J Comput Aided Mol Des (2006) 20:647–671
DOI 10.1007/s10822-006-9087-6

ORIGINAL PAPER

PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results

Steven L. Dixon · Alexander M. Smondyrev · Eric H. Knoll · Shashidhar N. Rao · David E. Shaw · Richard A. Friesner

Received: 28 June 2006 / Accepted: 17 October 2006 / Published online: 24 November 2006

Summary We introduce PHASE, a highly flexible system for common pharmacophore identification and assessment, 3D QSAR model development, and 3D database creation and searching. The primary workflows and tasks supported by PHASE are described, and details of the underlying scientific methodologies are provided. Using results from previously published investigations, PHASE is compared directly to other ligand-based software for its ability to identify target pharmacophores, rationalize structure-activity data, and predict activities of external compounds.

Keywords Pharmacophore perception · 3D QSAR · 3D databases · Ligand-based design

Introduction

As a practical matter, computer-aided molecular design is frequently split into disciplines that focus on either structure-based or ligand-based techniques. When sufficient information is available or inferable about the structure of the biological target and its binding site, then it is possible to invoke a structure-based approach, wherein specific ligand–receptor interactions are studied to help identify new molecules with activity toward the target. If, however, knowledge about the structure of the target is limited, but a sufficient number of actives have already been identified, then ligand-based methods provide alternative ways of leveraging the available information into models that can help identify new actives.

While ligand-based design formally includes any number of computational methods that rely only on the structure of known and potential ligands, it has become largely synonymous with pharmacophore modeling [1, 2] and quantitative structure-activity relationships (QSAR) [3, 4]. Dating back several decades, numerous academic and pharmaceutical discovery labs have done and continue to do pioneering work in these areas [5–14], but there is also significant reliance on commercially available tools such as Catalyst [15, 16], DISCO [17], GASP [18] and CoMFA [19], in part because of the expertise and resources required to develop and maintain ligand-based software that is suitable for all manner of drug discovery projects.

The relative success of any ligand-based methodology may be attributed to a combination of factors, including demonstrated scientific validity, novelty, ease of use, and integration with other software and experimental workflows. Note, however, that favorable performance in these areas does not necessarily prevent the devaluation of a particular piece of commercial software when it fails to deliver satisfactorily in a critical investigation. As a result, researchers in drug discovery companies frequently resort to internal development to overcome perceived deficiencies in existing commercial software, or engage in an eclectic

S. L. Dixon (✉) · A. M. Smondyrev · E. H. Knoll · S. N. Rao · D. E. Shaw · R. A. Friesner
Schrödinger, Inc., 120 W. 45th St., 29th Floor, New York, NY 10036, USA
e-mail: dixon@schrodinger.com

E. H. Knoll · R. A. Friesner
Department of Chemistry, Columbia University, New York, NY 10027, USA

D. E. Shaw
D E Shaw & Co, 120 W. 45th St., 39th Floor, New York, NY 10036, USA

approach that emphasizes the particular strengths of each ligand-based tool at their disposal. Thus despite the decades of work in the field of ligand-based design, the associated problems are far from being solved [20], and the development of improved, more flexible tools is vital to the continued goal of accelerating drug discovery and controlling its costs.

Toward this end, we introduce PHASE, a comprehensive, self-contained system for pharmacophore perception, QSAR model development, and 3D database screening. We provide details of the underlying PHASE methodology, and strategies for tackling relevant, practical problems of varying difficulty. To illustrate the relative merits of this new system, we present results of head-to-head comparisons between PHASE and other widely used commercial ligand-based software packages.

PHASE methodology

Overview

An important consideration in designing pharmacophore-based software is the highly approximate nature of the models it frequently furnishes. While this does not call into question the basic premises upon which pharmacophore modeling is founded, it does underscore the importance of capitalizing on the knowledge and expertise of the user wherever possible. With this consideration in mind, PHASE was designed to provide a high degree of flexibility and feedback, emphasizing the user as an integral part of the pharmacophore development process. The ultimate goal is not to provide a single model that is deemed to be the best by some predetermined measure, but rather to suggest a set of plausible models that can be evaluated by diverse criteria whose relevance is assessed by the user.

Figure 1 summarizes the major tasks and workflows supported by PHASE. Both graphical and command line interfaces exist to provide access to an integrated suite of tools that facilitate the development of pharmacophore models and 3D QSAR models, and the creation and screening of 3D databases.

