J Comput Aided Mol Des (2012) 26:805–819
DOI 10.1007/s10822-012-9583-9

R-group template CoMFA combines benefits of "ad hoc" and topomer alignments using 3D-QSAR for lead optimization

Richard D. Cramer

Received: 22 March 2012 / Accepted: 14 May 2012 / Published online: 4 June 2012

Abstract Template CoMFA methodologies extend topomer CoMFA by allowing user-designated templates, for example the experimental receptor-bound conformation of a prototypical ligand, to help determine the alignment of training and test set structures for 3D-QSAR. The algorithms that generate its new structural modality, template-constrained topomers, are described. Template CoMFA's resolution of certain topomer CoMFA concerns, by providing user control of topological consistency and structural acceptability, is demonstrated for sixteen 3D-QSAR training sets, in particular the Selwood dataset.

Keywords 3D-QSAR · Topomers · Selwood · Template CoMFA · Ligand alignment

Introduction

With lead optimization being the most challenging phase of drug discovery [1],[1] the ability to rank the therapeutic promises among a project's candidate structures should be highly beneficial. However, during lead optimization the various measured pA50s, whose values must be predicted to meaningfully rank candidate structures, vary by little more than a log unit [2, 3]. Thus any actual benefit depends upon the average error in a pA50 prediction being no greater than a log unit. This demanding goal is most reliably met by 3D-QSAR studies [4].[2] However, the tedium and subjectivity of the various ad hoc ligand alignment methods that 3D-QSAR has long required is a major drawback.

The topomer methodology, originated to provide a very rapid and objective means of assessing 3D shape similarity [5, 6], also proved far more successful than 2D fingerprints in prospectively guiding "scaffold-hopping" during the earlier "hit-to-lead" phase of discovery [7]. The ensuing application of topomers as an automatic means of 3D-QSAR alignment [8] further yielded an enormous improvement to the convenience and objectivity of 3D-QSAR. Of course performance is what really matters, and so far "topomer CoMFA" 3D-QSAR models have also performed extremely well in actual lead optimization projects, yielding a standard error in prospective predictions of ~0.6 pA50 units over 144 measurements reported by four discovery organizations, a result whose unexpected relative precision has been rationalized statistically [9].[3] Any methodology which combines such a remarkable record of effectiveness with superior convenience and objectivity would seem worth trial in most lead optimization situations.

Yet some methodological concerns remain with topomers for 3D-QSAR alignment.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9583-9) contains supplementary material, which is available to authorized users.

R. D. Cramer (✉)
Tripos DE, 1699 South Hanley Road, St. Louis, MO 63144, USA
e-mail: Richard.Cramer@certara.com

[1] Lead optimization costs, per new drug introduction, are the highest of all, exceeding those of Phase II and III development because, being earlier, they generate more dead-ends and tie up capital for longer. More specifically, lead optimization accounts for 17 % of total R&D cost and around 50 % of discovery cost, and may be the 3rd largest opportunity area for overall R&D cost reduction.

[2] Challenges for its chief competitor in practice, binding free energy calculation, are most recently discussed by Stouch [4]

[3] Patents are pending on template-constrained topomers and their applications.

- Desire to employ all available structural information while generating a 3D-QSAR, notably any experimental receptor-bound conformations, when generating structural alignments.
- Challenge of producing a single 3D-QSAR from structurally diverse training sets believed to share a common structural target (exemplified by the R2 groups in the Selwood data set below).
- Possible structural inhomogeneities among the topomers generated from series of related structures. When setting the dihedral angle of any acyclic single bond A–B–C–D, the topomer protocol selects, from among candidate D atoms, the one attached to the largest count of more distant atoms. As one outcome, within a lead series including a phenyl ring bearing a key substituent and a variety of other substituents, that key substituent may "flip" by 180° depending on the relative size of the other substituents (exemplified by the R1 groups in the Selwood data set below).
- The unacceptable features that a topomer conformation occasionally includes, such as sterically overlapping atoms.

The major goal in developing the "Template CoMFA" methodologies was to provide means for a CADD specialist to address the concerns listed above, while retaining the objectivity and most of the convenience and speed of "topomer CoMFA". The template CoMFA idea is easily summarized. *To generate the 3D conformation of any candidate R-group, wherever template and candidate connectivities "match", directly assign the coordinates of the template atoms to the corresponding candidate atoms. To position the remaining "unmatched" candidate atoms, apply the topomer protocol.*

Practicable implementation of this idea requires "atom match" definitions that produce 3D similarities which agree with human expectations. Limiting such matches to strict correspondences between atom and bond types seems insufficient. For example, usually a cyclohexyl within a candidate R-group should assume a shape almost identical to that of a N-piperidinyl R-group corresponding within a template, despite the N-to-C difference. Yet not always, for perhaps in a particular lead subseries that piperidinyl nitrogen is inverted in the receptor-bound conformation. To provide an analyst with general means of satisfying a variety of preferences such as these, this initial template CoMFA idea has been extended in two ways:

- Multiple templates can be employed. Connectivity matching of a candidate will then be performed against each such template, with the template that provides the optimal set of connectivity-matching atoms becoming the one guiding that candidate's geometry.
- Connectivity matching is applied in two successive stages. The first requires an exact match of atom and bond types. The second allows "fuzzier" matching of atom and bond types, with a template-specific definition of "match" under moderate user control.

This template CoMFA idea is currently embodied in two distinct protocols, provisionally named "R-group" and "whole", with each believed better suited to lead optimization or hit-to-lead phases of discovery, respectively. Here we introduce the general algorithms that generate such template-constrained structures and present two illustrative applications of "R-group template CoMFA", with isolated ligands as the templates. Other studies and extensions, in particular "whole template CoMFA" with receptor-bound ligands as templates, will be described separately.

Methodology

R-group template CoMFA shares characteristic methodological steps with topomer CoMFA, which, having been detailed elsewhere [2, 8], will here simply be mentioned where appropriate.

R-group definition by "fragmentation". Once the templates and the QSAR training set structures have been selected, the first (and the only user-performable) operation in template (or topomer) CoMFA is specifying fragmentations of all those structures, by designating within each structure one or more acyclic bonds whose disconnections when applied to all template and training set structures should yield acceptably commensurate "stacks of R-groups". Each such R-group stack contributes its own separate set of fields, or "CoMFA column", to the PLS analysis and the resulting 3D-QSAR model, instead of the single set of fields and "CoMFA column" that represent intact structures in other 3D-QSAR methods.

Since in template (or topomer) CoMFA the R-groups are the only source of field descriptors, the user's primary goal in fragmentation should be to ensure that all structural variations within the training set are included within an R-group stack. For a series containing an invariant common core, that goal may be accomplished by disconnecting all the bonds that attach the varying side chains to that core. More often, training set structures are simply split into two sets of R-groups, by designating one acyclic bond within each for disconnection. Usually that bond connects central heavy atoms, but even structures lacking acyclic bonds between heavy atoms (e.g., steroids) can be fragmented, by disconnecting a C–H bond.

Template fragmentations, specified and performed in the same way (and, within the released implementation, during the same process) as training set fragmentations, are also

needed, to define the template 2D topologies that will be matched against each candidate's R-group 2D topologies, and thence the template atom coordinates that will be assigned to that candidate's matching atoms. At the same time, each template fragment is spatially positioned to superimpose its open valence onto the X-axis, just as for topomer-only generation but of course leaving its internal geometry unaffected.

Two reasons why R-group template (and topomer) CoMFA process R-group "stacks", rather than whole structures, may be of interest. One is methodological. Effective ligand alignment for 3D-QSAR requires compatibility in spatial positioning as well as internal 3D geometry. The open valence that defines any R-group provides a universal and unambiguous spatial positioning for a fragment, simply by its overlay onto an arbitrary fixed Cartesian vector. The other is the typical project goal during lead optimization, with the structural variations being considered usually limited to a few peripheral R-group attachment points, such that any 3D-QSAR need only consider those R-group variations. On the other hand, R-group template CoMFA ignores any possible effects of interactions of an R-group stack with either another R-group stack or a common scaffold.

Template CoMFA algorithms. To generate the template-constrained geometry of an arbitrary candidate structure, the two fundamental processes are:

1. Identifying the most numerous set of root-connected atoms within the candidate structure that appropriately match root-connected atoms within a template structure (the "root" in R-group template CoMFA being a cleavage bond).
2. Positioning all of the candidate atoms, by:
   - copying the coordinates of the matching template atoms.
   - transforming the existing coordinates of the non-matching atoms (as initially generated by a 3D model builder such as Concord) to produce conformationally acceptable attachments to the matching atoms
   - applying the topomer protocol to those non-matching atoms

These two fundamental processes occur sequentially and separately, the only information transferred being an array mapping each matching candidate atom to the corresponding template atom ID, so they will be described independently.

