB), relative to that for the literature alignments and especially that for "standard CoMFA" on topomer alignments (TopA). Unfortunately, the previously published hiv results (Figure 6 in ref 23) do not allow identification of two key structures whose potencies were greatly under-predicted by the published model. However, the huge difference between the TopA and TopB hiv prediction performances is no doubt caused by overfit in the TopA model (to include eight components instead of three). Thus, this influential data

set provides evidence favoring the more conservative criterion for PLS analysis, minimizing SDEP rather than maximizing q^2.

Searching the two-piece libraries for possible higher potency side chains, as predicted by the 15 "standard topomer CoMFA" models, generated the results summarized in Table 3. Its first "Compound Potencies" block provides a direct comparison between the potency of the most active compound in the original publication and that predicted for the most active combination of side chains found in the library search. In all but two of the 15 searches, there was found a combination of commercially offered reagents that promised higher potency than was reported in the original publications. (The two less successful searches were 5ht3, with its remarkably synergistic substituent effects as noted above, and ICEc, an alternative fragmentation to the successful ICEb search.) The average of this "accessible predicted potency enhancement" over all 15 searches is 1.28 log units (around 20× in potency). With the exclusion of 5ht3 and ICEc this average would increase to 1.75 log units. It should however be cautioned that, for four of these 13 higher potency predictions, the "best R1" and "best R2" groups found and reported are in combination topomerically dissimilar from the query R1+R2 by more than (a rather arbitrary cutoff of) 150 units.

The remainder of Table 3 presents supporting details for each of the 15 topomer CoMFA searches. Shown for R1 and R2 are the further activity cutoff and the number of R-groups exceeding that cutoff (to provide some impression of the size of a potential combinatorial library), and, for the single "best" potential R-group, its predicted partial potency (total contribution of this R group to the QSAR prediction), its similarity to the topomer CoMFA query, and its structure, side-by-side with the corresponding R-group within the most active structure reported in the original publication. (Although the division in this table between R1 and R2 may suggest otherwise, in fact the R1 and R2 searches are simultaneous.)

In comparing the structures of the "max literature" with the "best R found" in topomer CoMFA searching, two points should be kept in mind:

(i) The searching similarity criterion is to the average of all input structures and the potency predictions result from a 3D-QSAR model that was derived from all the input structures. The single R-group structure shown, even though taken from the most potent reported compound, cannot represent all that information very well.

(ii) The pool of candidate R-groups being searched was in this example almost completely restricted to commercially available reagents as sources. This restriction tends to generate more structural novelty and accessibility, but perhaps less structural credibility, than would ordinarily be acceptable in lead optimization projects.

To try to convey some general and deeper sense of the behavior of topomer CoMFA searching, a few "CoMFA contour maps" are shown as Figures 5−8. These particular maps were chosen as ones that most clearly illustrated how the 3D-QSAR would predict high potency for the R group shown, while viewed from either of two fixed, hence standardized, directions, both of which place the topomer-aligning attachment bond at the left and strictly parallel (or perpendicular) to all three viewing dimensions. Each map displays two orthogonal views of three overlaid objects:

(i) The topomerically aligned R group from the most active structure in the original publication, uniformly cyan (intense blue-green) in color (its 2D structure is the left-hand of the paired structures in the appropriate cell of Table 2);

