J Comput Aided Mol Des (2010) 24:865–878
DOI 10.1007/s10822-010-9379-8

QMOD: physically meaningful QSAR

Ajay N. Jain

Received: 25 May 2010 / Accepted: 3 August 2010 / Published online: 19 August 2010

Abstract Computational methods for predicting ligand affinity where no protein structure is known generally take the form of regression analysis based on molecular features that have only a tangential relationship to a protein/ligand binding event. Such methods have utility in retrospective rationalization of activity patterns of substituents on a common scaffold, but are limited when either multiple scaffolds are present or when ligand alignment varies significantly based on structural changes. In addition, such methods generally assume independence and additivity of effect from scaffold substituents. Collectively, these non-physical modeling assumptions sharply limit the utility of widely used QSAR approaches for prospective prediction of ligand activity. The recently introduced Surflex-QMOD approach, by virtue of constructing physical models of binding sites, comes closer to a modeling approach that is congruent with protein ligand binding events. A set of congeneric CDK2 inhibitors showed that induced binding pockets can be quite congruent with the enzyme's active site but that model predictivity within a chemical series does not necessarily depend on congruence. Muscarinic antagonists were used to show that the QMOD approach is capable of making accurate predictions in cases where highly non-additive structure activity effects exist. The QMOD method offers a means to go beyond non-causative correlations in QSAR analysis.

A. N. Jain (✉)
Department of Bioengineering and Therapeutic Sciences, Helen Diller Family Comprehensive Cancer Center, University of California, 1450 3rd Street, Room D373, MC 0128, P.O. Box 589001, San Francisco, CA 94158-9001, USA
e-mail: ajain@jainlab.org

Keywords QSAR · Ligand based modeling · Similarity · Docking

Introduction

In our initial paper reporting the Surflex Quantitative MODeling (QMOD) method for ligand-based binding affinity prediction, we showed accurate scaffold-independent affinity predictions on a particularly challenging structure–activity data set [1]. Using just 20 ligands of two relatively rigid scaffolds, accurate predictions were made on 35 molecules from related series as well as on 17 compounds of widely varying structural types. This was done by construction of a physical binding site made up of molecular fragments (a "pocketmol") such that the maximally active pose of each training ligand (measured using the Surflex-Dock scoring function) yielded a score close to the experimental pK_d. New molecules were flexibly fit into the pocket, and the maximal score was the predicted pK_d, with the corresponding pose being the prediction of binding mode.

Figure 1 illustrates the process on a set of CDK2 inhibitors in a recently published modeling study [2]. The process begins with structures and activities, develops a rough hypothesis for relative alignments of ligands (many per ligand), generates a diverse set of possible binding pocket fragments, and finally selects and refines a set of optimal fragments. Optimality describes both the fit of the model to binding activity data as well as the fit of ligands into the model: the model itself defines the preferred binding modes of the ligands. Building such models requires a method for model derivation where the objects to be modeled have multiple possible instantiations and where choice among these is dependent on the evolving model.

The Compass method was the first to make an iterative refinement paradigm that addressed this problem [3–5], and a formalization of this early work, termed multiple-instance learning [6], has found applications in many areas of machine learning. We have also used it in scoring function development for molecular docking [7–9].

There were four chief limitations of the initial QMOD approach. First, results for only a single target were shown, albeit a challenging one. Second, the computational approach to identifying pocket probe subsets (Step E from Fig. 1), was somewhat brittle, and, more importantly, required specification of a *single* preferred pose for each training ligand rather than choosing automatically from among the pool of many that exist for each ligand. Third, solutions to the pocket induction problem for a given set of training molecules are *numerous*, but we did not present a general method for model selection. Fourth, while we showed the relationship of our induced physical model to a modeled structure of 5HT1a, we were unable to make a direct comparison to a specific and relevant experimentally determined crystal structure of the target. This paper addresses all of these limitations as well as examining the theoretical basis for the superiority of physically sensible models over purely empirical QSAR approaches.

Most QSAR approaches derive a mathematical relationship between molecular descriptors and activity that is only tangentially related to the physical process of ligand binding. The implications of these limitations on model predictivity were highlighted by Johnson [10], focusing on the logical fallacy of assigning causality to correlated variables. In particular, it was suggested that "Reliable prediction of future compounds requires that the model have some basis in physical reality." There are two central limitations of most QSAR approaches relating to this issue of physicality. First, most QSAR methods make an implicit assumption that the effects of substituent changes at different positions on the same scaffold will be *strictly additive*, which is not physically realistic. Second, many such methods do not depend on a prediction of ligand binding mode, and even for those that do, the "predicted" ligand binding mode *does not* generally depend on the model or its parameters, which is also non-physical. All non-3D approaches share the second limitation (since they do not depend on molecular alignment at all), and those

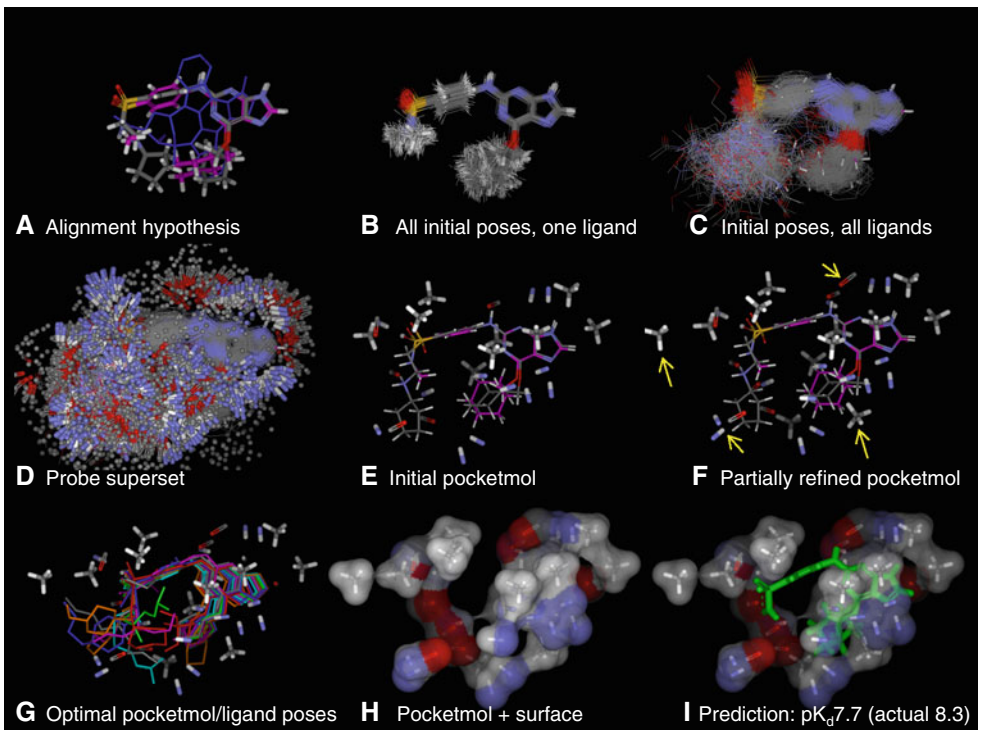

Fig. 1 Derivation of a pocketmol. Panel *A*: A 3D similarity-based alignment hypothesis of active ligands. *B*: Each training ligand is aligned to the hypothesis, resulting in 100–200 initial poses. *C*: Each training ligand has many poses, resulting in uncertainty as to where the interacting parts of the pocket might be. *D*: Interacting probes are placed that make a favorable interaction with at least one pose of each active training ligand. *E*: Activity data are used to identify a subset of probes that yield a good fit to binding data. *F*: Partial refinement of the pocketmol includes addition of new probes (some marked with *yellow arrows*) along with changes to ligand poses. *G*: Final refinement yields an optimal pocket with optimal poses for each training ligand. *H*: The final pocket forms a partial enclosure with hydrophobic and charged surfaces. *I*: New molecules are docked into the pocket and scored, yielding predictions of activity and binding mode

that are linear functions of molecular descriptors generally share the first. The most widely used approaches for 3D QSAR (CoMFA and related variants [11–14]) have both limitations. Multi-point quantitative pharmacophoric methods can theoretically address both issues [15], but they lack physically realistic detail in hydrophobic binding pocket shape. In a historical sense, the present work is also related to the pseudoreceptor concept, which addresses aspects of both limitations. This work includes that of Snyder and Rao [16], further refinements including Vedani [17], and the work of Zbinden with Vedani on PrGen [18]. See Tanrikulu and Schneider [19] for a review and the initial report of the QMOD approach for additional discussion [1].