Atom-matching. This graph-matching algorithm differs in several ways from those employed by familiar methodologies such as substructure searching and maximal common sub-graph (MCS):

- *Graph matching is rooted.* Every mapped atom must be connected by at least one unbroken path of mapped atoms to the open valence of the candidate TC topomer.
- *Graph matching is carried out in two successive stages*. As already mentioned, 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.
- *Graph matching is breadth-first rather than depth-first*. A breadth-first search within a chemical structure graph considers every atom that is separated by the same number of bonds from a starting atom, or in the same "layer", before considering any atom separated by a greater number of bonds, in the next "layer". One reason for choosing breadth-first over depth-first searching is that when a template (or topomer) CoMFA model is then sought, the steric field descriptors of topomers are "attenuated", such that the more rotatable bonds that separate an atom from the root, the smaller the steric influence of that atom. Atoms that are nearest to the root are thus the most influential, and a breadth-first search prefers those matches.

Figure 1 summarizes template CoMFA's breadth-first atom matching algorithm. This algorithm is executed twice, first in the "exact-atom-match" phase, and then in the "approximate-atom-match" phase. The two phases differ at only two places within Fig. 1:

1. In the definition of "atom match" —
   a. In the exact phase, a match between atom and bond states requires:
      i. Atom agreement (element identity, including hybridization and stereochemistry)
      ii. Bond agreement (type identity, with matches of amide to single, and aromatic to Kekule single/double also being acceptable, and agreement in ring membership, i.e., (a)cyclic to (a)cyclic)
   b. In the approximate phase, these atom, bond, and cyclic match requirements become approximate rather than exact and also become user-specifiable attributes of an individual template.
2. In the criteria for retaining alternative intermediate mappings, or "paths", as the possible permutations of candidate atoms onto template atoms are screened. A candidate atom may match more than one atom in a template. For example, within any phenyl ring, the 2-carbon and 6-carbon are indistinguishable so the 2-carbon in a candidate matches both 2- and 6-carbons in the template. If either ring is unsymmetrically

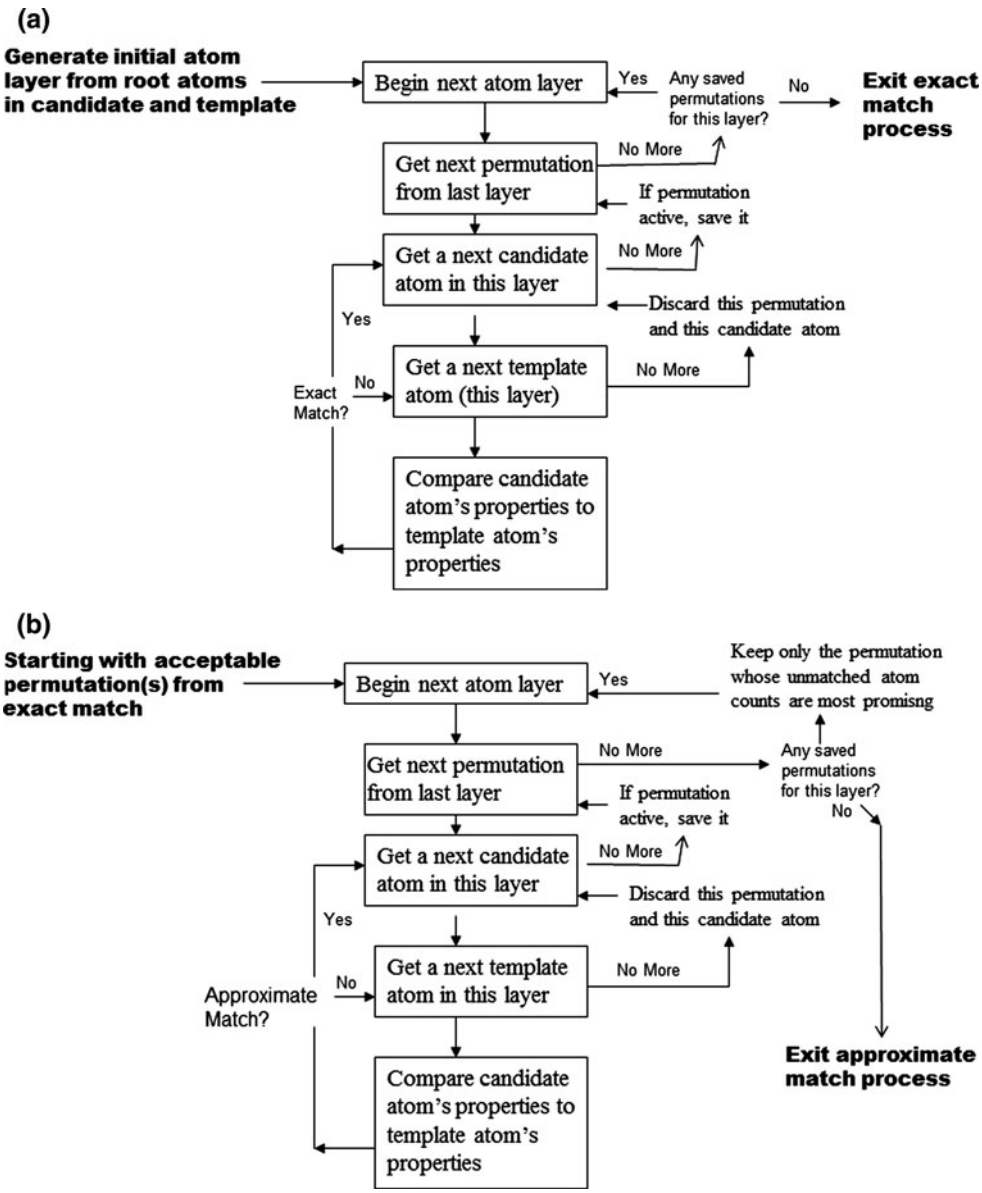
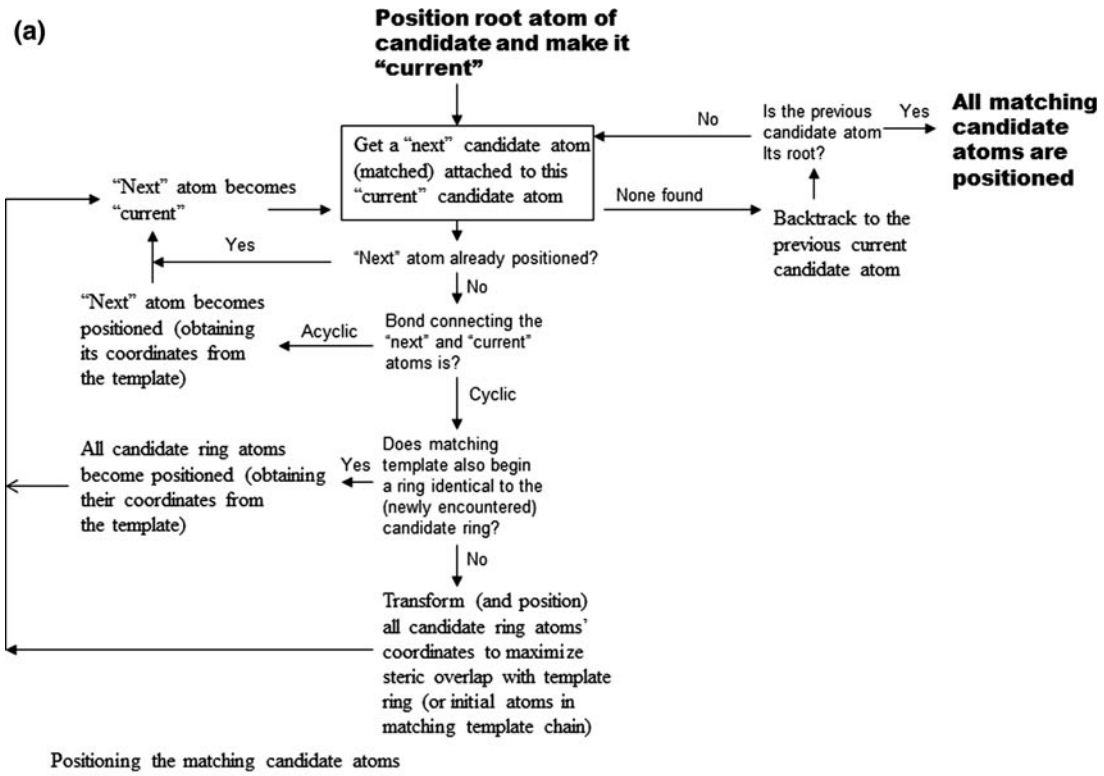


Fig. 1 Flow charts of the "exact" and "approximate" atom-mapping phases for generation of template-constrained topomers

substituted, then the path that leads to a particular substituent may eventually prevail, by containing more matching atoms. But with a breadth-first search, such a more distant differential match will not yet be detectable, so every acceptable path must be retained and processed in turn, until a single optimal atom matching can be identified. It is the definition of "until" that distinguishes the exact and approximate matching processes. In the exact match phase, every distinct mapping is processed until growth of every path in every mapping is blocked, with selection of the single atom mapping to be retained (the one having the maximal count of mapped atoms) being done only then. In the approximate match phase, because of a potentially much larger number of active mappings, the list of paths is pruned whenever a depth level is completed (when all possible matches to all paths in all mappings to the current level have been processed). The single mapping to be continued to the next layer will then be the one connected to the most heavy atoms in all the layers yet to be processed, because of being presumed the most likely winner if all mappings were to be continued (Fig. 2).