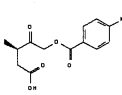
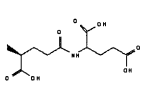
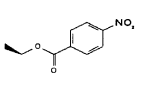
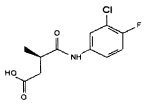
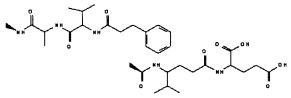
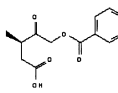
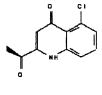
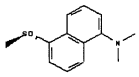
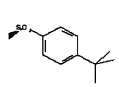
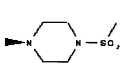
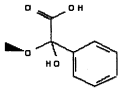
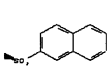
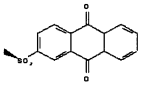
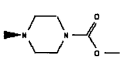
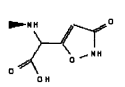
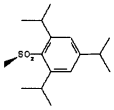
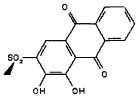
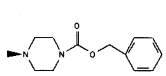
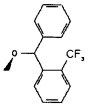
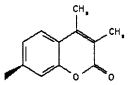
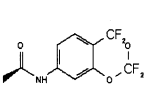
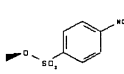
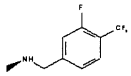
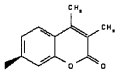
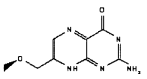
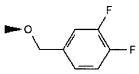
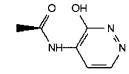
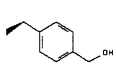
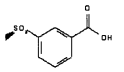
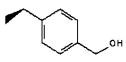
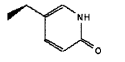
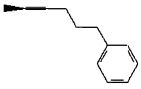
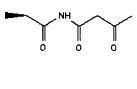
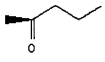
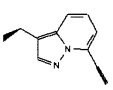
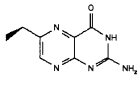
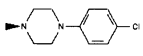
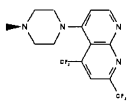
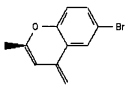
(ii) The topomerically aligned R group with the highest predicted activity from the topomer CoMFA search, colored by atom type (its 2D structure is the right-hand of the paired structures in the appropriate cell of Table 2);

(iii) The set of colored polyhedra, surrounding those lattice points where there is a very strong and consistent association between changes in activity and changes in the steric or electrostatic fields as exerted by the various R groups in their topomeric conformations. Color indicates the nature of the association, increased potency being favored by steric increases near green, steric decreases near yellow, increased negative charge (decreased positive charge) near red, and decreased negative charge (increased positive charge) near blue.

Figure 5, corresponding to the R2 group in the ICEc data set and the input 3D model overlay in Figure 1, illustrates many of the basic characteristics of topomer CoMFA QSAR. First, topomeric alignment places topologically similar main chains in almost identical locations. (So in this figure a bit of study may be needed to be certain which structure is responsible for a particular displayed atom). Second, the carboxylic acid side chain in the proposed side chain must have had a negligible effect on the potency prediction, because there are no nearby polyhedra. (Such an absence of polyhedra means that there was no structural variation affecting that spatial region within the data set, or that any variation in that region had no consistent effect on potency.) On the other hand, if such an unprecedented group had been much larger, the (dis)similarity penalty in topomer CoMFA searching would have been too large for its R-group to become an acceptable hit. Thus, one innate behavior of topomer CoMFA searching is to gently probe unexplored regions of space, in effect a conservative search for secondary binding pockets. Third, the two features in the proposed side chain most probably responsible for its enhanced potency prediction are the amide carbonyl oxygen (a strongly electronegative atom adjacent to several red polyhedra) and the meta chlorine (a bulky atom embedded in a green polyhedron). An attractive feature of topomer CoMFA searching is that such a relatively sophisticated SAR analysis takes place completely automatically and extremely rapidly.

Figures 6 and 7, corresponding to the R1 groups in thrombin and trypsin respectively and the single input 3D model overlay in Figure 2, may provide an interesting comparison with receptor docking studies, especially to workers more familiar than the author with these proteases' S3 binding pockets. In both cases, the forecasts of superior R1 potency appear to be simple consequences of superior positioning of bulk within green polyhedra, while in the case of thrombin (Figure 6) also evading a yellow polyhedron. So do the enzymes actually possess "empty binding pockets" or "key resi-

Table 3. Comparison of the Most Potent Structures from the Literature with the Best Compounds Found by "Topomer CoMFA Searching" of a Virtual Library Based on Commercially Available Fragments, for the Fifteen 3D QSAR Literature Studies Repeated with Topomeric CoMFA