Figure 2 shows that the first assumption is false and illustrates why making ligand poses dependent on models might offer a means to avoid the assumption in the first place. The four muscarinic antagonists shown were synthesized as part of the same effort for developing a treatment for urinary incontinence [20, 21]. While two single changes from the parent compound yielded improvements over a full log unit in K_d, the combination of the two changes was *worse* than either of the singly substituted compounds. One simple physical explanation, that the pocket is too small to fit the largest of the four compounds easily, is *beyond* the explanatory capability of many QSAR methods. This represents an *anti*-additive effect. Generalizing this issue further, consider the case of a rigid protein and a ligand whose substituents have minor effects on ligand conformational energetics. The best possible outcome in the case of two separately favored substituents on a common scaffold is that their combined effect is exactly additive in pK_d. However, for this to be true, the two derivative with single substituents must have the same preference for the position of the central scaffold when bound to the protein. In general, this is not to be expected; even small changes in substitution pattern yield some variation in scaffold alignment. Non-additive behavior is a natural and common consequence of the physical interplay between ligand variants and a protein binding site. The ability to model and predict such effects is a natural by-product of the Surflex-QMOD approach, since it constructs a physical binding pocket that is analogous to a protein active site.

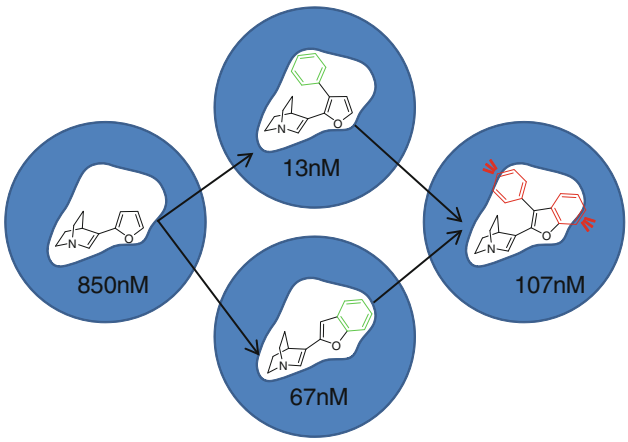

Fig. 2 Four muscarinic antagonists whose activity pattern shows a strong non-additive effect. A change from the furan to 3-phenyl or to the benzofuran yielded a significant improvement in K_d. However, combining the two changes yielded a compound with poorer K_d than either single change. One simple explanation for the effect is that the disubstituted compound is slightly too large to accommodate both substituents, with the corollary effect that each mono-substituted compound has slightly different preferences in binding mode

This paper reports improvements to the QMOD technique and expands the set of validation cases to include a typical QSAR data set and a more challenging one. The former consisted of 80 congeneric CDK2 inhibitors, split between 30 for training and 50 for testing. This set offered the ability to consider the relationship between induced models and experimentally determined protein binding pocket structure. The latter consisted of 25 muscarinic antagonists, split between 22 for training and 3 for testing the highly non-additive structure–activity effect shown in Fig. 2. For the CDK2 set, the primary model showed a mean error of prediction of 0.4 log units (approx. 0.5 kcal/mol), and highly significant rank correlations were obtained (Kendall's Tau 0.77, $p \ll 0.01$, by permutation analysis). Of the top 10 predicted test ligands, 7 of the *bona fide* top 10 were identified ($p \ll 0.01$, by exact binomial). One surprising aspect of model-building for CDK2 was that models that were congruent to the active site were *not* significantly more predictive than those that were not. However, this was true only for molecules within the chemical series used for model construction. When considering a diverse set of CDK2 inhibitors, the more geometrically accurate model was more predictive. For the muscarinic case, the primary model accurately ranked the potencies of the three substituted furans shown in Fig. 2.

The methods and results presented here by no means represent a "solution" to the 3D QSAR problem. However, the Surflex-QMOD approach can be seen to be both a practical and theoretical improvement upon the status quo in a field built historically upon correlative analysis that has been premised on non-causative observations.

All data and computational protocols are available for download (see http://www.jainlab.org for details).

Methods and data

The following describes the molecular data sets, computational methods, computational procedures, and quantification of performance.

Molecular data sets

Two sets of ligands were used. The first, illustrated in Fig. 3, consisted of 80 CDK2 inhibitors, ranging in pK_i from 4.0 to 8.3. These were split randomly into a training set of 30 and testing set of 50 inhibitors. All molecules were N2, O6 substituted guanines and were the subject of a recent modeling study [2]. In addition, for some model-building, staurosporine was also used (structure shown in Fig. 1), in order to yield a more accurate representation of the absolute configuration of the ligands when bound to CDK2. In these cases, the activity of staurosporine was specified as being greater than a pK_i of 7.0. In addition, a set of 67 PDB co-crystal structures of CDK2 bound to non-covalent inhibitors was identified from Binding MOAD [22] and were mutually aligned in order to provide a direct comparison between QMOD-generated models and the actual CDK2 binding site under normal conformation variation.

The second set, illustrated in Fig. 4, consisted of all furan-based quinuclidinene muscarinic antagonists from two structure–activity studies, with the addition of two benzofuran compounds for testing non-additive predictions [20, 21]. The activity range was pK_d 5.0–8.0. These

