JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

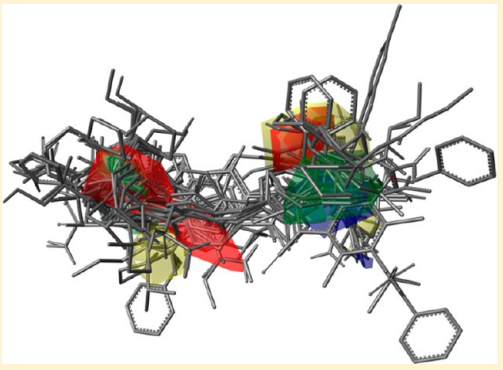
Template CoMFA: The 3D-QSAR Grail?

Richard D. Cramer* and Bernd Wendt

Tripos DE, 9666 Olive Blvd, St. Louis, Missouri 63132

Supporting Information

ABSTRACT: Template CoMFA, a novel alignment methodology for training or test set structures in 3D-QSAR, is introduced. Its two most significant advantages are its complete automation and its ability to derive a single combined model from multiple structural series affecting a biological target. Its only two inputs are one or more "template" structures having 3D coordinates that share some Cartesian space, as may result from X-ray crystallography or pharmacophoric hypothesis, and one or more connectivity-only SAR tables associated with a common target. Template CoMFA also overcomes the major disadvantages of both existing 3D-QSAR alignment methodologies, specifically the tedium and subjectivity of familiar ad hoc approaches, and the awkwardness, occasional physicochemical heresies, and structural scope limitations of the purely topomer approach. The template CoMFA algorithms are described, and two of its application classes are presented. The first class, general models of binding to factor Xa and P38 map kinase, uses crystallographic structures as templates, with the encouraging result that the statistical qualities of each of these two combined models are equivalent to those of their constituent individual series models. The second, 15 data sets originally collected for validation of topomer CoMFA, with arbitrary structures as templates, confirms that the modeling power of template CoMFA resembles that of its predecessors.



INTRODUCTION

Differences among ligand biological activities, which computational drug designers are expected to understand and predict, are primarily caused by spatially dependent differences among their possible noncovalent interactions with a biological target. Such spatially dependent noncovalent interaction differences may be approximated by differences in ligand fields, in particular steric and electrostatic, as sampled over a lattice or grid. These premises motivated the creation of 3D-QSAR methodologies, in particular CoMFA ("comparative molecular field analysis"),[1] and indeed, numerous publications, usually including successful prospective potency predictions,[2,3] continue to encourage use of this approach.

However, 3D-QSAR success requires informative alignments of ligands, meaning the prescription of a conformer and an orientation for each training or test structure. Such alignments are extremely tedious to generate, and more or less subjective. Usually, the only available guide is the cross-validated r^2, or q^2, of the resulting model, a noisy index whose value can depend as much on overall training set composition as on individual alignments. Thus, the most widely accepted "gold standard" for 3D-QSAR alignment of a ligand quickly became its experimental crystallographic bound structure, despite its high cost and its unavailability for candidate structures yet to be synthesized.

More recently, there has been an increasing suspicion that the most informative alignments of a set of ligands should focus only on the structural differences among those ligands.[4−6] Why? *Because the only intended difference among the inputs to a biological assay, and thus the only causative source of differences among observed potencies, are those structural differences, more exactly the effects of those differences on molecular field differences.*

One alignment approach that focuses the effect of structural differences onto molecular field differences is the topomer protocol.[7] Indeed "topomer CoMFA"[8] has proven to be a robust alignment methodology, for example yielding an average SDEP (standard error of prediction) over 144 structures prospectively synthesized by four different discovery organizations of 0.6 pIC50 units,[6,9,10] and is also exceptionally rapid and simple, and completely objective, to practice. Furthermore, the rapidity of topomer alignments enables "inverse 3D-QSAR," the capability of searching the largest databases for highly promising structures and R groups.[2]

Yet topomer CoMFA has unattractive characteristics. Its training and test set structures become sets of monovalent fragments, or "R groups," not complete structures, which are a bit clumsy in themselves and also require all structures to share a characteristic acyclic single bond, whose cleavage generates those sets of monovalent fragments. Therefore, training and test set scopes are limited to homologous series. Moreover, the conformation of any topomer will seldom agree with any whole-ligand X-ray coordinates and may even contain physicochemical heresies such as sterically overlapping atoms, reducing the credibility of the resulting CoMFA model in ways that the practitioner is powerless to influence.

Received: November 25, 2013
Published: January 19, 2014

   dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

This publication introduces a new 3D-QSAR alignment protocol, "template CoMFA."[11,12] Template CoMFA removes all of these unattractive characteristics of topomer CoMFA, while retaining all of its virtues. And, more importantly, template CoMFA provides two major new advantages. First, template CoMFA alignment is by default completely automatic. Its only inputs are one or more "templates," whose superimposed 3D atomic coordinates might have crystallographic, pharmacophoric, or simple energy minimization sources, and an SAR table, the familiar collection of "2D" (connectivity-only) structures paired with their observed potencies in some assay of interest. All the template CoMFA operations themselves—model development, visual inspection, potency prediction, and "inverse QSAR" searching—then require only simple "do it" buttons. The second, frequently even greater, advantage is that there are no intrinsic limitations on the number or variety of structures that a single template CoMFA model may describe. In particular, a single 3D-QSAR model describing multiple lead series affecting the same target can be produced simply by positioning each of the templates within the same coordinate system and merging the SAR tables. Thus template CoMFA's applicability exactly addresses perhaps the most common CADD hit-to-lead or lead optimization responsibilities within drug discovery projects—to use all available information, typically a few crystal structures or a pharmacophore model, accompanied by many more SAR observations, to rapidly identify the most promising structures to acquire and test next, based on a structurally interpretable model.

But first, how effective are template CoMFA models? This publication includes two studies exemplifying the above-mentioned advantages, the automatic application of template CoMFA, in that very common project situation with the few X-ray structures accompanied by SAR data representing multiple structural series. Each of these studies compares a collection of template CoMFA models for highly diverse templates and multiple structural series acting at a common target with their single "combined" template CoMFA model. A third study explores the applicability of template CoMFA when crystallographic or pharmacophoric data are unavailable and provides a benchmark for comparing the effectiveness of template CoMFA with its predecessors.

METHODS

The template CoMFA idea is easily summarized. *To generate the 3D conformation of any candidate structure, wherever template and candidate connectivities "match," directly "copy" the coordinates of the template atoms to the corresponding candidate atoms. To position the remaining "unmatched" candidate atoms, apply the topomer protocol.*

Template CoMFA also addresses a further alignment issue, the positioning of that 3D conformation. All the previous topomer-centric protocols, which manipulate only R-group fragments, position such a fragment by overlay of its "fragmentation bond" (open valence) onto a fixed Cartesian coordinate. For template CoMFA, this "fragmentation bond" becomes an "anchor bond," and the positioning of an anchor bond within a "candidate" (training or test set) structure is its overlay onto the "best-matching" bond within a template structure. In the automatic mode of template CoMFA, used for these studies, for each candidate, a pairing of anchor bonds is identified as that best matching any candidate bond with any bond in any of the templates. Alternatively, the user can manually specify the bond that must be "best-matched" within each of the templates. Thus template CoMFA positioning of a "candidate" (training set or test set) structure combines two consecutive steps: selection of the anchor bond within a candidate, including an associated "best mapping" of candidate atoms onto atoms of a template, followed by the actual template-guided positioning of that candidate's individual atoms.