	Cmpd Potencies[a]		R1 Topomer CoMFA searching results						R2						
Dataset		best	R1 candidates		Best R1 found		best R1 structures		R2 candidates		Best R2 found		best R2 structures		
Name	max lit[b]	R1+R2[c]	cutoff	#	p.pot[d]	sim[e]	max literature	best R1 found	cutoff	#	p.pot[d]	sim[e]	max literature	best R2 found	Intcpt
ICEc	6.11	5.8	2.0	53	2.6	142			1.5	4	2.1	109			1.1
ICEb	6.11	8.4	3.0	207	3.4	98			2.0	4	2	83			3.0
thrombin	8.38	10.1	4.0	13	4	87			4.0	2	4.1	106			2.0
trypsin	7.70	8.7	3.0	3	3.1	103			3.0	12	3.4	91			2.2
factorXa	6.05	8.1	2.5	46	2.9	112			2.0	3	2.1	129			3.1
MAOa	7.90	10	1.0	88	1.6	117			2.0	71	3.2	65			5.2
MAOb	8.94	12.1	1.5	12	1.9	128			4.0	28	4.6	54			5.6
hiv	8.51	10.1	2.0	63	2.2	74			1.5	115	1.9	64			6.0
a2a	8.81	10.4	1.0	799	2.6	95			0.0	51	0	23			7.8
d4	9.21	11.2	1.0	1098	2	68			3.0	60	4.5	126			4.7
flav	9.00	10.4	1.5	213	1.9	118			2.0	78	3.2	101			5.3
cannab	3.17	5.6	1.0	23	1.3	138			0.5	101	1.2	138			3.1
ACEest	7.68	8.2	0.0	208	0.6	104			3.0	158	4	62			3.6
5ht3	12.09	9.5	1.0	14	1.2	114			2.0	29	2.5	116			5.8
rvtrans	9.22	10.1	2.0	103	2.9	141			2.0	20	2.7	84			4.5

[a] Potency is defined as −log(IC50) throughout this table. [b] "max lit(erature)" refers to the most active structure reported in the original publication. "exp" is its experimental potency and "QSAR" is its potency calculated by the "standard topomer CoMFA model". [c] Best R1+R2 is the sum of the partial potencies over the "best" R1, the "best" R2, and the intercept. [d] Partial potency of the best R group found. [e] Similarity, in topomeric units including shape and feature differences, to the average of the CoMFA input structures.

dues" in these green or yellow regions of space? This frequently asked question is not very relevant for topomer CoMFA, because topomer conformations depend solely on ligand topology and thus will only coincidentally complement any actual receptor binding pocket. However, it was encouraging to discover that a compound containing the suggested anthraquinone side chain (trypsin R1) had in fact already been found to have the second highest trypsin-binding affinity among the 88 values reported, despite the mild tautology (since the anthraquinone was thus one input into the trypsin model derivation).

The final overlay, in Figure 8, illustrates a much more complex situation, for the R1 group in the rvtrans data set. Here the topologies of the known R1 and of the suggested replacement are completely different, and the replacement also lacks almost all the thymidyl features. Nonetheless, the overlay shows how there may be enough subtler similarities in overall shape that the two groups might yet fit into the same cavity. The higher predicted potency of the suggested structure seems in part a result of the benzyl group overlapping with the only green polyhedron while skirting all of the yellow ones. The region around the original thymidyl contains few polyhedra, despite fairly extensive structural variation among the overlaid input topomers, evident in Figure 3. Such an observation implies that the large existing amount of structural variation must not have had a consistently interpretable affect on potency. It is also apparent that there will not be many reagents commercially available with enough structural complexity to resemble the entire original R1. Breaking the original structure into three or more pieces would have yielded many more similar and perhaps higher scoring hits, which would then all have the same general topology as the original structure instead of the more speculative structure shown.