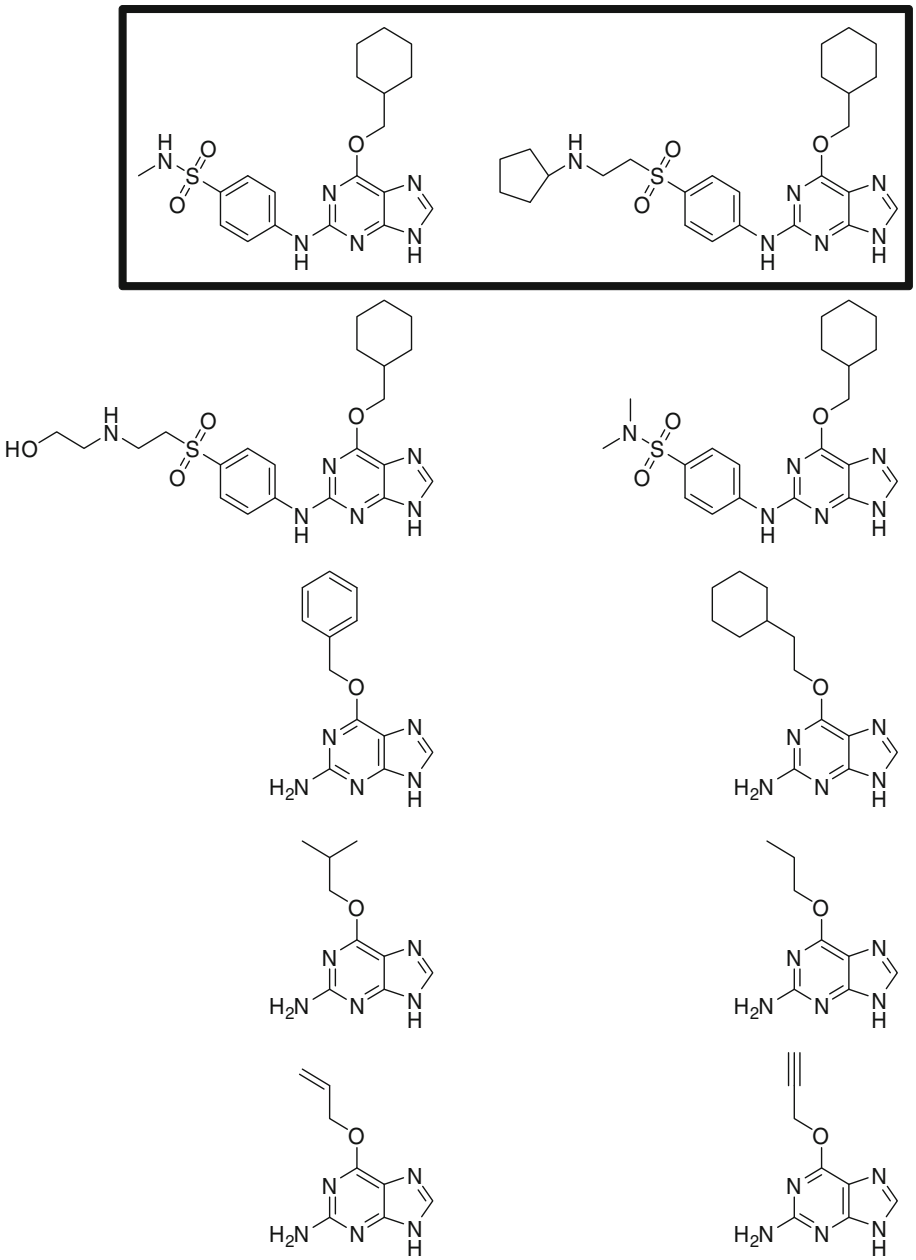

Fig. 3 Examples of typical ligands from the training set for CDK2. All compounds were N2, O6 substituted guanines. Activity ranged from pK_i of 8 (*top row*) down to 4 (*bottom row*). The *top* two compounds (*boxed*) were used in the initial alignment hypotheses. Test compounds were of similar structural variety (training and test compounds were partitioned 30/50 randomly from a full set of 80)

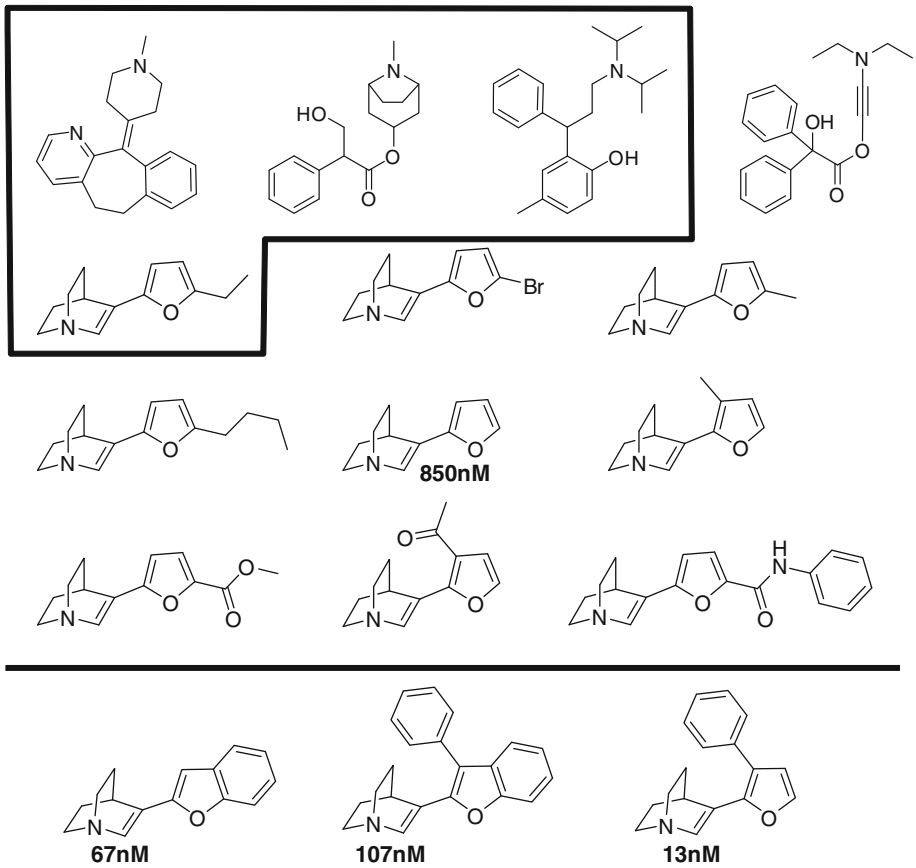

Fig. 4 Examples of ligands from the training set of muscarinic antagonists (*top*) and the test compounds (*bottom*). The *top row* include (*left* to *right*): azatadine, atropine, tolterodine, and oxybutynin (all of which were indicated to have $pK_d > 7.0$). The *boxed* molecules formed the initial alignment hypothesis. The test ligands included a substituted furan more active than any training furan (*bottom right*) as well as one showing strong non-additive behavior (*bottom middle*)

compounds were synthesized as part of an effort to produce a new muscarinic antagonist with reduced dry-mouth side-effects that resulted in the drug tolterodine [23]. To simulate a typical drug discovery effort focused on a single scaffold (here the furan-based antagonists), models were constructed using known, potent ligands (competitive scaffolds) along with 22 from the furan series. At the time, oxybutynin was a competing therapeutic, atropine was one of the earliest known muscarinic antagonists, and azatadine offered a relatively rigid example of a potent (but non-selective) muscarinic antagonist. The three substituted furans shown at the bottom of Fig. 2 were used to test the final model. Of note, the 3-phenyl was the most potent of the series, and (as discussed above) the phenyl-substituted benzofuran was much less active than one would expect based on the activity of the other two test compounds. The split of 22 training and 3 testing ligands was done specifically to illustrate the potential for accurate predictions of highly non-additive effects that depend on molecular alignment.

All ligand structures as well as preparation protocols are available for download (see http://www.jainlab.org for details). Additional details regarding computational procedures for training ligand alignment, model induction, and testing of novel ligands follows.

Computational methods

The core computational methods for the Surflex-QMOD approach have been reported previously [1], and the basic steps are summarized in Fig. 1. The present work makes a significant improvement to the initialization of a pocketmol and offers a validation procedure to help select from multiple possible models, and these will be described in detail. Overall, there are five steps to construct and employ a physical binding pocket for activity prediction (a "pocketmol"):

1. Generation of an initial set of alignments for each training ligand.
   a. **Input**: Structures of ligands, with 2 or 3 chosen to serve as the seed alignment hypothesis.
   b. **Output**: At least one pose for each training ligand, all of which are plausible within the same mutual alignment. Typically, there are 100 poses per ligand.
2. Generation of an initial set of molecular probes to form the binding pocket.

   a. **Input**: All poses (optionally limited to some maximum number) for each active training ligand, from the pool from Step 1.
   b. **Output**: A large set of molecular probes surrounding the ligands, where each probe makes a near-optimal interaction with at least one active ligand's pose. Typically, there are 1,000–3,000 such probes.

3. Selection of an optimal minimal pocketmol followed by augmentation to improve the fit to data.
   a. **Input**: The set of probes from Step 2, *all poses* for each training ligand (actives and inactives), and activity values for each ligand specified as exact values or inequalities.
   b. **Intermediate**: A minimal set of probes such that nominal interaction scores against this set lie within a specified accuracy (using the highest scoring pose for each ligand).
   c. **Output**: A refined set of probes, a refined set of ligand poses (using the refined probes), and a small set of additional probes that improve pocketmol performance given the new ligand poses.

4. Refinement of the pocketmol by modifying probe positions interleaved with refining ligand poses.
   a. **Input**: The output pocketmol from Step 3, the full set of ligand poses for all training ligands, and the molecular activities.
   b. **Output**: A refined pocketmol with refined ligand poses such that further local optimization of ligand poses against the pocketmol yields little change in scores and where the final scores are close to the experimentally measured ones.

5. Testing of new putative ligands within the pocketmol:
   a. **Input**: A new molecular structure, the final pocketmol, and a selection of optimal poses for training molecules for use in alignment of new molecules.
   b. **Output**: Predicted score and pose alternatives for the new ligand using a procedure analogous to docking ligands into a protein active site.

This procedure will yield a single model, but Steps 1 and 3 each generate ranked lists of putative alignments (Step 1) and putative initial minimal probe sets (Step 3). For a particular alignment, several different probe sets may yield models that fit the training data well. Further, several such alignments may all yield plausible models, so there must be some means to choose from among multiple models based only on training data. The following two sections will describe the new pocketmol initialization approach as well as a model selection approach.