As mentioned above, feedback from collective experience in applying template CoMFA within a variety of discovery projects is likely to encourage alterations and extensions to these algorithms, most likely in the form of user options. An interesting example is that its strictly one-to-one atom matching algorithm does not always provide an acceptable 3D similarity. Consider the two situations depicted in Fig. 3, rather frequent possibilities considering the heterocyclic variations that medicinal chemists



Positioning the matching candidate atoms

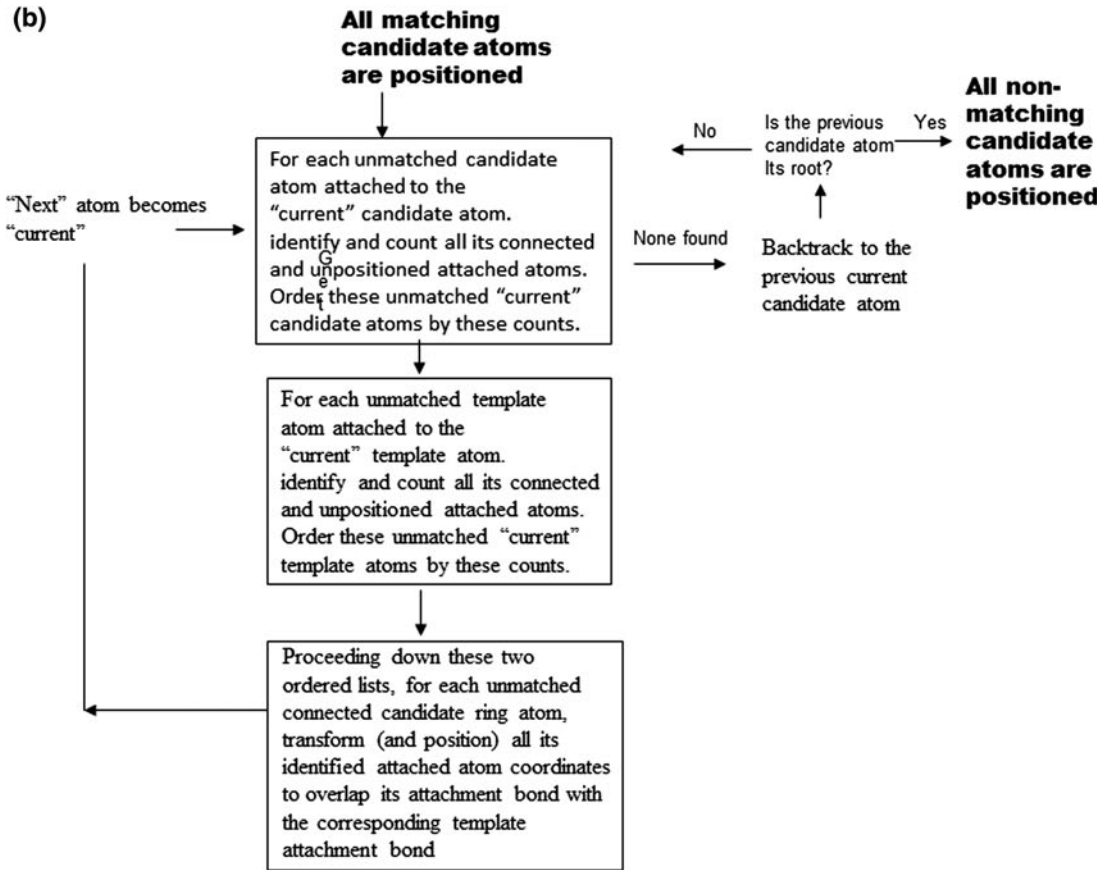


Fig. 2 Flow charts of the "mapped" and "unmapped" atom-positioning phases for generation of template-constrained topomers

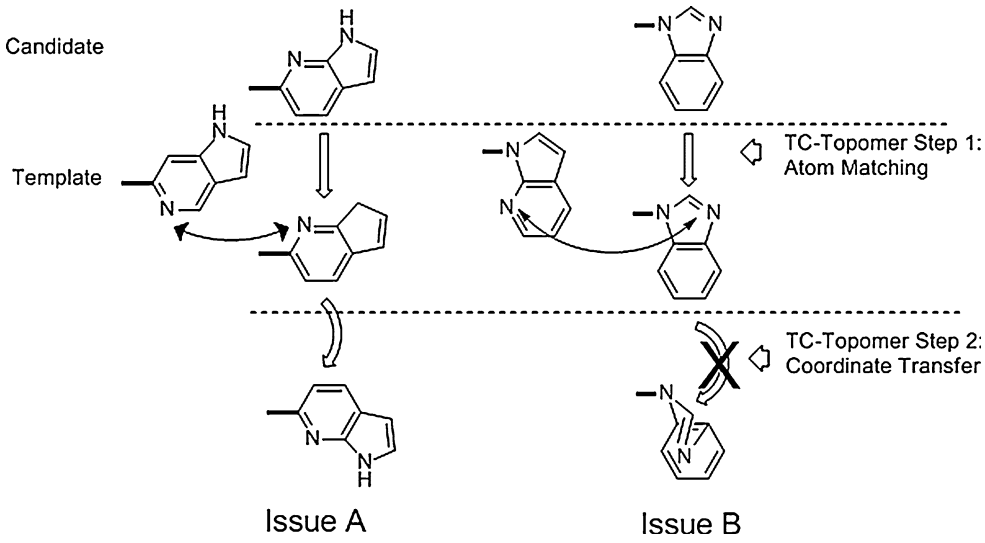


Fig. 3 Illustrations of two possible atom-mapping issues during generation of template-constrained topomers

frequently explore. This strictly one-to-one atom matching protocol insists on a candidate alpha- or beta-nitrogen atom being mapped to a template alpha- or beta-nitrogen atom, as shown by the double-ended arrows, even though alternative mappings from nitrogen to carbon yield much higher 3D similarities. Yet not necessarily the 3D similarity desired, because a close steric overlap of hydrogen-bond accepting nitrogens might be pharmacophorically preferred, as the resultant (bottom) structure for Issue A in Fig. 3 depicts. In this case a user can induce his preferred mapping by providing multiple templates. However the Issue B mapping in Fig. 3 can never yield a viable ring conformation when the coordinates of the matching template atoms are copied to the candidate. The conformational monster that must result is suggested by the bottom structure for issue B. To prevent such unworkable outcomes, a loose geometrical constraint has been added. When an atom match is being evaluated, the dihedral angle that would result within the candidate is compared with the corresponding dihedral in the template (both following their shortest path back to the roots). Denoting these dihedrals by A–B–C–D, whenever the B–C and C–D bonds in both candidate and template are both cyclic, and the A–B–C–D dihedrals differ by more than 90°, the candidate-to-template atom match of atom D's is not allowed.

Generating a template-cosntrained conformation. The second phase of template-constrained (TC) structure generation, the positioning of its atoms, determines, as with any 3D-QSAR protocol, the non-covalent field intensities that then become structural descriptors. For 3D-QSAR, self-consistency is proving to be the primary goal in structure generation, whether template-constrained or not. The valence geometry of a structure of interest, or "candidate", specifically its bond lengths, valence angles, and ring dihedral angles, is generated by a 3D-builder such as Concord or Corina (any open valences having been blocked temporarily). In R-group template or topomer CoMFA, this structure is then positioned by placing its (reopened) valence marker or "root" at the Cartesian coordinate origin and pointing the open valence bond along the positive X axis. The remaining positional degree of freedom, the dihedral of the substituent(s) on the root atom about the open valence bond, is determined by the template. (As noted above, every template R-group will already have been positioned as just described, except that the internal geometry of the template R-group of course remains unaffected.)

