Challenging the gold standard for 3D-QSAR: template CoMFA versus X-ray alignment

Bernd Wendt · Richard D. Cramer

Abstract X-ray-based alignments of bioactive compounds are commonly used to correlate structural changes with changes in potencies, ultimately leading to three-dimensional quantitative structure–activity relationships such as CoMFA or CoMSIA models that can provide further guidance for the design of new compounds. We have analyzed data sets where the alignment of the compounds is entirely based on experimentally derived ligand poses from X-ray-crystallography. We developed CoMFA and CoMSIA models from these X-ray-determined receptor-bound conformations and compared the results with models generated from ligand-centric Template CoMFA, finding that the fluctuations in the positions and conformations of compounds dominate X-ray-based alignments can yield poorer predictions than those from the self-consistent template CoMFA alignments. Also, when there exist multiple different binding modes, structural interpretation in terms of binding site constraints can often be simpler with template-based alignments than with X-ray-based alignments.

Keywords 3D-QSAR · Alignment · Template · X-ray · CoMFA · CoMSIA

Introduction

3D-QSAR is one of a few CADD methodologies that is intended to combine the two major approaches, structure-based, with its emphasis on physical models, and ligand-based, with its emphasis on statistical models [1]. In principle, a CoMFA model could expand the interpretability of an individual X-ray structure by visually superimposing its summary of the ligand SAR—or an X-ray structure could suggest structural extrapolations most likely to exploit a promising 3D-QSAR trend [2–4]. However, in practice, existing methods for aligning ligand structures for 3D-QSAR have been so tedious and subjective as to discourage such methodological synergies.

The X-ray structure of the biological target receptor site has long been a highly regarded method of 3D-QSAR alignment, with each of the training set structures being aligned into that site by a docking calculation [5, 6]. Although the structural relevance of this method contributes some assurance, particularly if that X-ray structure includes one of the training set ligands, the well-known uncertainties of docking pose selection introduces subjectivity into any resulting 3D QSAR model [7–9].

A recently proposed alternative is template CoMFA, an automatic and general method for generating seemingly informative and robust 3D-QSAR alignments [10]. Its inputs are one or more template ligands (which may of course be extracted from X-ray structures) and one or more training sets. Each training set structure, or "candidate", is aligned by identifying (manually or automatically) the most similar template structure. Then, wherever template and candidate connectivities "match", the coordinates of the template atoms are directly "copied" to the corresponding candidate atoms. To position the remaining "unmatched" candidate atoms, a topomer protocol [11] is applied. Template CoMFA can be completely objective. However, those of a candidate's atoms that are topomerically aligned may have physicochemically troubling locations, for example by clashing with receptor atoms.

A further alternative, generating experimental X-ray models for every ligand in a 3D-QSAR training set in the

B. Wendt (✉) · R. D. Cramer
Certara, Martin-Kollar-Str. 17, 81829 Munich, Germany
e-mail: bwendt@tripos.com; bernd.wendt@certara.com

target, addresses both the uncertainties of docking alignments and the physicochemical heresies of template CoMFA alignments. This approach is very costly, and also cannot be used for predictions of structures being considered for synthesis. Nevertheless, it might reasonably be considered as a "gold standard" for 3D-QSAR alignment, against which the performance of other alignment approaches, such as template CoMFA, should be compared.

Three such X-ray-for-every-training-ligand data sets have provided us the opportunities for such a comparative validation of template CoMFA. We have considered two specific questions:

How comparable are the statistical qualities of their 3D-QSAR models?

How comparable are the interpretabilities of their 3D-QSAR models, in providing guidance for synthetic selections?

Methods

Data sets

Three data sets with experimentally derived ligand poses were taken from protein–ligand complexes determined by X-ray crystallography [12]. The data sets span a diverse range of targets with Urokinase -a serine protease, the Protein-tyrosine-phosphatase-1B (PTP-1B)—a phosphatase and the Checkpoint Kinase 1 (Chk-1)—a serine/threonine protein kinase. Also, the ligand structures of the data sets are diverse, spanning low molecular fragment structures up to lead series compounds with different distributions of charged or uncharged compounds (see Ref. [3] for a detailed analysis).

X-ray-based alignment

According to supplementary material in Ref. [12], a publically available protein–ligand structure has been chosen to use as a reference to produce a common alignment for the compounds against each target. The protein superpositions to the reference structure served as a guide to bring the ligand alignments into a common reference frame, which preserves each ligand's relative position in the binding pocket. The following structures were used to produce the alignments (identified by PDB ascension code): for Urokinase, 1OWK; for PTP1-B, 1NZ7; and for Chk1, 2YWP. Typical RMSD values varied between 0.95 and 1.2 A for Urokinase; 0.5 and 1.5 A for PTP-1B; 0.4 and 0.8 for Chk-1 when all heavy atom positions are used.

Template-alignment

Template CoMFA (TC) [10] is a hierarchical assembly of two concepts: topomer alignments; and template alignments. To generate the 3D conformation of any "candidate structure", wherever template and candidate connectivities "match", the coordinates of the template atoms are directly assigned to the corresponding candidate atoms. To position the remaining "unmatched" candidate atoms, the topomer protocol is applied.

TC-alignment starts with assignment of an anchor bond in the template structure. The anchor bond should be chosen as a potentially unique bond surrounded by a unique neighborhood since the search for connectivity matches starts from there. All structures in the data set or subset of the data set are required to share this bond in order to be aligned.

3D-QSAR

CoMFA and CoMSIA [13] models were generated in a pre-released version of Sybyl-X 2.1. Region definitions were created automatically to extend at least 4 Angstroms beyond every molecule of all the structures in their X-ray-based as well as template-based alignments. Thus the same region file has been used for all alignments and methods. For each of the three data sets, a set of two region files with grid spacings of 2 and 1 Angstroms was created. Only steric and electrostatic fields were calculated. In order to achieve optimal prediction accuracy, partial charges were calculated for each molecule using the MMFF94 force field [14]. Cross-validation in the form of Leave-one-out (LOO) was carried out to determine the statistical significance and the robustness of the resulting models.

External validation

Data sets for external validation of the template CoMFA models were collected from www.bindingdb.org (accessed on 27th April 2014). The resulting structures from an initial search for the respective target (e.g. Urokinase) were filtered for records with pKi or Ki values and further restricted to include one of the individual scaffolds (e.g. amidine-, aniline or guanidine substructures for Urokinase, oxalylarylaminobenzoic-acid and isoxazole-carboxylate for PTP-1B as well as tricycles, ureas and diazepines for Chk1). Finally, a duplicate check was performed to confirm that all external structures were unique with respect to the training set. Alignment of the structures was achieved using the same procedure as for the training sets, i.e. using the same template structures and anchor bonds.

Results

Urokinase

Urokinase is a trypsin-like serine protease that catalyzes the cleavage of the Arg-Val bond in Plasminogen. Typical inhibitors of the urokinase possess a positively charged anchoring group such as amidines and guanidines to mimic Arginine. The set of inhibitors of urokinase in this study consists of three series corresponding to three different anchoring groups. The data set is composed of 51 amidines, 4 guanidines plus 15 anilines. The anilines were included in the original data set despite lacking the charged anchoring group. Three structures have been omitted from the original data set since these do not belong to any of the principal scaffolds.

For template-based alignment the selection of template compounds was based on compounds with a maximally extended structure and high potency in order to cover most of the interaction features with the receptor. We have chosen structures 25, 47 and 51, as templates for the amidine, aniline and guanidine subsets, respectively. These template structures were used to align all 70 structures of the data set (Fig. 1).

The alignments shown in Fig. 2 exemplify the fluctuations of common atom positions for the experimental structures separated by their template structure. Structures overlaid using the template method revealed a highly consistent alignment with almost no fluctuations in the positions of invariant atoms.
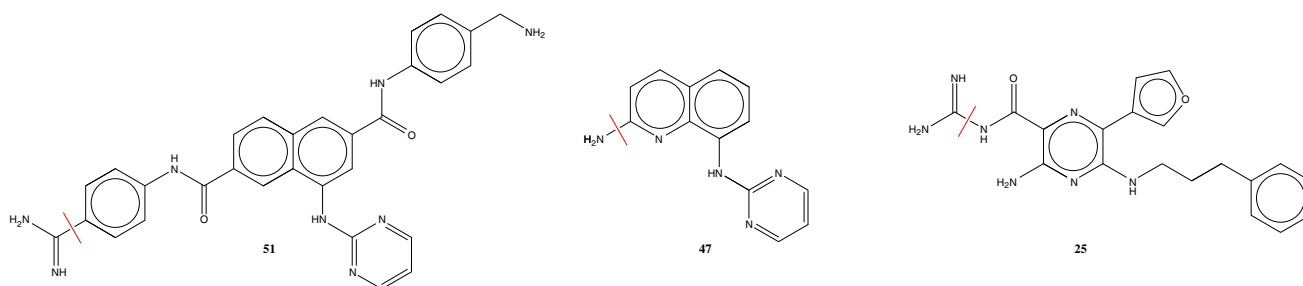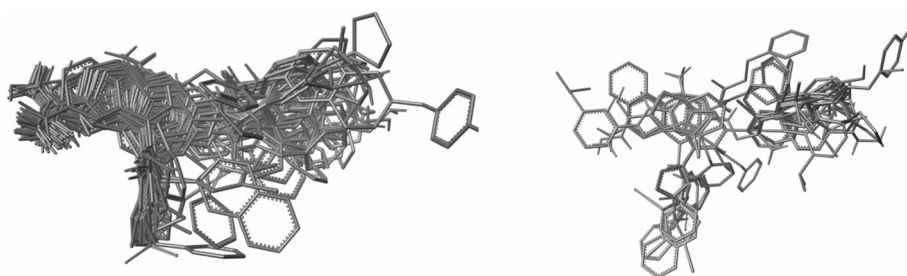
Experimental binding mode

As a prerequisite for tight binding to urokinase the inhibitors usually possess a positively charged anchoring group that forms a salt bridge to Asp189 carboxylate in the S1 pocket. The non-charged anilines are not able to interact as strongly and therefore cannot bind as tightly as the amidines and guanidines. As a consequence the fluctuations are larger among this subset of compounds with several different binding modes merging seamlessly. The 4 guanidine-type of inhibitors although few in numbers also show several binding modes, with 1 Angstrom movements of the central aryl group, an unusual conformation of an aminopyrimidine group in 5-position, and a couple of quite different conformations of sidechains in 4-position of the guanidino-aryl core. In addition, aniline-type inhibitors show a small spread of potencies with values ranging between 4 and 5.6 and a standard deviation of 0.5. Therefore focus will be on the amidine-type of inhibitors. The majority of amidine-type of inhibitors are constructed from a naphthyl-core with variations in 6-, 7-, and 8-position [15–17].

- Variations in the 7-position have only little effect on the compound potency with the sidechains projected into the solvent.
- Phenyl amide groups in 6-position are favored with the amide bond positioned over the active site Ser195 being involved in a hydrogen-bonding network. The phenyl group is oriented in a coplanar arrangement with the naphthyl group and creates hydrophobic contacts to His57. Methylamino groups in the para-position of the

Fig. 1 Template structures selected for the Urokinase data set. Anchor bonds for template-based alignment are indicated with *red bars*

Fig. 2 Alignments of the 70 structures in the Urokinase data set, with X-ray-based alignments on the *left* and template-based-alignments on the *right*

phenyl group form hydrogen bonds to Asp60 carboxylate.

- Amino-linked heteroaromatic rings represent the most favorable substitution of the 8-position of the napthyl-core, with nitrogens of the amino-linker and ortho-positions in the pyrimidine forming hydrogen bonds with Gly216 and Gln192.