Pocketmol initialization

The input to the pocketmol initialization is a pool of poses for each training ligand (typically 100 each) as well as a pool of pocketmol probes (typically 1,000–3,000). Our initial approach had relied upon a numeric programming method that forced a choice of a *single* pose for each ligand prior to yielding a set of probes, and it also made use of an ad hoc procedure for covering areas of the pocketmol that were too open [1]. The current approach makes use of a greedy algorithm that optimizes an objective function that minimizes error in computed molecular activity and also minimizes probe density. Deviation in computed activity is quantified as mean-squared difference between computed and experimental pK_d for all training molecules. Density is quantified based on the positions of probes of like type. For each pair of like probes, the contribution to the density penalty is a constant weight (default 2.0) multiplied by the amount that the RMSD between the probes is closer than a set threshold (defaults are 3.5 for steric probes and 1.5 for polar probes).

Given a particular probe set, interaction scores for each molecule are computed by considering each of the molecule's poses and taking the maximal computed interaction score. So, a "perfect" pocketmol will yield the experimental pK_d for each training molecule and will have no probe pairs that are closer than the selected RMSD thresholds. The algorithm for greedily growing an optimal probe set makes multiple random starts (default 100), each time identifying N probes (default 2) to initialize a pocketmol. Then an iterative process begins, seeking to add a single probe to the existing pocketmol that makes the largest improvement in the objective function. The process stops when no probe can be added that improves the objective function. In Fig. 1, this corresponds to the selection that takes place between Panels D and E.

Typically at this point, the nominal computed molecule scores deviate from experimental ones by less than 0.5 log units. However, if one now optimizes the poses of the ligands in order to maximize their interactions with the pocketmol, deviations increase to values of 2–6 log units. This occurs for two reasons. First, probe positions are identified based on the static initial alignments of the ligands. Second, the probe subset that comprises the pocketmol is selected based on the static initial ligand alignments. There is no reason to believe that either the composition of the pocketmol or the precise positions of its probes are truly optimal when considering ligand movement. Inactive molecules, in particular, when given the opportunity to make optimal interactions with the pocketmol can "evade" the constraints of the physical pocket. In order to ensure that the pocketmol can capture the important interactions (both positive and negative), two short iterations of probe refinement followed by pose refinement are carried out. This results in slightly

modified positions of the initial pocketmol as well as an increased ligand pose pool that includes optimal poses with respect to this modified pocketmol. The probe addition procedure is run again, adding probes that improve the combined objective function. In Fig. 1, Panel F shows the modified pocketmol along with the added probes (some highlighted with arrows). The probe addition step typically adds 10–25% more probes to the initial pocketmol.

Model selection

The most critical aspect governing the predictive behavior of the final model is the ligand pose pool and initial pocketmol that result from the preceding step. As mentioned above, many such combinations will yield models that have a good fit to the training data, some of which may look rather different from one another, especially if they result from divergent initial alignment hypotheses. The models will vary in predictive ability, even if they nominally fit the training data equally well. In order to select which final model to use for prospective predictions, QMOD makes use of a modified cross-validation procedure, constructing multiple fully refined models (default 6), where each model is refined using a different subset of the training ligands. In each iteration, the training ligands not used for model refinement are tested against the refined model. The set of such training set predictions is used to identify the fully trained model that is most likely to be predictive. To validate this approach, multiple models were constructed for the 5HT1a set from the original QMOD publication using the methods described above. Of 23 models constructed (each with its own sixfold cross-validation), the top 8 selected by Kendall's Tau rank correlation from the training set validation identified all of the top 5 most predictive models on the blind test set ($p = 0.01$).

Two details bear further discussion. This is *not* a true cross-validation. The full set of training ligands is used to identify the initial probe subset and to produce the augmented pocketmol that serves as input to this validation procedure. So, information from the "holdout" ligands has been used, in part, prior to full refinement and testing of the holdouts. This is necessary since one must evaluate the specific probe set that will be used for final model building. Were the holdout molecules fully held out, the resulting pockets would have different compositions, and the results of the validation procedure would have little bearing on the predictivity of a pocket with a new composition. This contamination effect is ameliorated by restricting the pose pool at the start of each validation model refinement to the naïve poses from the initial similarity-based ligand alignments, thus reducing the amount of information leakage from the full training set into the training subsets. This procedure is only used as a heuristic for model selection, so the information leakage does not contaminate predictions outside the training set.

Computational procedures

Detailed scripts for generating the results presented here are available in the data archive associated with this paper. Surflex-QMOD version 1.011 and Surflex-Sim version 2.512 were used. A single command ("`sf-qmod.exe runsetup SetupFile`") produces a script that will generate initial alignment hypotheses, full alignments of training ligands, results for internal validation tests, and final pocketmols. The setup file contains information on pathnames to training ligands and their activities, which ligands to use for hypothesis generation, and modifications to default parameters for model building if desired. A typical example is shown here:

```
# This is a QMOD run setup file.
QMODPath:../Paper-v1011/Code/Surflex-QMOD-
 v1011/sf-qmod.exe
SIMPath:../bin/surflex-sim-v2512.exe
RunPrefix: runa
# Required: Training molecule path. This file
 contains pathnames to
# training molecules, activity constraints, and
 pKd as follows:
#    mols/m5.mol2 = 7.5
# By convention, the first N mols in the list will
 be used as the to
# produce the hypothesis for alignment.
TrainPath: TrainMols
# Required: Number of mols for hypothesis.
NHypoMols: 2
# The parameters that control initial probe
 placement:
Srms: 0.7
Prms: 0.2
NPoses: 5
Minact: 6.5
# The parameters that control pocketmol model
 generation:
ModelProbeTries: 4
ModelSrms: 1 3.5
ModelPrms: 1 1.5
ModelPenRms: 1 2.0
# Initial refinement control:
InitEpoch: 10
InitIter: 2
# Cross validation for model selection:
Xval: 6
```

The final pocketmols can be selected on the basis of the validation results. Testing on a list of new ligands requires a single command ("`sf-qmod.exe scorepocketmol TestLigList align-targets.mol2 pocket-mol.mol2 logscore`"). This procedure makes use of an all-atom optimization procedure, so ligand strain enters into the pose optimization process explicitly and prevents excessive deformation of ligands. The final scores represent the intermolecular interaction energy along with entropic fixation terms and do not include internal strain directly.

Results and discussion

The goals of the Surflex-QMOD approach include accurate prediction of ligand activity, but they also include prediction of ligand binding modes and rationalization of activity across chemical scaffolds that allows information learned from one series to be transferred to another. The CDK2 set offers a means to further validate predictive capacity and to explore the relationship between learned models and actual experimentally determined enzyme structures. The muscarinic set offers a means to assess a pocketmol's ability to extrapolate and to see whether construction of physical models can address the fundamental problem of non-additivity and the interplay of molecular pose with predicted activity.

Performance on CDK2: ligand-based modeling with staurosporine

Using the initial alignment shown in Fig. 1 (which was the single highest scoring alignment hypothesis), three models were generated. The models resulted from the top three initial probe sets from the pocketmol initialization procedure described earlier. All three models yielded highly significant rank correlations in the sixfold cross-validation tests (ranging from 0.53 to 0.66), and all three showed excellent predictive performance on the test set (Tau of 0.62–0.77). Figure 5 shows plots of the performance for the model that both fit the training data best as well as yielded the best performance on the test data. This was the model derived from the single best scoring initial pocketmol. All three models were similar in composition and in terms of the preferred binding poses of the training ligands; in what follows the model corresponding to Fig. 5 will be discussed.