Pharmacophore models may be created manually using a single *reference ligand* structure, or through an automated procedure, wherein common pharmacophores are exhaustively perceived among a group of actives, then scored according to various geometric and heuristic criteria, yielding a set of ranked pharmacophore hypotheses. This scoring procedure may rely on information from just the actives, or it may incorporate data from inactives as well.

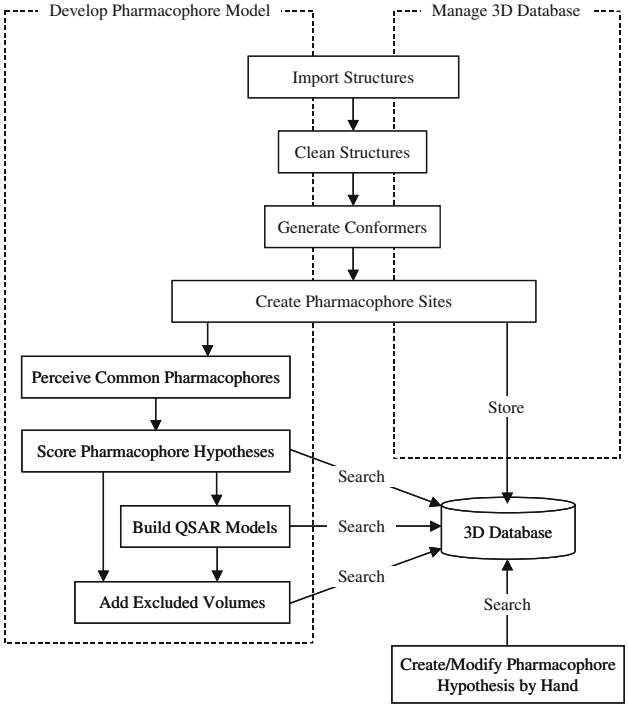

Fig. 1 PHASE Workflows

When experimental activities are known, a 3D QSAR model may be created for each hypothesis, using ligand structures that are aligned to the associated pharmacophore on at least three points. A QSAR model may consider the entire ligand structure, or just the pharmacophoric features that can be mapped to the hypothesis. The former case is appropriate for congeneric series with limited flexibility, whereas the latter is recommended for datasets with significant diversity and/or flexibility.

A given hypothesis may be augmented with a set of excluded volume spheres, to map out regions of space that cannot be occupied by ligands that bind with high affinity. The locations of those spheres can be assigned manually, or through a variety of automated techniques that consider the space occupied by actives and inactives, or the space occupied by the receptor to which a reference ligand is bound.

Ultimately, a pharmacophore model can be used to search a 3D database to identify additional molecules that satisfy the hypothesis. If the model satisfactorily embodies characteristics that are critical for ligand binding, the *hits* from a database search should possess a greater than average probability of being active.

Throughout these workflows, users are presented with an array of configurable parameters that may be

assigned to reflect general or specific a priori knowledge, to improve the tractability of the problem at hand, or to explore a wide range of models that are consistent with the known data. Because each dataset presents its own characteristic challenges, there is no universal set of PHASE parameters that is optimal or even appropriate for every situation. However, as we shall detail, there are basic rules of thumb that can be applied to make intelligent decisions about how to approach most problems of practical interest.

Preparing ligands

Before undertaking the tasks of pharmacophore model development and 3D database creation, low-energy, 3D structures must be available for each molecule of interest. Accordingly, PHASE incorporates a structure cleaning step utilizing LigPrep [21], which attaches hydrogens, converts 2D structures to 3D, generates stereoisomers, and, optionally, neutralizes charged structures or determines the most probable ionization state at a user-defined pH. PHASE also allows for the importation of 3D structures prepared outside its own workflow.

Because one does not generally know the structure that a given molecule will adopt if and when it binds to a target protein, it is customary to represent each molecule as a series of 3D structures that sample the thermally accessible conformational states. For purposes of pharmacophore model development, PHASE provides two built-in approaches, both of which employ the MacroModel conformational search engine [22].