The atom positioning process for a template-constrained structure is summarized in Fig. 2.

1. Proceeding depth-wise from the root, the coordinates of each candidate atom found in the candidate-to-template atom mapping list (this mapping as mentioned being the only information transmitted from the atom-mapping phase) are "copied" from the corresponding template atom. "Copied" is quoted because the valence geometries of the matched atoms may not satisfactorily coincide, particularly within ring systems. Currently such coordinate discrepancies are resolved in favor of the existing candidate within cyclic topologies, but in favor of the template for alicyclic topologies. Partially mapped ring systems pose the greatest challenge, because conformational differences among the unmapped ring system atoms may seriously distort the relative geometry of the mapped ring system atoms. Currently this issue is addressed by retaining the candidate's entire ring system geometry, including any pucker, while adjusting the dihedral of the ring system attachment bond (the one arriving from the root) to overlay the bond formed by the first- and second-encountered in-ring atoms in the candidate onto the corresponding template bond.

2. Next, those "unmapped" candidate atoms whose coordinates were not determined during step 1 are positioned. Again proceeding depth-first and atom-by-atom through the directly mapped and the retained ring candidate atoms, each attached and unmapped atom, along with its more distant unmatched attachment atoms, is transformed to regenerate the attaching bond, inheriting its length and valence angles from the candidate. Wherever there are multiple unmapped atoms attached to the same mapped atom, both template unmapped-from and candidate unmapped-to attachments are ordered by descending heavy atom count and then assigned in that order.
3. Finally the topomer protocol [6], but leaving existing stereochemistry unchanged, is independently applied to each of the unmapped attachments, starting with the dihedral of the bond attaching a mapped atom to its unmapped neighbor.

Usage of template-constrained topomers by R-group template CoMFA. Subsequently, R-group template CoMFA is algorithmically and operationally indistinguishable from topomer CoMFA. Thus, from a user's perspective, R-group template CoMFA behaves as a simple variant of topomer CoMFA, the only difference being a requirement for an additional first step, the designation of at least one template structure.

However, topomer CoMFA also is not yet widely used, and so brief mention of its most distinctive characteristics seems appropriate.

- Conceptually, the most important distinction from other 3D-QSAR workflows, one that a user needs to keep very much in mind, is that all of its outputs are based on R-groups, not complete structures. For example, instead of the familiar contour display, there will result separate and entirely independent contours for each R-group stack (as illustrated for example by Figs. 8, 9).
- Operationally, the automated generation of 3D-QSAR alignments, whether by the template or the topomer-only protocol, and whether for constructing a 3D-QSAR or for then performing 3D-QSAR predictions, multiplies the ease, speed, and objectivity of a 3D-QSAR exercise to an extent that is difficult to appreciate until it is experienced.
- Finally, this 3D automation also enables "R-group virtual searching", which as the name suggests within large structural databases seeks R-groups that promise to confer both the desirable traits of high 3D similarity and high predicted potency.

There are a few differences between R-group template and topomer CoMFA, the most important of course being the multiple lattices and "CoMFA columns" and an associated need to think in terms of "R-group contributions" to potency rather than absolute potencies.

Results

The first of the two studies to be presented here shows how R-group template CoMFA addresses two of the four topomer CoMFA concerns and also compares the effects of a systematic variation of template geometries on the resulting models. The second expands the comparison of R-group template CoMFA and topomer CoMFA model performances to the fifteen data sets used in the original validation of topomer CoMFA. Except as noted all studies were carried out within pre-release versions of SYBYL-X 2.1.

Choice of the template structure(s) is obviously a major decision whenever template CoMFA is performed. It is expected that the experimentally determined structure of a receptor-bound ligand will be strongly preferred as a template for 3D-QSAR of a structural series, whenever available. Of course, discovery projects often pursue several such receptor-bound ligand structures, drawn from different structural series. In this situation, if a single 3D-QSAR including all the series is desired, a yet to be presented "whole template CoMFA" methodology would be needed. However, the current work, description of the underlying template generation methodology, is illustrated instead by a systematic if brief "ligand-based" study, of the effects of template variation on the resulting 3D-QSAR models. Thus for the following studies, template structures were generated by various manipulations of the Concord-generated conformations of one or more ligand structures chosen from the training set series.

The first set is the "Selwood dataset" of antifilarial antimycin analogs [10], well-known among developers of 2D-QSAR methodologies, because the large number of acceptable 2D QSARs that multiple regression using familiar descriptors discovers within this set has made it a benchmark for comparing 2D-QSAR protocols [11–13]. The 31 structures in the Selwood data set are based on a variously substituted 2-hydroxy-benzoyl moiety, all but one derivatized by amide formation between the benzoyl moiety and either a 4-phenoxyaniline or a simple linear alkylamine. In the other structure replacement of the amine by a phenyl group forms a benzophenone rather than an amide. Half of these structures appear in the left "Grid" panel of Fig. 4. The overall composition of the Selwood set is also conveyed by a "structure similarity map", displayed in the upper right-hand panel of Fig. 4, based on Tanimoto dissimilarities among "2D-fingerprints". Manually added lines connect each of three points in the structure similarity map to the corresponding structure within the "Grid'. A similarity structure map highlights overall structural distributions, here the dominance of the 4-diphenyloxyaniline and the singular benzophenone. Such an impression can be particularly helpful whenever a QSAR study is begun, especially with training sets often now numbering in the hundreds or even

thousands rather than dozens. The third panel in Fig. 4 is discussed below.

A 3D-QSAR study of the Selwood dataset has already been reported [14], applying topomer CoMFA with each structure split at the amide (or corresponding) bond. The statistical quality of this topomer CoMFA is satisfactory (as repeated in the top five numerical entries in the "Topomer" column of Table 1). Nevertheless its overlaid "R-group stacks", shown in Fig. 5, raise concerns. (It may be worth note that in the 3D representation of any topomer, unless otherwise stated the open valence bond will by convention be horizontal, blue, and positioned toward the upper left of its representation.) Considering first the benzoyl or R1-group stack in Fig. 5, it surely would be preferred that the 2-hydroxy moiety found in every Selwood structure occupies the same lattice region in every R1-group. Yet the topomer protocol does not produce this result, because the "D atom" defining the dihedral angle that positions these hydroxyl groups is identified only as the attached atom maximizing the count of all its more distant attached heavy atoms. The hydroxyl group location thus "flips" by 180°, depending entirely on size variations among the other attachments to the phenyl ring.

The topomer alignments among the R2-group stack in Fig. 5, including the phenoxyanilines, n-alkylamines, and phenyl, reflect a somewhat different concern. Here the topomer protocol directs phenoxyanilino and n-alkylamino groups at almost right angles to one another. Such a collective disarray seems most unlikely to resemble the relative disposition of these R2-groups when bound to any actual receptor.

R-group template CoMFA allows the analyst to impose any appropriate spatial correspondences to such R-group stacks. To superimpose structurally common features, such as all the hydroxyls in the Selwood R1 groups, a single template structure that deploys that common feature in the preferred conformation will usually suffice. To bring structurally diverse groups, such as the phenoxyamino, n-alkylamino, and phenyl Selwood R2 groups, into some desired correspondence, a separate template for each group is necessary, each template then determining the geometries of the corresponding fragments in the most similar training or test set structures, as presented above. Each template would presumably be a ligand known to have the biological activity of interest, and probably a training set member. Templates must currently be provided as complete structures rather than fragments, as the more convenient and natural input form in most situations.