Figure 6 illustrates the congruence between the pocketmol and the structure of CDK2. Of 67 mutually aligned structures of CDK2 bound to inhibitors, three (1H01, 1E9H, and 1H00) showed particularly close correspondence to the derived pocketmol. Two of the three important hinge-binding elements were closely matched, and the aspartic acid critical for explaining the activity of the amines in the modeled chemical series was also closely matched. The hydrophobic surface comprised of two leucine residues was mimicked by two hydrophobic methane probes in the pocketmol. Other probes served to construct a surface that was concordant in many respects to the experimentally determined structures. However, Fig. 7 shows that the portion of the pocketmol that was congruent to CDK2 was not sufficient to adequately enclose the series being modeled. A small number of probes did not correspond to any structural aspects of CDK2's binding pocket variants. Consequently, some known inhibitors would simply not fit the induced model of the binding site (cyan area marked with an arrow in Fig. 7). This is a primary limitation of the QMOD approach. Lack of diversity in training data, particularly with respect to ligand scaffolds, will generally be reflected by degeneracies in the induced physical models.

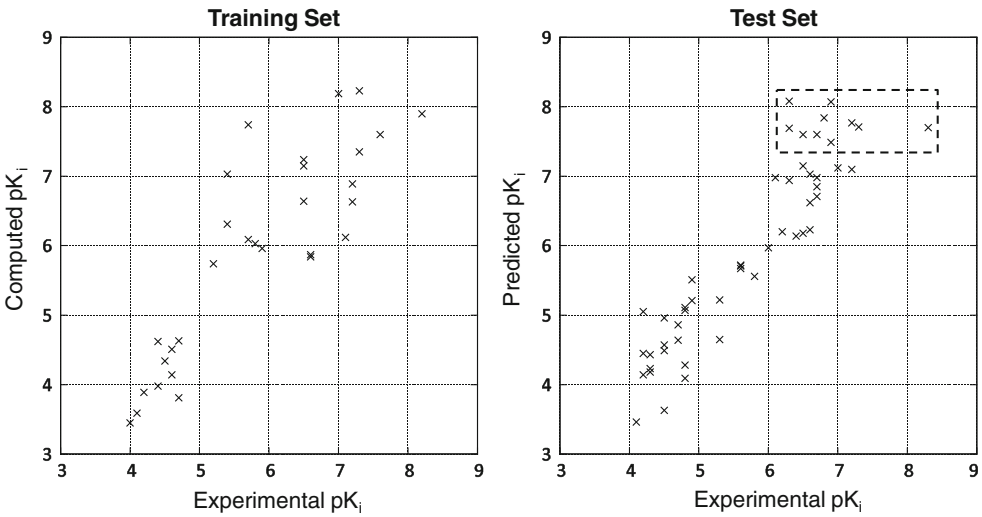

Fig. 5 Plots of training performance (*left*) and test molecule performance (*right*). The top 10 predicted test molecules (of 50 total) included 7/10 of the most potent compounds. The *top* half of the predicted test compounds were correctly identified. Kendall's Tau rank correlation for the training set fit was 0.68, with a mean error of 0.56. For the test set, Tau was 0.77 ($p \ll 0.01$) and mean error 0.43

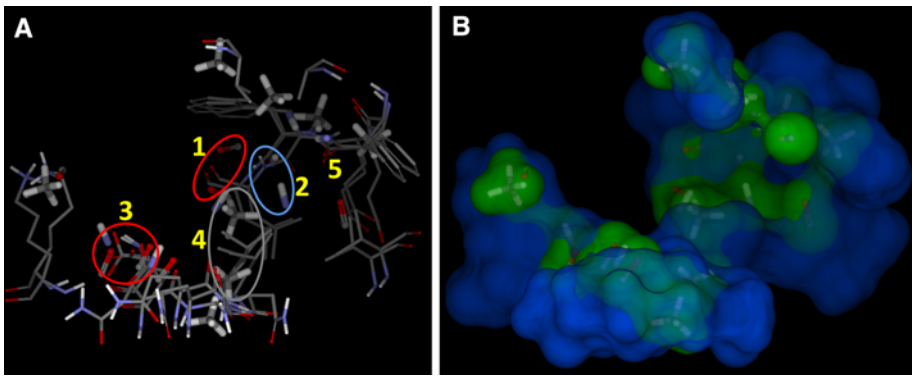

Fig. 6 Comparison of commonalities between CDK2 conformational variants (*thin sticks* at *left*, *blue* surface at *right*) and the derived pocketmol (*thick sticks*, *green* surface). In particular, there is very good correspondence between polar probes and two hingebinding elements: the carbonyl and N–H of LEU-83 (marked as *1* and *2*). There is also good correspondence to the carboxylate of ASP-86 (marked as *3*). There is also good correspondence between two steric probes and the sidechains of LEU-83 and LEU-134 (marked as *4*). Note that the hinge-binding aspect of GLU-81 is missed (marked as *5*)

Parsimony and confidence

In the foregoing, we discussed the use of cross-validation for model selection. In the case of CDK2, in contrast to experiments on more structurally diverse data sets, all of the derived models yielded excellent performance in internal validation tests. This raises the question of how one might choose from among models that seem equally likely to be predictive. Since the approach includes predictions of optimal training molecule poses, we can assess the extent that a model is quantitatively parsimonious. That it, if two molecules have similar activity, and if it is possible to explain their binding in geometrically similar ways, we should prefer models that make that choice over models that do not. This is expressible in terms of a weighted sum of pairwise similarities of all final ligand poses, where molecule pairs with similar activity receive higher weight than those with different activity values. Figure 8 shows the final optimal training ligand poses from three models derived from the initial alignment shown in Fig. 1. By eye, it is clear that the top model is the most coherent in terms of commonality of binding modes. Using a quantitative measure of parsimony (scale 0–1, with 1 being most parsimonious), the scores were 0.76, 0.72, and 0.69 (top to bottom, respectively). This measurement captures the intuitive notion of which model seems the most "compact" in an explanatory sense and also correctly identified the model with the best performance on the test set (Tau of 0.77).

A related idea allows one to define a simple measure of confidence in a prediction on a new molecule using molecular similarity. If a predicted pose of a test molecule *based purely on the pocketmol* is quantitatively similar to the optimal pose of a training ligand, one has higher confidence in a prediction than otherwise. For the 67 diverse CDK2 ligands with known bound configuration, 14 had high confidence (similarity greater than 0.7 to an optimal pose of a training molecule). Among these 14, the rank correlation of predicted compared with experimental pK_i was 0.44 ($p = 0.02$), but the average error of prediction was 1.0 log units (significantly worse than within the congeneric test series). The mean RMS deviation of the best of the top 5 poses for each of these 14 ligands was 2.1 Å. Among the 30 ligands with moderate to high confidence (similarity $\geq$0.60), rank correlation was lower (Tau of 0.30) but still statistically significant ($p < 0.01$). The mean RMS deviation for these 30 ligands was 2.5 Å. Among the 37 ligands with low confidence (similarity $<$0.60), mean RMS deviation was 5.5 Å, the rank correlation was lower still (Tau of 0.09), and the correlation was not statistically significant. Note that all test ligands from the substituted guanines had high-confidence predictions.

There is a clear relationship between the similarity-based measure of confidence and the predictive performance of the QMOD pocketmol, both in terms of quantitative affinity prediction and binding mode prediction. Figure 9 shows examples of 4 molecules from this set, including two with scaffold variations beyond the substituted guanines of the training set. In each case, while the predicted pose was not perfect in any case, the correct correspondence of ligand and protein/pocket interactions was present, which would support structure-guided ligand design. It is important to note that the scaffold variation exhibited in Fig. 9 is still significant, and reliable predictive generalization in 3D QSAR of this magnitude would be useful in lead optimization.

