3D QSAR Methods: Phase and Catalyst Compared

David A. Evans,[†] Thompson N. Doman,[‡] David A. Thorner,[†] and Michael J. Bodkin*,[†]

Eli Lilly and Company Ltd., Lilly Research Centre, Erl Wood Manor, Sunninghill Road, Windlesham, Surrey, GU20 6PH, England, and Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, DC 1930, Indianapolis, Indiana 46285

Received January 10, 2007

The programs Phase and Catalyst HypoGen are compared for their performance in determining three-dimensional quantitative structure−activity relationships. Eight sets of compounds with measured activity were collected from the public literature and partitioned into suitable training and test sets by an automated procedure. A range of models is built with each program, and suggested parameter variations are investigated. The models are assessed by their ability to predict the activity of compounds in the test set, and it is demonstrated that the performance of Phase is better than or equal to that of Catalyst HypoGen, with the data sets and parameters used here. Additionally, compounds in two of the data sets are overlaid on crystallographic structures of similar ligands in complex with the target receptor, in order to guide pharmacophore generation by the two programs, but the resulting models do not perform better.

1. INTRODUCTION

Discovering three-dimensional pharmacophores which can explain the activity of a series of ligands is one of the most significant contributions of computational chemistry to drug discovery.[1] Several programs exist which aim to elucidate such pharmacophores from a set of active compounds, including DISCO,[2] Catalyst,[3] GASP,[4] Phase,[5] and Galahad.[6] A crucial part of the operation of these programs is the conformational exploration and alignment of these compounds on possible common pharmacophores. This step enables three-dimensional "ligand-based design", meaning that the features of a molecule which are important for activity can be investigated in the absence of a receptor structure.

Three-dimensional quatitative structure−activity relationship (3D QSAR) programs such as the HypoGen module of Catalyst,[7,8] Phase,[5] CoMFA,[9] and CoMSIA[10] aim further to use three-dimensional alignments of molecules to quantitatively predict the activity of compounds, on the basis of QSAR models fitted to the activity of an aligned training set of molecules.

Such 3D models are more easily interpretable than 2D descriptor or fingerprint-based QSAR models, making it easier to suggest new compounds for synthesis. It should also be possible to make connections from such activity models to structure-based design, either to add more information to overlays for the construction of a pharmacophore model[11] or to use a pharmacophore to assist in the refinement of protein homology models.[12,13]

One of the most widely used packages for pharmacophore elucidation is Catalyst.[14] The HypoGen module of Catalyst generates "hypotheses" (possible pharmacophores) by matching features between conformer libraries of active compounds. The hypotheses can then be used to predict the activity of compounds, on the basis of how well the features in compounds match the hypotheses.[7,8] Catalyst has been widely used to create QSAR models (for example studies, see refs 13 and 15−17). Phase[5,18] is a more recently developed pharmacophore modeling package. It follows a hypothesis generation step, with a grid-based 3D QSAR method, in which the grid positions of atoms in molecules overlaid to the hypotheses are correlated to their activities using a partial-least-squares (PLS) fitting approach. The main stages of both methods are summarized in Figures 1 and 2.

The work reported in this study attempts a direct comparison of the activity prediction of Phase and Catalyst HypoGen. We aimed to select sets of compounds with measured activity values which are publicly accessible but still resemble data which might be used in the lead-optimization stage of a drug discovery project. Previous case studies often have limited test sets of compounds and rely on the internal consistency of models as a measure of their reliability. In assessing the methods, we focus instead on the accuracy of predicted activities on test sets of comparable size to the training set.

2. METHODS

2.1. Data Set Preparation. In total, eight data sets from published studies were collected. Three of the data sets from the publication of Sutherland et al.[19] were used: BZR, COX2, and DHFR. Four further sets were created from the binding Db database,[20] either by selecting compounds which all had a measured K_i binding constant (CDK2 and QC) or measured IC50 values reported in the same publication (CDK1 and VEGFR2). The original publications containing this data are referenced in Table 1. The published QC data included several low molecular weight compounds, for example, ethanolamine, which were not suitable for the study, and so only compounds of molecular weight > 150 were included. The ET_A data set recently studied with Catalyst[21] was also incorporated, in order to compare to previous data. The

* Corresponding author e-mail: m.bodkin@lilly.com.
[†] Eli Lilly and Company Ltd., Surrey, England.
[‡] Eli Lilly and Company, Indianapolis, Indiana.

10.1021/ci7000082 CCC: $37.00
Published on Web 05/04/2007

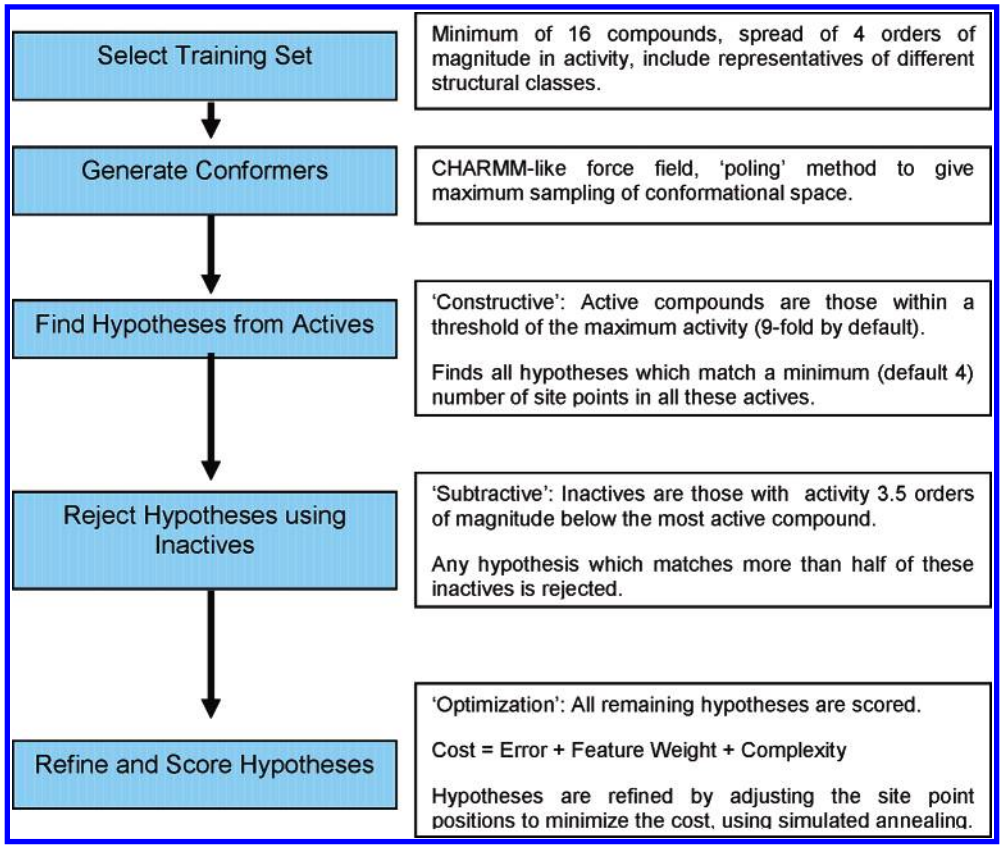

Figure 1. Summary description of Catalyst HypoGen methology.[47,48]