CoMFA and CoMSIA models

All 3D-QSAR models show a high degree of consistency, irrespective of the chosen method and grid resolution. In terms of predictivity the models based on the template-based-alignment are the more robust, with q^2-values between 0.66 and 0.69 for template CoMFA compared to q^2-values between 0.53 and 0.58 for the X-ray-based alignment. Electrostatic contributions dominate all models with a fraction of 66 % over steric contributions with a fraction of 33 % (Fig. 3).

The chemical interpretation of the 3D-QSAR models of the two alignments are remarkably similar. However, on closer inspection there are subtle differences that are associated with fluctuation of related atom positions in the X-ray-based-alignment. The green contours around the two aromatic rings attached to the 6- and 8-position of the naphthyl core occupying hydrophobic pockets and therefore represent attractive steric interactions. For the template-based alignment these contours are tightly bound to the ring volumes, however for the X-ray-based-alignment the green contours are shifted away from the ring centers and are more blurred. Similarly, the yellow contours representing repulsive steric interactions around the 7-position are very cloudy and extended for the X-ray-based-alignment but very precisely located for the template-based alignment. For the template-based alignment there are two additional yellow contours around the anchoring group and the 3- and 4-position of the naphthyl-group. These contours result from a subset of inhibitors with an aniline or guanidine anchor group adopting different binding modes (see Fig. 4) as well as benzothiophene-amidines. Since all of the compounds belonging to this subset show low to medium potencies (pKi between 4.0 and 6.5) the steric regions they occupy are interpreted as disfavorable. These

Fig. 3 Contour Maps of steric (*top*) and electrostatic contributions (*bottom*) of CoMSIA-models: *left hand side* based on X-ray-derived ligand poses; *right hand side* based on template CoMFA. The crystal structure of 1SQA is shown with a pocket surface
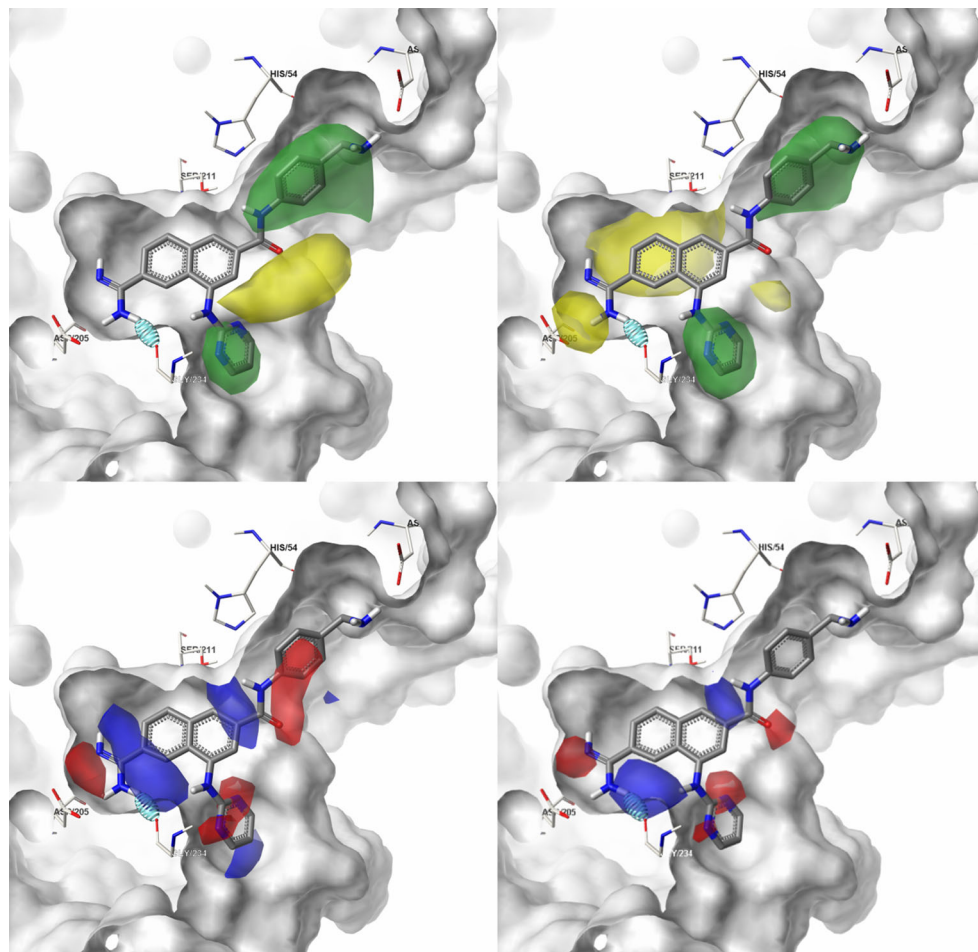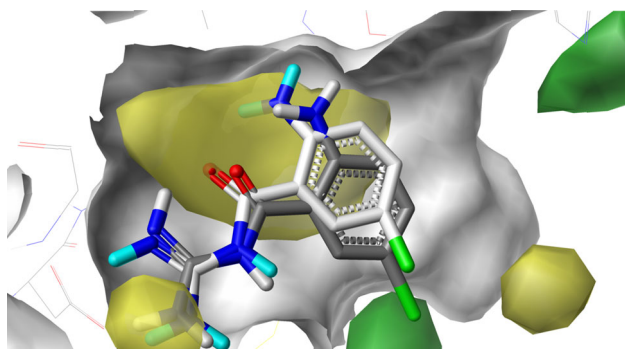
Fig. 4 Contour Maps of steric contributions from the template-CoMSIA-model: compound 17 is shown in their X-ray-ligand pose (*black*) and template-based alignment (*white*). The crystal structure of 1SQA is shown with a binding pocket surface

yellow contours are missing in the X-ray-based alignment. B-factors on guanidine compound 35 (PDB-code: 1FL5) vary between 30 and 44 and indicate a higher flexibility of this type of ligands.

Contours displaying the electrostatic contributions in Fig. 3 also show a high degree of similarity when comparing the X-ray- with the template-based alignment. Blue contours between one of the amidine-nitrogens and the nitrogen of the amino-linker in 8-position of the naphthyl core refer to eletropositive interactions with the carbonyle of Gly216. Red contours around the ortho-positions of the

amino-pyrimidine group in the 8-position of the naphthyl-ring and around the other nitrogen of the amidine-group are highly similar. However a red contour at the carbonyl position of the amide-bond attached to the 6-position of the naphthyl-ring is precisely positioned for the template CoMFA alignment whereas for the X-ray based alignment this contribution is smeared over to the phenyl ring, again a consequence of the greater fluctuations in this part of the X-ray based alignments (see also Fig. 5).

An interesting situation to examine is when a compound is too large for the binding pocket, as is the case for compound 68 (see Fig. 6). In the X-ray based alignment the structure is shifted more than 1 Angstroms outside the binding pocket with its long sidechain in 6-position bent towards the solvent in order to avoid collision with the protein wall of the S1′ pocket. The same sidechain in its template-based alignment, ignorant of the pocket boundary, clashes with the protein wall. Nevertheless the difference in best-fit potencies is rather small: 6.9 versus 7.1 against an actual value of 6.9.

PTP1B

From the 110 original structures of the set of phosphatase inhibitors, the 72 structures with an oxalylarylamino-benzoic acid scaffold and 24 structures with an isoxazole scaffold have been used in this analysis.

Fig. 5 Amidine structures of the Urokinase data set in their X-ray- (*left*) and template-based alignment (*right*). Nitrogen and oxygen positions are shown as atom-colored balls
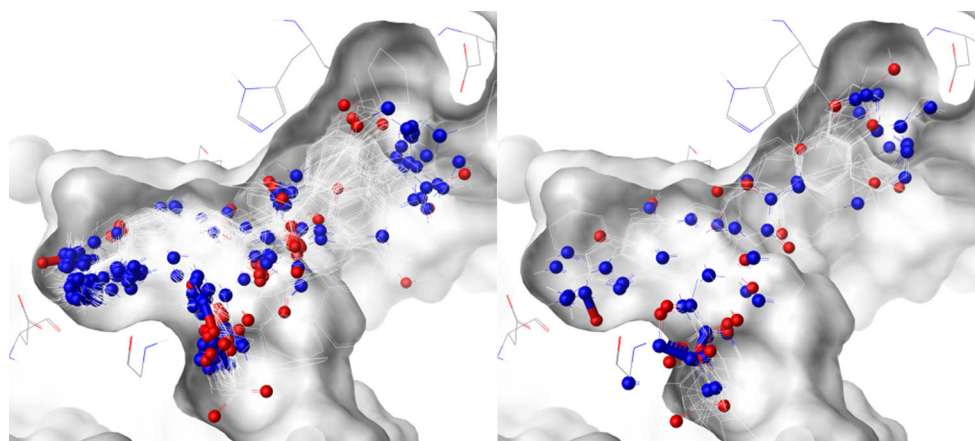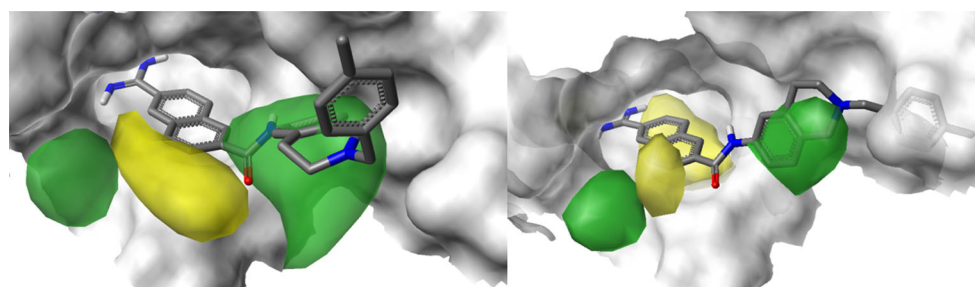

Fig. 6 Compound 68 in its X-ray pose (*left*) and template-based pose (*right*) inside the binding pocket of Urokinase (1SQA) with steric contours from their respective CoMSIA models

Structures 94 and 99 were selected as templates for the formamido and isoxazole series, being the most extended structures having the highest potencies.

Again the X-ray-based alignments show much greater fluctuations in atomic positions than do the template-based alignments.

Experimental binding mode

PTP-1B has two adjacent binding sites for phosphate groups, a catalytic site and a non-catalytic phosphotyrosine binding site (site 2) [18]. The two sets of PTP-1B inhibitors, oxalylarylaminobenzoic-acids and isoxazole-carboxylates, occupy both sites, however they bind to different conformations of the protein: whereas the former binds to the open form the latter binds to the closed conformation with respect to the WPD loop [19–21]. As a result these two subsets show different binding modes for the catalytic site.

oxalylarylaminobenzoic-acids:

- The oxamic acid and benzoic acid binding to the catalytic site form a hydrogen-bonding network with Ser216, Ala217, Ile219 and Arg221, Cys215 resp.
- The nitrogens of the two amide-groups in the linker form hydrogen-bonds to the ASP48 carboxylate
- The oxamic-acid aryl occupies the hydrophobic pocket, with its substituents in the 2- and 3-positions generating contacts to Tyr46 and Val49

isoxazole-carboxylates:

- The isoxazole carboxylate forms bidentate H-bonds with Arg221
- Isoxazole and phenyl rings occupy the hydrophobic pocket with close contacts to Phe182 and Tyr46

For both subsets the binding of the salicylate group to the secondary site is essentially the same:

- Salicylate-ether forms an H-bond with Gln262
- Salicylate-hydroxyl forms H-bonds with Arg24 and Arg254
- Salicylate-carboxylate forms H-bonds with Tyr20, Arg254

The statistical quality of the CoMFA and CoMSIA models for the X-ray-based alignments and the template-based alignments are highly consistent and comparable with q^2-values around 0.78 and SDEP values around 0.6 (see Table 2). Interestingly, the model complexity is rather high with 5–6 components, indicating a lot of exceptions from general rules. In fact, analysing the two subsets on their own yielded robust models with fewer components. The model complexity stems from the different protein conformations where the sidechains and backbone atoms of Trp179, Pro180 and Asp181 change positions to adopt an active open and an inactive closed conformation. The closed conformation restricts the size of the catalytic site and as a consequence the steric interactions of bound ligands can switch from attractive to repulsive when comparing an oxalylarylaminobenzoic-acid with those of an isoxazole-carboxylate.