Anchor Bond Selection. In the automatic mode, to reduce the combinatorial number of bond comparisons and the resulting computational time, in both template and candidate structures every potential anchor bond must first meet one of several criteria:

- one of the atoms defining the bond must not be carbon or
- the bond type must be double or triple or
- one of the atoms defining the bond must be in a ring and attached to at least three non-hydrogen atoms

Additional information is generated for every such bond, of two types. The first type of information locates that bond with respect to any ring network within the overall structure, the ring network being simplified with all its atoms as carbons and all its bonds as single, and including only cyclic bonds, bonds connecting ring systems, and, if the potential anchor bond is acyclic, extended to include the path connecting the bond to that ring network. The second type of information is localized to that bond and includes its type and cyclicality, the types of the two bond-defining atoms, and variably weighted counts of the heavy atoms directly attached to the defining atoms.

Automatic selection of the optimal anchor bond combinatorially compares every template bond meeting one of these criteria with every candidate bond meeting one of these criteria. The process for evaluating all these anchor bond pairings then has two stages. The first stage generates a scored list of potential anchor bond pairings by, on the one hand, recognizing the desirable identities between their ring network positionings and, on the other hand, summing undesirable differences between their various localized properties. Only bond pairings whose score exceeds a user-defined threshold become finalist pairings. If no such bond pairings exist, alignment of the candidate fails.

In the second stage, the following path-matching procedure is applied to each of the finalist anchor bond pairings. The best-matching anchor bond is the one generating the best (in these studies, the most numerous) mapping of candidate atoms to atoms within a specified template. The graph-matching algorithm used to identify such a best-mapping path of candidate atoms differs in several ways from those employed by familiar cheminformatics tools such as substructure searching and maximal common subgraph (MCS):

- *Atom matching is rooted.* Every mapped atom must be connected by at least one unbroken path of mapped atoms to the anchor bond.
- *Atom matching is carried out in two successive stages.* The first stage requires exact matches of atom and bond types, and the second extends the first atom mapping by relaxing the requirements for a match.
- *Atom matching is breadth-first rather than depth-first.* A breadth-first search within a chemical structure graph processes every atom that is separated by the same and smallest number of bonds from either of the atoms defining the anchor bond, or in the same "layer," before

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

processing any atom separated by a greater number of bonds, in the next "layer."

- *Hydrogen atoms are excluded.*

The key data structure manipulated during anchor bond selection is the "atom map," an array indexed by candidate atom ID, whose elements record specific mappings of those candidate atom IDs to template atom IDs, plus a template identifier and a status flag indicating this atom map's current "atom matching stage" (first, second, or finished). The result of anchor bond selection, the only information that is reported to the second step of candidate alignment, is simply the "best" such atom map found.

In more detail, the initial layer of this breadth-first atom-matching process identifies, within an atom map, each of the two connected atoms in a possible candidate anchor bond with one of the atoms defining a possible anchor bond in one of the templates. Each subsequent layer begins with a preprocessing, to identify all the potentially acceptable matches of atoms in the new layer of the current candidate to atoms in the new layer of the current template, considering atom and bond types and the connectivities of atoms in this layer to the matched atoms within the previous layer. The central process comprises three nested loops, first through all the still-active atom maps accumulated from processing previous layers, next through each of the previous-layer atoms within the candidate, and finally through all possible permutations of preprocessing-acceptable (new layer) candidate atoms attached to this current previous candidate atom mapped onto specific preprocessing-acceptable template atoms. During the innermost loop, every permutation over a previous-layer candidate atom that identifies at least one more acceptable new-layer candidate-to-template atom match is evaluated and, if acceptable, recorded. (Acceptability criteria include ring closure requirements and the handling of new-layer atom counts that differ between candidate and template.) When processing of all active previous-layer atoms within an atom map (the middle loop) terminates, atom maps are generated (or updated) for every possible combination of the newly recorded previous-layer permutations. If no new permutations have been recorded for an atom map, one of its two "stages" finishes. A finished first stage map becomes second stage, and with the resulting relaxation of atom-matching constraints, all of its preceding layers are immediately reviewed for any newly acceptable atom matches. A finished second stage atom map becomes inactive.

A snapshot intended to illustrate these processes is given, starting in Figure 1. The anchor bond pairing being evaluated is between the orange bonds, specifically connecting atom 5 in the candidate to 1 in the template and 8 in the candidate to 5 in the template. The first layer atoms, colored cyan, 1, 9, and 24 in the candidate and 20 and 6 in the template, have already been processed and thus are the current "previous layer," yielding two active and "first stage" atom maps, the first matching candidate atoms 9 and 24 with template atoms 20 and 6, respectively, and the second the other permutation, matching 9 and 24 with template atoms 6 and 20, respectively. The second layer atoms, colored magenta, are the "new layer" currently being processed. During preprocessing, it is recognized that candidate atoms 15 and 22 and template atoms 7 and 15 are each connected to the same first layer atom and so must be permuted in every candidate-to-template comparison in which they are considered. Also, a permutation that includes a "match" of a candidate atom to no atom in the template is valid if the permutation also includes a conventional atom-to-atom match, as in this illustration happens during processing of the second of the two "first stage" atom maps, when the two attachments to candidate atom 24 are matched against the single attachment to template atom 20. When this second layer processing ends, from the discussion so far, six distinctive atom maps would be anticipated. These are shown in Figure 2, with double-headed arrows indicating the matches between second-layer atoms. However, a further practical constraint, that all newly defined dihedrals within a candidate ring be cyclically compatible with the correspondingly mapped dihedrals in a template ring,[11] immediately discards all but the first and last of these six maps.

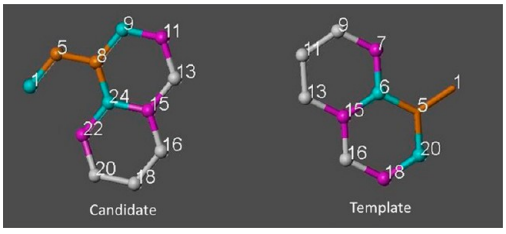


Figure 1. Starting position for generation of the second layer in a trial mapping of atoms in a candidate structure to atoms in a template structure. The anchor bonds whose pairing is being evaluated are orange. See the text for explanation.

The entire atom-matching process is completed when all atom maps have become inactive. The single highest-scoring atom map is returned to the caller, where in these studies this score is simply the number of matching atoms, with first stage matching atoms counted twice.

Template-Driven Candidate Alignment. The conformation of any candidate originates from application of a program such as Concord or Corina to its connection table, fixing in particular its bond lengths, valence angles, and ring geometries. The selected anchor bond of the candidate is superimposed on the selected anchor bond of the selected template. Template driven alignment is then a depth-first connectivity traversal of the candidate, starting from one of the atoms defining the selected anchor bond and handling each ring system as a unit when its first cyclic atom is encountered. At any branch point, attached atoms are processed in a hierarchical order, cyclic atoms before noncyclic atoms surrounding mapped atoms before unmapped atoms. Then, considering candidate acyclic atoms first, if an acyclic atom is mapped, its coordinates are usually just copied from the matching template atom. An exception arises if candidate and template atom types differ; the candidate bond length and valence angle are then preserved, and only the dihedral is taken from the template. Whenever an acyclic atom is instead unmapped, it and all its more distant (and necessarily unmapped) atoms are repositioned as a group, by transforming the coordinates of all these grouped atoms to produce a conformationally appropriate bond between unmapped and mapped atoms. Wherever there is more than one acyclic unmapped atom attached to a mapped atom, the unmapped atoms are treated in order of decreasing group size.