Discussion

The central finding is that the automatic, receptor-ignorant, topomeric alignment procedures yielded CoMFA of overall quality little worse than the individually context-sensitive alignments underlying the published studies. Such a result seems more or less of a complete surprise, as it tends to contradict much of the conventional wisdom about CoMFA methodologies and their sensitivity to a "correct" alignment. Therefore, it is important to summarize the supporting evidence.

(i) The consistency of acceptable quality among the individual topomer CoMFA models. On the basis of 15 useful results in 15 tries, it seems reasonable to assert that topomer CoMFA is very likely to be productive in any situation where more complex alignment procedures would also be successful. (Perhaps it should be explicitly stated here that there has been no invisible "filtering" of unfavorable results. All topomer CoMFA reanalyses of literature data that were attempted are reported here.)

(ii) In particular, the consistent accuracy of the topomer CoMFA potency predictions. One might be concerned that the more or less arbitrary nature of topomer alignments produces artifactual correlations, perhaps by grouping together structures of similar classes in such a way that leave-one-out cross-validation gives an inflated estimate of statistical quality. But if topomer CoMFA correlations are predominantly artifactual, it is difficult to understand how their predictions could be so consistently satisfactory.[32]

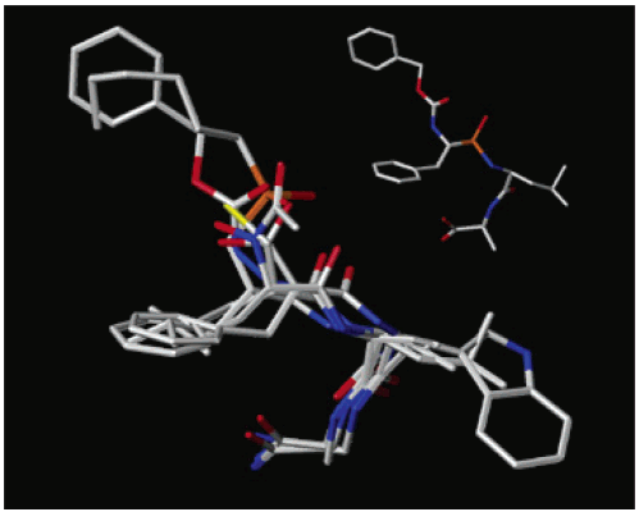


Figure 12. Overlaid results from docking several members of a combinatorial library, illustrating the resulting lack of superposition among the atoms in the common core.

(iii) Any bias in selecting the 15 test cases was toward, rather than away from, the most widely admired literature alignment protocols. In particular, given the widespread assumption of the optimality of receptor docked alignments for CoMFA, it may be noted (in Table 1) that seven of the 15 cases (ICEc, ICEb, thrombin, trypsin, factor Xa, hiv, acest) used docking to a known crystal structure as their primary 3D modeling strategy. Of course, like any other 3D modeling procedure, receptor docking may not always be well done, but this would seem an argument in favor of a completely objective if context-ignorant alignment methodology, such as topomers.

How is it possible that such a context-ignorant and general alignment procedure can compete successfully with the various individually crafted literature alignment procedures? In summary, it currently is believed both that the literature procedures may not be as good as they seem and also that the topomer procedure may not be as naive as it may seem, for purposes of 3D-QSAR generation and prediction. The following supporting observations are intended merely as illustrative alternates to the dominant paradigms, not as formal hypotheses inviting validation studies.

Consider first the literature procedures, starting with docking because of its wide acceptance and its prevalence among these 15 cases. One weakness of docking as an alignment procedure for 3D-QSAR is that in general it will not represent identically the contributions of fragments that are structurally identical throughout a series, such as common cores. Figure 12 illustrates this issue. The main view is of a few docked structures from a combinatorial library, all having the common core $C(=O)NHCH(CH_2Ph)SO_2NHCH$, as can be seen somewhat more clearly within the single structure in the upper right corner. Evidently, the common core has been shifted around in the docked conformation, and these shifts must produce significant changes in the steric and electrostatic fields at lattice points nearby.