Performance on CDK2: modeling without staurosporine

In the foregoing, we established that CDK2 models constructed with the benefit of staurosporine in addition to the

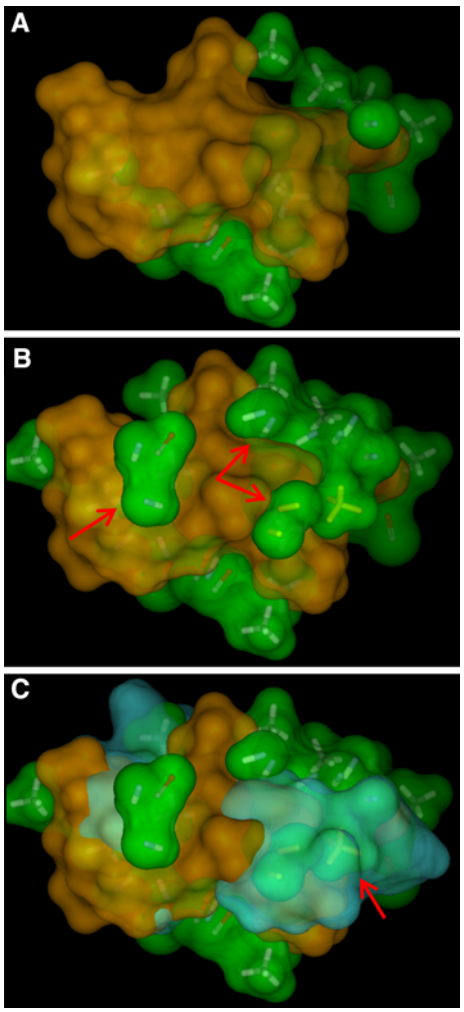

Fig. 7 Panel *A* shows the portion of the learned pocketmol (*green*, with probes in sticks) with close correspondence to the CDK2 active site along with the surface of the union of training ligand final poses (*orange*). Panel *B* shows the entire pocketmol, which includes several probes that *did not* correspond closely with any CDK2 conformational variants (highlighted with *red arrows*), and these can be seen to help "lock down" the training ligands. However, as can be seen in panel *C*, the envelope of diverse CDK2 inhibitors with known bound structure (*cyan* surface) intersects with volume that was incorrectly learned on the bases of a single congeneric series (*red arrow*)

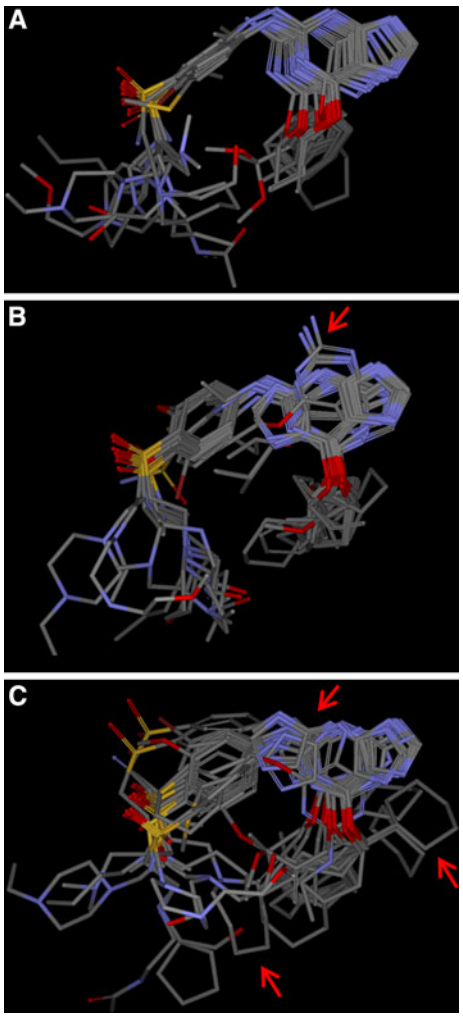

Fig. 8 The three panels show final poses of the 30 training ligands from models of decreasing parsimony, all based on the initial alignment from Fig. 1. The model in panel *A* (parsimony score of 0.76/1.0) shows a single binding mode, panel *B* shows alternate binding modes for the guanine (parsimony of 0.72), and panel *C* shows multiple binding modes for both the guanine and for the pendant substitutions (parsimony 0.69). The most parsimonious model was more predictive on the test set than the other two

guanine-based series were predictive both in terms of affinity and binding mode. The presence of staurosporine forced a choice of the absolute configurations available to the QMOD procedure, since the protonated amine of staurosporine has a fixed geometric relationship to its hinge-binding moieties. Models were also constructed by omitting staurosporine but using otherwise identical procedures. Figure 10 shows the alignment hypothesis along with performance on the 50 molecule test set. The best alignment corresponded to an extended conformation, which is clearly incorrect in terms of what is known about the bound state of these ligands (see 2G9X in Fig. 9). However, the performance on the congeneric series of test molecules was roughly equivalent to that shown earlier. The reason for the apparent lack of dependence of predictive performance on model "correctness" stems from the fact that the test molecules fall into the same chemical series as the training molecules, so the *relative* alignment of the predicted compounds is very similar independent of the absolute configuration of the model. That is, the correspondence of parts among ligands and their relative relationship to entities in the pocketmols is similar.

However, when considering predictions on the diverse set of 67 molecules for which bound configurations and binding affinities were known, predictions using the

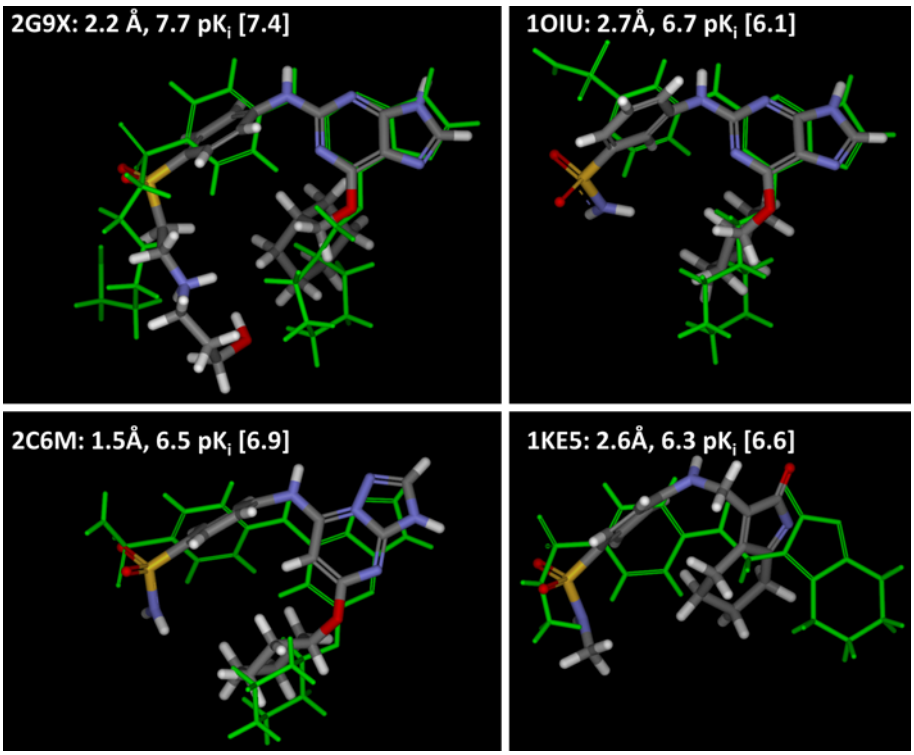

Fig. 9 The four panels show the best predicted and known bound conformations of four of the ligands for which crystal structures were available (labels are PDB-code: RMSD, predicted activity [experimental activity]). These were predicted with high confidence, based on molecular similarity to training ligands. They included two novel scaffolds: a triazolopyrimidine (*bottom left*) and an oxindole-based compound (*bottom right*). Note that the incorrect orientation of the oxindole of 1KE5 was influenced by the conservative enclosure of the binding site (see Fig. 7)