Turning to cyclic atoms, when a candidate cyclic system is first encountered and the new atom is mapped, and if the mapped template atom is also the first encountered in a ring system, the two ring systems are compared in their entirety. If the ring systems are identical (and therefore completely mapped), all the template ring atom coordinates are copied to the candidate. Otherwise, the ring system atoms are only

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

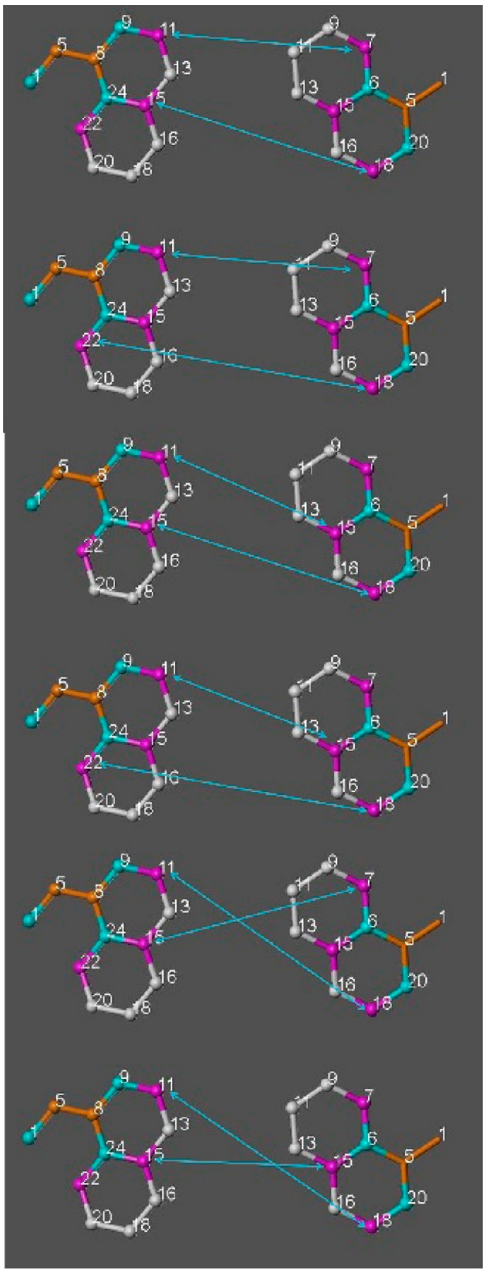


Figure 2. The six second layer mappings resulting from the step initiated in Figure 1. See the text for explanation.

partially mapped, and because differences among the unmapped ring system atoms may seriously distort the relative geometry of the mapped ring system atoms, the ring system will be copied as though entirely unmapped.

The coordinates of hydrogen atoms are also usually simply copied from the template, with the exception arising for any atom having counts of attached hydrogens that differ between template and candidate, in which case the hydrogens are positioned using standard rules.

Topomer Alignment Finishing Protocol. The overall goal of objectively generating candidate conformations requires a deterministic process in every step. Topomer alignments, originally developed as a diversity metric for the design of combinatorial screening libraries,[13] with the explicit goal of localizing steric field differences to variations in molecular connectivity, were soon found to be effective for predicting biological similarities[15] and later for generating 3D-QSAR alignments.[8] These properties well suit the topomer protocol for the final disposition of the nonmatching atoms in a template CoMFA candidate alignment. To do so, the topomer protocol is applied to each of the above-mentioned "unmapped groups," adjusting as encountered each acyclic dihedral angle, each indeterminate stereocenter, and each ring flip, as detailed in a previous publication.[7]

Template CoMFA Applications. The two target classes which on 6/10/2013 had the most entries at the URL http://www.bindingdb.org/validation_sets/index.jsp, as "affinities" to factor Xa and map kinase P38 alpha, provided the training sets used for the first two of the three studies. Their ligand templates with a shared frame of reference were obtained from the downloaded .pdb structures by least-squares superimposition of five alpha carbons followed by deletion of all nonligand atoms. The SAR data were the "2D" .sdf files downloaded from the same bindingdb.org URL. These ligand template overlays and their "2D" connectivities appear in Figure 3. The 3D alignments for these connection tables were derived following the algorithms described above, with completely automatic anchor bond identification, by an unreleased version of SYBYL-X 2.1 that included template CoMFA as an implementation written by the senior author, mostly in the C language. The requisite CoMFA lattice or "region" is automatically generated as a 2 Å grid whose *X*, *Y*, and *Z* coordinates extend 4 Å beyond the maximum and minimum *X*, *Y*, and *Z* among all atomic coordinates. The number of model components that minimized the lowest leave-one-out cross-validation standard error of prediction (SDEP) was chosen for the final model. For each of the two targets, models were derived individually, for each of the templates with their associated SAR tables, and collectively, by combining all their templates and SAR data into single tables.

The third study applied template CoMFA to the 15 data sets used to initially validate topomer CoMFA.[8] Its workflows are the same as those for the first two studies, except for the initial selections of templates. With no crystallographic or pharmacophoric information as guides for most of the sets, the selections of their template structures are arbitrary, and more or less subjective. To maintain objectivity, a selection rule was invented. We simply chose the MMFF-minimized Concord model of the highest molecular weight structure in a data set as its template. Then, recognizing the possibility that such a biggest structure might happen to be a rather unrepresentative template for many of the other structures, we added a second template, the lowest molecular weight structure in the template-aligned conformation that had resulted from the previous largest-structure run, and repeated the analyses. Each analysis included both template(s) in its training set.

RESULTS

The statistical parameters for the factor Xa and the P38 map kinase models are given in Tables 1 and 2, with the individual models labeled by the PDB identifier for their template source, and the collective model at the bottom. From left to right, the columns contain the counts of unprocessed structures, the training set count, the standard deviation of the measured bioactivities, the leave-one-out cross validation q^2 and SDEP, the number of components that minimized SDEP, and the conventional, r^2 and s for the final model.

The most remarkable feature in both of these tables is the contrast between the satisfactory q^2 for the combined model