But in fact the only structural changes among these ligands, therefore the only true causes of any changes in biological activity, are among the side chains, far distant from the common core. During the PLS analysis of all the field changes, the relevant "signal" from the side chain changes will tend to be obscured by the "noise"[33] arising from the consequential shifts of the common core atoms.

When docking is not possible, it is common practice to select CoMFA conformations by minimization of the isolated ligand structures, or by a "field fitting" procedure of minimizing overall differences in the shapes of the surrounding fields. As in docking, however, neither minimization nor field-fitting does of itself attempt to preserve the 3D identity of invariant structural features such as common cores, and so the same noise-obscuring-signal tendencies will exist. For practitioners of "standard CoMFA", the major implication of this interpretation would be to ensure that any series-common fragments overlap exactly whenever any CoMFA alignment protocol has been completed.[17]

The topomer alignment procedures could be described as taking this general rule "align identical structural fragments identically" to the next stage of "and also try to align similar structural fragments similarly". The more that these objectives can be achieved, the relatively stronger will be the influence of the truly dissimilar features upon the matrixes of field differences that underlie a CoMFA analysis.

Several observers have also noted that the topomer definition rules tend to produce the same "fully extended" conformations that are often observed in the bound conformations of endogenous ligands. Perhaps any such conformational mimicry is not a complete coincidence, given that the ligand recognition mechanisms must themselves have evolved to maximize the selectivity of a biological response. On the other hand, the topomer overlays shown here (Figures 1 to 3) already suggest that resemblance of a set of topomers to any overlay of docked side chains will seldom be very great.

Which structural fragments actually are aligned by topomer CoMFA evidently depends on the only user-adjustable input, the choice of fragmentation bond(s). In principle, at least for Case 2 series lacking an extended common core, there are $a \times b \times c... \times n$ possible combinations of fragmentations to be considered (a, b, ... n being the number of acyclic single bonds in the first, second, and nth (last) structure in the series). In practice, the choices have so far been so obvious that users of this technology have typically been identifying the bond(s) to be broken in each structure automatically, by substructure search. The lone example of multiple fragmentations in the current study (comparing ICEc with ICEb) does not suggest unacceptable sensitivity of topomer CoMFA results to the fragmentation choice.

Topomer CoMFA operations are very fast, not really worth the bother of formal process time measurements. For example, on standard single processor SGI workstations, the topomer CoMFA analyses described here seldom required as much as a minute to finish, including 3D model generation, and the subsequent topomer CoMFA searches never failed to complete overnight. Of more practical and greater importance, the almost complete objectivity of the individual steps (excluding fragmentation, and that so far only in principle) also greatly simplifies and accelerates decision-making based on topomer CoMFA results.

As the central guide within an accelerated lead optimization project, the topomer CoMFA methodology[34] promises several significant advantages over alternative approaches. The clearest of these are its speed, convenience, and complete objectivity, and the enormity of the structural space that may be explored throughout the optimization phase. Also noteworthy is its consistent performance to date in providing a useful accuracy of potency forecasts, already over an unusually large range and number of data sets. It is further hoped that encouraging prospective applications of topomer CoMFA, within lead optimization collaborations currently underway, can in due course be fully reported. To strengthen support for "parallel track" lead optimization, efforts to incorporate ADME(T) considerations into topomer searching (beyond existing criteria based on the Lipinski Rule of 5 and "undesirable structural fragments") have also begun.

Acknowledgment. Michael Lawless and Bernd Wendt are particularly thanked for early application and valuable feedback in development of the topomer CoMFA technology, as are Robert Clark and certain of the referees for exceptionally thorough and constructive critiques of the manuscript.

References

(1) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(2) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3030.