Chemical interpretation of the contours of the X-ray-based and template based CoMFA and CoMSIA models lead to similar conclusions. Red contours indicating favored negative charge interactions around the salicylic-carboxylate and blue contours favoring positive charge interactions around the salicyl-hydroxyl group are more diffuse for the X-ray-based alignment, with the blue contour for the template-based alignment being particularly precise. The same is true for the amide-groups in the linker part of the structure, the red contours for the carbonyl-oxygen and the blue contour for the nitrogen are more precise for the template-based alignment, the latter contour being more diffuse since it covers the positive charge interactions of both amide groups pointing to Asp48. Interpretation of the electrostatic contours around the catalytic site is more difficult: there are many isolated contours seen for the X-ray-based alignment resulting from the two different binding modes for the oxalylarylamino-benzoic acids and isoxazole-carboxylates and their fluctuating positions. The inconsistent superpositions of carboxylate groups among the different structures give rise to more complex rules resulting in numerous red and blue contours.

Chemical interpretation of the template-based alignment in this region is much clearer. The red contours around the oxalyl-group clearly indicate the H-bond interactions with donors of Ser216, Ala217 and Ile219. No contours appear around the terminal benzoic acid of the oxalylaryl and the isoxazole-carboxylate since the atom positions of these groups are invariant and present in every structure. In addition the isoxazole compounds are in general much weaker in potency so do not contribute significantly to the model; apart from a blue contour where the isoxazole carboxylate is positioned, no other specific contours are seen (Figs. 7, 8).

The steric contours shown in Fig. 9 look rather similar with green contours around the salicyl-group, parts of the linker and the hydrophobic pocket in the catalytic site. Again, the contours for the X-ray poses are more cloudy compared to the template-based alignment. Yellow contours indicating repulsive steric interactions are different. There is a large yellow contour around the linker in the X-ray-based alignment related to several individual structures missing the pharmacophore for binding to the secondary site with deviations from the optimal conformation of the linker group.

In general, yellow contours in the template-based alignment are more distinctive and enable an
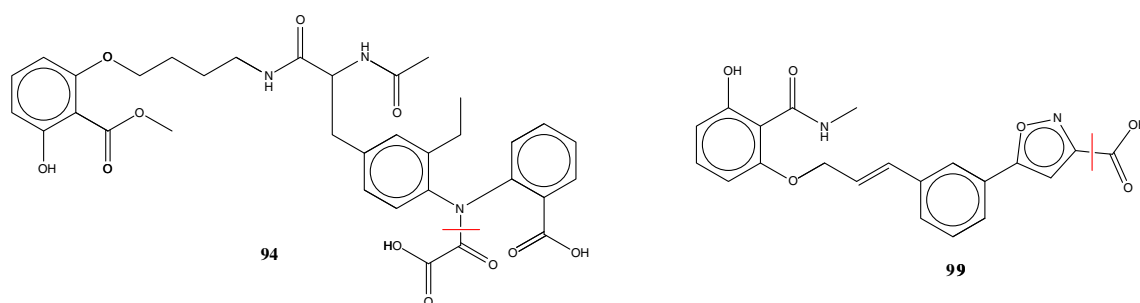
Fig. 7 Template structures selected for the PTP-1B data set. Anchor bonds for template-based alignment are indicated with *red bars*

Fig. 8 Alignments of the 96 structures of the PTP-1B data set, with X-ray-based alignments on the *left* and template-based alignments on the *right*
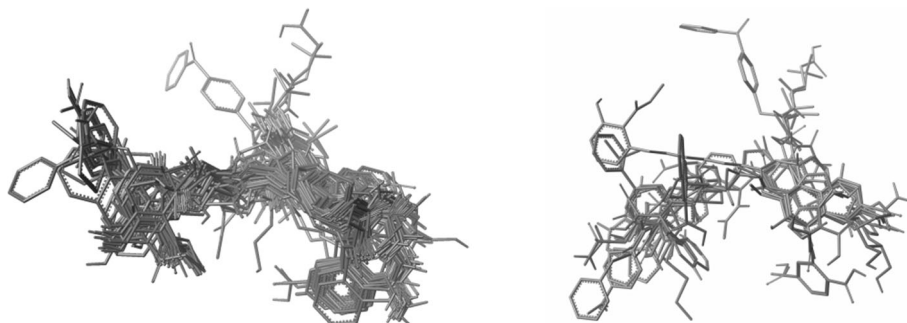
Fig. 9 Contour Maps of electrostatic (*top*) and steric contributions (*bottom*) of CoMSIA-models: *left hand side* based on X-ray-derived ligand poses; *right hand side* based on template CoMFA. The crystal structure of 1PH0 is shown with a pocket surface
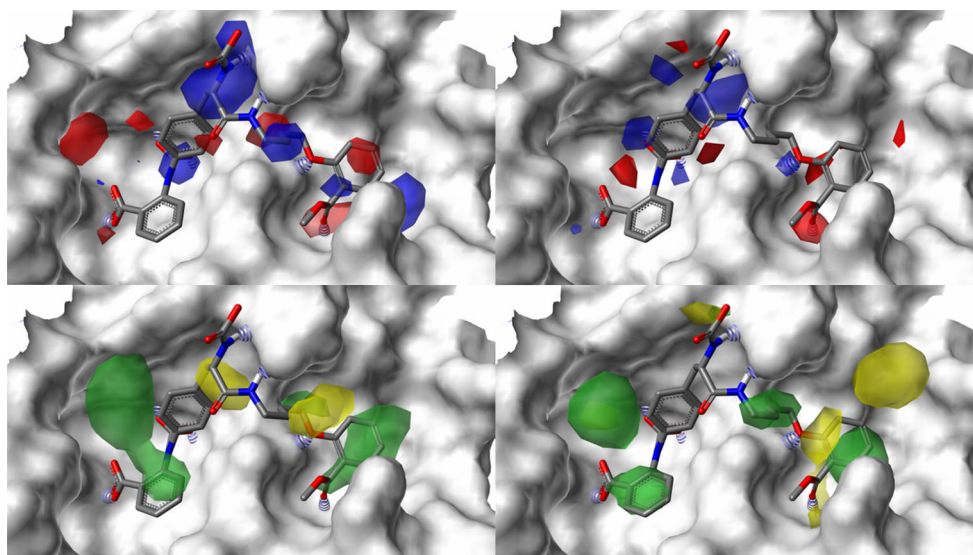
understanding of the actual structure–activity relationships. For example, a yellow contour at the far end of the green contours surrounding the preferred naphthyl-core is worth noting. This contour seems to represent the boundary of the pocket as exemplified with compound 27 in Fig. 10. In the actual X-ray pose compound 27 moves about half an Angstrom out of this pocket with the piperidine ring adopting a more streched conformation to avoid collision with the amine groups of Lys116 and Lys120.

Another yellow contour around substituents in the 5-position of the salicylate-group indicates repulsive steric interactions in that region of the binding pocket. Red electrostatic contours directly at the 5-position imply that single heteroatoms are favored. This results from compounds 4 or 110 with chlorine substituents in that position having high potencies. However larger hydrophobic groups as isopropoxy (compound 39) or phenyl (compound 47, see also Fig. 11) are disfavored since these substituents will interfere with the polar pocket boundary, namely residue Asp29.
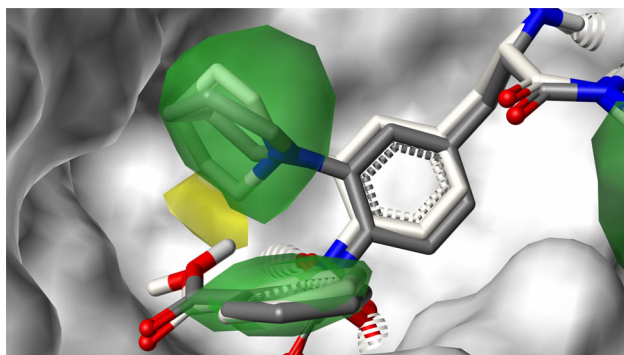
Fig. 10 Compound 27 in its X-ray pose (*black carbons*) and template-based pose (*white carbons*) overlaid with steric contours from the template-based CoMSIA-model

Fig. 12 Compound 26 in its X-ray pose (*black carbons*) and template-based pose (*white carbons*) overlaid with steric contours from the template-based CoMSIA-model

Fig. 11 Compound 47 in its X-ray pose (*black carbons*) and template-based pose (*white carbons*) overlaid with steric contours from the template-based CoMSIA-model

template-based alignment is oriented towards a clash with Tyr46. This flipping of naphthyl-groups is generally associated with lower potencies, explaining the yellow contour around the flipped part of the naphthyl core alluding to repulsive steric interactions. However, in the template-based alignments, all compounds retain the naphthyl-conformation, while the disposition of this group in the protein conformations varies greatly. Without knowledge about the dynamic opportunities that exist for the structure to flip this two-ring system, the predictions of the template based alignments in this region are much less accurate (predicted potency of compound 1 is 4.6 versus an actual value of 3.8). Perhaps, such a localized weakness of a template-based 3D-QSAR might even be suggestive of major local conformational changes among ligands,worthy of X-ray investigation?

### Chk1-kinase

The data set of Chk1-Kinase inhibitors is very diverse with regards to the variation sof the substituents attached to the scaffolds. Analysis focussed on the three series around a diazepin core with 37 structures, around tricyclic pyrazoles with 29 structures and around pyrimidine-ureas with 23 structures.

Selection of templates revealed structures 49, 26 and 84 for the diazepine, trycyclic pyrazole and pyrimidine-urea subsets based on extended structures and high potencies.

Structure- and template-based alignments show the typical difference of magnitude in fluctuations of the invariant atom positions.

Experimental binding modes

All three subsets show different binding modes therefore are described separately:

Another yellow contour at the tert-butoxy group of the structure indicates repulsive steric interactions of bulky groups in this part of the structure (see Fig. 12). Comparison of compound 26 in its protein-bound pose with its template-aligned pose shows that the tert-butoxy group is flipped around the carboxyl group to avoid close contacts with Asp48. The corresponding contour of the model from the X-ray-based alignment does not allow for such a clear interpretation of this structure–activity relationship. In fact, for several X-ray structures flexible end groups of this sidechain are not completely resolved. Those atoms that are resolved show b-factors in the range of 30–50 proving the high flexibility of this solvent exposed sidechain.