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

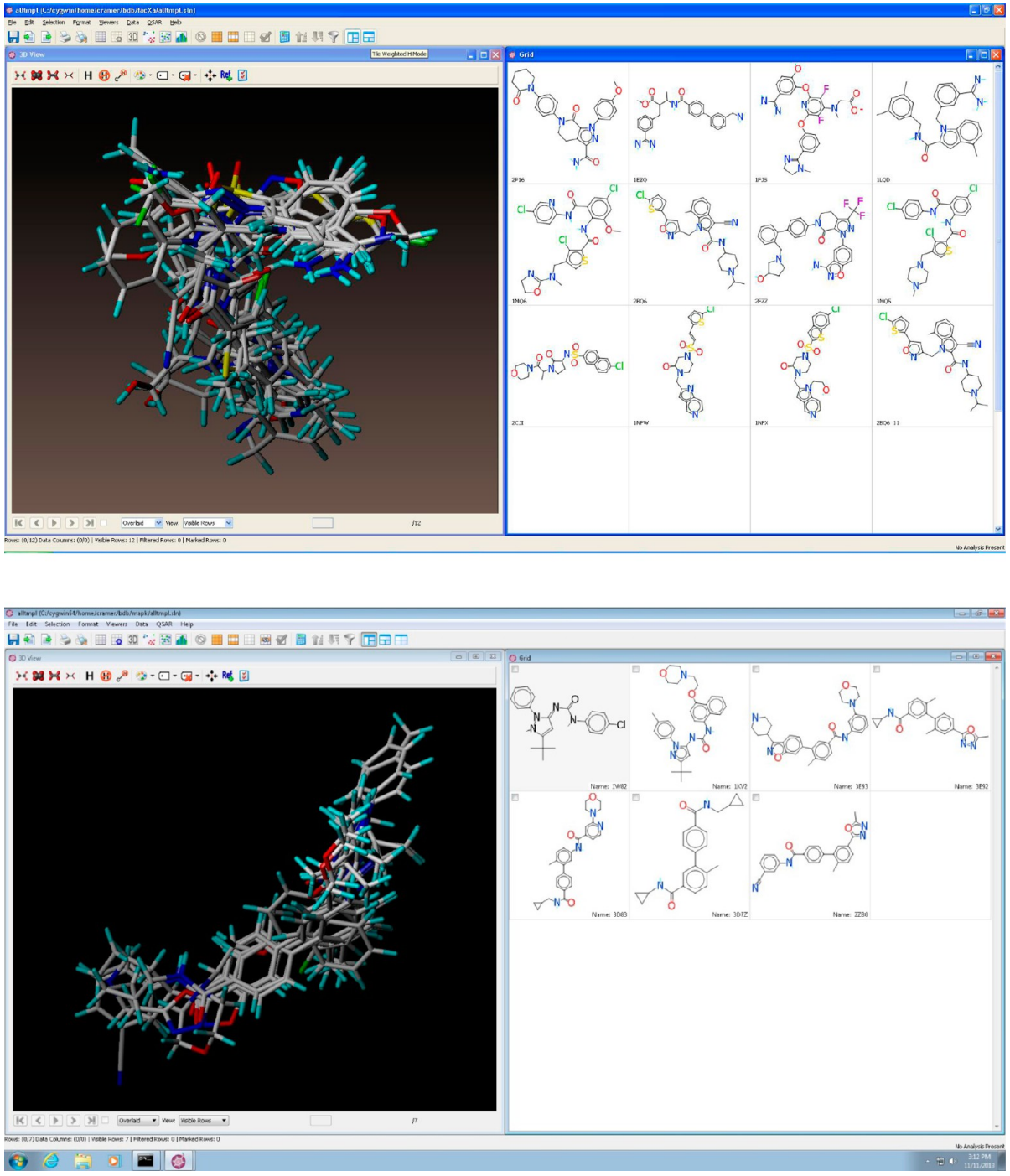


Figure 3. Overlaid crystallographic structures and "2D" connectivities of the templates used to build the combined template CoMFA models. The top panel contains the factor Xa templates; the bottom, p38.

and the mostly unacceptable q^2 for the individual models. However, notice should also be taken of another trend for most of the individual models, the contrast of these unacceptable q^2 with the good quality of their best-fit statistics (i.e., models with few components yielding high r^2).

To provide supporting visual impressions, several of the individual models are depicted, in Figure 4 for factor Xa and Figure 5 for p38, each labeled with the ID of the .pdb that provided its template. Each image pair shows the stack of automatically aligned training set structures, surrounded by the familiar steric (left) or electrostatic (right) sdev*coeff model contours, scaled so that the "blobs" in total enclose 75% of the final (not the cross-validated) model-correlated potency variance. Steric increase is favorable in green blobs and disfavored within yellow blobs; negative (positive) charge is favored within red (blue) blobs. For each target, the viewing directions are identical, but the viewing locations and image magnifications vary slightly.

Evidently, these automatic template CoMFA alignments do consistently limit the 3D structural variation within a training set to its connectivity variation, with the locations of structurally identical moieties within each training set being perfectly superimposed. Presumably, the field variations (not displayed) are also greatest in the regions surrounding the

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

Table 1. Factor Xa Model Statistics:

				LOO cross-valid			best fit	
PDB entry	# out	n	sdev	q^2	SDEP	# cmp	r^2	s
1EZQ	7	26	0.64	0.023	0.71	2	0.756	0.35
1FJS	0	21	1.56	0.628	1.00	2	0.888	0.55
1LQD	0	27	0.90	−0.585	1.19	2	0.379	0.74
1MQ5	0	22	0.55	0.254	0.50	2	0.697	0.32
1MQ6	0	46	0.79	0.474	0.60	4	0.876	0.29
1NFW	1	8	0.99	−1.270	1.80	2	0.505	0.84
1NFX	1	15	1.06	−0.843	1.62	2	0.507	0.84
2BQ6	0	30	0.96	0.655	0.61	4	0.966	0.19
2CJI	6	35	0.89	−0.027	0.93	2	0.562	0.61
2FZZ	0	21	0.50	−0.239	0.59	2	0.633	0.32
2P16	0	19	1.06	0.632	0.70	3	0.901	0.37
combined	15	270	1.30	0.570	0.87	6	0.757	0.66

Table 2. P38 Map Kinase Model Statistics:

				LOO cross-valid			best fit	
PDB entry	# out	n	sdev	q^2	SDEP	# cmp	r^2	s
1KV2	0	26	1.06	0.024	1.29	5	0.962	0.25
1W82	0	14	1.17	0.274	1.01	2	0.796	0.53
2ZB0	0	26	0.79	0.579	0.59	2	0.907	0.28
3D7Z	0	13	0.76	−0.649	1.12	2	0.543	0.59
3D83	0	15	0.78	−0.282	0.95	2	0.735	0.43
3E92	0	25	0.81	−0.004	0.85	2	0.691	0.47
3E93	0	12	0.76	0.206	0.75	2	0.828	0.35
combined	0	131	1.41	0.558	0.96	5	0.840	0.58

locations of 3D structural variation, explaining the strong tendency for the contours of the models (displayed) to also surround the locations of 3D structural variation. More interesting, then, are the contours that do not surround locations of 3D structural variation and greatest field variation (for example, .pdb's 2BQ6 and 3D72). These most probably result from a more diffuse 3D-QSAR, reporting on the one hand that, say, steric bulk here is generally desirable but on the other that, closer up, the steric fields of individual training set structures as currently aligned have contradictory effects on potency.

Visual impressions of the two combined models appear in Figures 6−9 for factor Xa and Figures 10−12 for p38. All are based on CoMSIA field assemblies, which explicitly separate hydrogen-bonding from electrostatics and hydrophobicity from sterics and whose potential forms lack discontinuities. Table 3 shows that the model statistics for these CoMSIA and CoMFA field assemblies are indistinguishable. As usual, the contours displayed in these figures represent only the steric and electrostatic effects. However, CoMSIA contours tend to be more precisely focused, which aids the comparison of contour maps with protein cavity structures.

These visual impressions illustrate the structure−activity relationships of the combined series of factor Xa inhibitors and p38 inhibitors. The crucial interactions as highlighted by the CoMSIA contours are in full agreement with typical binding modes of inhibitors of these two targets and are thus representative visualizations of their whole sets. For factor Xa, the combined templates and training sets include classical inhibitors with a basic group occupying the S1 pocket as well as novel inhibitors with an arylhalide anchor. The contours in the S1-pocket displayed in Figure 8 reveal the positions of the two alternative anchoring groups for factor Xa inhibitors. The blue contour in the ortho position marks the region of the basic group of classical factor Xa inhibitors with the amidine-nitrogens interacting with Asp189. The red contour in the para position of the aryl ring marks the region for the arylhalide interacting with Tyr228. Red contours at the other ortho position correspond to a preferred region for heteroatoms, potentially acting as hydrogen-bond acceptors, as shown by the phenolic group interacting with Ser195.