The first approach involves a rapid torsion angle search followed by minimization of each generated structure using either the MMFFs [23] or OPLS_2005 [24] force field, with implicit GB/SA or distance-dependent dielectric solvent model. The torsion search samples ring conformations, invertible pseudo-chiral nitrogens, and all rotatable bonds within a *core* region, which includes everything from the center of a molecule out to, but not including, the last rotatable bond along each path. Torsion angles for these terminal groups are varied either one-at-a-time or simultaneously, according to the user's preference. As torsions are sampled, a truncated force field is applied to identify dihedral minima, and the overall energy of each structure is estimated by combining dihedral potentials with internal ring energies. Structures with high estimated energies are eliminated, as are structures with close non-bonded contacts.

As an alternative to the rapid torsion search, conformational space may be explored through a combination of Monte–Carlo Multiple Minimum (MCMM) [25] sampling and Low Mode (LMOD) conformational searching [26]. Monte–Carlo sampling provides excellent global coverage of the potential energy surface, while low mode searching facilitates effective treatment of local regions with complex and/or problematic characteristics, such as saddle points. Although more computationally demanding, the mixed MCMM/LMOD approach is among the most powerful and generally applicable conformational searching methods currently available.

Whether the initial search utilizes torsion sampling or mixed MCMM/LMOD, the minimized structures that are ultimately obtained are filtered through a user-defined relative energy window, typically 5–10 kcal/mol, and a redundancy check, where any two structures within 1 kcal of each other are deemed to be equivalent if all corresponding pairs of heavy atoms in the two structures are within a user-defined distance, typically 1–2 Å. By varying these parameters in conjunction with the maximum number of conformers initially sampled, any desired level of conformational coverage may be achieved.

Most drug-like molecules are adequately represented by anywhere from tens to hundreds of conformers, but molecules with greater than about 10 rotatable bonds may require thousands of conformers to achieve thorough coverage. While PHASE does not impose any preset limits on the number of conformers stored, the computational demands that accompany large, highly flexible ligands can make certain steps in the workflow very expensive, and sometimes infeasible. Thus it is important to recognize that not every problem can be approached at the most rigorous level; rather, one must occasionally employ more approximate strategies and accept that the solutions obtained will be less precise.

Practical considerations are particularly important when it comes to 3D databases. While the aforementioned conformer generation techniques are routinely applied to develop PHASE pharmacophore models, they are usually impractical for all but the smallest of 3D databases. Depending upon the size and flexibility of the molecules being studied, thorough conformational sampling with full minimization may require up to several minutes per molecule. This translates to years of CPU time to create a 3D database of one million compounds. While the task may be split across multiple CPUs, considerable computing resources would still be required for an extended period of time.

To address the demands of creating conformers for large 3D databases, PHASE provides an option to simply use the structures produced by rapid torsion

sampling. Although the conformers are not fully minimized, the filters applied during the sampling process effectively eliminate unrealistic structures that are far away from any local minima on the potential energy surface. This approach allows thorough coverage of conformational space with only modest computational requirements. Running on a 1.7 GHz Pentium 4 processor, conformers can typically be created for 1–2 database molecules each second, and scaling is nearly linear when multiple processors are utilized.

Creating pharmacophore sites

For purposes of pharmacophore model development, each ligand structure is represented by a set of points in 3D space, which coincide with various chemical features that may facilitate non-covalent binding between the ligand and its target receptor. These *pharmacophore sites* are characterized by type, location and, if applicable, directionality. In accordance with the most commonly cited explanations of ligand–receptor binding, PHASE provides six built-in types of pharmacophore features: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobe (H), negative ionizable (N), positive ionizable (P), and aromatic ring (R). In addition, users may define up to three custom feature types (X, Y, Z) to account for characteristics that don't fit clearly into any of the six built-in categories.

By default, a hydrogen bond acceptor site is positioned on a surface-accessible atom that carries one or more donatable lone pairs, and a vector attribute is assigned to each idealized hydrogen bond axis, according to the hybridization of the acceptor atom, Fig. 2a. Likewise, a hydrogen bond donor site is centered on each donatable hydrogen atom, and a single vector feature is directed along its idealized hydrogen bond axis, Fig. 2b. As an alternative to this ligand-centric convention, users may choose to represent acceptors and donors as pure projected points, located at the complementary positions on a theoretical binding site, Fig. 2c–d. The projected point approach does not incorporate vector character into the site definition, so as to permit the situation where two ligands form hydrogen bonds to the same receptor atom, but from different locations and directions.