Hydrophobic interactions with Tyr46 are important for ligand-binding. In the X-ray-structure of the protein–ligand-complex of compound 1 with PTP-1B (1NO6) the naphthyl-ring is flipped to the other side to better align with the tyrosine residue (see Fig. 13). The same flip of an aromatic ring is seen with compound 30: the compound in the protein-bound pose flips the indole-ring for optimal interactions with Tyr46 whereas the methyl-indole in the

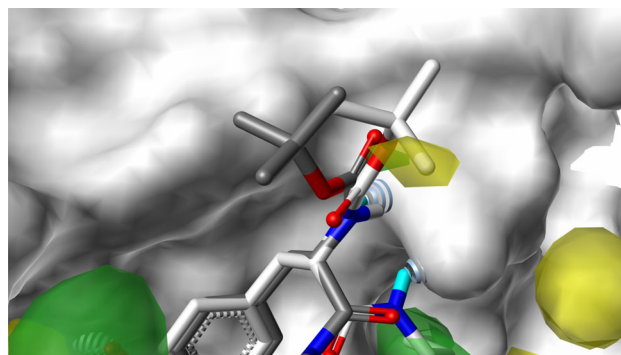**Fig. 13** Alignments of compound 1 (*right*) and compound 30 (*left*) of the PTP-1B data set, with X-ray poses in *black* and template-based alignments in *white*, contours of the CoMSIA model of the X-ray-based alignment are shown
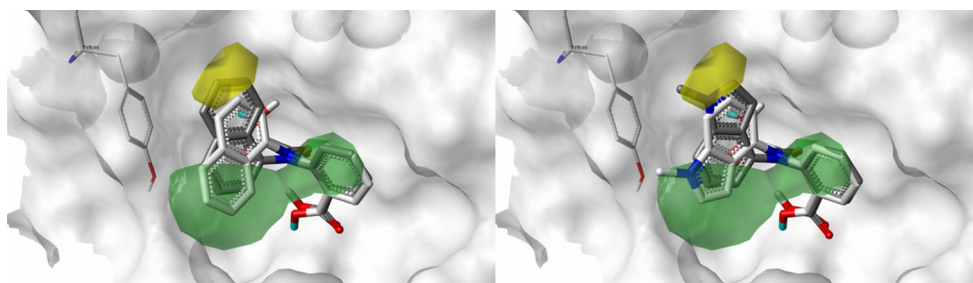
Diazepines [22]:

- The cyclic amide forms hydrogen bonds between the backbone carbonyl and amide nitrogen of Cys87 and anchors the ligand to the hinge region
- Polar group in the sidechain attached to the 4-position of the A-ring forms a hydrogen bond with Lys38
- Substituents on the B-ring point into solvent

Tricycles [23]:

- The two nitrogens of the pyrazole-ring constitute H-bonding partners for the carbonyl of Glu85 and the amide nitrogen of Cys87
- Biphenyl-hydroxyl attached to the 3–position of the pyrazole ring forms hydrogen bonds with Glu55 and Asn59
- Substituents attached to the phenyl-ring of the tricycle are projected into the solvent

Ureas [24]:

- The amine group of the urea attached to the pyrazine ring and the carbonyl of the urea form the anchoring

hydrogen bonds to carbonyle of Glu85 and the amide nitrogen of Cys87

- The cyano-nitrogen in 5-position of the pyrazine forms a hydrogen bond with Lys38.
- The 2-position of the phenyl and the 6-position of the pyrazine point into the ribose pocket. These two positions were used to form a macrocyle and to restrict conformation of the ligand
- Substituents in 5-position of the phenyl point into the solvent
- The nitrogen at the 4-position of the pyrazine forms water-mediated hydrogen bonds with Asn59 (Figs. 14, 15)

3D-QSAR-models

All models, regardless of the alignment, the method, or grid resolution, are statistically weak with low q^2-values and large SDEPs. This weakness is mainly caused by the urea subset that shows a skewed activity distribution coupled with non-linear and non-additive relationships, e.g. the presence or absence of a single amino-function causes a potency shift of 3 log units, the introduction of a nitrile-
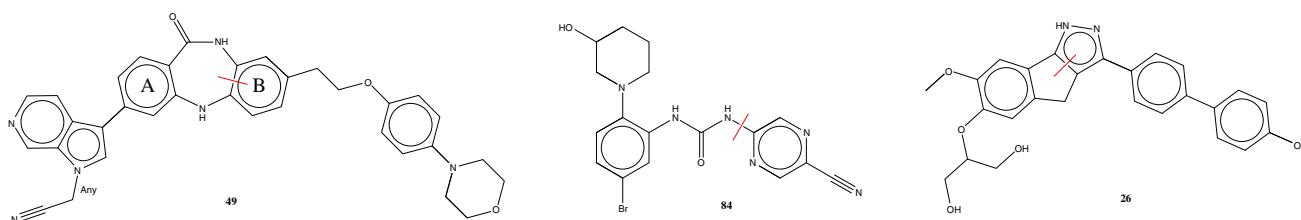
Fig. 14 Template structures selected for the Chk1 data set. Anchor bonds for template-based alignment are indicated with *red bars*

Fig. 15 X-ray-based alignments for 89 structures of the Chk1 data set are shown on the *left* compared with template-based alignments on the *right*
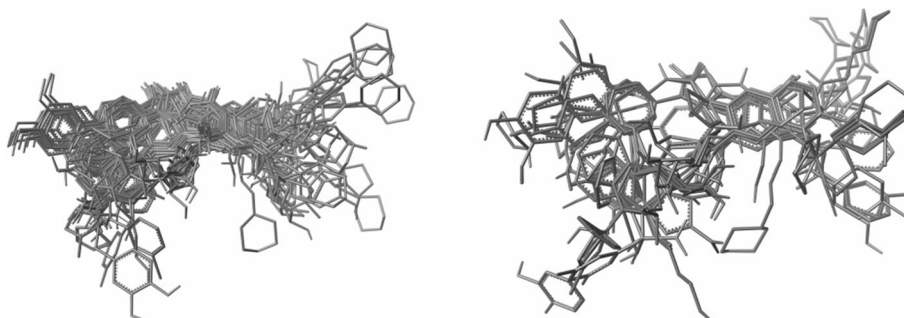
Fig. 16 Contour Maps of steric (*top*) and electrostatic (*bottom*) contributions of CoMSIA-models with Grid resolution of 1A: *left hand side* based on X-ray-derived ligand poses; *right hand side* based on template CoMFA, compound 26 together with the crystal structure of 2E90 with a pocket surface is shown
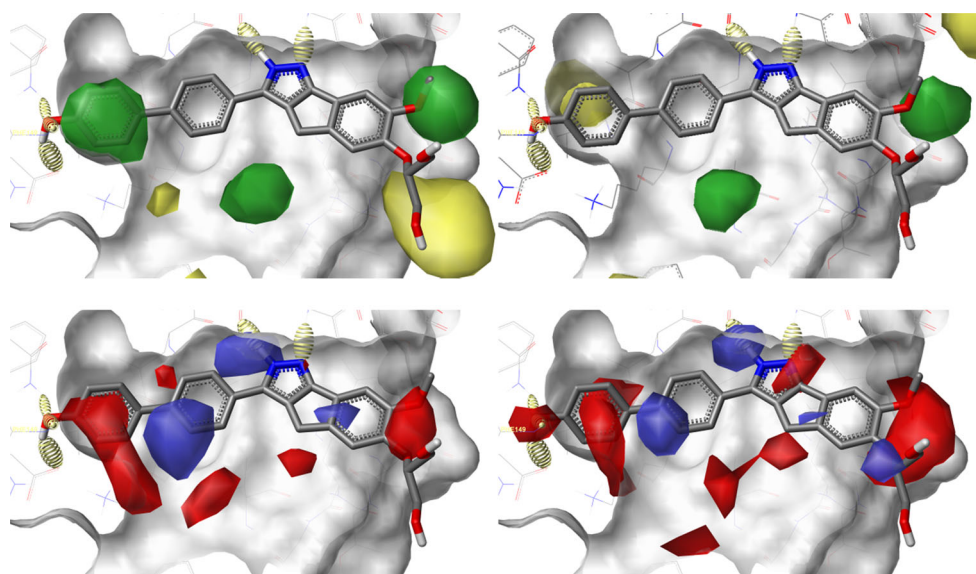
group leads to potency shifts larger than 1 log unit in both directions.

Chemical interpretation of models with poor statistical quality does not deserve much attention. Nevertheless, the electrostatic contours of the template-based alignment clearly show both of the two anchoring hydrogen-bonding interactions. The contours of the X-ray-based alignment omits the interaction with amine of Cys87, because of fluctuations in the corresponding acceptor position in the ligands. Heteroatoms directly attached to the phenyl ring are favored as indicated by the red contours. Conformations of the larger and highly flexible sidechains pointing into the solvent produce inconclusive steric contours (Fig. 16).

Analysis of fluctuations of interaction energies

The 3D-QSAR models for all 3 data sets proved that chemical interpretation is simpler with template CoMFA aligned structures. Presence of electrostatic and steric contours in the display of a 3D-QSAR model are indicative of a correlation between the biological potencies of compounds with the interaction energy values of grid points of the same compounds in particular spatial areas. The question is whether this correlation is stronger for the template CoMFA based models versus the X-ray based models. It is evident that the interaction energies for equivalent atom positions within the X-ray alignments will show a certain degree of variation whereas the interaction energies for equivalent atom positions in the template CoMFA alignments will not show any variation at all. But does this variation have any impact on the correlation between potencies and interaction energies, in other words, is variation of interaction energy values destroying correlation? The template CoMFA model of the Chk-1 data set

(see Fig. 16) presented a red electrostatic contour around the position of the pyrazole nitrogen of the tricycle compound revealing it's donor function forming a hydrogen bond to the amine of Cys87. This red contour is missing from the same representation of the corresponding X-ray based model. We have analysed the individual interaction energies of the grid points surrounding this nitrogen. Figure 17 shows the variation of interaction energies on a particular grid point close the pyrazole nitrogen. As expected, most of the values of the template-CoMFA aligned structures show almost no variation at all. Here, interaction energies are aligned to three discrete values for each of the three individual subsets as this is the only difference in that spatial location. Since electrostatic interactions are long-range, there are still some variations though. The regression line in Fig. 17 shows that with increasing negative electrostatic interaction energies on this grid point the potency increases, confirming that a higher negative charge leads to a stronger favourable hydrogen-bonding interaction. For the template CoMFA aligned structures the distance of the donor atom to this grid point differs for the three subsets, however this difference originates from the orientation of the three template structures used for aligning all structures of the Chk-1 data set and is exactly the same for any structure within its subset. A similar regression line can be seen in the plots for all the other grid points surrounding the pyrazole nitrogen. Obviously, this consistency in correlation is recognized by the PLS-statistics and results in the electrostatic contour shown in Fig. 16.

The interaction energy values of the structures of the X-ray-based alignment show a lot of variation, both within their subsets and overall. The regression line is rather flat, meaning there is no significant correlation
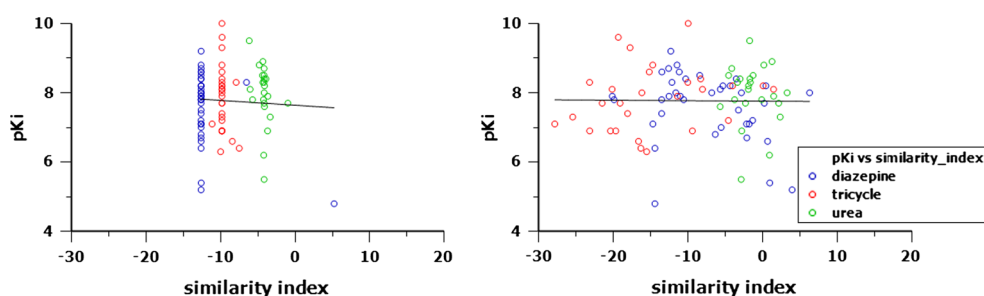
Fig. 17 Plot of potencies for a selected grid point (x = −4; y = 12; z = −19) versus calculated electrostatic CoMSIA similarity indices for the structures of the CHK1 data set. Values calculated for the X-ray structures are on the *left* and for template-based aligned structures on the *right*. Linear regression lines were calculated and are displayed

between the interaction energy values and the potencies of the compounds. However, plots of other grid points around the pyrazole nitrogen show similar correlations as we have seen for the grid points of the template-based alignment. The PLS-statistics is able to identify a similar trend, however, due to the disconnected correlation among those grid points no electrostatic contour results.