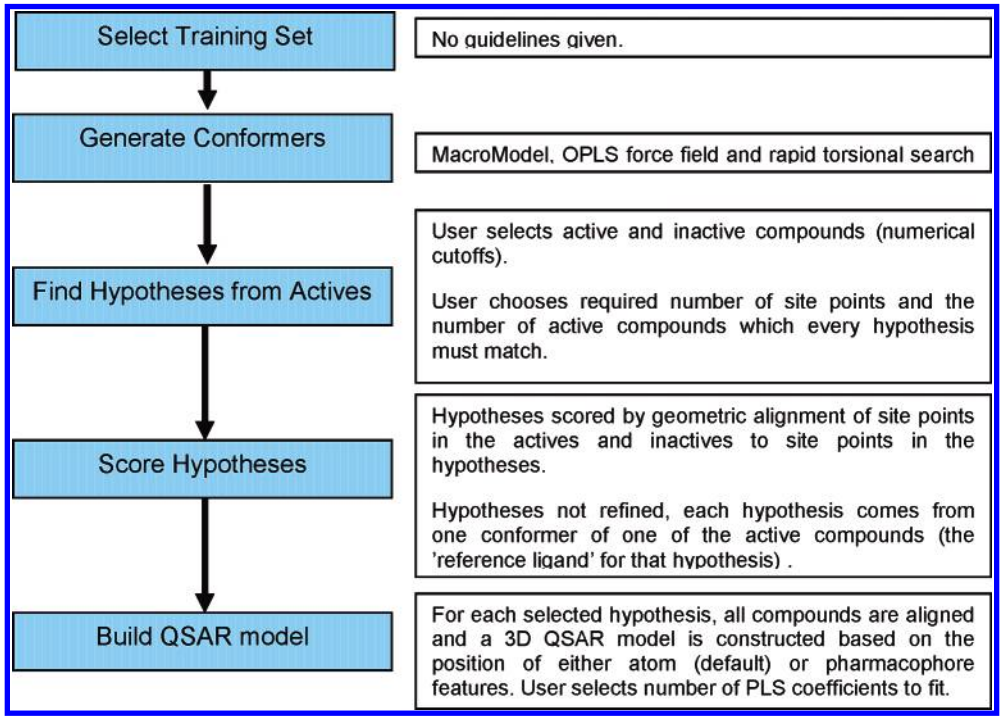

Figure 2. Summary description of Phase methodology.[18]

Molconvert program[22] was used to produce 3D structures.

2.2. Training and Test Set Selection. Since both Phase and Catalyst involve conformation searching and overlay of multiple structures to generate hypotheses, there is a practical limit on the number of compounds which can be used as a training set. We aimed to produce training sets of 20−40 compounds, and test sets of similar size to assess the predictivity of the models. As will be discussed in the Results section, we focus on the results for the test set rather than the internal consistency of the training set.

In order to remove subjective bias from the procedure, we employed an in-house program "CatScan" to select the training sets from the data sets in an automated fashion. CatScan is designed to roughly obey the recommended

Table 1. Data Sets Used in This Study[a]

protein receptor	n_{orig}	n_{train}	n_{test}	p*A* range	MPS	references
cyclin dependent kinase 2 (CDK2)	52	19	33	4.70−8.70	0.62	37−40
glutaminyl cyclase (QC)	63	35	28	2.36−7.22	0.47	41, 42
cyclin dependent kinase 1 (CDK1)	79	23	56	3.00−7.62	0.64	43, 44
vascular endothelial growth factor receptor 2 (VEGFR2)	63	23	40	4.30−8.52	0.77	45
endothelin A (ETA)	61	19	42	4.85−10.15	0.77	21
cyclooxygenase-2 (COX2)	467	43	38	4.00−9.00	0.62	19
dihydrofolate reductase (DHFR)	756	38	35	0.81−8.63	0.73	19
benzodiazepine receptor (BZR)	405	41	32	3.60−9.47	0.73	19

[a] n_{orig} is the total number of compounds sourced from the literature; n_{train} and n_{test} are the number of compounds included in the training and test sets, respectively, in the QSAR model building and testing. p*A* range is the activity range in the data set (combined training and test set). MPS is the mean pairwise similarity between all pairs of compounds in the data set, measured by the Tanimoto similarity coefficient between the MACCS fingerprints, as calculated by the MOE package (version 2005.06).[46]

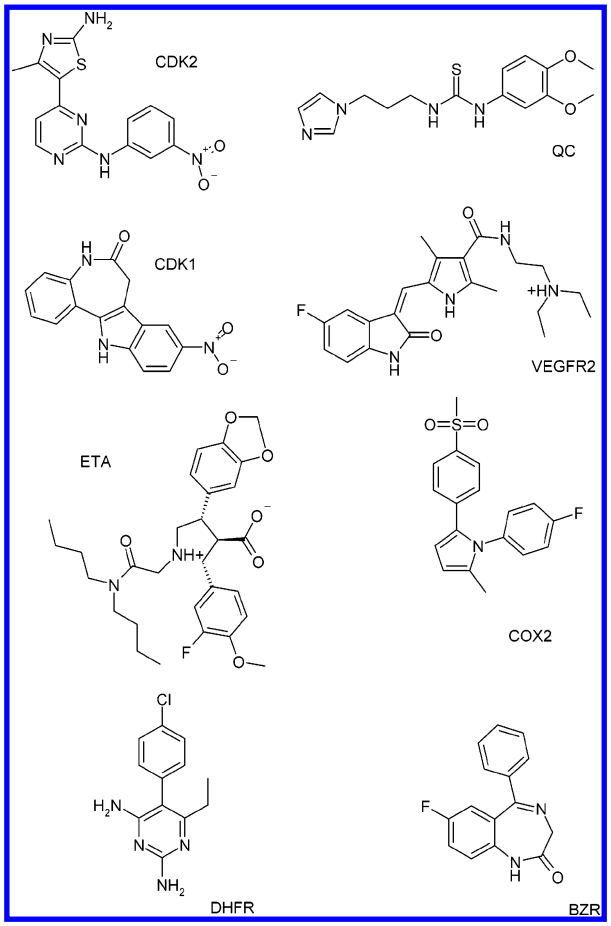

Figure 3. The most active compound in each data set, in the protonation state used in the model, shown as a representative.

"rules" for Catalyst model building,[14,23] that is, to select a series of compounds with a broad activity range (4 orders of magnitude is recommended) where no pair of compounds is included in which both are similarly active and similar structurally.

The following sequence of operations within CatScan was used:

1. Compounds are binned into activity classes, based on the order of magnitude.

2. For pairs of compounds within an activity bin which are more similar than a given threshold t_1 (defined by Tanimoto similarity with an in-house structural fingerprint, default 0.85), retain only one member of the pair.

Table 2. Features Input to Catalyst for Each Data Set[a]

data Set	features input to Catalyst
CDK2	H RA HBAl HBD PI
QC	H RA HBAl HBD PI
CDK1	H RA HBAl HBD PI
VEGFR2	H RA HBAl PI PC
ETA	H RA HBAl PI NI
COX2	H RA HBAl HBD PI
DHFR	H RA HBAl PI PC
BZR	H RA HBAl HBD PI

[a] *H* = HYDROPHOBIC, RA = RING_AROMATIC, HBA1 = HB_ACCEPTOR_lipid, HBD = HB_DONOR PI = POS_IONIZABLE, NI = NEG_IONIZABLE, PC = POS_CHARGE. In all cases, the allowed number of each feature could vary between 0 and 5.