Furthermore, if a ligand is to become a conformational template for other ligands, it seems reasonable for that ligand

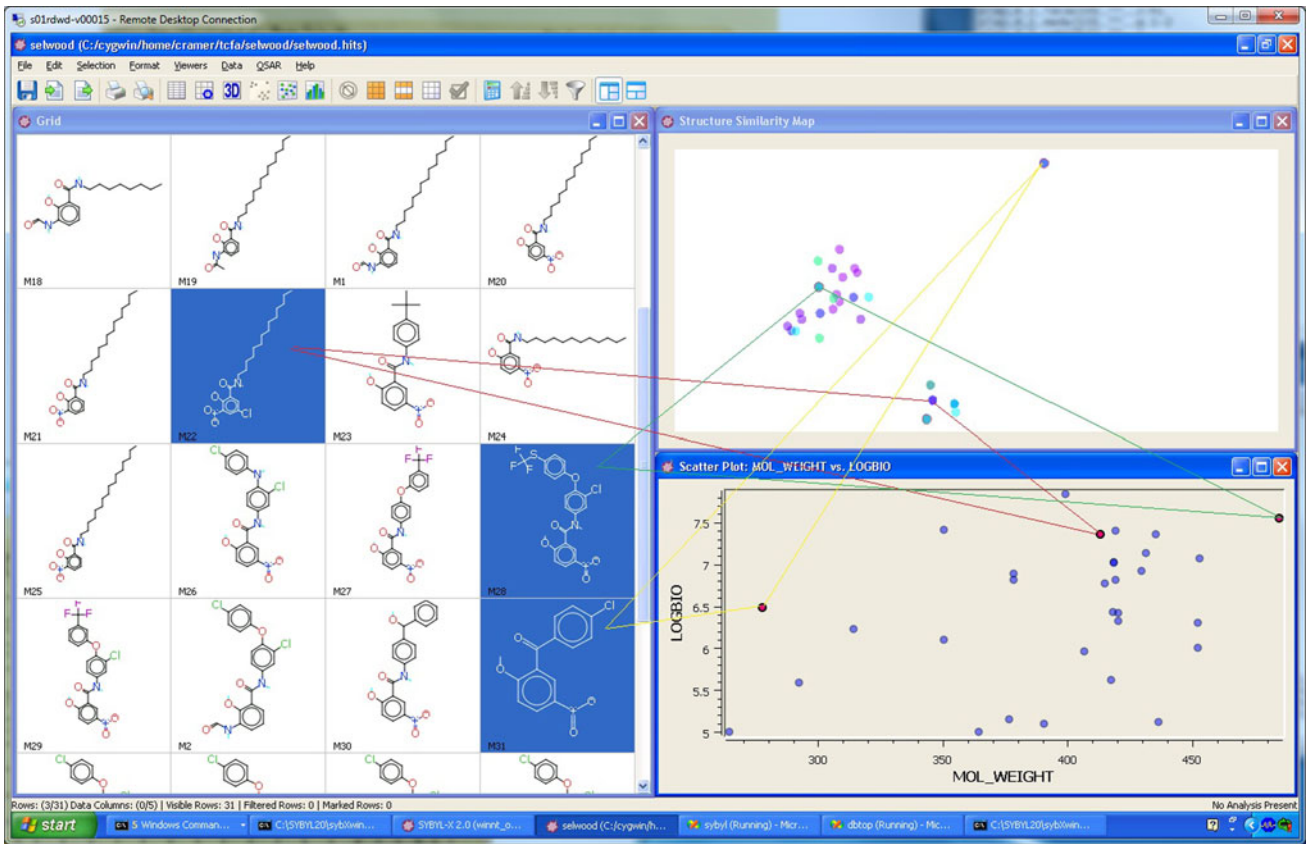


Fig. 4 Distribution of structures within the Selwood training set, illustrating the selection of the three template structures

Table 1 Comparative summary of 3D-QSAR statistical parameters from four different alignment approaches to the Selwood data set

	Template CoMFA			
	Topomer CoMFA	Maximum Overlap	Surflex-Sim	Minimum Enthaply
Statistical properties				
# Components	5	4	8	9
LOO q^2	0.457	0.533	0.646	0.697
LOO sdep	0.670	0.610	0.570	0.540
r^2	0.921	0.896	0.979	0.991
S	0.250	0.290	0.140	0.090
r^2 of model coefficients				
Steric field terms				
Topomer	1.000			
Template: max overlap	0.722	1.000		
Template: SURFLEX/SIM	0.380	0.163	1.000	
Template: min enthalpy	0.470	0.522	0.344	1.000
Electrostatic field terms				
Topomer	1.000			
Template: max overlap	0.731	1.000		
Template: SURFLEX/SIM	0.383	0.174	1.000	
Template: min enthalpy	0.465	0.514	0.351	1.000
Model properties				
intercept	7.22	5.77	5.80	5.67
R1 contributions				
lowest	−0.77	−1.10	−0.75	−1.38
highest	1.09	0.81	−0.24	1.05
sdev	0.51	0.40	0.48	0.53
R2 contributions				
lowest	−2.27	−1.17	−1.73	−1.53
highest	0.08	1.72	0.94	1.21
sdev	0.68	1.05	0.86	0.79

See text for model definitions and details

to be structurally representative, strongly active, and relatively large. To apply these criteria when selecting the three specific templates to represent the R2 variety within the Selwood set, the two right hand graphs in Fig. 4 were utilized. "Sweeping" each of the three clusters within the "Structure Similarity Map" (upper right) highlighted the points representing those structures within the "Scatter Plot" (lower right) of molecular weight (size) versus biological potency. Points closest to the upper-right-hand corner of the "Scatter Plot" then represent the most appropriate individual structures, according to the these criteria of size and potency. In Fig. 4 the structures of the three ligands thus

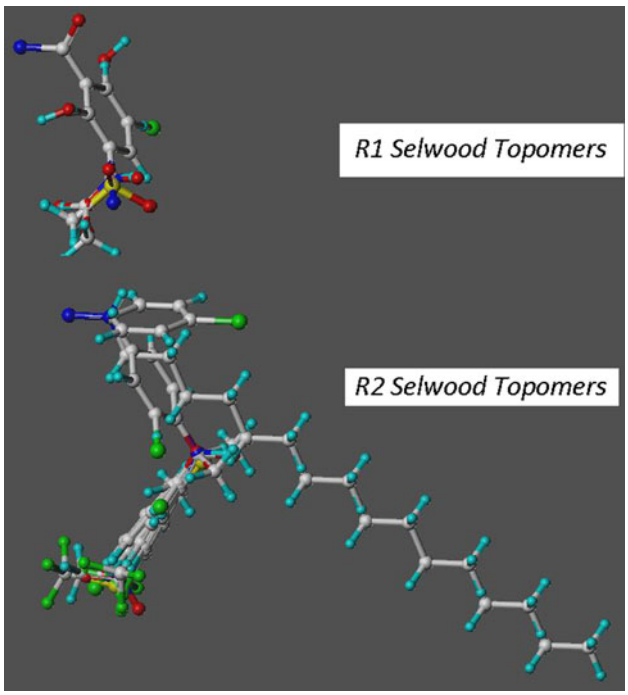


Fig. 5 "Rgroup stacks" of the purely topomeric conformations for the 31 members of the Selwood training set

selected as templates for this exercise are connected to their corresponding data points by lines forming three triangles. Their 2D structures appear more clearly in Fig. 6.

Appropriate conformations must then be generated for each of the three template structures, preferably in an objective way. Three alignment goals, representing two major precepts of optimality in ligand superposition, seemed worth survey. All started from Concord models. The first goal, maximally overlapping template conformations, was fulfilled by superimposing their 2-hydroxy-benzoyl fragments, and then manually adjusting successive dihedrals in the linear alkyl template to maximize the overlap of alkyl carbons onto the aromatic carbons in the phenoxyanilino template (but also maximizing 1-, 6- steric repulsions). Figure 7 depicts its outcome, with the top structure including the alkylamine chain after these torsional adjustments. A second "hybrid goal" sought a balance between low ligand strain enthalpy and high overlap among the three template structures by simultaneously minimizing both criteria with a suitable program, here Surflex-Sim's best-scoring result [15]. Finally, taking low ligand strain enthalpy of the individual templates as the only alignment goal, the three template conformations were generated by applying the MMFF force field to their Concord-generated conformations. This triad of approaches thus includes the two most common precepts for ligand superposition, ranging from "best overlap, enthalpy indifferent" (denoted by Maximum Overlap), through a tradeoff

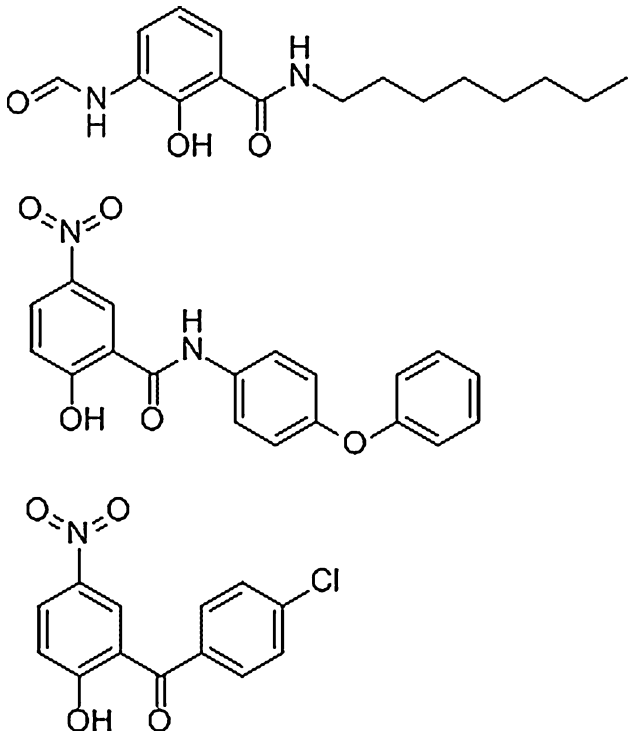


Fig. 6 The three structures selected from the Selwood training set to become its templates

between overlap and enthalpy (Surflex-Sim), to "overlap indifferent, enthalpy optimal" (Minimum.Enthalpy).