External validation

The predictability of the template CoMFA models was validated using external test sets assembled from the Binding Database (see Table 4). For the Urokinase model resulting predictive-R^2 values between 0.7 and 0.76 were slightly superior to internally LOO-validated Q^2-values in the range of 0.66–0.69. Results for the PTP-1B model showed a contrary picture, predictive R^2-values between 0.57 and 0.6 were slightly lower than internally validated Q^2 values between 0.75 and 0.79. Predictive-R^2 values for the Chk1-model showed larger fluctuations with values ranging from 0.09 to 0.52. This compares with rather weak values from internal validation between 0.15 and 0.31. In summary, the external validation proved that predictability of statistically significant template CoMFA models is excellent.

Comparisons of field sources and grid resolutions

The leave-one-out crossvalidated correlation coefficients (see q^2-values in Tables 1, 2, 3, 4, 5 and "Appendix") showed little dependence on field source or grid resolution when models from the same data set and alignment method were compared. Variations of q^2-values are larger for the models based on X-ray ligand poses. There is almost no difference in the predictive abilities of CoMFA and CoMSIA models for the template CoMFA alignments so it is up to the user to choose one over the other. In fact, a difference between the template CoMSIA models can only be detected at the third decimal place. Since the CoMSIA models are easier to interpret visually all 3D-QSAR figures in this paper were based on CoMSIA models.

Our results agree with larger studies published by Mittal recently [25] who concluded that there is only a minimal effect of altering the lattice spacing and that there are no significant differences in predictive ability of using the standard CoMFA and CoMSIA molecular fields.

Comparison of 3D-QSAR predictivity with the original approach

In the original publication of the data sets used in this study the authors have applied a high-throughput formulation of the

Table 1 Urokinase model statistics

Alignment	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
				q^2	SDEP		r^2	S
X-ray	CoMFA	2	70	0.53	0.76	2	0.73	0.58
		1	70	0.56	0.73	2	0.74	0.56
	CoMSIA	2	70	0.58	0.72	2	0.73	0.58
		1	70	0.58	0.72	2	0.73	0.58
TC	CoMFA	2	70	0.69	0.62	2	0.78	0.51
		1	70	0.69	0.62	2	0.78	0.52
	CoMSIA	2	70	0.66	0.64	2	0.76	0.54
		1	70	0.66	0.64	2	0.76	0.54

Grid, grid resolution in A

Table 2 PTP-1B model statistics

Alignment	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
				q^2	SDEP		r^2	S
X-ray	CoMFA	2	96	0.79	0.57	5	0.92	0.35
		1	96	0.78	0.58	5	0.92	0.34
	CoMSIA	2	96	0.75	0.62	5	0.91	0.37
		1	96	0.75	0.62	5	0.91	0.37
TC	CoMFA	2	96	0.79	0.57	6	0.94	0.31
		1	96	0.78	0.59	6	0.94	0.31
	CoMSIA	2	96	0.78	0.58	5	0.91	0.37
		1	96	0.78	0.58	5	0.91	0.37

Grid, grid resolution in A

Table 3 Chk-1 model statistics

Alignment	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
				q^2	SDEP		r^2	S
X-ray	CoMFA	2	89	0.34	0.79	5	0.86	0.36
		1	89	0.28	0.82	5	0.86	0.37
	CoMSIA	2	89	0.14	0.89	2	0.55	0.64
		1	89	0.14	0.89	2	0.55	0.64
TC	CoMFA	2	89	0.29	0.81	3	0.7	0.53
		1	89	0.31	0.8	3	0.7	0.52
	CoMSIA	2	89	0.18	0.88	4	0.69	0.53
		1	89	0.18	0.88	4	0.69	0.53

Grid, grid resolution in A

Table 4 Results from external validation

	N	Q^2	RMSE
Urokinase			
CoMFA(1A)	32	0.72	0.74
CoMFA(2A)	32	0.70	0.74
CoMSIA(1A)	32	0.76	0.69
CoMSIA(2A)	32	0.76	0.69
PTP-1B			
CoMFA(1A)	27	0.60	0.53
CoMFA(2A)	27	0.58	0.54
CoMSIA(1A)	27	0.57	0.55
CoMSIA(2A)	27	0.57	0.55
CHK-1			
CoMFA(1A)	8	0.09	0.42
CoMFA(2A)	8	0.52	0.31
CoMSIA(1A)	8	0.26	0.38
CoMSIA(2A)	8	0.52	0.31

Q^2 = Predictive-r^2

RMSE root-mean-square error

Table 5 Comparison of Pearson correlation coefficients, R, for ligand-potency prediction using htMM-PBSA and leave-one-out cross-validated predictions using 3D-QSAR models versus experimental data for each target

	Urokinase	PTP-1B	Chk-1
Muchmore			
MM-PBSA	0.78	0.84	0.42
MW	0.59	0.85	0.23
X-ray			
CoMFA(1A)	0.71	0.89	0.52
CoMFA(2A)	0.71	0.89	0.59
CoMSIA(1A)	0.76	0.87	0.40
CoMSIA(2A)	0.76	0.87	0.40
TC			
CoMFA(1A)	0.82	0.87	0.55
CoMFA(2A)	0.83	0.89	0.54
CoMSIA(1A)	0.79	0.87	0.35
CoMSIA(2A)	0.79	0.87	0.35

MW molecular weight

molecular mechanics with Poisson–Boltzmann surface area (htMM-PBSA) to estimate relative binding potencies on the ligands [12]. They calculated the Pearson product-moment correlation coefficient as the quality measure of their predictions. These values have been recalculated to reflect the smaller composition of the trainings sets for the three proteins used in this study and correlations coefficients were in the range of 0.42–0.84. This compares with Pearson product-moment correlation coefficients derived from leave-one-out

cross-validated predictions with values between 0.4 and 0.89 and 0.35–0.89 for the X-ray-based and template-based 3D-QSAR models respectively (see Table 4). This modest 3D-QSAR superiority may be significant. Of course, as a practical matter, the contrast between the huge effort of generating the predictions for htMM-PBSA (the authors reported about 200 min per structure) and the low effort of generating a template CoMFA model from a full data set, which takes <5 min, seems highly noteworthy.

Quality of the X-ray data

The quality of the X-ray data showed target-dependent variation. Published PDB-structures were solved within resolution values between 1.6 and 3.1 A, typical for protein–ligand structures with PTP-1B and Chk-1 structures at the lower end and Urokinase structures at the higher end of resolution range. Most of the R-values for the X-ray structures of Urokinase and PTP-1B were reported to be below 0.25, however the majority of X-ray structures for Chk-1 showed R-values beyond this threshold indicating uncertainties with the location or identity of protein or ligand atoms [26] (see Fig. 18). Although corresponding R_{FREE} values were below the threshold of 0.40 [27] for almost all X-ray structures (2 exceptions) the differences between R_{FREE}- and R-values were larger than the ideal threshold of 0.05 [26].

The ligand atoms of the X-ray structures have been examined in detail. Among the X-ray structures for Urokinase the terminal bromine of a bromo-propylether sidechain in 8-position of the naphtyl-core of MOL3 was not resolved. For 7 ligand structures of the PTP-1B data set flexible end groups such as n-butyl, methoxy or ethoxy were not resolved. Deviations from their 2D structures were identified for 21 ligands of the Chk-1 crystal structures. Among those 11 diazepine ligands showed small artifical substituents attached to aryl rings such as CO, CN, OC or NO2. Chlorine substituents on 7 urea ligands were either unresolved or two alternative positions were recorded for the same substituent. There were 2 triazole ligands with CN groups attached to aryl rings. All of the artifactual atom positions have been removed before analysis. None of the unresolved atom positions have been added to the structures.

It turned out that the impact of ligand structure deviations on the QSAR statistics are numerically minor. However, as we have seen with the extended sidechains of oxalylarylaminobenzoic-acid-compounds like t-butoxy (see Fig. 12), when alkyl-groups sticking into the solvent were unresolved, their negative impact on repulsive steric interactions, in this particular sense, hydrophobic interactions, are missing from the crucial structure–activity relationships, this may lead to wrong conclusions.

Discussion

There are numerous references in the literature where CoMFA models derived from template-based alignments are compared with a variety of alternative alignment methods like docking or pharmacophore-based alignments [28]. However, an overall consensus could not be established. In early studies, DePriest reported models of similar quality when he compared template-based alignments with computationally decuded alignments [29]. However, at around the same time, Waller revealed that template-based structures yield superior models [30]. In later studies, Doweyko reported distorted conclusions when he compared template-based models with models developed using binding site constraints [31]. Just recently, Tuccinardi presented favorable results for a multitemplate based model over a docking-based model [32]. Our article is the first publication of a study where the differences between physically realistic X-ray-ligand poses and highly consistent template-based alignments are systematically explored. We have applied a novel template-based method, Template CoMFA (TC), for generation of ligand-centric alignments rooted to bioactive conformations from X-ray-ligand poses. Based on our positive experiences with the topomer CoMFA method [33–35], a protocol that yields highly consistent alignments but with some physically unrealistic conformations, we hypothesized [36] that inconsistent alignments (i.e. X-ray-based alignments) can lead to weaker 3D-QSAR models when compared with models derived from consistent alignments (i.e. template-based alignments). Overall, our

Fig. 18 Distribution of R- and R_{FREE}-values for all X-ray structures of the original data sets. Recommended thresholds are indicated by *reference lines*
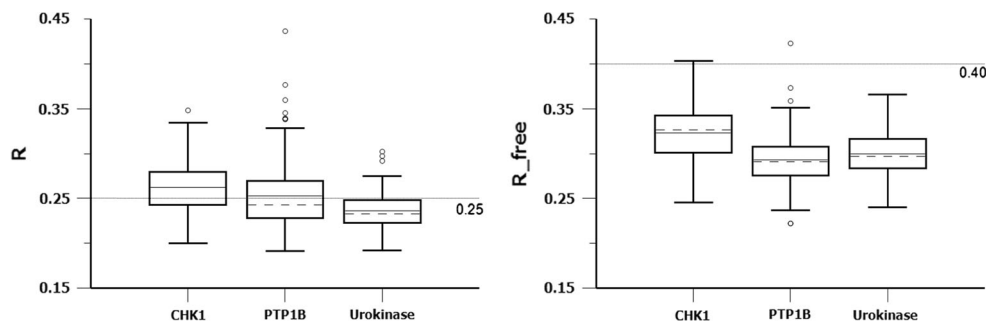
current findings confirm our hypothesis, with the few exceptions noted.

In general, depending on the composition of the data set, partial least squares regression can extract predictive models out of multiple conformations and orientations of identical or similar functional groups, though often by implicitly representing numerous local models. One example is the flipped naphthyl ring within the set of PTP-1B inhibitors. In the case of experimental ligand poses, this behavior makes sense since it better reflects physical realities; different compounds are adopting different binding modes. But on the downside, these realities give rise to blurred images of protein–ligand interactions which can make chemistry guidance harder to extract. In situations where binding site constraints are inconsistent, concurrently enabling two alternative conformations of the same molecular feature, the template-based model will be challenged and may give rise to poor predictions. Nevertheless, in the course of a project, as soon as experimental evidence for a different binding mode becomes available, the template CoMFA method can be updated by including this experimental structure as an additional template. Then all chemically more closely related structures in the training set, i.e. those having more matching atoms with the new template, will be aligned on this new template structure as opposed to chemically less related template structures that will be aligned using the original template(s).