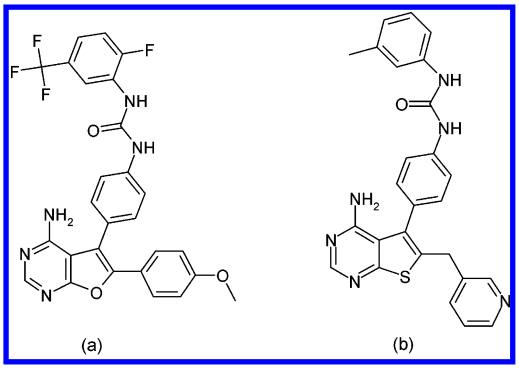

Figure 4. (a) Ligand from crystallized complex (PDB ID: 1YWN). (b) Example ligand from VEGFR2 data set used in this study.

3. If a list of >50 compounds remains, further cluster the compounds with single-linkage clustering, using chemically advanced template search fingerprints[24] and the Tanimoto coefficient. Consider each cluster at threshold t_2 (default 0.65) as a possible training set.

4. For any cluster larger than 35 compounds, remove compounds with unmarked chirality, then choose the most rigid seven compounds per order of magnitude in activity.

5. Check whether each cluster (or reduced cluster if processed in step 4 above) has > 15 members and spans at least 4 orders of magnitude in activity. If no valid training sets are found, repeat from step 2 with smaller values of t_1. If more than one valid training set is found, use the set with the most members.

For the larger data sets (COX2, BZR, and DHFR) we ran CatScan twice, the first time to select the training set and the second time on the remaining compounds to select the test set. For the other, smaller, data sets, CatScan was run

Table 3. The Highest R^2 Value from the Top Ten Scoring Hypotheses for Each System, Program, and Parameter Set[a]

	Phase						Catalyst					
system	1	2	3	4	5	best	1	2	3	4	5	best
CDK2	−0.34	−0.27	**0.06**		−0.20	**0.06**	−0.47	−0.24	**−0.09**	−0.26	−0.57	**−0.09**
QC	0.11	0.04	0.13	**0.20**	0.03	**0.20**	0.08	−0.03	−0.03	0.18	**0.22**	**0.22**
CDK1	0.37		*0.46*	*0.46*	***0.48***	***0.48***	0.08	0.00	−0.29	**0.32**	−0.08	**0.32**
VEGFR2	0.18	−0.69	0.17	0.17	**0.20**	**0.20**	**−0.31**	−0.59	−0.34	−0.49	−0.37	**−0.31**
ETA	***0.46***	0.17	0.18	0.17	0.24	***0.46***	−0.05	−0.62	−0.47	**0.16**	−0.09	**0.16**
COX2	0.18	0.13	0.18	**0.23**	0.18	**0.23**	−0.03	**0.23**	−0.03	0.13	−0.04	**0.23**
DHFR	**0.34**	0.26	0.27	0.27	0.26	**0.34**	0.19	**0.22**	0.19	0.19	0.22	**0.22**
BZR	0.26	0.11	***0.48***	***0.46***	***0.43***	***0.48***	0.09	0.08	0.09	0.17	0.10	0.17
mean	0.20	−0.03	0.24	0.25	0.20	0.31	−0.05	−0.12	−0.12	0.05	−0.08	0.12

[a] Values are in italics when $R^2 > 0.4$. Blank cells indicate that the program failed to find and score any hypotheses.

Table 4. The Highest r_p Value from the Top Ten Scoring Hypotheses for Each System, Program, and Parameter Set[a]

	Phase						Catalyst					
system	1	2	3	4	5	best	1	2	3	4	5	best
CDK2	0.32	**0.62**	0.39		0.37	**0.62**	0.46	0.53	0.51	**0.58**	0.46	**0.58**
GlutCyc	***0.78***	0.52	0.60	*0.73*	0.55	***0.78***	*0.74*	*0.76*	***0.77***	*0.74*	*0.70*	***0.77***
CDK1	0.65		***0.78***	***0.78***	***0.78***	***0.78***	0.36	0.39	0.35	**0.64**	0.39	**0.64**
VEGFR2	**0.67**	0.41	0.62	0.62	0.63	**0.67**	0.61	0.61	**0.63**	0.61	0.60	**0.63**
ETA	**0.69**	0.59	0.56	0.60	0.57	**0.69**	*0.76*	*0.74*	***0.78***	*0.76*	*0.75*	***0.78***
COX2	0.45	0.41	0.45	**0.49**	0.45	**0.49**	0.17	**0.44**	0.17	0.41	0.22	**0.44**
DHFR	**0.58**	0.52	0.52	0.52	0.51	**0.58**	0.54	0.57	0.55	**0.57**	0.53	**0.57**
BZR	0.52	0.39	***0.71***	0.68	0.67	***0.71***	0.45	0.43	0.42	0.34	**0.47**	**0.47**
mean	0.58	0.43	0.58	0.55	0.57	0.66	0.50	0.51	0.48	0.57	0.49	0.59

[a] Values are in italics when $r_p > 0.7$. Blank cells indicate that the program failed to find and score any hypotheses.

once to select the training set and all the remaining compounds formed the test set. More details of the data sets are given in Table 1, and representative compounds are shown in Figure 3.

2.3. Phase Methodology. Phase, version 2.0.212, was used for pharmacophore eludication and QSAR model building.[18] For each data set, we ran the Phase procedure with five different parameter sets. The command line interface to Phase was used.

1. Ligands were processed with the LigPrep program to assign protonation states appropriate for pH 7, using the "ionizer" subprogram.[18] Conformer generation was carried out with the MacroModel ligand torsion search[18] and default parameters. The default pharmacophore feature definitions were used in site generation. With the activity (K_i or IC50) defined as A, and $pA = -\log_{10} A$, active ligands were defined as those with $pA > pA_{max} - 3.0$, where A_{max} is the maximum activity in the training set. Inactive ligands were defined as those with $pA < pA_{max} - 3.0$. Hypotheses were generated by a systematic variation of the number of sites (n_{sites}) and the number of matching active compounds (n_{act}). With $n_{act} = n_{act\_tot}$ initially (n_{act_tot} is the total number of active compounds in the training set), n_{sites} was varied from 7 to 3 until at least one hypothesis was found and scored successfully. If this failed, then n_{act} was decreased by one and the n_{sites} cycle repeated. The scoring was done using the default parameters and the score weighted to include consideration of the alignment of inactive compounds, again with default parameters. All hypotheses successfully generated and scored in this manner were then used to build atom-based 3D QSAR models with 1−3 PLS factors. Statistics on the correlation of predicted with actual activity were collated for the top ten scoring hypotheses, by the default hypothesis scoring functions.