The three sets of three template structures resulting from these approaches appear in Fig. 8. Particularly take note that, although these structures are overlaid, with R-group template CoMFA the relative orientations of template structures have no influence whatsoever on the 3D-QSAR model. Instead, with the first steps in template preparation being fragmentation and positioning of each resulting R-group by open valence superposition, it is only the conformations of these R-groups, their transformed internal coordinates, that remain from the input template structures to influence the conformations of the 3D-QSAR training or test set structures.

The Selwood training set was submitted to R-group topomer CoMFA using each of these three sets of template structures, the individual structures being fragmented as described above. The statistical and other numerical properties of the three resulting 3D-QSAR models are listed in Table 1, along with those of the topomer model for comparison. Their structural implications, in the form of the familiar contour maps superimposed on their training set TC-topomers appear in Figs. 9 and 10, as the R1 and R2 stacks.

Here are the important results from these three R-group template CoMFA models of the Selwood training set.

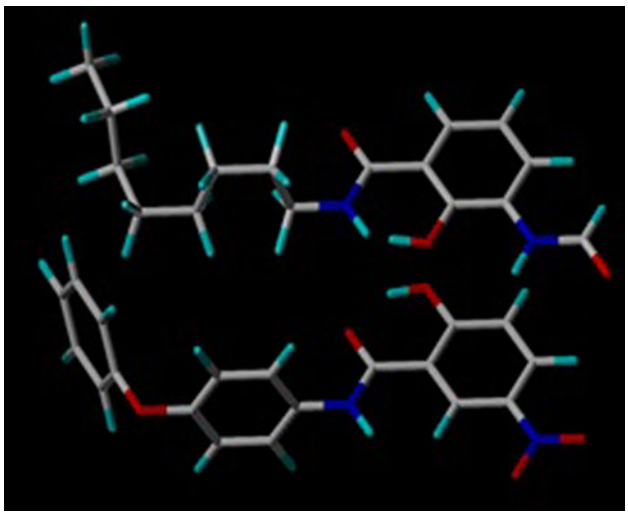


Fig. 7 Two Selwood template structures, with the torsion angles in the upper structure adjusted to "maximally overlap" its alkyl carbons with the aromatic carbons in the lower structure

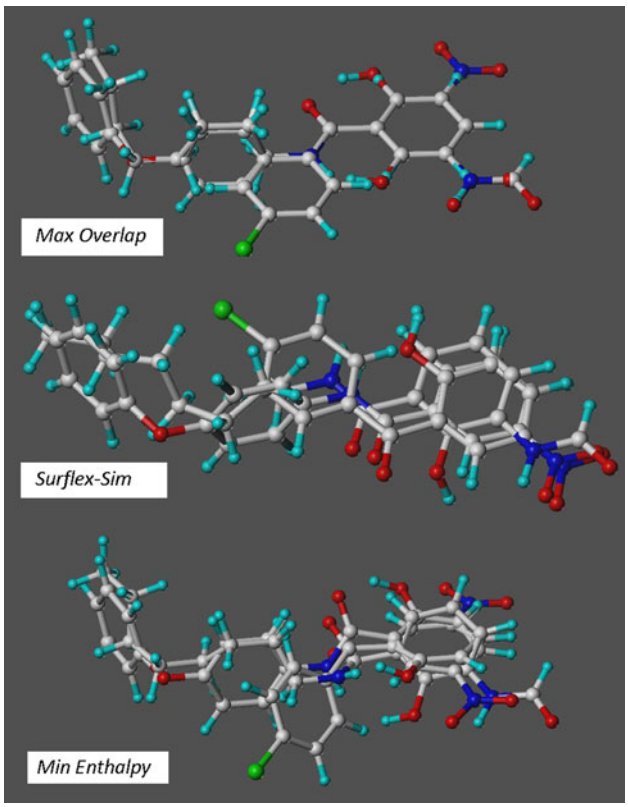


Fig. 8 Template conformations of the three selected templates for each of the three alignment approaches

- The template constrained topomer generation algorithm performed as intended. Specifically, within the three R1 stacks of Fig. 9 the 2-hydroxyl groups are superimposed, and within the three R2 stacks of Fig. 10 the dispositions of phenoxyanilino- and alkylamino- groups

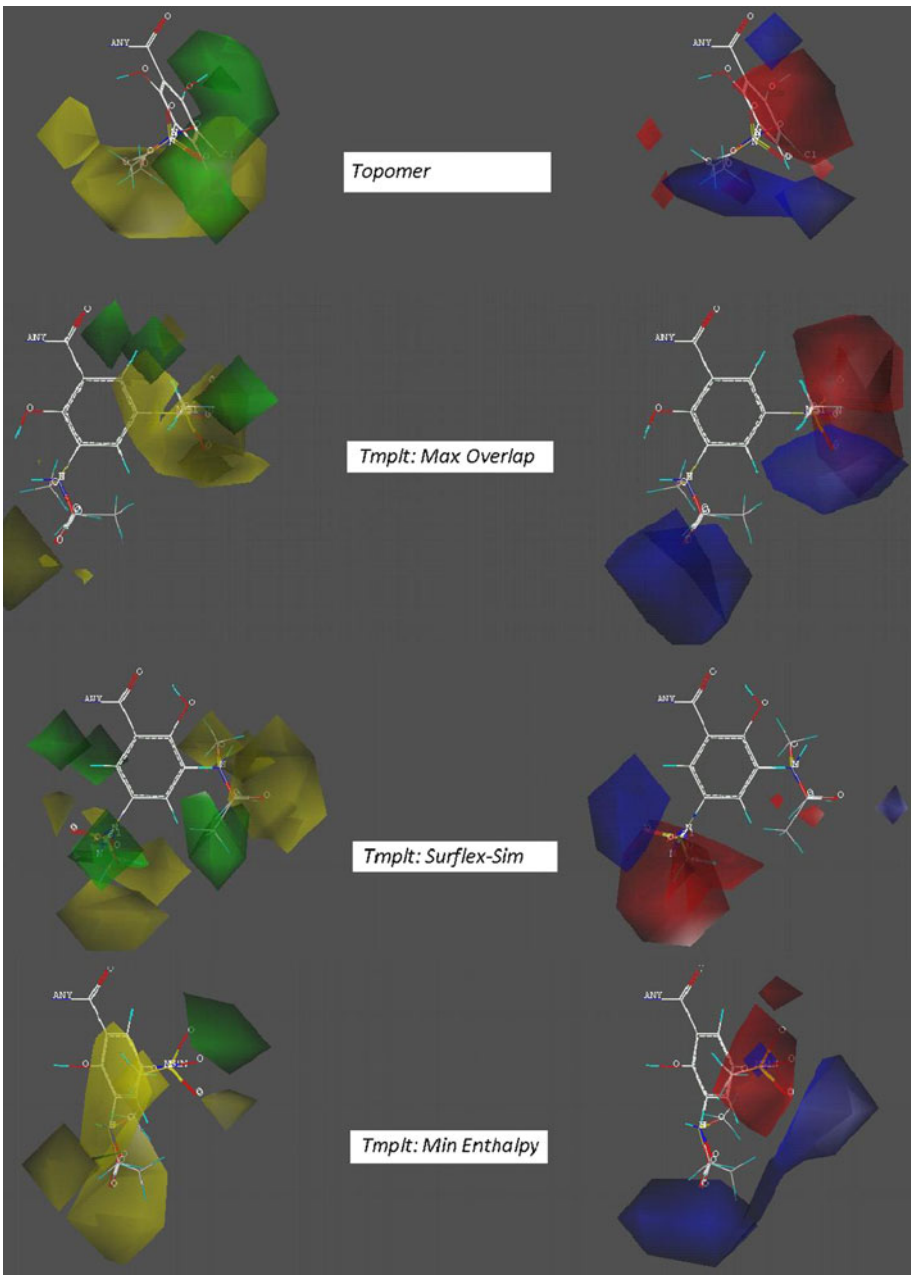


Fig. 9 Overlay of 3D-QSAR model contours and aligned training set structures for the R1 groups of the Selwood training set, for each of the three alignment approaches

follow the conformation of their 2D-most-similar template. (The template R1s and R2s are included within these stacks, because all these templates were originally taken from the training set.)