(3) According to one fairly recent publication, "In the papers cited in SCI from 1989 to Dec. 7, 2000, the total number of papers with the keyword "CoMFA" are more than 5000.". Zhu, L. L.; Hou, T. J.; Chen, L. R.; Xu, X. J. 3D QSAR Analyses of Novel Tyrosine Kinase Inhibitors Based on Pharmacophore Alignment, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1032−1040.

(4) Kubinyi, H. *3D QSAR in Drug Design: Theory, Methods, and Applications*; ESCOM: Leiden, The Netherlands, 1993.

(5) Kubinyi, H.; Folkers, G.; Martin, Y. C. 3D QSAR in Drug Design: Recent Advances. *Perspect. Drug Discovery Des.* **1998**, *12/13/14*, 3−338.

(6) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis CoMFA): 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(7) Kubinyi, H.; Folkers, G.; Martin, Y. C. 3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 3−398.

(8) Bursi, R.; Grootenhuis, P. D. G. Comparative molecular field analysis and energy interaction studies of thrombin-inhibitor complexes. *J. Comput. Aided Mol. Des.* **1999**, *13*, 221−232.

(9) Sao, S.-S.; Karplus, M. Evaluation of designed ligands by a multiple screening method: Application to glycogen phosphorylase inhibitors constructed with a variety of approaches. *J. Comput. Aided Mol. Des.* **2001**, *15*, 613−647.

(10) Klebe, G.; Mietzner, T.; Weber, F. Different approaches toward an automatic structural alignment of drug molecules: Applications to sterol mimics, thrombin, and thermolysin inhibitors. *J. Comput. Aided Mol. Des.* **1994**, *8*, 751−778.

(11) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure−activity relationship of human immunodeficiency virus (I) protease inhibitors: 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37*, 2206−2215.

(12) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: A comparison of CoMFA models based on deduced and experimentally determined active-site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(13) Muegge, I.; Podolgar, B. L. 3D-Quantitative Structure Activity Relationships of Biphenyl Carboxylic Acid MMP-3 Inhibitors: Exploring Automated Docking as Alignment Methodol. *Quant. Struct.-Act. Relat.* **2001**, *20*, 215−222.

(14) Hopfinger, A. J.; Burke, B. J.; Dunn, W. J., III A generalized formalism for three-dimensional quantitative structure−activity relationship using tensor representation. *J. Med. Chem.* **1994**, *37*, 3768−3774.

(15) Other approaches exist that do promise direct 3D database searching with potency predictions, such as pseudo-receptor modeling, the Catalyst suite, and post-3D-searching CoMFA. However, these are limited in speed, in range of applicability, and in the number of supporting examples.

(16) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.

(17) A referee notes that CoMFA alignment by the overlap of common cores was independently advocated by Robert S. Pearlman, in connection with the combinatorial library 3D building program Combilibmaker. This approach would of course be applicable only to Case 1 data sets. Similar advice had also been given much earlier by the current author (Cramer, R. D., DePriest, S. A., Patterson, D. E., Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D-QSAR*. H. Kubinyi, Ed.; ESCOM: Leiden, 1993). However none of this previous advice on CoMFA alignments has been systematically evaluated.

(18) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *6*, 1010−1023.

(19) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer Similarity Searching of Conventional Databases. *J. Mol. Graph. Model.* **2002**, *20*, 447−462.

(20) Kulkarni, S. S.; Kulkarni, V. M. Three-Dimensional Quantitative Structure−Activity Relationship of Interleukin 1-b Converting Enzyme Inhibitors: A Comparative Molecular Field Analysis Study. *J. Med. Chem.* **1999**, *42*, 373−380.

(21) Bohm, M.; Sturtzebecher, 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.

(22) Gnerre, C.; Catto, M.; Leonetti, F.; Weber, P.; Carrupt, P.-A.; Altomare, C.; Carotti, A.; Testa, B. Inhibition of Monoamine Oxidase by Functionalized Coumarin Derivatives: Biological Activities, QSARs, and 3D-QSARs. *J. Med. Chem.* **2000**, *43*, 4747−4752.