2. This step is identical to step 1, except a pharmacophore-based 3D QSAR model was used.

3. This step is identical to step 1, except the active ligand threshold was tightened to $pA > pA_{max} - 1.0$.

4. This step is identical to step 3, except the activity of the "reference ligand" was used in hypothesis scoring with a weighting of 1.0. Initially, within Phase, every conformer of every active ligand is a potential hypothesis, and the "Find Hypotheses from Actives" step (Figure 2) selects the best hypotheses, leaving each surviving hypothesis associated with a reference ligand.

5. This step is identical to step 3, except the alignment cutoff of ligands to hypotheses was decreased from 1.2 to 0.8.

A slight modification to this procedure was used for the ET_A set. Because of the flexible alkyl chains in these compounds, the original conformer libraries were very large and resulted in several hypothesis generation steps failing to complete in a reasonable time. The conformers for each ligand were therefore clustered using the XCluster program.[18] Clustering was done using the heavy atom Cartesian coordinates and the clustering level with the minimum value of the reordering entropy selected in each case.[25]

2.4. Catalyst Methodology. Catalyst, version 4.11, was used.[14] For each data set, we ran the HypoGen procedure with five different parameter sets. The command line interface to Catalyst was used.

1. Conformer generation was carried out with the "Best" algorithm, and default parameters. In hypothesis generation, the default feature dictionary was used, and for each data set, five feature types were chosen using an automated procedure based on SMARTS string matching (a feature of the CatScan program). The feature types are shown in Table 2. All other parameters were left at their default values,

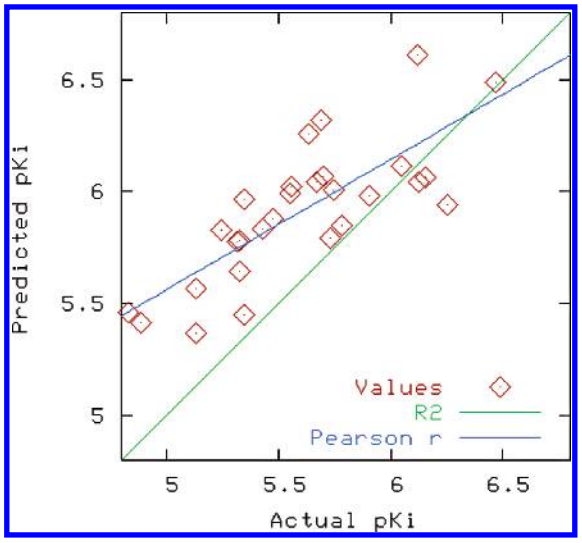

Figure 5. Predicted versus actual pK_i values for the Phase model with the highest r_p. The data points would need to be on the R^2 line for a R^2 value of 1.0, and on the r_p line for $r_p = 1.0$. $r_p = 0.78$ and $R^2 = 0.02$ for this model (another Phase hypothesis obtained with these parameters achieved an R^2 of 0.20, as indicated in Table 3).

including a minimum feature distance of 2.97 Å and activity uncertainty values for all ligands of 3.0. The ten best scoring hypotheses were refined and then used to estimate the activities of the test set, with "best" fitting of the test compounds to the hypothesis, and correlation statistics with the experimental activities collated.

2. This step is identical to step 1, except that the minimum feature distance was reduced to 1.5 Å.

3. This step is identical to step 1, except that the weights and tolerances of the hypothesis features were allowed to vary (VariableWeight 1 and VariableTolerance 1).[26]

Table 5. The Highest R^2 Value from the Top Ten Scoring Hypotheses for Each System and Parameter Set When Ligands Overlaid in Catalyst Are Input to Phase, as Described in Section 2.5[a]

	Phase					
system	1	2	3	4	5	best
CDK2	−0.32	−0.25	−0.11		**−0.09**	**−0.09**
QC	**0.15**	0.03	0.14	0.14	0.11	**0.15**
CDK1	*0.45*	0.10	***0.46***	***0.46***	*0.45*	***0.46***
VEGFR2	−0.23	−0.89	**0.07**	**0.07**	0.01	**0.07**
ETA	−0.17	0.06	**0.05**	**0.05**	**0.05**	**0.06**
COX2	0.23	0.07	**0.24**	**0.24**	0.23	**0.24**
DHFR	**0.23**	0.16	0.15	0.15	0.08	**0.23**
BZR	0.16	0.21	***0.42***	0.22	0.20	***0.42***
Mean	0.06	−0.09	0.18	0.19	0.13	0.19

[a] Values are in italics when $R^2 > 0.4$. Blank cells indicate that the program failed to find and score any hypotheses.

4. This step is identical to step 1, except that the minimum feature distance was reduced to 1.5 Å and the weights and tolerances of the hypothesis features were allowed to vary (VariableWeight 1 and VariableTolerance 1).

5. This step is identical to step 1, except one excluded volume feature was permitted to be added.[27]

2.5. Significance of Atom-Based Scoring Grid. To help determine the reasons for any different performance between the programs, we carried out further Phase runs on the overlays generated by Catalyst. The best performing Catalyst parameter set on average was set 4 (see Results section). Therefore, for each data set, the training and test set ligands overlaid onto the Catalyst hypothesis which gave the highest Pearson r (r_p) value from parameter set 4 were input into Phase as the sole conformer per ligand.

2.6. Use of Structural Data. The Protein Data Bank (PDB) ID 2C5P structure contains the most active ligand

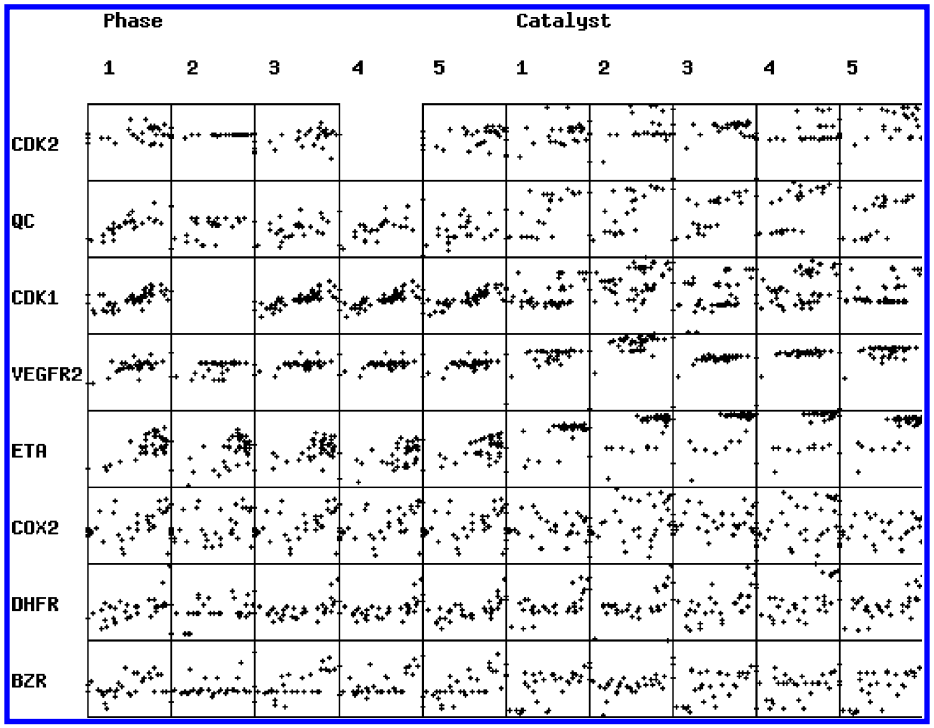

Figure 6. Predicted (y axis) versus actual (x axis) pK_i values for the models with the highest r_p. This is a graphical representation of the data used to derive the correlation coefficients in Table 4. The axes for each model are scaled to fit the experimental data for each set on the x axis and scaled to fit all the predicted values from every model for a set on the y axis.