In summary, we report QSAR models derived from X-ray-based alignments to be less balanced and less robust, which is not the outcome that many readers would have anticipated.

We suggest that template- and X-ray-based alignments are complementary methods for 3D-QSAR. With regards to the uncertainties of certain ligand atoms that we have seen in several of the X-ray structures, whenever electron density is missing for determination of the location of ligand atoms, a comparison with a template-based alignment of the same ligand structure can elucidate the orientation of unresolved parts and can serve as reference to mend the incomplete ligand structure. This can be crucial for development of a 3D-QSAR model as incomplete structures may cause inconsistencies that can lead to wrong interpretations of the model.

Over a full project lifetime from hit-to-lead to lead optimization, the template-based models initially depend on rather under-informed alignments because of lack of experimental structures, while an X-ray structure will suggest larger structural changes that may greatly enhance potency. As a series is developed, erratic template-based models may highlight structures where changes to the common binding mode seem to have occurred, worthy of X-ray investigation. Eventually the template-based models will converge to physically realistic models, providing rapid and robust predictions valuable for efficient lead optimization.

Appendix

See Tables 6, 7, 8, 9, 10, 11 and 12.

Table 6 Urokinase subset model statistics

Subset	Align	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
					q^2	SDEP		r^2	S
Amidines	X-ray	CoMFA	1	51	0.47	0.67	2	0.72	0.48
			2	51	0.46	0.67	2	0.72	0.49
		CoMSIA	1	51	0.46	0.68	2	0.68	0.52
			2	51	0.46	0.68	2	0.68	0.52
	TC	CoMFA	1	15	0.63	0.57	3	0.81	0.41
			2	15	0.61	0.57	2	0.73	0.47
		CoMSIA	1	15	0.6	0.58	3	0.8	0.41
			2	15	0.6	0.58	3	0.8	0.41
Anilines	X-ray	CoMFA	1	15	0.18	0.45	1	0.62	0.31
			2	15	0.21	0.44	1	0.64	0.3
		CoMSIA	1	15	0.04	0.49	1	0.53	0.34
			2	15	0.04	0.49	1	0.53	0.34
	TC	CoMFA	1	15	0	0.5	1	0.5	0.35
			2	15	0.02	0.49	1	0.51	0.35
		CoMSIA	1	15	−0.06	0.51	1	0.48	0.36
			2	15	−0.06	0.51	1	0.48	0.36

Table 7 PTP-1B subset model statistics

Subset	Align	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
					q^2	SDEP		r^2	s
Oxalyl	X-ray	CoMFA	2	72	0.7	0.57	4	0.92	0.3
			1	72	0.71	0.57	4	0.92	0.29
		CoMSIA	2	72	0.67	0.6	4	0.91	0.32
			1	72	0.67	0.6	4	0.91	0.32
	TC	CoMFA	2	72	0.64	0.63	4	0.89	0.35
			1	72	0.67	0.6	4	0.89	0.34
		CoMSIA	2	72	0.65	0.62	4	0.86	0.39
			1	72	0.65	0.62	4	0.86	0.39
Isoxazoles	PDB	CoMFA	2	24	0.67	0.46	3	0.96	0.17
			1	24	0.67	0.47	3	0.95	0.18
		CoMSIA	2	24	0.61	0.49	2	0.89	0.26
			1	24	0.61	0.49	2	0.89	0.26
	TC	CoMFA	2	24	0.65	0.45	1	0.74	0.39
			1	24	0.64	0.46	1	0.73	0.4
		CoMSIA	2	24	0.64	0.46	1	0.74	0.39
			1	24	0.64	0.46	1	0.74	0.39

Table 8 Chk1 subset model statistics

Subset	Align	Metric	Grid	n	LOO cross-valid		# Cmp	Best fit	
					q^2	SDEP		r^2	s
Diazepines	X-ray	CoMFA	2	37	0.45	0.77	3	0.92	0.3
			1	37	0.46	0.76	3	0.92	0.29
		CoMSIA	2	37	0.35	0.82	2	0.87	0.36
			1	37	0.35	0.82	2	0.87	0.36
	TC	CoMFA	2	37	0.38	0.81	3	0.81	0.46
			1	37	0.37	0.82	3	0.8	0.46
		CoMSIA	2	37	0.29	0.87	3	0.82	0.44
			1	37	0.29	0.87	3	0.82	0.44
Tricycles	X-ray	CoMFA	2	29	0.65	0.61	5	0.99	0.1
			1	29	0.67	0.59	5	0.99	0.09
		CoMSIA	2	29	0.54	0.68	4	0.96	0.19
			1	29	0.54	0.68	4	0.96	0.19
	TC	CoMFA	2	29	0.59	0.64	4	0.94	0.25
			1	29	0.53	0.67	3	0.86	0.36
		CoMSIA	2	29	0.56	0.67	5	0.98	0.16
			1	29	0.56	0.67	5	0.98	0.16
Ureas	X-ray	CoMFA	2	23	0.14	0.85	2	0.75	0.46
			1	23	0.13	0.86	2	0.74	0.47
		CoMSIA	2	23	0.02	0.89	1	0.57	0.59
			1	23	0.02	0.89	1	0.57	0.59
	TC	CoMFA	2	23	0.02	0.89	1	0.53	0.61
			1	23	0.22	0.81	2	0.71	0.49
		CoMSIA	2	23	−0.04	0.91	1	0.54	0.6
			1	23	−0.04	0.91	1	0.54	0.6

Table 9 Fraction of variance

Target	Subset	Metric	Grid	Alignment	Org_Var
CHK1	ALL	Comfa	1	TC	0.88
CHK1	ALL	Comfa	1	X-ray	0.88
CHK1	ALL	Comfa	2	TC	0.88
CHK1	ALL	Comfa	2	X-ray	0.88
CHK1	ALL	Comsia	1	TC	0.88
CHK1	ALL	Comsia	1	X-ray	0.88
CHK1	ALL	Comsia	2	TC	0.88
CHK1	ALL	Comsia	2	X-ray	0.88
CHK1	DIAZEPAMS	Comfa	1	TC	0.95
CHK1	DIAZEPAMS	Comfa	1	X-ray	0.95
CHK1	DIAZEPAMS	Comfa	2	TC	0.95
CHK1	DIAZEPAMS	Comfa	2	X-ray	0.95
CHK1	DIAZEPAMS	Comsia	1	TC	0.95
CHK1	DIAZEPAMS	Comsia	1	X-ray	0.95
CHK1	DIAZEPAMS	Comsia	2	TC	0.95
CHK1	DIAZEPAMS	Comsia	2	X-ray	0.95
CHK1	TRICYCLES	Comfa	1	TC	0.82
CHK1	TRICYCLES	Comfa	1	X-ray	0.82
CHK1	TRICYCLES	Comfa	2	TC	0.82
CHK1	TRICYCLES	Comfa	2	X-ray	0.82
CHK1	TRICYCLES	Comsia	1	TC	0.82
CHK1	TRICYCLES	Comsia	1	X-ray	0.82
CHK1	TRICYCLES	Comsia	2	TC	0.82
CHK1	TRICYCLES	Comsia	2	X-ray	0.82
CHK1	UREAS	Comfa	1	TC	0.73
CHK1	UREAS	Comfa	1	X-ray	0.73
CHK1	UREAS	Comfa	2	TC	0.73
CHK1	UREAS	Comfa	2	X-ray	0.73
CHK1	UREAS	Comsia	1	TC	0.73
CHK1	UREAS	Comsia	1	X-ray	0.73
CHK1	UREAS	Comsia	2	TC	0.73
CHK1	UREAS	Comsia	2	X-ray	0.73
PTP1B	ALL	Comfa	1	TC	1.45
PTP1B	ALL	Comfa	1	X-ray	1.45
PTP1B	ALL	Comfa	2	TC	1.45
PTP1B	ALL	Comfa	2	X-ray	1.45
PTP1B	ALL	Comsia	1	TC	1.45
PTP1B	ALL	Comsia	1	X-ray	1.45
PTP1B	ALL	Comsia	2	TC	1.45
PTP1B	ALL	Comsia	2	X-ray	1.45
PTP1B	FORMAMIDOS	Comfa	1	TC	1.01
PTP1B	FORMAMIDOS	Comfa	1	X-ray	1.01
PTP1B	FORMAMIDOS	Comfa	2	TC	1.01
PTP1B	FORMAMIDOS	Comfa	2	X-ray	1.01
PTP1B	FORMAMIDOS	Comsia	1	TC	1.01
PTP1B	FORMAMIDOS	Comsia	1	X-ray	1.01
PTP1B	FORMAMIDOS	Comsia	2	TC	1.01
PTP1B	FORMAMIDOS	Comsia	2	X-ray	1.01

Table 9 continued

Target	Subset	Metric	Grid	Alignment	Org_Var
PTP1B	ISOXAZOLES	Comfa	1	TC	0.54
PTP1B	ISOXAZOLES	Comfa	1	X-ray	0.54
PTP1B	ISOXAZOLES	Comfa	2	TC	0.54
PTP1B	ISOXAZOLES	Comfa	2	X-ray	0.54
PTP1B	ISOXAZOLES	Comsia	1	TC	0.54
PTP1B	ISOXAZOLES	Comsia	1	X-ray	0.54
PTP1B	ISOXAZOLES	Comsia	2	TC	0.54
PTP1B	ISOXAZOLES	Comsia	2	X-ray	0.54
UK	ALL	Comfa	1	TC	1.17
UK	ALL	Comfa	1	X-ray	1.17
UK	ALL	Comfa	2	TC	1.17
UK	ALL	Comfa	2	X-ray	1.17
UK	ALL	Comsia	1	TC	1.17
UK	ALL	Comsia	1	X-ray	1.17
UK	ALL	Comsia	2	TC	1.17
UK	ALL	Comsia	2	X-ray	1.17
UK	AMIDINES	Comfa	1	TC	0.79
UK	AMIDINES	Comfa	1	X-ray	0.79
UK	AMIDINES	Comfa	2	TC	0.79
UK	AMIDINES	Comfa	2	X-ray	0.79
UK	AMIDINES	Comsia	1	TC	0.79
UK	AMIDINES	Comsia	1	X-ray	0.79
UK	AMIDINES	Comsia	2	TC	0.79
UK	AMIDINES	Comsia	2	X-ray	0.79
UK	ANILINES	Comfa	1	TC	0.21
UK	ANILINES	Comfa	1	X-ray	0.21
UK	ANILINES	Comfa	2	TC	0.21
UK	ANILINES	Comfa	2	X-ray	0.21
UK	ANILINES	Comsia	1	TC	0.21
UK	ANILINES	Comsia	1	X-ray	0.21
UK	ANILINES	Comsia	2	TC	0.21
UK	ANILINES	Comsia	2	X-ray	0.21
UK	GUANIDINES	Comfa	1	TC	0.3
UK	GUANIDINES	Comfa	1	X-ray	0.3
UK	GUANIDINES	Comfa	2	TC	0.3
UK	GUANIDINES	Comfa	2	X-ray	0.3
UK	GUANIDINES	Comsia	1	TC	0.3
UK	GUANIDINES	Comsia	1	X-ray	0.3
UK	GUANIDINES	Comsia	2	TC	0.3
UK	GUANIDINES	Comsia	2	X-ray	0.3