- The four 3D-QSAR equations, derived either only from the topomer rules or from the three template sets, have generally similar statistical quality. There is a modest increase in statistical fit, though associated with an increase in the number of PLS components, [4] with the template conformation sets that increased physico-chemical fidelity and decreased overlap.

[4] Large "# component" values may raise concerns about over-fitting, especially when accompanied by unreasonably low "SDEP" values. However, from PLS over-fitted and unstable models are much less of a practical risk than from other common algorithms such as multiple regression, because PLS operates on blocks of descriptors rather than individual columns. The usual effects of additional components on a PLS model are increasingly minor refinements, seldom having any effect on overall "statistical significance". Therefore, in the standard topomer CoMFA implementation as used in these studies, during leave-one-out cross-validation, component extraction ends only when the resulting SDEP value first increases. Of course the analyst may then truncate the "#components" to a smaller value, but in these studies, such a necessarily subjective decision seemed inappropriate.

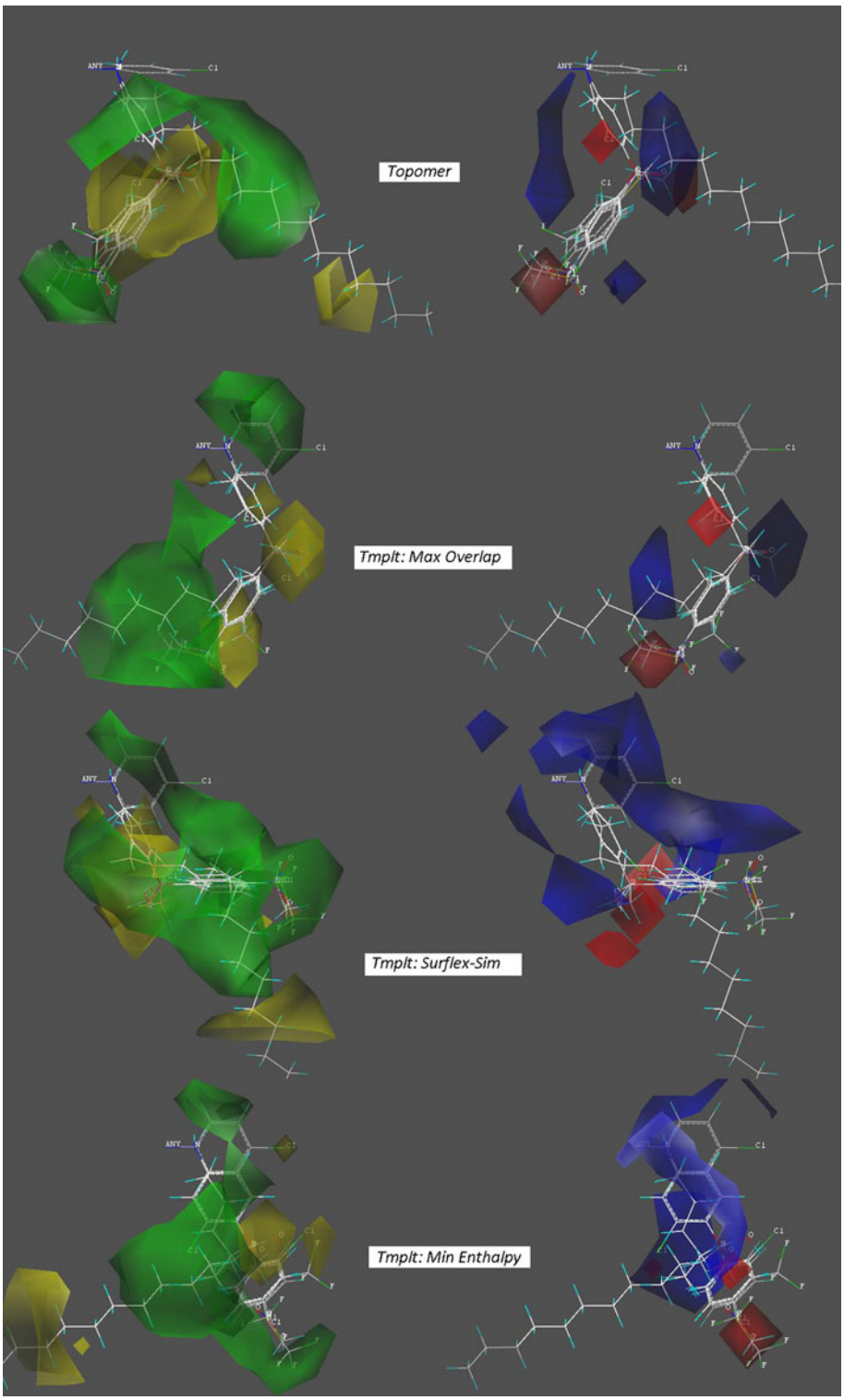


Fig. 10 Overlay of 3D-QSAR model contours and aligned training set structures for the R2 groups of the Selwood training set, for each of the three alignment approaches

- Yet the four 3D-QSAR model equations themselves are rather dissimilar. This dissimilarity is apparent either qualitatively, by inspection of the contours in Figs. 9 and 10, or quantitatively by the modest correlation values within the block "r^2 of Model Coefficients" in Table 1. (Each of the sets of model coefficients underlying these correlation values was extracted from a standard RGVS-intermediate *.dat* file using a Python script.)

However, these dissimilarities among the four model equations must, to some perhaps large degree, simply

reflect the alignment differences among the templates (since each term arises from a specific lattice point and Figs. 9 and 10 also show how the variety among template alignments changes the identities of the most adjacent and affected lattice points). This idea is supported by the very similar correlation values of corresponding steric and electrostatic field terms in Table 1. If so, the predictions from these four apparently disparate equations may be unexpectedly similar.

Template CoMFA was then applied to the fifteen data sets originally used to validate topomer CoMFA, then selected randomly from those to which a 3D-QSAR method had already been successfully applied [8]. Fragmentations were the same as shown in Table 1 of that same publication. For each of these fifteen R-group template CoMFA studies, a single template structure was used, here simply the most active of the training set structures. Two protocols were then applied to each of these singleton templates, the Concord conformation and its MMFF-minimized conformation. Table 2 compares these two template-based alignment approaches (respectively labeled as Cncd and MinH) with those of topomer CoMFA and the original 3D-QSAR models (labeled Lit), in terms of five statistical properties: the leave-one-out cross-validated q^2, the underlying SDEP, the number of PLS components for which SDEP first reached a minimum; the conventional r^2; and errors of prediction for any test set provided by the original publication. At the bottom of Table 2 is a summary line, reporting the average or total, over the applicable training sets, of that statistical property for its 3D-QSAR alignment approach.

The results in Table 2 are easily summarized. As was originally reported, the "Lit" models, from the original 3D-QSAR publications, yielded statistical parameters somewhat better than the corresponding topomer ("Tpmr") CoMFA models. Not a surprising result, inasmuch as in the Lit studies the alignments of individual structures could be adjusted as many times as desired, while the topomer rules are completely inflexible, producing a single alignment outcome that either works or does not. Topomer-based alignments thus substitute objectivity and effortlessness, especially when predicting, for moderately superior training set statistics. However, it may be surprising that with the two template conformational protocols, the average statistical properties of the resulting 3D-QSAR models do not significantly differ, either from each other or from those of the topomer-only alignments. Although there is a hint of template alignment superiority over purely topomeric alignment, in a slightly smaller average "true prediction" error, the magnitude of that difference is statistically insignificant. Yet this lack of improvement in statistical quality with the more realistic alignment protocols is consistent with the hypothesis [9] that self-consistency is more important for 3D-QSARalignments than physicochemical realism,

Discussion and conclusions

The goal in creating template-constrained topomers, as a second general alignment methodology for 3D-QSAR, was to address the four concerns existing with the first such method, topomer CoMFA, as listed in the introduction. The Selwood data set study establishes that the current implementation of template CoMFA resolves the two specific concerns and suggests that the other two concerns have also been substantially addressed.