Table 6. The Highest r_p Value from the Top Ten Scoring Hypotheses for Each System and Parameter Set When Ligands Overlaid in Catalyst Are Input to Phase, as Described in Section 2.5[a]

	Phase					
system	1	2	3	4	5	best
CDK2	0.03	**0.52**	0.17		0.25	**0.52**
QC	0.58	0.50	**0.58**	**0.58**	**0.58**	**0.58**
CDK1	*0.70*	0.42	***0.73***	***0.73***	*0.70*	***0.70***
VEGFR2	0.47	0.36	**0.54**	**0.54**	0.51	**0.54**
ETA	0.59	**0.65**	0.58	0.58	0.58	**0.65**
COX2	0.49	0.31	**0.49**	**0.49**	**0.49**	**0.49**
DHFR	**0.51**	0.48	0.43	0.43	0.39	**0.51**
BZR	0.46	0.52	***0.70***	0.51	0.50	***0.70***
Mean	0.48	0.47	0.53	0.55	0.50	0.59

[a] Values are in italics when $r_p > 0.7$. Blank cells indicate that the program failed to find and score any hypotheses.

from the CDK2 set bound to the CDK2 receptor in the ATP binding pocket. The conformers generated for this compound by the Phase (MacroModel) and Catalyst procedures were overlaid onto the crystallized conformation. For each program, the conformer with the lowest root-mean-squared deviation (RMSD) in the heavy atom coordinates was found, in order to assess whether the conformer generation had included the active conformation.

To further investigate the importance of good conformer generation, we constructed overlays of all the compounds in data sets onto crystal conformations for the two data sets for which relevant structures containing the target protein exist. For the CDK2 set, we used the ligand from the 2C5P structure, as above, and for the VEGFR2 set, we used the PDB ID 1YWN structure, which contains a ligand closely related in structure to those in the data set (see Figure 4).

Overlays were constructed using the ROCS program,[28] with both shape and chemical feature matching. The crystal conformation was used as a rigid template to overlay all the data set compounds. The best overlay for each compound in the training set was input into the Phase and HypoGen programs as the sole conformer at the hypothesis generation stage, therefore forcing the programs to use these overlaid conformations to construct the hypotheses. The same five parameter sets as in the original runs were employed in the two programs, and the quality of the hypotheses assessed by predicting the activities for the test set was as before.

3. RESULTS

The results for this analysis are presented in Tables 3 and 4. The highest test set correlation coefficient from the ten hypotheses evaluated with each of the parameter sets is presented for each system. For Phase, the highest correlation from the three maximum PLS factors used is reported. Table 3 contains the R^2 values, which can be interpreted as the fraction of the variance in the experimental test set activities which is explained by the model. Obtaining the optimal value of R^2 (1.0) would require that the predicted values correlate with the actual values with a slope of 1.0 and zero offset. A less stringent measure is the r_p value, given in Table 4. An optimal value of r_p can be obtained with any value of the slope and any offset. Full definitions of R^2 and r_p are given in the Appendix.

For the purposes of 3D QSAR modeling, a significant r_p alone might indicate a useful model, insofar as it suggests that a hypothesis and associated features and/or grid-based coefficients may partially explain the trend in the data and so form a basis for further molecular modeling work.

For both programs, it was observed that the hypothesis with the highest correlation coefficient was not in general the best scoring hypothesis, by "cost" in the case of Catalyst or the hypothesis scoring function in Phase. Therefore, an important recommendation from this study is to directly test the predictivity of as many hypotheses as possible to find the most successful one, rather than relying on the scoring of the hypotheses alone.

When the optimal correlation coefficients are compared, the results indicate better performance for the Phase models than that for the Catalyst models, with the exception of the ET_A set (which was previously published as a Catalyst example).

Concerning the selection of Phase parameters, we note that the pharmacophore-feature-based QSAR models (set 2) perform consistently worse than the atom-based models, with the exception of the CDK2 system. On average, better R^2 values are obtained with sets 3, 4, and 5, suggesting that restricting the number of active compounds used to build hypotheses is advantageous. However, we note that the differences are small and that optimal r_p values are obtained with parameter set 1 in four systems (QC, VEGFR2, ETA, and DHFR), indicating that the best choice of parameters is system-dependent.

For Catalyst, the best parameter set on average is set 4, involving variable feature weights and tolerances and a minimum feature spacing lower than the default. However, as for Phase, the differences are not large, and for some sets, the best results are obtained with the other parameter sets. Indeed, the biggest variation in results with a particular program is not between parameter values, but between data sets. The best Phase models are obtained with the CDK1 and BZR data sets, which both contain relatively rigid compounds.

It is also interesting that a good r_p value (0.78) was obtained for the QC test set, despite the best R^2 value being

Table 7. The Highest r_p Value from the Top Ten Scoring Hypotheses for Each System, Program, and Parameter Set, with the Data Set Compounds Overlaid onto a Crystallized Conformation of a Ligand in or Closely Related to the Data Set[a]

	Phase						Catalyst					
system	1	2	3	4	5	best	1	2	3	4	5	best
CDK2[b]	0.40	0.57	0.28		0.28	0.57	0.55	0.56	0.55	0.55	0.53	0.56
CDK2[c]	0.38	0.32				0.38	0.54	0.55	0.58	0.55	0.51	0.58
VEGFR2	0.48	0.30	0.43	0.55	0.52	0.55	0.58	0.58	0.58	0.59	0.61	0.61

[a] Blank Cells Indicate That the Program Failed to Find and Score Any Hypotheses. [b] All the compounds in the original CDK2 data set were used. [c] Only the 45 anilino-pyrimidine compounds used.

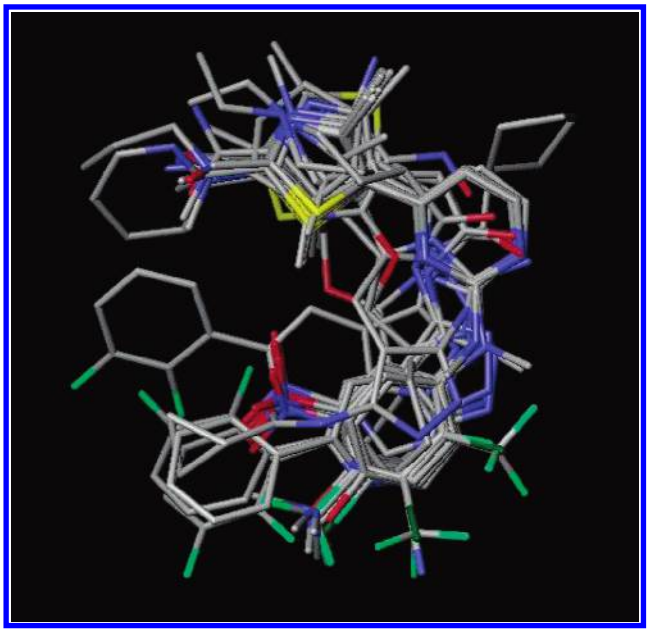

Figure 7. ROCS overlay of CDK2 training set compounds onto the ligand conformation from the 2CP5 PDB structure.