Target	Latent variable									
	1 R^2	2	3	4	5	6	7	8	9	10
CHK1	0.31	0.49	0.7	0.74	0.8	0.87	0.89	0.91	0.93	0.94
CHK1	0.29	0.6	0.68	0.76	0.86	0.9	0.94	0.97	0.99	0.99
CHK1	0.28	0.47	0.7	0.73	0.81	0.85	0.88	0.91	0.92	0.94
CHK1	0.28	0.59	0.68	0.76	0.86	0.89	0.93	0.96	0.98	0.99
CHK1	0.27	0.4	0.61	0.69	0.81	0.84	0.86	0.89	0.91	0.93

Table 9 continued

Target	Latent variable									
	1 R^2	2	3	4	5	6	7	8	9	10
CHK1	0.29	0.55	0.64	0.76	0.84	0.88	0.92	0.94	0.96	0.97
CHK1	0.27	0.4	0.61	0.69	0.81	0.84	0.86	0.89	0.91	0.93
CHK1	0.29	0.55	0.64	0.76	0.84	0.88	0.92	0.94	0.96	0.97
CHK1	0.46	0.72	0.81	0.9	0.93	0.95	0.98	0.99	1	1
CHK1	0.66	0.79	0.92	0.96	0.98	0.99	1	1	1	1
CHK1	0.44	0.72	0.8	0.9	0.93	0.95	0.98	0.99	1	1
CHK1	0.65	0.79	0.92	0.96	0.98	0.99	1	1	1	1
CHK1	0.39	0.73	0.82	0.88	0.91	0.94	0.96	0.98	0.99	0.99
CHK1	0.66	0.87	0.92	0.96	0.98	0.99	0.99	1	1	1
CHK1	0.39	0.73	0.82	0.88	0.91	0.94	0.96	0.98	0.99	0.99
CHK1	0.66	0.87	0.92	0.96	0.98	0.99	0.99	1	1	1
CHK1	0.59	0.83	0.87	0.94	0.98	0.99	0.99	0.99	1	1
CHK1	0.54	0.69	0.95	0.98	0.99	1	1	1	1	1
CHK1	0.59	0.82	0.86	0.93	0.96	0.98	0.99	0.99	0.99	1
CHK1	0.53	0.7	0.95	0.98	0.99	1	1	1	1	1
CHK1	0.54	0.8	0.93	0.96	0.98	0.99	0.99	1	1	1
CHK1	0.55	0.82	0.94	0.96	0.99	1	1	1	1	1
CHK1	0.54	0.8	0.93	0.96	0.98	0.99	0.99	1	1	1
CHK1	0.55	0.82	0.94	0.96	0.99	1	1	1	1	1
CHK1	0.53	0.72	0.75	0.81	0.88	0.95	0.98	0.99	1	1
CHK1	0.37	0.75	0.92	0.96	0.99	1	1	1	1	1
CHK1	0.52	0.71	0.76	0.83	0.88	0.95	0.98	0.99	1	1
CHK1	0.38	0.74	0.91	0.95	0.99	1	1	1	1	1
CHK1	0.54	0.67	0.73	0.78	0.89	0.96	0.97	0.98	0.99	0.99
CHK1	0.57	0.79	0.92	0.96	0.98	0.99	1	1	1	1
CHK1	0.54	0.67	0.73	0.78	0.89	0.96	0.97	0.98	0.99	0.99
CHK1	0.57	0.79	0.92	0.96	0.98	0.99	1	1	1	1
PTP1B	0.7	0.76	0.83	0.89	0.91	0.94	0.95	0.97	0.98	0.98
PTP1B	0.69	0.79	0.82	0.9	0.92	0.95	0.97	0.98	0.99	0.99
PTP1B	0.7	0.77	0.83	0.89	0.92	0.94	0.95	0.97	0.97	0.98
PTP1B	0.69	0.79	0.82	0.9	0.92	0.95	0.97	0.98	0.99	0.99
PTP1B	0.7	0.75	0.85	0.89	0.91	0.94	0.95	0.96	0.97	0.98
PTP1B	0.69	0.75	0.81	0.89	0.91	0.94	0.95	0.96	0.97	0.98
PTP1B	0.7	0.75	0.85	0.89	0.91	0.94	0.95	0.96	0.97	0.98
PTP1B	0.69	0.75	0.81	0.89	0.91	0.94	0.95	0.96	0.97	0.98
PTP1B	0.6	0.74	0.84	0.89	0.92	0.94	0.96	0.97	0.97	0.98
PTP1B	0.58	0.71	0.87	0.92	0.94	0.97	0.98	0.99	0.99	1
PTP1B	0.6	0.75	0.84	0.89	0.92	0.93	0.95	0.96	0.97	0.98
PTP1B	0.59	0.71	0.87	0.92	0.94	0.96	0.98	0.98	0.99	0.99
PTP1B	0.58	0.73	0.82	0.86	0.9	0.92	0.94	0.96	0.97	0.98
PTP1B	0.58	0.7	0.86	0.91	0.93	0.94	0.96	0.97	0.97	0.98
PTP1B	0.58	0.73	0.82	0.86	0.9	0.92	0.94	0.96	0.97	0.98
PTP1B	0.58	0.7	0.86	0.91	0.93	0.94	0.96	0.97	0.97	0.98
PTP1B	0.74	0.83	0.92	0.93	0.96	0.97	0.97	0.98	0.98	0.99
PTP1B	0.69	0.92	0.96	0.98	0.99	0.99	1	1	1	1
PTP1B	0.73	0.83	0.92	0.94	0.96	0.97	0.98	0.98	0.99	0.99

Table 9 continued

Target	Latent variable									
	1 R^2	2	3	4	5	6	7	8	9	10
PTP1B	0.68	0.91	0.95	0.98	0.99	0.99	1	1	1	1
PTP1B	0.74	0.84	0.92	0.94	0.96	0.97	0.98	0.98	0.99	0.99
PTP1B	0.68	0.89	0.96	0.97	0.98	0.99	1	1	1	1
PTP1B	0.74	0.84	0.92	0.94	0.96	0.97	0.98	0.98	0.99	0.99
PTP1B	0.68	0.89	0.96	0.97	0.98	0.99	1	1	1	1
UK	0.63	0.78	0.85	0.9	0.92	0.94	0.96	0.97	0.98	0.98
UK	0.57	0.74	0.83	0.9	0.94	0.96	0.98	0.99	0.99	1
UK	0.62	0.78	0.86	0.89	0.92	0.94	0.95	0.96	0.97	0.98
UK	0.56	0.73	0.84	0.89	0.93	0.96	0.97	0.99	0.99	0.99
UK	0.61	0.76	0.85	0.9	0.92	0.94	0.95	0.95	0.96	0.96
UK	0.56	0.73	0.79	0.85	0.88	0.91	0.93	0.95	0.96	0.97
UK	0.61	0.76	0.85	0.9	0.92	0.94	0.95	0.95	0.96	0.96
UK	0.56	0.73	0.79	0.85	0.88	0.91	0.93	0.95	0.96	0.97
UK	0.59	0.73	0.81	0.87	0.92	0.94	0.96	0.97	0.98	0.98
UK	0.55	0.72	0.83	0.91	0.95	0.97	0.98	0.99	1	1
UK	0.6	0.73	0.81	0.86	0.91	0.93	0.95	0.96	0.97	0.98
UK	0.56	0.72	0.83	0.92	0.96	0.98	0.98	0.99	0.99	1
UK	0.58	0.72	0.8	0.87	0.9	0.92	0.93	0.94	0.95	0.95
UK	0.52	0.68	0.76	0.84	0.88	0.91	0.94	0.95	0.97	0.98
UK	0.58	0.72	0.8	0.87	0.9	0.92	0.93	0.94	0.95	0.95
UK	0.52	0.68	0.76	0.84	0.88	0.91	0.94	0.95	0.97	0.98
UK	0.5	0.68	0.81	0.89	0.93	0.97	0.98	1	1	1
UK	0.62	0.87	0.95	0.99	1	1	1	1	1	1
UK	0.51	0.67	0.8	0.91	0.93	0.97	0.99	0.99	1	1
UK	0.64	0.88	0.96	0.99	1	1	1	1	1	1
UK	0.48	0.7	0.78	0.83	0.85	0.88	0.91	0.94	0.96	0.98
UK	0.53	0.75	0.93	0.97	0.99	1	1	1	1	1
UK	0.48	0.7	0.78	0.83	0.85	0.88	0.91	0.94	0.96	0.98
UK	0.53	0.75	0.93	0.97	0.99	1	1	1	1	1
UK	0.99	1	1	1	1	1	1	1	1	1
UK	0.97	1	1	1	1	1	1	1	1	1
UK	0.99	0.99	1	1	1	1	1	1	1	1
UK	0.95	1	1	1	1	1	1	1	1	1
UK	0.98	1	1	1	1	1	1	1	1	1
UK	0.98	1	1	1	1	1	1	1	1	1
UK	0.98	1	1	1	1	1	1	1	1	1
UK	0.98	1	1	1	1	1	1	1	1	1

Table 10 External validation—predictions for Urokinase test set

Name	pKi	Pred_comfa_1	Pred_comfa_2	Pred_comsia_1	Pred_comsia_2
BindingDB_50091818	4.5229	4.4029	4.3789	4.4065	4.4065
BindingDB_92286	4.301	4.6647	4.6372	4.6779	4.6779
BindingDB_92300	4.301	4.5132	4.5065	4.5458	4.5458
BindingDB_92324	5.1379	6.2318	6.1631	6.1354	6.1354
BindingDB_92310	4	4.6338	4.6681	4.4357	4.4357
BindingDB_92312	4	4.5234	4.543	4.2243	4.2243
BindingDB_92307	5	5.3097	5.3885	5.2578	5.2578
BindingDB_92303	4.301	5.3891	5.5293	5.3093	5.3093
BindingDB_92295	5.0953	6.1574	6.142	6.0451	6.0451
BindingDB_92323	4	4.9328	4.8319	5.2974	5.2974
BindingDB_92311	4	4.7198	4.7686	4.5932	4.5932
BindingDB_92288	5.3468	5.9573	5.924	6.2143	6.2143
BindingDB_92290	5.0605	5.2813	5.3333	5.6678	5.6678
BindingDB_92301	6.1772	6.1052	6.1329	6.0755	6.0755
BindingDB_92313	7.7447	7.9149	7.936	7.7327	7.7327
BindingDB_92297	5.1051	6.8762	7.0216	6.5295	6.5295
BindingDB_92314	5.8268	7.3539	7.4949	6.9848	6.9848
BindingDB_92304	7.4815	7.1758	7.0656	7.0625	7.0625
BindingDB_92291	4.3979	4.8921	4.9875	4.5582	4.5582
BindingDB_92298	5.4473	5.4385	5.3555	5.4504	5.4504
BindingDB_92292	9.0458	8.2948	8.2854	8.0136	8.0136
BindingDB_92319	7.3768	6.6754	6.6192	6.9173	6.9173
BindingDB_92305	6.5376	6.4496	6.4883	6.5155	6.5155
BindingDB_92318	6.6517	6.6116	6.674	6.6065	6.6065
BindingDB_92308	7.1805	6.4824	6.5265	6.5229	6.5229
BindingDB_92320	7.0269	6.8392	6.8151	6.9894	6.9894
BindingDB_92322	5.9031	6.6953	6.6342	6.8773	6.8773
BindingDB_92296	6.1141	6.8357	6.9036	6.8165	6.8165
BindingDB_92302	7.7959	6.9393	6.8893	6.9997	6.9997
BindingDB_92321	8.0458	6.7353	6.8341	7.0683	7.0683
BindingDB_92316	7.6778	7.2872	7.3701	7.463	7.463
BindingDB_92315	6.1096	6.5951	6.5931	6.5018	6.5018