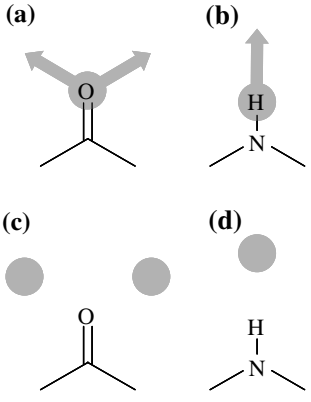

Fig. 2 Hydrogen bond acceptor and donor mappings based on the use of vector features (**a**, **b**) and pure projected points (**c**, **d**)

Hydrophobic sites are assigned using a procedure that has been described in detail previously [15], so we provide only an overview here. Rings, isopropyl groups, t-butyl groups, various halogenated moieties, and chains as long as four carbons are each treated as a single hydrophobic site. Chains of five or more carbons are broken into smaller fragments containing between two and four carbons ($5 \rightarrow 2 + 3$, $6 \rightarrow 3 + 3$, $7 \rightarrow 3 + 4$, etc.), and each fragment is designated as a separate hydrophobic site. The location $\mathbf{r}_H$ of a given hydrophobic site is a weighted average of the positions of the non-hydrogen atoms in the associated fragment:

$$\mathbf{r}_H = \frac{\sum_i s_i t_i \mathbf{r}_i}{\sum_i s_i t_i} \quad (1)$$

Here, s_i is the solvent-accessible surface area of atom i, computed using a probe radius of 1.4 Å, and t_i is a hydrophobicity factor that ranges between 0 and 1. Polar atoms (O, N, S) are assigned a hydrophobicity of 0, whereas carbons and halogens at least three bonds from any polar atom receive a value of 1. Intermediate hydrophobicities are assigned to carbons and halogens when polar atoms are within a distance of two bonds. Figure 3 illustrates hydrophobic sites for a variety of fragments.

Negative and positive ionizable sites are modeled as a single point located on a formally charged atom, or at the centroid of a group of atoms over which the ionic charge is shared, Fig. 4. Because most commonly

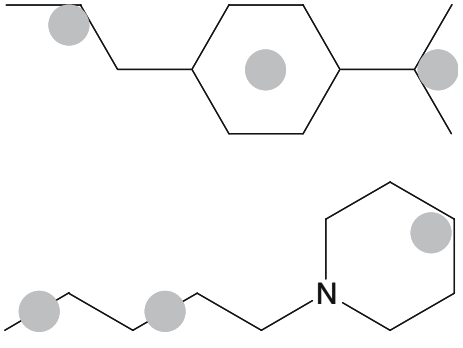

Fig. 3 Hydrophobic feature mappings

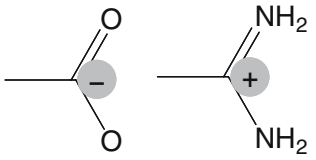

Fig. 4 Negative ionizable and positive ionizable feature mappings

occurring ionizable centers are automatically recognized by PHASE, the structures provided need not be ionized explicitly. While minimized ionic structures may differ somewhat from their neutral counterparts, the locations of the ionizable sites are usually quite similar, so the impact on pharmacophore models is only minor.

Finally, if the user so chooses, aromatic rings may be distinguished from other hydrophobic groups, and designated as a separate type of pharmacophore feature (i.e., ''R'' rather than ''H''). In that case, a single site is placed at the centroid of each aromatic ring, and a two-headed vector normal to the plane of the ring is associated with the site, Fig. 5. Unlike acceptors and donors, an aromatic feature cannot be represented as a pure projected point.

The rules that are applied to map the positions of pharmacophore sites are known as *feature definitions*, and they are represented internally by a set of SMARTS [27] patterns. As such, the rules may be customized by adding new SMARTS patterns and/or by ignoring one or more of the built-in patterns. When a custom feature type (X, Y, Z) is created, any number of SMARTS patterns may be used to define precisely which structural characteristics should be recognized as occurrences of that feature.

Perceiving common pharmacophores

In the absence of reliable co-crystallographic data, pharmacophore discovery must ultimately rely on the active analog approach [28] to identify common characteristics that provide hypotheses to explain ligand–receptor binding. Various algorithms [16–18, 29] have been developed to address the problem of common pharmacophore perception, with varying levels of approximation, assumption, and heuristics to improve the tractability of the task. In contrast to these methods, PHASE employs an exhaustive analysis of k-point pharmacophores culled from the conformations of a set of actives, and identifies all spatial arrangements of pharmacophore features that are shared by those molecules, according to a user-defined matching tolerance.