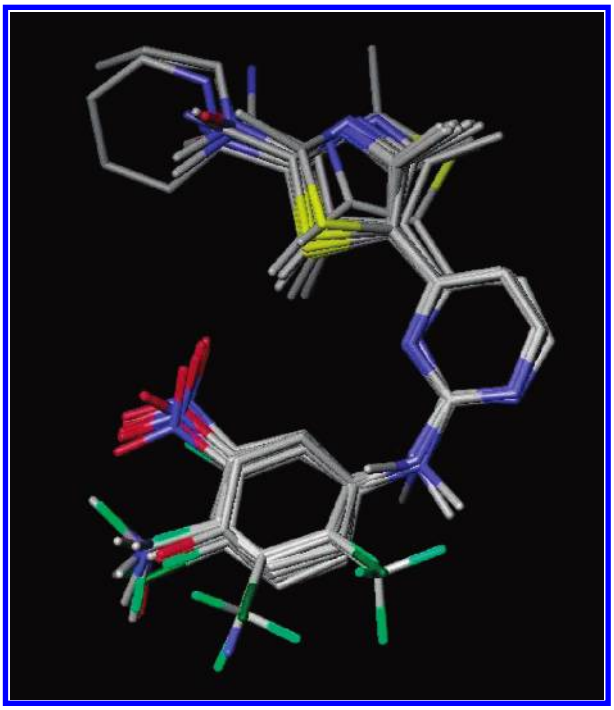

Figure 8. ROCS overlay of aniline−pyrimidine CDK2 training set compounds onto the ligand conformation from the 2CP5 PDB structure.

poor (0.20) (see Figure 5). The compounds in this data set are more structurally diverse (mean pairwise similarity, MPS = 0.47) and flexible than the CDK1 and BZR sets; so, it is encouraging that a relatively good model can still be found in this case.

The ET_A compounds produce the best Catalyst models. A comparison of the best hypothesis in this study with that published by Cucarull-Gonzalez et al.[21] shows that different hypothesis features were found (3 H, 1 HBAl, and 1 PI here; 2 H, 1 HBAl, 1 PI, and 1 NI in ref 29). The hypotheses are not identical, which is unsurprising given that a different

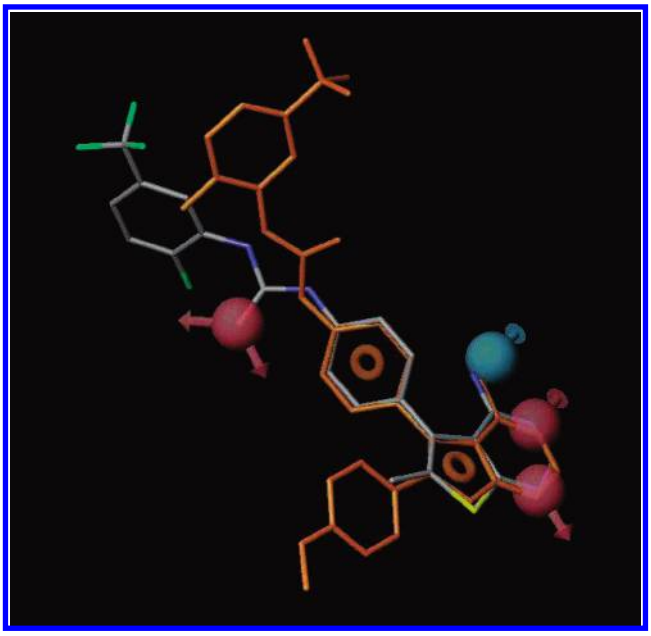

Figure 9. The most active VEGFR2 data set compound (CPK colors) overlaid with the best Phase hypothesis and the 1YWN crystal conformation.

selection of training set compounds was made and differing parameters were used, but the predictivity of our best model ($r_p = 0.78$) was comparable to that reported by the previous authors ($r_p = 0.7247$).

The predictions for the best r_p model are shown as scatterplots in Figure 6, to give a more visual impression of the quality of the results.

3.1. Significance of Atom-Based Scoring Grid. The results in terms of R^2 and r_p values for the different Phase parameter sets (as above) are shown in Tables 5 and 6. A comparison to Tables 3 and 4 shows that there are relatively few significant differences in the performance of Phase when Catalyst-generated overlays are used as opposed to Phase-generated overlays. This result indicates that the improved performance of Phase over Catalyst is due largely to the use of the grid-based 3D-QSAR model. The exception to this finding is the results for the ET_A set, which show significantly worse performance in R^2 predictivity when the Catalyst overlays are used in Phase. The fact that better r_p coefficients are obtained from Catalyst for this system indicates that an overlay which is good for one scoring scheme is not necessarily good for another, which raises the question of which scoring scheme is likely to select overlays closest to the actual experimental binding mode.

3.2. Use of Structural Data. As described in section 2.6, the most active ligand from the CDK2 data set has been crystallized in complex with the CDK2 receptor (PDB ID: 2C5P). The smallest heavy-atom RMSD of the conformers found by MacroModel for Phase to the crystal structure was 2.4 Å; for Catalyst, it was 3.1 Å. These RMSD values are worse than the typical performance of such conformer-generation algorithms.[30]

The training set compounds were overlaid on the crystal conformers and input to hypothesis generation as described in section 2.6, and the results on the test set are shown in Table 7. There is no improvement in the correlation statistics in comparison with the original results, indicating that

poor conformational sampling may not be the key issue in the poor performance of the programs in these two cases.

The overlaid training set compounds are shown in Figure 7. For the CDK2 set, it is apparent that there is sufficient diversity in the compounds to make the overlays ambiguous, and subsets of the compounds binding in different modes are a possibility. We therefore repeated the model building but restricted the test and training set compounds to contain only members of a particular structural class (the anilinopyrimidines as depicted in Table 1). The overlay is shown in Figure 8, but again the results, as shown in Table 7, were not an improvement. This result further indicates that for some data sets there are limitations to the activity prediction methods beyond conformational sampling and compound overlay. A recent paper applying Catalyst to a CDK2 data set reports more success, but only in terms of discriminating active from inactive compounds, rather than correlation statistics as used here.[27]

For the VEGFR2 set, we compared the conformation predicted by the best original hypothesis (Phase parameter set 1, $r_p = 0.67$) with the crystal conformation from the complex (PDB ID: 1YWN; Figure 9). It is clear that part of the molecule is overlaid correctly, but the urea group has an incorrect orientation. A comparison to the crystallized structure shows that in the complex the NH groups of the urea interact with the E883 carboxylate group, an interaction which could not be predicted by a ligand-based approach.

As for the CDK2 data set, the training set compounds were overlaid onto the crystal conformation with ROCS, and the overlaid conformers were used as input to Phase and Catalyst. The results are shown in Table 7. The Catalyst model shows a slight improvement in r_p, and the Phase model is unimproved.