In the front view of the factor Xa pocket (Figure 9), green CoMSIA contours indicate favored steric interactions with the S3 pocket, where typically heterocyclic side chains (red contours) are sandwiched between residues Tyr99 and Phe174. The yellow contour at Phe174 corresponds to disfavored steric interactions of side chains with either too large or nonplanar geometries. The S2 pocket is rather narrow in factor Xa as reflected by the yellow contour between residues His57, Gln192, and Ser195 showing that this region should be unoccupied. The red contour in this position highlights the position of a heteroatom that potentially builds a hydrogen bond with a crystal water in this position. Green and blue contours at the entrance to the S1 pocket correspond with a region that should be occupied by hydrogen-bond donating groups.

For p38, the combined series include type-I inhibitors binding to the active form of the kinase as well as type-II inhibitors binding to the inactive form, where these two forms are distinguished by movement of the DGF motif from an inward (DFG-in) to an outward (DFG-out) position. The SAR effects of the invariant regions of these two binding modes can be observed in the CoMSIA model (Figure 12) as the green and yellow contours on the left-hand side of the pocket indicating the pocket boundaries, and the red contour corresponding to key interactions with the hinge region

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

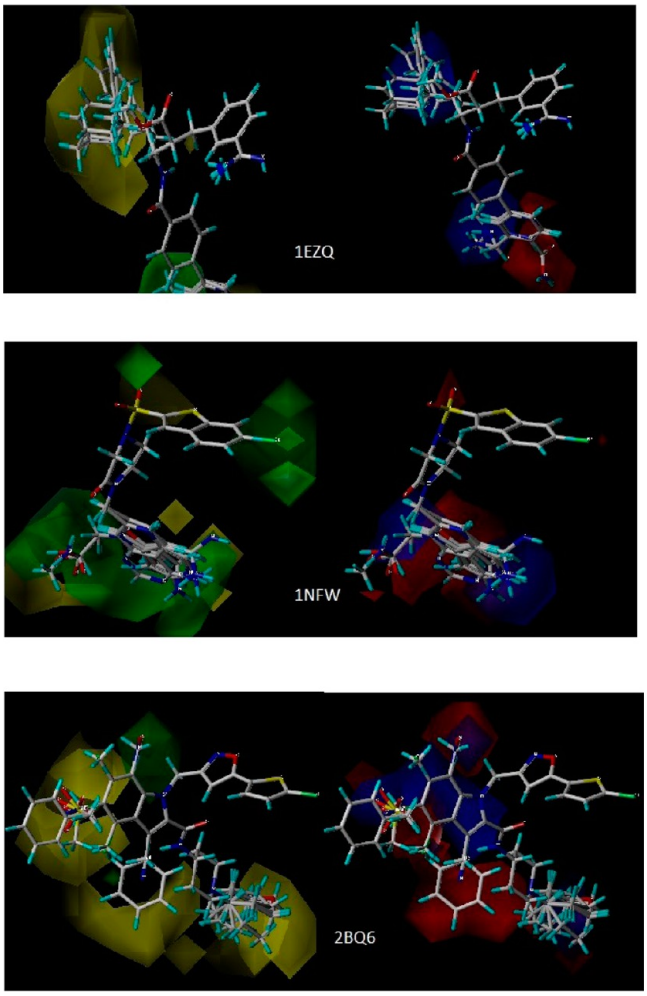


Figure 4. Aligned training set and model contours from three individual template CoMFA models of factor Xa.

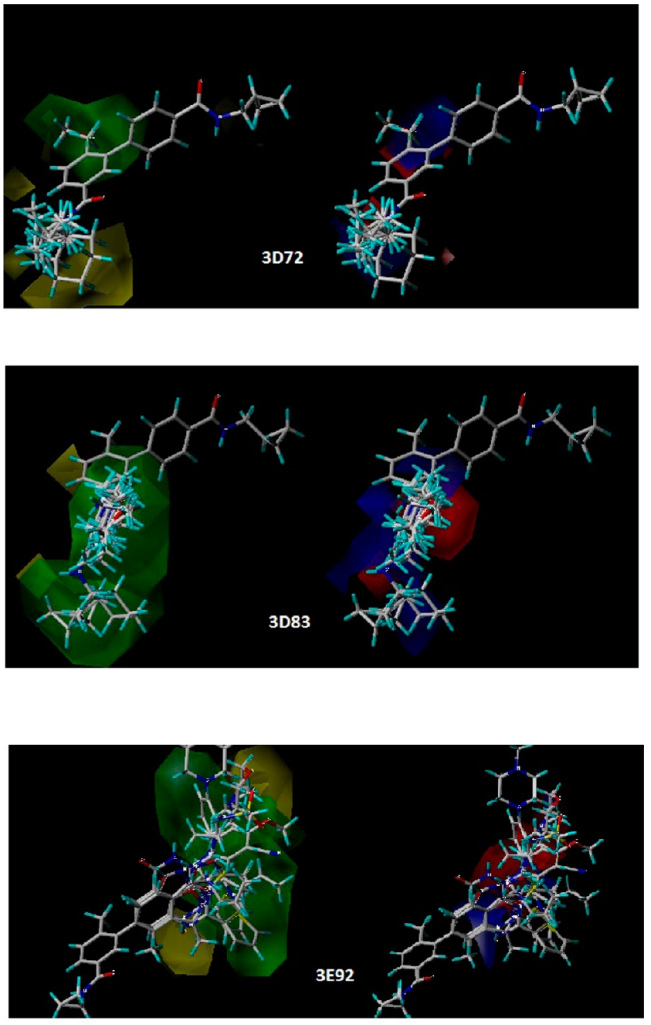


Figure 5. Aligned training set and model contours from three individual template CoMFA models of p38.

marking the positions of the hydrogen bond to Met109. The SAR effects of the variant binding modes appear with the green and yellow contours on the right-hand side: the yellow contour corresponds to the pocket boundary of the active form of p38 with Phe169 in the DGF-in position (left picture) and the green contour indicating steric space to be occupied when Phe169 moved into the DFG-out conformation (p38-inactive form, shown on the right picture). There is another blue contour around the linker group and another red contour in the hydrophobic pocket indicating the preferred nature of the atoms forming these groups: hydrocarbons and heteroaryls, respectively.