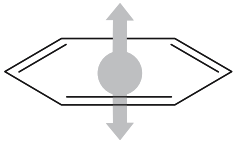

Fig. 5 Aromatic ring feature mapping

By default, PHASE will look for pharmacophores that are common to all actives, but this condition can be relaxed so that a common pharmacophore need match only a subset of the actives. The ligands that are matched may vary from one common pharmacophore to another, so the user does not have to choose any particular subset of ligands, just the number. Henceforth, we shall say that a common pharmacophore must match a minimum required number of actives, where the minimum number is set by the user.

Note that partial matching is not considered in this process, so a k-point pharmacophore must be matched on all k sites by a minimum required number of actives. As such, there is no deliberate construction of a ''union-of-features'' pharmacophore model that's intended to cover strong, moderate, and weak binders, or molecules that bind in different modes. Rather, the requirement of full k-point matching across different subsets of actives provides an automated means of identifying pharmacophore models that explain one or more unique, high-affinity binding modes.

Common pharmacophores are perceived using a tree-based partitioning technique that groups together similar pharmacophores according to their intersite distances, i.e., the distances between pairs of sites in the pharmacophore. Thus a k-point pharmacophore is represented by a vector of n distances, where $n = (k \cdot (k-1))/2$. Each intersite distance d is filtered through a binary decision tree, as illustrated in Fig. 6. This particular tree has a depth of four and partitions distances on the interval (0, 8] into terminal nodes that are 1 Å wide. By filtering all n distances in this manner, the pharmacophore is assigned to an n-dimensional box, whose sides are equal in length to the terminal node width, which we denote as ε.

All pharmacophores of a given *variant* (i.e., a particular combination of the feature types A, D, H, etc.) that are mapped into the same terminal box are considered to be similar enough to facilitate identification of a common pharmacophore. So if each of a minimum required number of actives contributes at least one pharmacophore to a particular box, then that box represents a common intersite distance pharmacophore, and each member of the box is a candidate pharmacophore hypothesis. Because the boundaries of

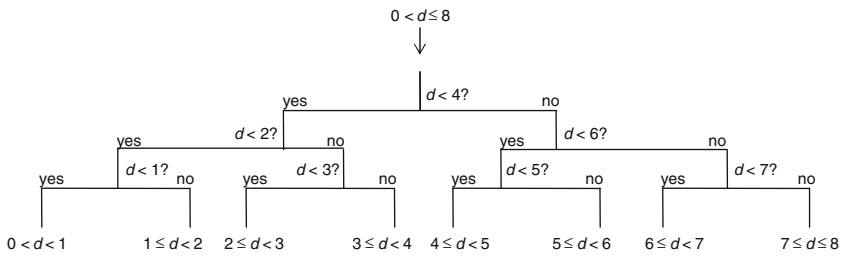

Fig. 6 Illustration of the tree-based partitioning technique used to identify common pharmacophores based on intersite distances. This tree has a depth of four, and partitions distances on the interval (0, 8] into terminal nodes that are 1 Å wide

a box are artificial, pharmacophores in neighboring boxes may also be counted. Thus if a terminal box contains at least one pharmacophore, it can still give rise to a common pharmacophore if the minimum required number of actives is represented when the pharmacophores from that box and its direct neighbors are pooled. Such boxes are said to survive the partitioning procedure, while all others are eliminated.

The efficiency of the algorithm lies in the fact that the tree need not be fully traversed in order to eliminate boxes. At each branch point, pharmacophores are split into left and right subspaces according to whether the applicable intersite distance d is less than or greater than some bisecting value β. In addition, pharmacophores in the left subspace and for which $\beta - d \leq \varepsilon$ are flagged as neighbors of the right subspace, because they would ultimately be neighbors in that dimension if partitioned into terminal boxes of width ε. Conversely, pharmacophores in the right subspace and for which $d - \beta \leq \varepsilon$ are flagged as neighbors of the left subspace. If all pharmacophores are partitioned to the β level in a particular dimension, and it is found that one of the subspaces, when combined with its neighbor region, fails to contain pharmacophores from the minimum required number of actives, then the entire subspace and all its subtrees can be safely eliminated from further consideration. This frequently occurs at a fairly shallow depth in the overall tree, eliminating a tremendous number of computations and comparisons.