4. CONCLUSIONS

To summarize briefly the performance of both programs, "good" ($r_p > 0.7$ or $R^2 > 0.4$) models were obtained for four of the eight data sets for Phase and two of the eight for Catalyst. We acknowledge that there are very many further variations of training set definition and parameter set selection that an expert user could employ to potentially obtain better results with either program on a given data set. The aim of this study is to evaluate these programs within a rapid, automated workflow. Such an approach is often necessary within a lead optimization project, where new compounds are added to the data sets regularly and there is only a short amount of time to construct models if they are to influence the discovery process.

With the choices of data set and parameter reported here, we find that the Phase program performs better than Catalyst, in terms of average R^2 correlation between experimental and predicted activities, and comparably in terms of r_p correlation. For both programs, we found it advantageous to build and test models from at least the top ten scoring hypotheses, as there was little correlation between hypothesis score and predictivity. Within Phase, we found that the atom-based grid QSAR model generally performed better than the pharmacophore-based grid, and by using overlays from Catalyst to build Phase grid QSAR models, we found evidence that the better performance of Phase on these data sets was due to the use of the grid technique.

A recent similarly automated comparison of descriptors for QSAR on different data sets demonstrated that the best performing descriptors used in a PLS model could obtain a model with predictive $R^2 > 0.5$ on 15−30% of sets, depending on the method used to select the training set.[31] A comparison with the results presented here would indicate that neither Phase nor Catalyst perform as well on average as the best fingerprint-descriptor-based QSAR methods in terms of R^2 predictivity alone.

In two systems (CDK2 and VEGFR2), models obtained from Phase and Catalyst when training set compounds are input overlaid onto the crystal structure do not show a marked improvement over models generated in standard operation from conformer libraries. This result indicates that inadequate conformer generation is not always responsible for the poor performance of these methods. There are several reasons why 3D QSAR methods might fail even if compounds appear to be consistently overlaid in an atom- or feature-based comparison. There may be insufficient compounds in the training set, which is typically more of an issue for three-dimensional methods because of the computational cost of conformational analysis and examining possible overlays. Different compounds might interact with differing sets of receptor features, meaning that no single pharmacophore can explain all the data. Flexible receptors can also adopt multiple conformations when binding to different ligands, again meaning that a single, rigid pharmacophore is inappropriate. Finally, one must also consider that the free energy of ligand binding depends on the free state of the ligand as well as the bound state modeled by the pharmacophore. The BZR and CDK1 data sets in this study, for which relatively good models could be constructed, contain relatively rigid compounds where there is unlikely to be a significant change in entropy on binding.

Receptor flexibility is a critical issue for structure-based design methods,[32] and the free energy of the unbound ligand is also a problematic term for docking scoring functions.[33] Typical examples of docking successes involve virtual screening, where the scoring function is employed mainly to discriminate between active and inactive compounds.[1] The performance of scoring functions in predicting the absolute activity of compounds is typically disappointing.[34] The current situation of ligand-based pharmacophore methods may be similar, and several recent reports have focused on virtual screening applications,[35,36] rather than predictive QSAR studies. We hope that the results presented here will inform further development of automated 3D QSAR methodology.

ACKNOWLEDGMENT

The authors thank Sandra Milutinovic, Dr. Howard Broughton, and application scientists and developers from Accelrys, Inc. and Schrodinger, LLC. for helpful suggestions and technical assistance.

APPENDIX: CORRELATION FORMULAS

The R^2 coefficient used here is calculated as

$$R^2 = 1 - \frac{\sum_i^{n_{test}} (x_i^{pred} - x_i^{exp})^2}{\sum_i^{n_{test}} (x_i^{exp} - \langle x^{exp} \rangle)^2}$$

where x_i^{exp} and x_i^{pred} are the experimental and predicted activity of compound x.

The Pearson r coefficient r_p is calculated via

$$r_p^2 = \frac{[\sum_i^{n_{test}} (x_i^{pred} - \langle x^{pred} \rangle)(x_i^{exp} - \langle x^{exp} \rangle)]^2}{\sum_i^{n_{test}} (x_i^{pred} - \langle x^{pred} \rangle)^2 \sum_i^{n_{test}} (x_i^{exp} - \langle x^{exp} \rangle)^2}$$

We note that R^2 can vary between +1.0 (ideal model) and $-\infty$, where negative values can be taken to mean that the error in the predicted values is larger than the variance of the experimental data about the mean. r_p ranges between +1.0 (perfect correlation) and −1.0 (perfect anticorrelation).

REFERENCES AND NOTES

(1) Clark, D. E. What has Computer-Aided Molecular Design Ever Done for Drug Discovery? *Expert Opin. Drug Discovery* **2006**, *1*, 103−110.

(2) Martin, Y. Distance Comparisons (DISCO): A New Strategy for Examining 3D Structure−Activity Relationships. In *Classical and 3D QSAR in Agrochemistry*; Hansch, C., Fujita, T., Eds.; American Chemical Society: Washington, DC, 1995; pp 318−329.

(3) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations among Molecules. *J. Chem. Inf. Model.* **1996**, *36*, 563−571.

(4) Jones, G.; Willett, P.; Glen, R. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(5) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647−671.

(6) Richmond, N.; Abrams, C.; Wolohan, P.; Abrahamian, E.; Willett, P.; Clark, R. GALAHAD: 1. Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567−587.

(7) Kurogi, Y.; Guner, O. F. Pharmacophore Modelling and Three-Dimensional Database Searching for Drug Design Using Catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035−1055.

(8) Li, H.; Sutter, J.; Hoffman, R. HypoGen: An Automated System for Generating 3D Predictive Pharmacophore Models. In *Pharmacophore Perception, Development and Use in Drug Design*; Guner, O., Ed.; International University Line: La Jolla, CA, 2000.

(9) Cramer, R. D.; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(10) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(11) Jalaie, M.; Erickson, J. Homology Model Directed Alignment Selection for Comparative Molecular Field Analysis: Application to Photosystem II Inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 181−197.

(12) Evers, A.; Gohlke, H.; Klebe, G. Ligand-Supported Homology Modelling of Protein Binding-Sites Using Knowledge-Based Potentials. *J. Mol. Biol.* **2003**, *334*, 327−345.

(13) Tafi, A.; Bernardini, C.; Botta, M.; Corelli, F.; Andreini, M.; Martinelli, A.; Ortore, G.; Baraldi, P. G.; Fruttarolo, F.; Borea, P. A.; Tuccinardi, T. Pharmacophore Based Receptor Modeling: The Case of Adenosine A_3 Receptor Antagonists. *J. Med. Chem.* **2006**, *49*, 4085−4097.

(14) *Catalyst*, version 4.11, Accelrys, Inc.: San Diego, CA, 2006.

(15) Bureau, R.; Daveu, C.; Lemaitre, S.; Dauphin, F.; Landelle, H.; Lancelot, J. C.; Rault, S. Molecular Design Based on 3D-Pharmacophore. Application to 5-HT_4 Receptor. *J. Chem. Inf. Model.* **2002**, *42*, 962−967.

(16) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of Conformational Analysis Techniques To Generate Pharmacophore Hypotheses Using Catalyst. *J. Chem. Inf. Model.* **2005**, *45*, 461−476.

(17) Debnath, A. K. Pharmacophore Mapping of a Series of 2,4-Diamino-5-deazapteridine Inhibitors of *Mycobacterium avium* Complex Dihydrofolate Reductase. *J. Med. Chem.* **2002**, *45*, 41−53.