These results have several significant implications.

- The superior accuracies of 3D-QSAR predictions generally, and the superior facility and objectivity of topomer CoMFA specifically, can now be further supported and/or validated by imposing as much agreement with other structural information, theoretical and experimental, as may be desired.
- An oft-stated advantage of 3D-QSAR, yet one realized only with difficulty, has been its potential for modeling more structurally diverse training sets. The Selwood example, particularly its R2 results, shows how template CoMFA can simplify many such investigations. Whole template CoMFA will greatly broaden this capability.
- Most 3D-QSAR publications treat model alignments and receptor-bound conformations of a training set as effective equivalents, for example proposing successful alignments as therefore representing receptor bound conformations, or offering the agreement of model contours with receptor features as validation of the model. Yet the evidence supporting such an equivalence is mixed (why then do the topomer-only alignments succeed?). Template CoMFA offers a convenient means for further investigation of this assumption.
- For a given training set, there is very little dependence of a 3D-QSAR's statistical quality on its alignment protocol (as long as that protocol is sufficiently self-consistent [9]). Nevertheless, large differences in the model term coefficients, usually obvious in contour plot visualizations such as Figs. 9 and 10, can exist, with perhaps significant impacts on both potency predictions and similarity scores.

Less obviously, and following from the last bullet point, template CoMFA may also help address a recurring and fundamental, yet seldom acknowledged, issue—is my latest QSAR prediction trustworthy enough to guide a critical project decision? On average 3D-QSAR seems to offer the strongest published record for accuracy in pA50 prediction

Table 2 Statistical parameters of model derivation and the external prediction errors, for the fifteen 3D QSAR literature studies and their repetitions with three different topomer-based CoMFA alignments

Dataset		# cpds	CoMFA model construction												CoMFA prediction				
			LOO x-val q^2				LOO x-val SDEP				# Components				RMS "true prediction" error				
ID	Name		Lit	Tpmr Only	Cncd Tmpl	MinH Tmpl	Lit[b]	Tpmr Only	Cncd Tmpl	MinH Tmpl	Lit	Tpm On	Tpm Tmpl	MnH Tmpl	# cpds	Lit[b]	Tpmr Only	Tpmr Tmpl	MinH Tmpl
1	ICEc	36	0.630	0.380	0.550	0.464	0.82	0.96	0.84	0.88	6	3	5	2	9	0.57	0.77	0.74	0.68
2	ICEb	38	0.630	0.481	0.603	0.531	0.82	0.94	0.80	0.84	6	5	3	2	10	0.55	0.60	0.48	0.39
3	Thrombin	72	0.687	0.498	0.454	0.464	0.59	0.75	0.80	0.80	4	3	6	8	16	0.67	0.72	0.70	0.68
4	Trypsin	72	0.629	0.667	0.627	0.655	0.56	0.52	0.58	0.57	5	4	11	12	16	0.52	0.43	0.43	0.40
5	FactorXa	72	0.374	0.255	0.455	0.474	0.52	0.57	0.48	0.49	3	4	3	7	16	0.28	0.33	0.35	0.29
6	MAOa	71	0.440	0.597	0.581	0.580	1.03	0.89	0.91	0.91	2	3	3	3					
7	MAOb	71	0.430	0.597	0.503	0.516	1.25	0.89	1.21	1.20	2	3	2	2					
8	hiv	25	0.680	0.389	0.552	0.557	0.57	0.85	0.74	0.74	3	3	5	5	7	0.82	1.12	1.12	1.13
9	a2a	78	0.541	0.323	0.356	0.357	0.56	0.69	0.68	0.68	4	3	4	3	23	0.67	0.69	0.58	0.58
10	d4	29	0.739	0.710	0.733	0.407	0.73	0.71	0.53	0.96	7	5	5	2					
11	flav	38	0.752	0.777	0.407	0.715	0.48	0.47	0.96	0.54	4	3	2	5	4	0.34	1.19	1.00	1.06
12	cannab	61	0.592	0.439	0.430	0.433	0.57	0.69	0.68	0.68	4	3	1	1	6	0.45	0.66	0.50	0.50
13	ACEest	41	0.937	0.842	0.723	0.719	0.35	0.57	0.77	0.20	4	3	4	8	7	0.41	0.47	0.43	0.53
14	5ht3	61	0.645	0.423	0.417	0.417	1.19	1.68	1.67	1.67	5	5	4	4					
15	rvtrans	82	0.837	0.841	0.742	0.725	0.57	0.57	0.73	0.75	4	6	5	6	19	0.79	0.60	0.65	0.61
	Total/Avg	847	0.636	0.548	0.542	0.534	0.71	0.78	0.83	0.79	4	4	4	5	133	0.55	0.69	0.63	0.62

[a] For ICEc, ICEb, and a2a, the individual prediction values were read from the graphs in Figs. 3, 6 and 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 q2. Other values were taken directly from the tables

among all CADD approaches. Yet, with unexpected SAR discontinuities caused by some undetected "activity cliff" [16] always being possible, how great is the risk that a single prediction of a critical pA50 by 3D-QSAR will be seriously mistaken? Neither "ad hoc" nor topomer methods have encouraged exploration of the alignment protocol, the most likely source of such "black swan" [17] disappointments from 3D-QSAR, because of the impractically high resource demands of any individual "ad hoc" alignment trial or the intentional rigidity of the topomer protocol. With template CoMFA, varying the template structures and conformations should readily afford a variety of objectively reproducible alignment protocols and 3D-QSAR outcomes. Potency predictions that survive diverse alignment protocols seem the most likely to successfully guide a project to its goals and, if nevertheless unsuccessful, the most unambiguously informative about the cause of an unexpected "activity cliff".

It may have been noticed that all of these example applications of R-group template CoMFA were for training sets to which 3D-QSAR had already been successfully applied. What then might the success rate for R-group template CoMFA be with other training sets? In this connection, perhaps it might be informally added that the success rate for 3D-QSAR model generation, from trying topomer CoMFA on roughly twenty training sets as randomly supplied during site visits, has so far exceeded 75 %.

While the remarkable prospective predictive accuracies so far reported for topomer CoMFA, within actual discovery projects [9], would also seem encouraging, private discussions at these sites often surface two concerns that may understandably impede its trial:

- The topomer structures (conformation, tautomeric state) may be uncomfortably inconsistent with other highly trustworthy information, such as experimental receptor-bound conformations or established conformational criteria.
- Perhaps related to a historical tendency for more automated CADD approaches to be over-promoted and then to under-perform, topomer CoMFA's minimal user-adjustable input, whilst guaranteeing the objectivity and consequent reproducibility of its results, may be regarded as a drawback.

This new 3D structural modality, template-constrained topomers, as alignments for 3D-QSAR, seemingly allows resolution of these concerns, while retaining the benefits that topomer CoMFA already provides, particularly during the lead optimization phase.

Acknowledgment It is a great pleasure to thank Bernd Wendt for calling attention to the didactic features of the Selwood data set.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindberg SR, Schacht AL (2010) Nat Rev Drug Disc 9:203–214
2. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi F (2008) J Chem Inf Model 48:2180–2195
3. Doweyko A (2007) J Comp Aided Drug Des 18:587–596
4. Stouch TR (2012) J Comp Aided Drug Des 26:125–134
5. Cramer RD, Clark RD, Patterson DE, Ferguson AM (1996) J Med Chem 39:3060–3069
6. Jilek RJ, Cramer RD (2004) J Chem Inf Comp Sci 44:1221–1227
7. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RD (2004) J Med Chem 47(6777):6791
8. Cramer RD (2003) J Med Chem 46:374–389
9. Cramer RD (2012) J Comp Aided Drug Des 25:197–201
10. Selwood DL, Livingston DJ, Comley JCW, O'Dowd AB, Hudson AT, Jackson P, Jandu KS, Rose VS, Stables JM (1990) J Med Chem 33:136–142
11. Nicoletti O, Gillet VJ, Fleming PJ, Green DVS (2002) J Med Chem 45:5069–5080
12. Kubinyi H (1994) Quant Struct Act Relat 13:285–294
13. So S-S, Karplus M (1996) J Med Chem 39:1521–1540
14. Wendt B, Cramer RD (2008) J Comp Aided Drug Des 22: 541–551
15. Jain AN (2004) J Med Chem 47:941–961
16. Maggiora GM (2006) J Chem Inf Model 46:1535
17. Taleb NN (2007) The Black Swan: the impact of the highly Improbable, Random House, ISBN 978-1-4000-6351-2