Because the PHASE algorithm incorporates a binary decision tree, comparisons to the recursive partitioning-based SCAMPI method [13, 14] naturally arise. However, the mechanics and goals of these two approaches are fundamentally different. PHASE filters pharmacophores through a series of nodes whose left and right branches correspond to whether or not a particular intersite distance in the pharmacophore is less than some boundary value. By contrast, SCAMPI filters *compounds* through a series of nodes whose left and right branches correspond to whether or not a particular pharmacophore is contained within the compound. The goal of PHASE is to exhaustively identify pharmacophores that are common to a set of actives; the goal of SCAMPI is to identify pharmacophores that best separate compounds into more active and less active groups.

Scoring pharmacophores with respect to actives

Each box that survives the partitioning procedure described in the previous section contains a set of pharmacophores that are highly similar in the space of intersite distances. Any member of a surviving box may in fact constitute a common pharmacophore, but additional steps must be taken to verify that this is actually the case. First of all, the intersite distance representation does not distinguish mirror images, so least-squares site-to-site alignments must be done among the surviving pharmacophores to eliminate false positives. Second, even if mirror image effects are absent, least-squares alignments may reveal poor superpositions of one or more site points or vector features, calling for the elimination of certain pharmacophores, or even an entire box.

The process of scoring with respect to actives is designed to filter out these sorts of inappropriate pharmacophores and to identify within each box a top-ranked representative, henceforth referred to as the pharmacophore hypothesis for that box. Hypotheses are assigned a score comprised of geometric and heuristic factors that can be weighted according to the user's preference. At this point only information from the actives is used; a subsequent procedure may be invoked to consider information from inactives and adjust hypothesis scores accordingly.

Each pharmacophore from a surviving box is treated temporarily as a *reference* in order to assign a score. Accordingly, all other non-reference pharmacophores from that box and its neighboring boxes are aligned, one-by-one, to the reference pharmacophore, and the quality the alignments are measured using two criteria: (1) the root-mean-squared deviation (RMSD) in the site point positions and (2) the average cosine of the angles formed by corresponding pairs of vector features (acceptors, donors and aromatic rings). These factors are combined with separate weights to yield a

combined site + vector score for each non-reference pharmacophore i that's been aligned to the reference:

$$\begin{aligned}\text{Site\_Vector\_Score}_i = {} & w_{\text{site}}\text{Site\_Score}_i \\ & + w_{\text{vector}}\text{Vector\_Score}_i,\end{aligned} \tag{2}$$

where

$$\text{Site\_Score}_i = 1 - \text{RMSD}_i/\text{cutoff}_{\text{RMSD}} \tag{3}$$

$$\text{Vector\_Score}_i = \frac{1}{n_{\text{v}}}\sum_{j=1}^{n_{\text{v}}}\cos\theta_{ij} \tag{4}$$

The parameters w_{site}, w_{vector}, and $\text{cutoff}_{\text{RMSD}}$, and are user-adjustable (with default values of 1.0, 1.0 and 1.2, respectively), n_{v} is the number of vector features in the hypothesis, and θ_{ij} is the angle between the jth vector feature in the non-reference pharmacophore and the corresponding vector feature in the reference pharmacophore. Note that if a non-reference ligand contributes more than one pharmacophore, the one yielding the highest value of Eq. 2 is selected. The total Site_Vector_Score for a given reference pharmacophore is the average value of Eq. 2, evaluated over all pairs of selected alignments.

In principle, a reference pharmacophore could yield a satisfactory average value for Eq. 2, even though it contains one or two poor individual alignments. To avoid this situation, PHASE will exclude a reference pharmacophore as a potential hypothesis if, for any individual alignment, RMSD_i exceeds $\text{cutoff}_{\text{RMSD}}$, or Vector_Score_i falls below $\text{cutoff}_{\cos\theta}$, which is another user-defined parameter (with default value 0.5). In addition, individual tolerances may be applied to the positions of the superimposed pharmacophore features, so that any single poorly aligned feature will be grounds for rejection, even if RMSD_i does not exceed $\text{cutoff}_{\text{RMSD}}$.