The results from applying template CoMFA to the 15 data sets originally used to validate topomer CoMFA appear in Table 4. Four alignment approaches are compared, with a block of four columns for each statistical metric: whatever method the original publication used, the pure topomer method, and template CoMFA, using as template(s) either the largest structure or the largest and the smallest structure. The statistical metrics are the leave-one-out cross-validated q^2, the associated SDEP, the number of components that minimized SDEP, and the SDEP of test set predictions whenever such a test set was provided in the original publication. The two columns labeled "# count" contain the test and training set counts. The "# out" column provides the count of unaligned training set structures,

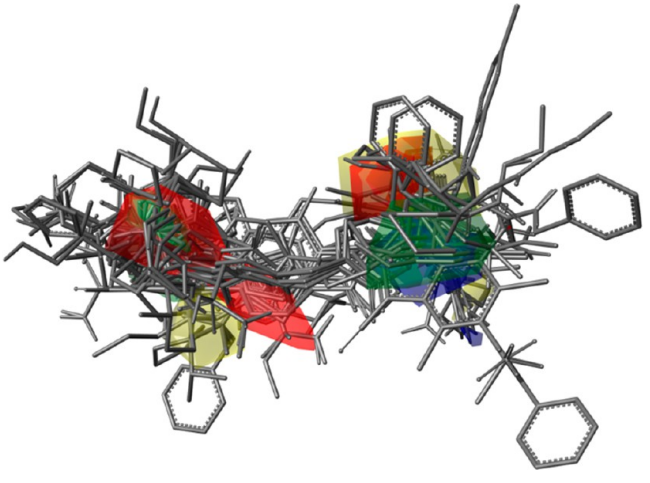


Figure 6. All 270 factor Xa inhibitors in their automatic template alignments with CoMSIA contours from the combined model.

with the usual reason being Concord failure. The "# w sml" column needs further description. When both largest and smallest structures become templates, either can provide the "best matching" anchor bond for a particular training set structure. This column reports the count of a training set's

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

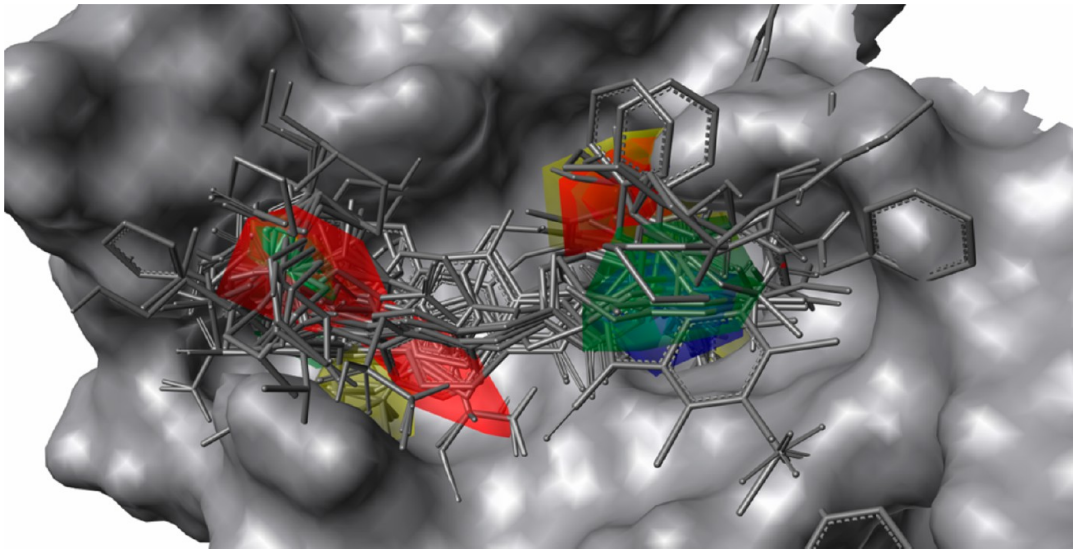


Figure 7. All 270 factor Xa inhibitors in their automatic template alignments shown inside the factor Xa binding pocket (PDB: 1MQ6) with CoMSIA contours from the combined model.

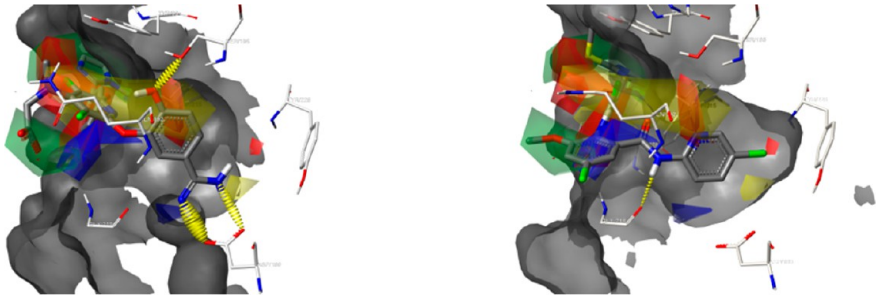


Figure 8. The S1-pocket of factor Xa (PDB: 1MQ6 [left], 1FJS [right]) with contours from the combined CoMSIA model.

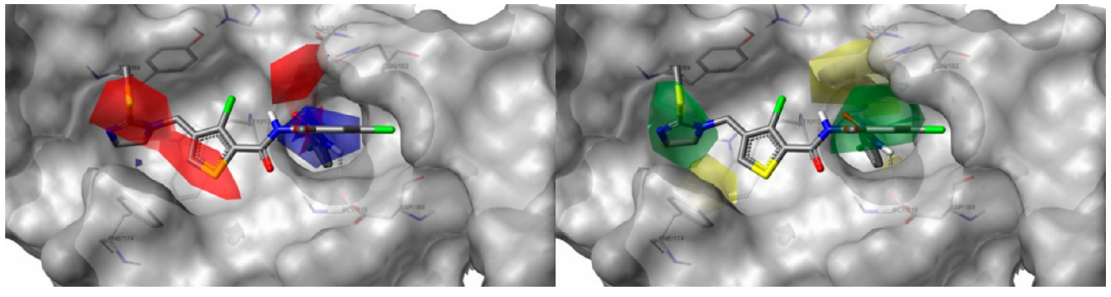


Figure 9. Front view of the factor Xa binding pocket (PDB: 1MQ6) with steric contours on the left and electrostatic contours on the right from the combined CoMSIA model.

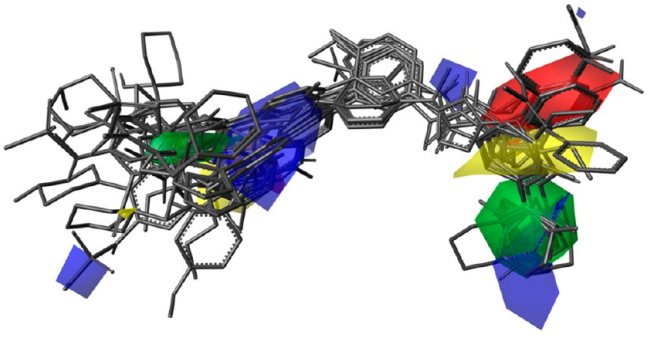


Figure 10. All 270 p38 inhibitors in their automatic template alignments with CoMSIA contours from the combined model.

structures that better matched the smallest structure; the count of better matches to the larger structure is then of course the difference between "# count" and "# w sml" values.

The averages of these statistical indices and the totals of the counts, in the bottom row of Table 4, indicate that in general the modeling power of template CoMFA using arbitrary templates differs little from that of topomer CoMFA. The moderately superior indices for the originally published 3D-QSAR models ("lit.") can reasonably be attributed to the possible alignment tunings of individual structures. Comparing the one- and two-template CoMFA models, the numerous near-zero values in the "# w sml" column support the expectation that larger templates tend to be more effective. Indeed, adding the second template yielded a few models with

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

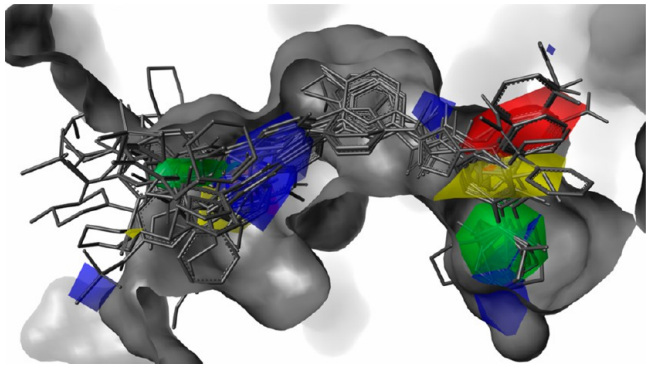


Figure 11. All 131 p38 inhibitors in their automatic template alignments shown inside the p38 binding pocket (PDB: 1KV2) with CoMSIA contours from the combined model.

somewhat weaker indices, because the additional alignment possibilities that might then be evaluated as "better" could decrease the focus on connectivity similarity within the training sets.