(23) Schaal, W.; Karlsson, A.; Ahlsen, G.; Lindberg, J.; Andersson, H. O.; Danielson, U. H.; Classon, B.; Unge, T.; Samuelsson, B.; Hulten, J.; Hallberg, A.; Karlen, A. Synthesis and Comparative Molecular Field Analysis (CoMFA) of Symmetric and Nonsymmetric Cyclic Sulfamide HIV-1 Protease Inhibitors. *J. Med. Chem.* **2001**, *44*, 155−169.

(24) Rieger, J. M.; Brown, M. L.; Sullivan, G. W.; Linden, J.; Macdonald, T. L. Design, Synthesis, and Evaluation of Novel A_{2A} Adenosine Receptor Agonists. *J. Med. Chem.* **2001**, *44*, 531−539.

(25) Lanig, H.; Utz, W.; Gmeiner, P. Comparative Molecular Field Analysis of Dopamine D4 Receptor Antagonists Including 3-[4-(4-Chlorophenyl)piperazin-1-ylmethyl]pyrazolo[1,5-*a*]pyridine (FAUC 113), 3-[4-(4-Chlorophenyl)piperazin-1- ylmethyl]-1H-pyrrolo-[2,3-*b*]pyridine (L-745,870), and Clozapine. *J. Med. Chem.* **2001**, *44*, 1151−1157.

(26) Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.-F.; Wong, B. L.; Jiang, H.; Chen, K. 3D-QSAR Model of Flavonoids Binding at Benzodiazepine Site in $GABA_A$ Receptors. *J. Med. Chem.* **2001**, *44*, 1883−1891.

(27) Tetko, I. V.; Kovalishyn, V. V.; Livingstone, D. J. Volume Learning Algorithm Artifical Neural Networks for 3D QSAR Studies. *J. Med. Chem.* **2001**, *44*, 2411−2420.

(28) Sippl, W.; Contreras, J.-M.; Parrot, I.; Rival, Y. M.; Wermuth, C. G. Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395−410.

(29) Bureau, R.; Daveu, C.; Baglin, I.; Santos, J. S.-D.; Lancelot. J.-C.; Rault, S. Association of Two 3D QSAR Analyses. Application to the Study of Partial Agonist Serotonin-3 Ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 815−823.

(30) Hannongbua, S.; Nivesanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative Structure−Activity Relationships of HEPT Derivatives as HIV-1 Reverse Transcriptase Inhibitors, Based on Ab Initio Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848−855.

(31) The possible inversion of chiral atoms and the neglect of known stereoisomers, in the context of a shape similarity descriptor, need additional explanation. Every potentially chiral atom must in a 3D model necessarily possess an absolute but completely arbitrary (unless externally specified or energetically determined) chirality. Similar structures will have similar shapes only if these arbitrary chiralities are geometrically standardized by appropriate inversions. However, to avoid producing nonsensical ring geometries, the topomer protocol allows only acyclic chiral atoms to be inverted. Quite recently a protocol has been introduced for topomerically standardizing ring pucker as well, by choosing between the two equivalent reflections of such ring geometries. Why then are known stereoforms ignored? Whether a fragment structure is part of a query or a candidate fragment, because the far more numerous unassigned stereoforms will always have been standardized topomerically, any known stereocenter has a roughly 50% chance of being the nontopomeric stereoisomer. As a result, structurally identical fragments would be topomerically dissimilar and unrecognized 50% of the time. Faced with this very unattractive alternative outcome, it was agreed that known stereocenters would be structurally registered but ignored in topomer modeling. Racemic synthons thus become two distinct registered "substances" mapping to the same topomer.

(32) One such "quasi-Free-Wilson" QSAR may have been observed among other topomer CoMFA applications. There also appears to be a good diagnostic for such an artifactual correlation, a variance among the predicted or "test set" potencies that is much smaller than the variance of the "training set potencies" used to construct the QSAR model.

(33) Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137−145.

(34) All new technologies described are the subjects of international patent filings.

JM020194O