Table 11 External validation—predictions for PTP-1B test set

Name	pKi	Pred_comfa_1	Pred_comfa_2	Pred_comsia_1	Pred_comsia_2
BINDINGDB_13951	4.0315	4.2512	4.3475	4.3297	4.3297
BINDINGDB_13961	4.4559	4.7971	4.7398	4.6338	4.6338
BINDINGDB_13964	4.9586	5.2288	5.2913	5.13	5.13
BINDINGDB_13991	3.6655	4.9604	4.9526	4.8424	4.8424
BINDINGDB_13992	3.9136	4.1555	4.1435	3.9207	3.9207
BINDINGDB_13997	6.0362	5.2068	5.167	5.1302	5.1302
BINDINGDB_13968	5.0088	4.9471	4.9618	5.1929	5.1929
BINDINGDB_13971	5.9208	5.8273	5.7802	5.723	5.723
BINDINGDB_13969	5.041	5.1051	5.1809	5.3549	5.3549
BINDINGDB_13972	5.9208	6.0985	6.0381	5.9224	5.9224
BINDINGDB_13973	5.8239	6.0863	6.1127	5.9349	5.9349
BINDINGDB_15809	5.6021	5.7302	5.672	5.6459	5.6459

Table 11 continued

Name	pKi	Pred_comfa_1	Pred_comfa_2	Pred_comsia_1	Pred_comsia_2
BINDINGDB_15808	5.3279	5.8657	5.8751	5.7012	5.7012
BINDINGDB_15807	5.9208	5.831	5.8595	5.612	5.612
BINDINGDB_13974	5.699	6.2342	6.2142	6.1083	6.1083
BINDINGDB_15806	6.2676	5.921	5.931	5.9221	5.9221
BINDINGDB_50414197	5	4.4536	4.5241	5.0556	5.0556
BINDINGDB_15810	5.4685	6.131	6.2437	5.9105	5.9105
BINDINGDB_50131551	5.7959	6.8084	6.8828	6.7239	6.7239
BINDINGDB_50131553	6.8861	6.7093	6.7892	6.7052	6.7052
BINDINGDB_15814	6.3665	6.0919	6.0579	5.8804	5.8804
BINDINGDB_50326165	7.7447	6.2729	6.384	5.8005	5.8005
BINDINGDB_50131546	6.4318	7.031	7.1945	7.1043	7.1043
BINDINGDB_15813	6.6021	6.3088	6.257	6.0257	6.0257
BINDINGDB_15811	5.8861	6.3034	6.3616	6.1051	6.1051
BINDINGDB_50131548	6.2518	6.9354	6.9988	6.9106	6.9106
BINDINGDB_50131544	6.4089	6.9902	7.0159	6.8744	6.8744

Table 12 External validation—predictions for CHK-1 test set

Name	pKi	Pred_comfa_2	Pred_comsia_2	Pred_comfa_1	Pred_comsia_1
BindingDB_91824	6.2	7.4323	7.6415	7.4367	7.6415
BindingDB_91939	7.1	6.9972	7.0688	6.9881	7.0688
BindingDB_50221585_558	7.9	8.7781	8.2336	8.6838	8.2336
BindingDB_50414074	8.6003	8.159	8.3386	8.0912	8.3386
BindingDB_50414087	8	8.3337	8.0748	8.25	8.0748
BindingDB_91857	9.3002	7.6937	8.4548	8.0244	8.4548
BindingDB_50414248	8	7.566	8.1709	7.4656	8.1709
BindingDB_91861	7.3	6.6679	7.1161	6.7629	7.1161

References

1. Martin YC (1998). 3D QSAR: current state, scope, and limitations. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in drug design. Springer, Netherlands, pp 3–23
2. Sippl W (2002) Development of biologically active compounds by combining 3D QSAR and structure-based design methods. J Comput Aided Mol 16:825–830
3. Matter H, Kotsonis P, Klingler O, Strobel H, Fröhlich LG, Frey A, Pfleiderer W, Schmidt HH (2002) Structural requirements for inhibition of the neuronal nitric oxide synthase (NOS-I): 3D-QSAR analysis of 4-oxo-and 4-amino-pteridine-based inhibitors. J Med Chem 45:2923–2941
4. Weber A, Böhm M, Supuran CT, Scozzafava A, Sotriffer CA, Klebe G (2006) 3D QSAR selectivity analyses of carbonic anhydrase inhibitors: insights for the design of isozyme selective inhibitors. J Chem Inf Model 46:2737–2760
5. Hillebrecht A, Klebe G (2008) Use of 3D QSAR models for database screening: a feasibility study. J Chem Inf Model 48:384–396
6. Clark RD (2009) Prospective ligand-and target-based 3D QSAR: state of the art 2008. Curr Top Med Chem 9:791–810
7. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishof CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
8. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52:867–881
9. Damm-Ganamet KL, Smith RD, Dunbar JB, Stuckey JA, Carlson HA (2013) CSAR benchmark exercise 2011–2012: evaluation of results from docking and relative ranking of blinded congeneric series. J Chem Inf Model 53:1853–1870
10. Cramer RD, Wendt B (2014) Template CoMFA: the 3D-QSAR Grail? J Chem Inf Model 54:660–671
11. Jilek RJ, Cramer RD (2004) Topomers: a validated protocol for their self-consistent generation. J Chem Inf Comput Sci 44:1221–1227
12. Brown SP, Muchmore SW (2009) Large-scale application of high-throughput molecular mechanics with poisson–boltzmann surface area for routine physics-based scoring of protein-ligand complexes. J Med Chem 52:3159–3165

13. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

14. Mittal RR, Harris L, McKinnon RA, Sorich MJ (2009) Partial charge calculation method affects CoMFA QSAR prediction accuracy. J Chem Inf Model 49:704–709

15. Wendt MD, Rockway TW, Geyer A, McClellan W, Weitzberg M, Zhao X, Mantei R, Nienaber VL, Stewart K, Klinghofer V, Giranda VL (2004) Identification of novel binding interactions in the development of potent, selective 2-naphthamidine inhibitors of urokinase. Synthesis, structural analysis, and SAR of N-phenyl amide 6-substitution. J Med Chem 47:303–324

16. Nienaber VL, Davidson D, Edalji R, Giranda VL, Klinghofer V, Henkin J, Magdalinos P, Mantei R, Merrick S, Severin JM, Smith RA, Stewart K, Walter K, Wang J, Wendt M, Weitzberg M, Zhao X, Rockway T (2000) Structure-directed discovery of potent non-peptidic inhibitors of human urokinase that access a novel binding subsite. Structure 8:553–563

17. Wendt MD, Geyer A, McClellan WJ, Rockway TW, Weitzberg M, Zhao X, Mantei R, Stewart K, Nienaber V, Klinghofer V, Giranda VL (2004) Interaction with the S1β-pocket of urokinase: 8-heterocycle substituted and 6, 8-disubstituted 2-naphthamidine urokinase inhibitors. Bioorg Med Chem Lett 14:3063–3068

18. Puius YA, Zhao Y, Sullivan M, Lawrence DS, Almo SC, Zhang ZY (1997) Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. Proc Natl Acad Sci USA 94:13420–13425

19. Xin Z, Oost TK, Abad-Zapatero C, Hajduk PJ, Pei Z, Szczepankiewicz BG, Hutchins CB, Ballaron SJ, Stashko MA, Lubben T, Trevillyan JM, Jirouseka MR, Liu G (2003) Potent, selective inhibitors of protein tyrosine phosphatase 1B. Bioorg Med Chem Lett 13:1887–1890

20. Liu G, Xin Z, Pei Z, Hajduk PJ, Abad-Zapatero C, Hutchins CW, Zhao H, Lubben TH, Ballaron LJ, Haasch DL, Kaszubska W, Rondinone CM, Trevillyan TM, Jirousek MR (2003) Fragment screening and assembly: a highly efficient approach to a selective and cell active protein tyrosine phosphatase 1B inhibitor. J Med Chem 46:4232–4235

21. Liu G, Xin Z, Liang H, Abad-Zapatero C, Hajduk PJ, Janowick DA, Szczepankiewicz BG, Pei Z, Hutchins CW, Ballaron SJ, Stashko MA, Lubben TH, Berg CE, Rondinone CM, Trevillyan JM, Jirousek MR (2003) Selective protein tyrosine phosphatase 1B inhibitors: targeting the second phosphotyrosine binding site with non-carboxylic acid-containing ligands. J Med Chem 46:3437–3440

22. Wang LE, Sullivan GM, Hexamer LA, Hasvold LA, Thalji R, Przytulinska M, Tao ZF, Li G, Chen Z, Xiao Z, Gu WZ, Xue J, Bui MH, Merta P, Kovar P, Bouska JJ, Zhang H, Park C, Stewart KD, Sham HL, Sowin TJ, Rosenberg SH, Lin NH (2007) Design,

synthesis, and biological activity of 5, 10-dihydro-dibenzo [b, e][1, 4] diazepin-11-one-based potent and selective Chk-1 inhibitors. J Med Chem 50:4162–4176

23. Tao ZF, Li G, Tong Y, Chen Z, Merta P, Kovar P, Johnson E, Park C, Judge R, Rosenberg S, Sowi T, Lin NH (2007) Synthesis and biological evaluation of 4′-(6,7-disubstituted-2,4-dihydro-indeno[1,2-c]pyrazol-3-yl)-biphenyl-4-ol as potent Chk1 inhibitors. Bioorg Med Chem Lett 17:4308–4315

24. Teng M, Zhu J, Johnson MD, Chen P, Kornmann J, Chen E, Blasina A, Register J, Anderes K, Rogers CS (2007) Structure-based design of (5-Arylamino-2 H-pyrazol-3-yl)-biphenyl-2′, 4′-diols as novel and potent human CHK1 inhibitors. J Med Chem 50:5253–5256

25. Mittal RR, McKinnon RA, Sorich MJ (2009) The effect of molecular fields, lattice spacing and analysis options on CoMFA predictive ability. QSAR Comb Sci 28:637–644

26. Davis AM, Teague SJ, Kleywegt GJ (2003) Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. Angew Chem Int Ed 42:2718–2736

27. Brunger AT (1997) Methods Enzymol 277:366–396

28. Sippl W (2010) 3D-QSAR–applications, recent advances, and limitations. In: Recent advances in QSAR studies. Springer, Netherlands, pp 103–125

29. DePriest SA, Mayer D, Naylor CB, Marshall GR (1993) 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. J Am Chem Soc 115:5372–5384

30. Waller CL, Oprea TI, Giolitti A, Marshall GR (1993) Three-dimensional QSAR of human immunodeficiency virus (I) protease inhibitors. 1. A CoMFA study employing experimentally-determined alignment rules. J Med Chem 36:4152–4160

31. Doweyko AM (2004) 3D-QSAR illusions. J Comput Aided Mol 18:587–596

32. Tuccinardi T, Ortore G, Santos MA, Marques SM, Nuti E, Rossello A, Martinelli A (2009) Multitemplate alignment method for the development of a reliable 3D-QSAR model for the analysis of MMP3 inhibitors. J Chem Inf Model 49:1715–1724

33. Cramer RD (2011) Rethinking 3D-QSAR. J Comput Aided Mol 25:197–201

34. Wendt B, Uhrig U, Bös F (2011) Capturing structure–activity relationships from chemogenomic spaces. J Chem Inf Model 51:843–851

35. Wendt B, Mülbaier M, Wawro S, Schultes C, Alonso J, Janssen B, Lewis J (2011) 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 54:3982–3986

36. Cramer RD, Wendt B (2007) Pushing the boundaries of 3D-QSAR. J Comput Aided Mol 21:23–32