DISCUSSION

Desirable as the goal of a single robust QSAR model for all structures interacting in the same way with a biological target has always been, template CoMFA seems to be the first CADD methodology that systematically addresses that goal, and, as evidenced by the combined models for factor Xa and P38 map kinase, is succeeding. This is the most important finding from these studies. The resulting template CoMFA contour maps seem readily interpretable, in terms of the structural features of the binding site. Also encouraging is the near equivalence, in average statistical quality, of the 15 3D-QSAR models generated from the arbitrary templates to that of earlier 3D-QSAR models from those data sets.

In general, QSAR models derived from structurally more diverse training sets are expected to be less accurate and reliable than those from self-similar training sets. This expectation makes even more surprising the contrast in 3D-QSAR q^2 values, for the crystallographic template studies, between the individual (global) and the (local) combined training sets. Although it is tempting to stop discussion here and perhaps leave an impression that template CoMFA has romantically mystical (though then perhaps suspicious) properties, actually two factors unrelated to template CoMFA seem primarily responsible for this contrast. The first is that the variation in potency, appearing in the sdev columns of Tables 1 and 2, is greater for the combined training set than for most of the individual model training sets. Such a greater variation silently

Table 3. Statistics for the CoMFA and CoMSIA Combined Models

			LOO cross-valid			best fit	
method	# out	n	q^2	SDEP	# cmp	r^2	s
			factor Xa				
CoMFA	13	257	0.593	0.86	10	0.863	0.50
CoMSIA	13	257	0.571	0.88	6	0.718	0.71
			P38 map kinase				
CoMFA	1	130	0.564	0.80	6	0.830	0.50
CoMSIA	1	130	0.526	0.83	7	0.821	0.51

biases the q^2 of a model upward. In general, a more informative metric is the key component of the q^2 ratio, the standard deviation of the error of prediction (SDEP) during cross-validation, appearing in the SDEP columns of Tables 1 and 2. When the tabulated SDEPs of the individual models for each of these targets are averaged, weighting each model's contribution by its training set count (n), the resulting values (0.96 for P38, 0.89 for factor Xa) are almost identical numerically to the SDEPs of each combined model.

A possible second contributing factor is suggested by the remarkably large differences between the cross-validated and best-fit statistical metrics, for almost all of the individual models. If the cross-validation criterion is set aside, the high r^2 from very few components for most of the individual best-fit models attests instead to exceptionally high quality. QSAR cross-validation indices suffer whenever the potency of a single unusual structure greatly differs from the training set's mean potency. To yield acceptable cross-validation results, a training set must include a replicate for any unusual SAR observation. (Whether such caution should still be requisite today for 3D-QSAR, now that many thousands of publication attest to its robustness, is perhaps debatable.) However, the individual sets posted at www.bindingdb.org are often subsets extracted from the originally published SAR tables, and the purpose for which these sets were extracted, validation of high quality and so costly docking energy calculations, would seem to favor set compositions that would challenge cross-validation. For, if the purpose is to evaluate how well an FEP calculation handles an unusual binding configuration, a single example of that configuration should suffice.

In reality, then, the two combined models seem equivalent in statistical quality to the individual models, not superior. However this equivalence is highly encouraging. It might have been expected that the field and potency effects of structural variation on an individual lattice point would often be contradictory, depending on what template the structural variation was attached to. To the extent that such contra-

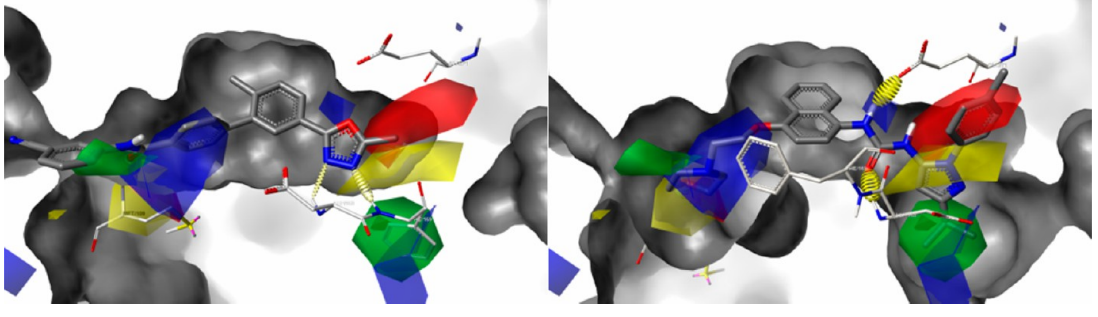


Figure 12. Binding pocket of p38 Map-kinase (PDB: 2ZB0 [left], 1KV2 [right]) with CoMSIA contours.

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

Table 4. Statistical Parameters of Model Derivation and the External Prediction Errors, for 15 3D QSAR Literature Studies and Their Repetitions with Three Other CoMFA Alignment Methodologies (See Text for Details)

			CoMFA model construction													CoMFA prediction				
					LOO x-val q^2				LOO x-val SDEP				# components				SDEP (RMS, prediction error)				
data set						Tpmr	Tmplte CfA			Tpmr	Tmplte CFA		Tpm	Tmplte CFA					Tpmr	Tmplte CFA	
ID	name	# cpds	# out	#w. sml	lit[b]	only	1 big	+1sml	lit[b]	only	1big	+1sml	lit	on	1	+1	# cpds	lit[b]	only	1 big	+1 sm
1	ICEc	36	0	0	0.630	0.380	0.361	0.361	0.82	0.96	0.97	0.97	6	3	3	3	9	0.57	0.77	0.70	0.55
2	ICEb	38	1	1	0.630	0.481	0.331	0.324	0.82	0.94	1.01	1.02	6	5	3	3	10	0.55	0.60	0.70	0.88
3	thrombir	72	0	1	0.687	0.498	0.598	0.598	0.59	0.75	0.69	0.69	4	3	7	7	16	0.67	0.72	0.94	0.84
4	trypsin	72	0	1	0.629	0.667	0.755	0.755	0.56	0.52	0.47	0.47	5	4	11	11	16	0.52	0.43	0.67	0.77
5	factor Xa	72	0	1	0.374	0.255	0.306	0.304	0.52	0.57	0.54	0.55	3	4	3	3	16	0.28	0.33	0.44	0.39
6	MAOa	71	0	3	0.440	0.597	0.445	0.450	1.03	0.89	1.03	1.03	2	3	2	2					
7	MAOb	71	0	3	0.430	0.597	0.445	0.474	1.25	0.89	1.28	1.31	2	3	2	8					
8	hiv	25	1	1	0.680	0.389	0.432	0.491	0.57	0.85	0.82	0.77	3	3	4	4	7	0.82	1.12	0.69	0.60
9	a2a	78	0	71	0.541	0.323	0.358	0.457	0.56	0.69	0.68	0.62	4	3	4	2	23	0.67	0.69	0.73	0.60
#	d4	29	0	2	0.739	0.710	0.767	0.767	0.73	0.71	0.67	0.67	7	5	6	6					
#	flav	38	0	5	0.752	0.777	0.779	0.779	0.48	0.47	0.47	0.47	4	3	4	4	4	0.34	0.61	1.44	1.22
#	cannab	61	0	25	0.592	0.439	0.583	0.465	0.57	0.69	0.60	0.69	4	3	4	6	6	0.45	0.66	0.94	0.75
#	ACEest	41	1	13	0.937	0.842	0.712	0.692	0.35	0.57	0.75	0.79	4	3	2	2	7	0.41	0.47	0.64	0.60
#	5ht3	61	6	21	0.645	0.423	0.262	0.442	1.19	1.68	1.95	1.76	5	5	4	8					
#	rvtrans	82	0	4	0.837	0.841	0.834	0.824	0.57	0.57	0.59	0.61	4	6	6	8	19	0.79	0.60	0.69	0.84
	total/avg	847	9	152	0.636	0.548	0.531	0.546	0.71	0.78	0.83	0.83	4	4	4	5	133	0.55	0.64	0.79	0.75