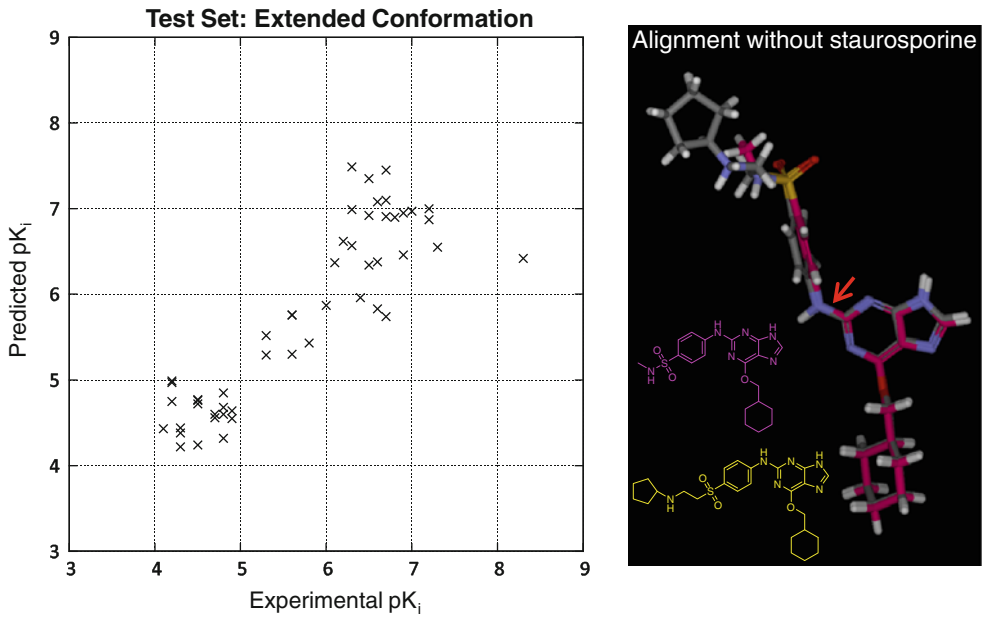

Fig. 10 Plot test molecule performance using an alignment hypothesis generated without the benefit of staurosporine. The primary difference in the ligand poses compared with the model shown previously involves a flip of the bond indicated by the *red arrow*. For the 50 molecule test set, Tau was 0.63 ($p \ll 0.01$) and mean error was 0.38

"extended" model (derived without staurosporine) are significantly worse than for the compact model. In 56/67 cases, predictions from the former model had larger errors than the latter one ($p \ll 0.001$ by exact binomial). The mean error for ligands predicted with low confidence for the extended model was 3.1 log units, but for the compact model was 1.6 log units. Interestingly, predictions for both models on high confidence ligands was statistically indistinguishable, but these ligands shared conformational flexibility in places that allowed maintenance of the correct correspondence of parts despite inaccurate absolute configurations (see Fig. 11).

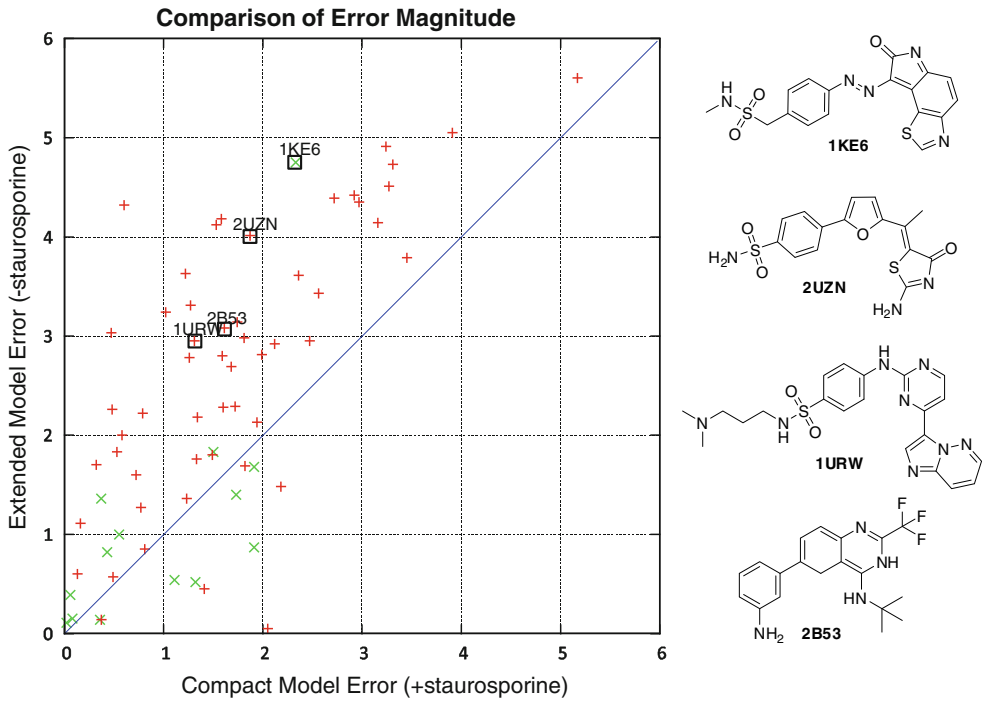

Fig. 11 The plot shows the comparison in absolute errors of prediction on 67 diverse CDK2 ligands between the pocketmol derived with the benefit of staurosporine (compact model) and the pocketmol that employed just the substituted guanines for learning. Green X's were ligands predicted with high confidence by the compact model, and *red* plus signs with low to medium confidence. Four cases were highlighted, all of which had binding mode prediction accuracies in the compact model of 2.5 Å RMSD or better. Regardless of confidence, 76% of the compact model's predictions were within 2 log units of experimental values, compared with just 45% of the extended model's predictions

Models that describe active site configurations that are more closely related to physical reality will generalize better to structurally diverse ligands than models with limited congruence to true binding sites. The practical aspect of this observation is that multiple active scaffolds should be used to help derive physically accurate models. In the case of CDK2, the addition of staurosporine was enough to make a large impact on model generalizability.

Performance on the muscarinic set

The muscarinic example shown in Fig. 2 is a hard and relevant case. The design of this test included multiple diverse active ligands (see Fig. 4) that were known at the time that the furan scaffold was being explored. Of eight models constructed, two had good internal validation performance (Tau of 0.5 in both cases), and the one with the highest parsimony was selected (0.80 vs. 0.72) for testing. Figure 12 shows this pocketmol along with predicted binding modes and activity values for the four compounds from Fig. 2. The 3-phenyl substituent yielded a predicted increase of 1.6 log units relative to the furan (experimental was an increase of 1.7). The benzofuran yielded a predicted increase of 1.2 (experimental: 1.1). The di-substituted compound showed a predicted *decrease* in activity compared with the 3-phenyl of 0.4 log units (experimental: 0.9) and the same predicted potency compared with the benzofuran (experimental: 0.2 log unit decrease). The ligand required a significant shift in alignment in order to fit into the pocket, preventing optimal interactions with the cluster of carbonyl probes that interact with the protonated amines in the series. If the model were to have made a linear additive prediction for this compound, the predicted activity would have been a pK_d of 8.4, incorrectly identifying it as the most active compound. The actual predicted activity was 1.4 log units less than the additive assumption would have dictated. The changes in substituents yielded predicted potencies that were accurate with alignments that were explanatory.

The presence of quartets of molecules like those in Figs. 2 and 12 is relatively rare in medicinal chemistry lead optimization data sets. Such examples will frequently lead to non-additive SAR, which by its nature requires modeling of the linkage between binding mode and predicted activity. Such quartets are *actively avoided* by a common strategy in lead optimization that is well-exemplified by the CDK2 guanine series. As can be seen in Fig. 3, the O6 substituent was systematically varied with *no* N2 substituent. Then, with the O6 substituent fixed (see Fig. 3: top four compounds), the N2 substituent was varied. No