(18) *Phase*, version 2.0.212; Schrodinger, LLC: New York, 2005.

(19) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Model.* **2003**, *43*, 1906−1915.

(20) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719−725.

(21) Cucarull-Gonzalez, J. R.; Laggner, C.; Langer, T. Influence of the Conditions in Pharmacophore Generation, Scoring, and 3D Database Search for Chemical Feature-Based Pharmacophore Models: One Application Study of ETA- and ETB-Selective Antagonists. *J. Chem. Inf. Model.* **2006**, *46*, 1439−1455.

(22) *Molconvert*, version 4.0.4; Chemaxon, Ltd.: Budapest, Hungary, 2005.

(23) Kim, S. G.; Yoon, C. J.; Kim, S. H.; Cho, Y. J.; Kang, D. I. Building a Common Feature Hypothesis for Thymidylate Synthase Inhibition. *Bioorg. Med. Chem.* **2000**, *8*, 11−17.

(24) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. 'Scaffold-Hopping' by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.

(25) *MacroModel XCluster Manual*, version 9.1; Schrodinger, LLC: New York, 2005.

(26) Sutter, J.; Guner, O.; Hoffman, R.; Li, H.; Waldman, M. Effect of Variable Weights and Tolerances on Predictive Model Generation. In *Pharmacophore Perception, Development and Use in Drug Design*; Guner, O. F., Ed.; International University Line: La Jolla, CA, 2000.

(27) Toba, S.; Srinivasan, J.; Maynard, A. J.; Sutter, J. Effect of Variable Weights and Tolerances on Predictive Model Generation. *J. Chem. Inf. Model.* **2006**, *46*, 728−735.

(28) *ROCS*, version 2.1; OpenEye Scientific Software: Santa Fe, NM, 2005.

(29) Cucarull-Gonzalez, J. R.; Laggner, C.; Langer, T. Influence of the Conditions in Pharmacophore Generation, Scoring, and 3D Database Search for Chemical Feature-Based Pharmacophore Models: One Application Study of ET_A- and ET_B-Selective Antagonists. *J. Chem. Inf. Model.* **2006**, *46*, 1439−1455.

(30) Bostrom, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.

(31) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.

(32) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.

(33) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein−Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.

(34) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; Lalonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(35) Langer, T.; Hoffman, R. D. Pharmacophore Modelling: Applications in Drug Discovery. *Expert Opin. Drug Discovery* **2006**, *1*, 261−267.

(36) Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A. The Discovery of New 11β Hydroxysteroid Dehydrogenase Type 1 Inhibitors by Common Feature Pharmacophore Modeling and Virtual Screening. *J. Med. Chem.* **2006**, *49*, 3454−3466.

(37) Wang, S.; Wood, G.; Meades, C.; Griffiths, G.; Midgley, C.; Mcnae, I.; Mcinnes, C.; Anderson, S.; Jackson, W.; Mezna, M.; Yuill, R.; Walkinshaw, M.; Fischer, P. M. Synthesis and Biological Activity of 2-Anilino-4-(1H-pyrrol-3-yl) Pyrimidine CDK Inhibitors. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4237−4240.

(38) Wang, S.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; Mcnae, I.; Wu, S. Y.; Mcinnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. 2-Anilino-4-(thiazol-5-yl)-

pyrimidine CDK Inhibitors: Synthesis, SAR Analysis, X-ray Crystallography, and Biological Activity. *J. Med. Chem.* **2004**, *47*, 1662−1675.
(39) Witherington, J.; Bordas, V.; Haigh, D.; Hickey, D. M.; Ife, R. J.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Ward, R. W. 5-Aryl-pyrazolo[3,4-b]pyridazines: Potent Inhibitors of Glycogen Synthase Kinase-3 (GSK-3). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1581−1584.
(40) Witherington, J.; Bordas, V.; Garland, S. L.; Hickey, D. M.; Ife, R. J.; Liddle, J.; Saunders, M.; Smith, D. G.; Ward, R. W. 5-Aryl-pyrazolo[3,4-b]pyridines: Potent Inhibitors of Glycogen Synthase Kinase-3 (GSK-3). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1577−1580.
(41) Buchholz, M.; Heiser, U.; Schilling, S.; Niestroj, A. J.; Zunkel, K.; Demuth, H. U. The First Potent Inhibitors for Human Glutaminyl Cyclase: Synthesis and Structure−Activity Relationship. *J. Med. Chem.* **2006**, *49*, 664−677.
(42) Schilling, S.; Niestroj, A. J.; Rahfeld, J. U.; Hoffmann, T.; Wermann, M.; Zunkel, K.; Wasternack, C.; Demuth, H. U. Identification of Human Glutaminyl Cyclase as a Metalloenzyme. Potent Inhibition by Imidazole Derivatives and Heterocyclic Chelators. *J. Biol. Chem.* **2003**, *278*, 49773−49779.
(43) Kunick, C.; Lauenroth, K.; Wieking, K.; Xie, X.; Schultz, C.; Gussio, R.; Zaharevitz, D.; Leost, M.; Meijer, L.; Weber, A.; Jorgensen, F. S.; Lemcke, T. Evaluation and Comparison of 3D-QSAR CoMSIA Models for CDK1, CDK5, and GSK-3 Inhibition by Paullones. *J. Med. Chem.* **2004**, *47*, 22−36.
(44) Leost, M.; Schultz, C.; Link, A.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Bibb, J. A.; Snyder, G. L.; Greengard, P.; Zaharevitz, D. W.; Gussio, R.; Senderowicz, A. M.; Sausville, E. A.; Kunick, C.; Meijer, L. Paullones are Potent Inhibitors of Glycogen Synthase Kinase-3beta and Cyclin-Dependent Kinase 5/p25. *Eur. J. Biochem.* **2000**, *267*, 5983−5994.
(45) Dai, Y.; Guo, Y.; Frey, R. R.; Ji, Z.; Curtin, M. L.; Ahmed, A. A.; Albert, D. H.; Arnold, L.; Arries, S. S.; Barlozzari, T.; Bauch, J. L.; Bouska, J. J.; Bousquet, P. F.; Cunha, G. A.; Glaser, K. B.; Guo, J.; Li, J.; Marcotte, P. A.; Marsh, K. C.; Moskey, M. D.; Pease, L. J.; Stewart, K. D.; Stoll, V. S.; Tapang, P.; Wishart, N.; Davidsen, S. K.; Michaelides, M. R. Thienopyrimidine Ureas as Novel and Potent Multitargeted Receptor Tyrosine Kinase Inhibitors. *J. Med. Chem.* **2005**, *48*, 6066−6083.
(46) *MOE*; Chemical Computing Group, Inc.: Montreal, Canada, 2006.
(47) Li, H.; Sutter, J.; Hoffman, R. HypoGen: An Automated System for Generating 3D Predictive Pharmacophore Models. In *Pharmacophore Perception, Development and Use in Drug Design*; Guner, O., Ed.; International University Line: La Jolla, CA, 2000.
(48) Kurogi, Y.; Guner, O. F. Pharmacophore Modelling and Three-Dimensional Database Searching for Drug Design Using Catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035−1055.

CI7000082