[a]For ICEc, ICEb, 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, a2a, flav, 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.

dx.doi.org/10.1021/ci400696v | J. Chem. Inf. Model. 2014, 54, 660−671

dictions occur, the statistical quality, interpretability, and potential predictive utility of a combined model would be degraded. However no such degradation occurred. The combined and individual models are generally reporting the same potency responses to field variations at individual lattice points, regardless of variation among the structures deploying those field variations. While further studies are obviously desirable, these data sets do represent the two largest and most diverse examples of multiple structural series with crystallographic examples bound to the same target, available on 6/10/2013. So it seems reasonable to suggest that 3D-QSAR models derived by template CoMFA from combined crystallographic observations and training sets can be trustworthy guides to decision making in similar project situations.

How reliable are the "true predictions" made from template CoMFA models? The average SE of 0.75 for the 133 "true predictions" directly reported herein, those associated with the 15 data sets having arbitrary templates, seems entirely satisfactory. A more interesting challenge is the thousands of factor Xa and P38 map kinase binding data points in public databases such as ChEMBL. To what degree can template CoMFA predictions based on the limited number of crystallographic templates, perhaps extended by pharmacophoric methods, be performed for such a variety of structures? Can a reliable prediction be meaningfully distinguished from an unreliable one? Consistent success in such "off-target prediction" would provide an extraordinary benefit to all of drug discovery, so the challenge is tempting.

It may be advantageous to perform both single series and combined template CoMFA models, particularly when 3D-QSAR goals include interpretation as well as prediction. Differences in contours between single series and combined template CoMFA analyses may also be instructive.

What about the other distinctive property possessed by template CoMFA alignment, its automation? Perhaps surprisingly, many CADD practitioners regard automated methodologies with suspicion, as a liability rather than a benefit. Certainly, considering the complexity of biological molecular interactions, it is desirable to seek experimental structural confirmation of an automatically derived model, and as the factor Xa and map kinase examples suggest, template CoMFA by its nature supports this activity. On the other hand, to prefer manual to automatic protocols for input operations, the 3D-QSAR alignments themselves, would seem to favor subjectivity over objectivity, a type of bias whose risk is obvious. Also the "big data" SAR arrays increasingly being generated by discovery projects can of course make impossible any 3D-QSAR effort entirely dependent on manual alignments.

Automatic alignment also has virtues beyond mere convenience that may not always be recognized. For example, the simple decisions of which structures to work on next involve cost–benefit trade-offs, between difficulty of synthesis and therapeutic (or informational) potential. Might such decision making be more effective if such costs and benefits could be simultaneously compared by the same individual (the synthetic chemist), fully informed by models automatically generated from the latest and most complete SAR and crystallographic data? The ability to include a vastly greater extent of chemistry space in such decision making,[15] also conferred by automatic alignment, would also seem highly beneficial. And for the CADD specialist, the simplicity of generating individual 3D-QSAR models, provided by automatic alignment, allows exploration of broader and probably more important methodological issues. For example, the automatic alignment afforded by (pure) topomer CoMFA has already made obvious the decisive importance of training set composition for success in 3D-QSAR modeling.[16−18]

Is template CoMFA the 3D-QSAR grail? Perhaps not, but the many significant advances it offers seem to justify its nomination.

■ ASSOCIATED CONTENT

Supporting Information

Includes all of the template CoMFA aligned structures referenced in the studies described above, in .sdf format. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: (505) 982-1055. E-mail: cramer@tripos.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The datasets previously assembled at www.bindingdb.org by Michael Gilson and associates were critical for enabling these and extraordinarily extensive validation studies.

■ REFERENCES

(1) 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*, 5939−5967.

(2) Cramer, R. D.; Cruz, P.; Stahl, G.; Curtiss, W. C.; Campbell, B.; Masek, B. B.; Soltanshahi, F. Virtual Screening for R-Groups, including Predicted pIC50 Contributions, within Large Structural Databases, Using Topomer CoMFA. *J. Chem. Inf. Model.* **2008**, *48*, 2180−2195.

(3) Doweyko, A. *J. Comput.-Aided Drug Des.* **2007**, *18*, 587−596.

(4) Cramer, R. D. Lead Hopping − and Beyond. *Expert Opin. Drug Discovery* **2006**, *1*, 311−321.

(5) Clark, R. D. A ligand's-eye view of protein binding. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 507−521.

(6) Cramer, R. D. Rethinking 3D-QSAR. *J. Comput.-Aided Drug Des.* **2011**, *25*, 197−201.

(7) Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for their Self-Consistent Generation. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1221−1227.

(8) Cramer, R. D. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003**, *46*, 374−389.

(9) Wendt, B.; Mulbaier, M.; Wawro, S.; Schultes, C.; Alonso, J.; Janssen, B.; Lewis, J. Toluidinesulfonamide Hypoxia-Induced Factor 1 Inhibitors: Alleviating Drug-Drug Interactions through Use of PubChem Data and Comparative Molecular Field Analysis Guided Synthesis. *J. Med. Chem.* **2011**, *54*, 3982−3986.

(10) Tresadern, G.; Bemporad, D.; Howe, T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *J. Mol. Graphics Modell.* **2009**, *27*, 860−870.

(11) Cramer, R. D. R-group template CoMFA combines benefits of "ad hoc" and topomer alignments using 3D-QSAR for lead optimization. *J. Comput.-Aided Drug Des.* **2012**, *26*, 805−819.

(12) The concept of "template CoMFA" was alluded to in ref 11, by the name of "whole template CoMFA" in distinction from "R-group template CoMFA." Although methodologically precise, these three-word names are verbally awkward, so in the future, these two methodologies will be called "template CoMFA" and "R-group CoMFA."

(13) 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.

(14) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. Lead-Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.
(15) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and Searching 10E20 Synthetically Accessible Structures. *J. Comput.-Aided Drug Des.* **2007**, *21*, 341−350.
(16) Cramer, R. D.; Wendt, B. Pushing the Boundaries of 3D-QSAR. *J. Comput.-Aided Drug Des.* **2007**, *21*, 23−32.
(17) Wendt, B.; Cramer, R. D. Quantitative Series Enrichment Analysis (QSEA): a novel procedure for 3D-QSAR analysis. *J. Comput.-Aided Drug Des.* **2008**, *22*, 541−551.
(18) Wendt, B.; Uhrig, U.; Bos, F. Capturing structure-activity relationships from chemogenomic spaces. *J. Chem. Inf. Model.* **2011**, *51*, 843−851.