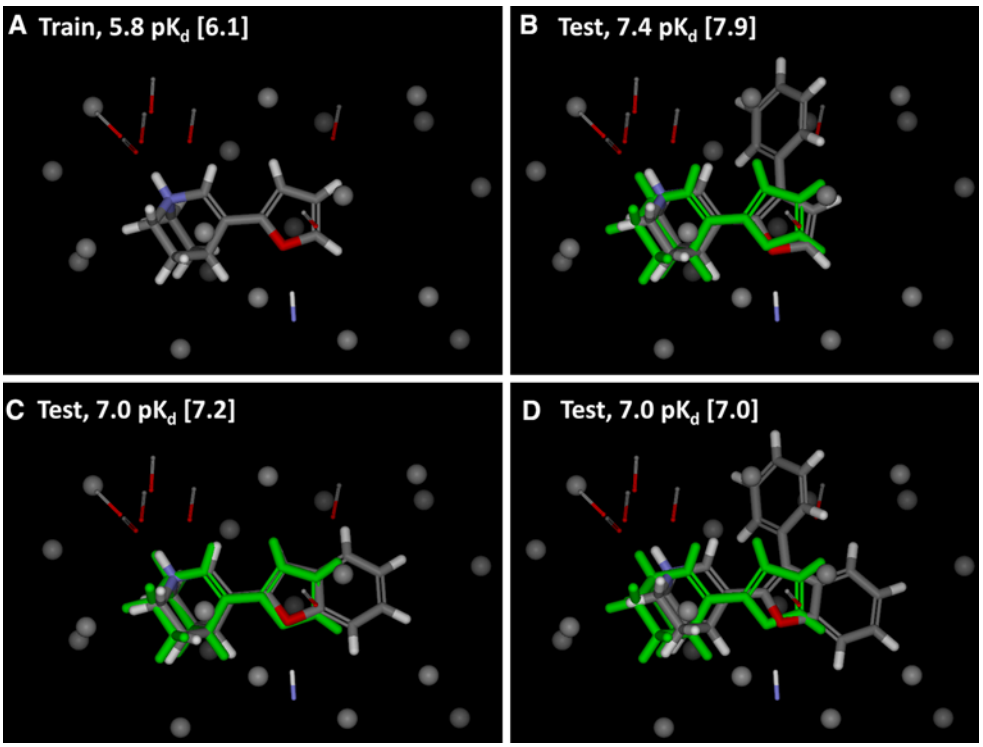

Fig. 12 The four panels show the predicted bound conformations of four muscarinic antagonists, three of which were used as tests for predictions involving non-additive substituent behavior. Predicted pK_d values are listed after an indication of whether a molecule was used for training (**a**) or testing (**b–d**) the model, and bracketed numbers indicate experimentally determined pK_d. For clarity, the pocketmol is shown with hydrogens on methane probes hidden

examples of "non-optimal" O2 substituents were tried with N2 substituents at all, the implicit assumption being that optimization at a single scaffold position can be done *independently* of other positions. This assumption is wrong, based both on first principles as well as the data shown in Fig. 2. A second, related, design strategy hinges on the incorrect presumption of additivity. Imagine that in a lead optimization exercise that the furan of Figs. 2 and 12 was discovered first, followed by synthesis of the benzofuran and the 3-phenyl benzofuran. Seeing that the phenyl group at position 3 marginally decreased activity, many chemistry teams *would not* support synthesis of the 3-phenyl furan, which is the most active compound in the series. The 3-phenyl modification was of no value in one particular context, but one cannot conclude that the 3-phenyl has no value in other contexts.

The QMOD method can learn non-additive behavior based on its representation of the QSAR model as a physically interpretable binding pocket as opposed to a purely mathematical construct linking descriptors to activities via a regression formula. Induction of such models is dependent on sufficient training data, especially in the sense of geometric diversity, in order to produce models close to those that are physically responsible for ligand binding.

Conclusions

The logical fallacy "*cum hoc ergo propter hoc*" (with this, therefore because of this) lies at the heart of the failure of many QSAR approaches to be prospectively predictive [10]. Simple statistical correlations between molecular features and molecular activities arise for many reasons, but many of these reasons are not related to the causal effects linking molecular structure to activity. The problem is made more serious since medicinal chemistry lead optimization often implicitly assumes that chemical substitutions on a scaffold will yield independent and additive effects. The assumption leads to serial optimization of substituents and to overgeneralization of context-specific information regarding the suitability of particular moieties. Consequently, many SAR data sets can be modeled retrospectively using purely correlative analyses, and therefore many QSAR methods appear to yield predictive models. Such models are challenged when there are notable non-additive effects within a single chemical series or when molecular conformation and alignment questions are non-trivial. Data sets that exhibit these effects are not difficult to find, but they are seldom seen, either in QSAR reports or in medicinal chemistry reports focused purely on structure–activity relationships.

The Surflex-QMOD approach addresses the physical linkage between activity model and molecular binding mode with pockets having detailed structure comparable to true protein binding sites. Because the model building process results in a model that *selects* ligand alignments based on mutual interaction, there is a direct correspondence between the physical process of protein/ligand binding and the act of prediction. Consequently, making use of multiple diverse ligand scaffolds is both possible and preferable to modeling congeneric series, leading to more accurate and predictive models. Notions of model parsimony and prediction confidence are intuitively related to physical notions of shared ligand binding modes and appear to bear directly on the quality of predictions. As the approach is further validated on more systems, generalization of the parsimony concept to become part of the model-building process is likely to improve the performance of Surflex-QMOD further.

Challenges remain, especially in cases where reliance on molecular similarity to guide initial ligand alignment results in *incorrect* relative alignments or when the protein pocket undergoes *extensive* rearrangement on binding different ligands. In the former case, the likely outcome is a model that essentially embeds a disjunction where one series of ligands productively interacts with one set of probes and the other series interacts with a different set. Predictions within each series may be reasonably accurate, but predictions for ligands that combine parts of each scaffold will likely not be. The hope is that such ligands could be used to provide data to refine the model. With respect to the latter issue, the approach explicitly models some receptor flexibility, with multiple probe positions covering a degree of conformational variation of the protein binding pocket. However, large rearrangements require both *detection* that the phenomenon is occurring and possibly multiple model pockets in order to appropriately predict activity. The multiple-instance formalism can easily model simultaneous construction and refinement of several models. But the bigger challenge may be to know when different subsets of ligands engender large changes in binding pocket geometry. This case, in an abstract sense, is quite similar to the problem of partially overlapping ligand binding modes, where detection of the problem may occur when poor predictions are made based on synthesis of hybrid ligands. In such cases, accurate prediction of activity may depend on a correct absolute relationship of ligand atoms with protein binding pocket atoms.

Acknowledgments The authors gratefully acknowledge NIH for partial funding of the work (Grant GM070481), Brian Goldman and Jonathan Weiss for pointing out the CDK2 data set and providing corresponding data, and Ann Cleves for comments on the manuscript. Dr. Jain has a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Tripos Inc. has exclusive commercial distribution rights for Surflex-Sim and Surflex-Dock, licensed from BioPharmics LLC.

References

1. Langham JJ, Cleves AE, Spitzer R, Kirshner D, Jain AN (2009) J Med Chem 52:6107
2. Alzate-Morales JH, Caballero J, Vergara Jague A, Gonzalez Nilo FD (2009) J Chem Inf Model 49:886
3. Jain AN, Dietterich TG, Lathrop RH, Chapman D, Critchlow RE, Bauer BE, Webster TA, Lozano-Perez T (1994) J Comput Aided Mol Des 8:635
4. Jain AN, Koile K, Chapman D (1994) J Med Chem 37:2315
5. Jain AN, Harris NL, Park JY (1995) J Med Chem 38:1295
6. Dietterich TG, Lathrop RH, Lozano-Perez T (1997) Artif Intell 89:31
7. Jain AN (1996) J Comput Aided Mol Des 10:427
8. Pham TA, Jain AN (2006) J Med Chem 49:5856
9. Pham TA, Jain AN (2008) J Comput Aided Mol Des 22:269
10. Johnson SR (2008) J Chem Inf Model 48:25
11. Cramer RD (2003) J Med Chem 46:374
12. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
13. Cramer RD, Wendt B (2007) J Comput Aided Mol Des 21:23
14. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130
15. Guner O, Clement O, Kurogi Y (2004) Curr Med Chem 11:2991
16. Snyder JP, Rao SN (1989) Chem Design Automation News 4:13
17. Vedani A, Zbinden P, Snyder JP (1993) J Recept Res 13:163
18. Zbinden P, Dobler M, Folkers G, Vedani A (1998) QSAR 17:122
19. Tanrikulu Y, Schneider G (2008) Nat Rev Drug Discov 7:667
20. Nordvall G, Sundquist S, Johansson G, Glas G, Nilvebrant L, Hacksell U (1996) J Med Chem 39:3269
21. Johansson G, Sundquist S, Nordvall G, Nilsson BM, Brisander M, Nilvebrant L, Hacksell U (1997) J Med Chem 40:3804
22. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Nucleic Acids Res 36:D674
23. Nilvebrant L, Gillberg PG, Sparf B (1997) Pharmacol Toxicol 81:169