Frequently, it is desirable to rely on certain heuristic measures when selecting a representative pharmacophore from a surviving box. Accordingly, PHASE supports the use of a conformationally independent property, such as $-\log K_i$, to bias the selection of reference ligands to favor those with higher activities, or higher values of whichever property is incorporated. The weighted property term is combined with Site_Vector_Score to yield an overall Reference_Score:

$$\text{Reference\_Score} = \text{Site\_Vector\_Score} + w_{\text{prop}}\text{Prop}_{\text{ref}} \tag{5}$$

Here, Prop_{ref} is the value of the conformationally independent property for the ligand contributing the reference pharmacophore. Observe that if the property is binary-valued (0/1), the score can be biased to effectively limit the selection of reference pharmacophores to a subset of actives for which the property is non-zero.

After all pharmacophores in a box have been treated as a reference, the one yielding the highest Reference_Score is selected as the hypothesis to represent that box. The ligand that contributes the reference pharmacophore is referred to as the *reference ligand* for that hypothesis. Note that the non-reference information is carried along with each hypothesis so that additional scoring may be performed using the best multi-ligand alignment identified at this stage.

Once hypotheses have been identified across all boxes, the user may wish to eliminate some of the lower scoring ones. Accordingly, a percentage cutoff may be applied to the Reference_Score so that, e.g., only hypotheses in the top 10% are retained. Further refinement may then be done using volume scoring, selectivity scoring, reference ligand relative conformational energy, and the number of actives matched.

Volume scoring measures how well each non-reference ligand overlays with the reference ligand, based on van der Waals models of the structures and taking into account all heavy atoms:

$$\text{Volume\_Score}_i = \text{Volume}_{i,\text{common}}/\text{Volume}_{i,\text{total}} \tag{6}$$

$\text{Volume}_{i,\text{common}}$ is the common or overlapping volume between ligand i and the reference ligand, and $\text{Volume}_{i,\text{total}}$ is the total volume occupied by both ligands. The overall Volume_Score for a hypothesis is the average obtained from applying the above formula to all non-reference ligands i. Volume_Score is then added to Reference_Score with its own adjustable weight (1.0 by default).

Selectivity is an empirical estimate of the *rarity* of a hypothesis, i.e., the fraction of molecules likely to match the hypothesis, regardless of their activity toward the receptor. The selectivity estimate is computed from the physical characteristics of a hypothesis, and it is defined on a logarithmic scale, so a value of 2 means that $1/10^2$ random, drug-like molecules would be expected to match the hypothesis. Higher selectivity is desirable because it indicates that the hypothesis is more likely to be unique to the actives. For a detailed description of how selectivity is estimated, the reader is referred to Appendix A. As with the other types of

scores, selectivity is added to the overall score with its own user-adjustable weight (1.0 by default).

Since a hypothesis is a pharmacophore that's observed in a particular conformation of some ligand, it is hoped that this computed structure will bear a close resemblance to the bound ligand structure. Although the bound and free ligands are certain to have different potential energy surfaces, it is generally assumed that the bioactive structure will correspond to a reasonably low point on the energy surface of the free ligand. Accordingly, it may be desirable to incorporate the reference ligand relative conformational energy $\mathrm{EConf}_{\mathrm{ref}}$ into the scoring process. To favor lower energies, $\mathrm{EConf}_{\mathrm{ref}}$ is added to the total hypothesis score with a negative weight.

Finally, the user may wish to assign higher scores to hypotheses that match a greater number of actives. This is relevant when the minimum required number of actives is smaller than the total number of actives. The reward comes in the form of $w^{M\text{-}1}$, where w is user-adjustable (1.0 by default) and M is the number of actives that match the hypothesis.

The total *active* score for a given hypothesis is then:

$$\begin{aligned}\mathrm{Active\_Score} = {} & \mathrm{Reference\_Score} \\ & + w_{\mathrm{volume}}\mathrm{Volume\_Score} \\ & + w_{\mathrm{selectivity}}\mathrm{Selectivity\_Score} \\ & - w_{\mathrm{conf}}\mathrm{EConf}_{\mathrm{ref}} \\ & + w_{\mathrm{match}}^{M-1}\end{aligned} \tag{7}$$

While this score provides an overall ranking of the hypotheses, it is not intended to imply that the top-scoring hypothesis is more correct than all the others. As indicated previously, PHASE provides a set of plausible models, and it is up to the user to examine any number of the higher scoring hypotheses to determine which are most useful and most consistent with any a priori information. To aid in this process, PHASE supports additional means for assessing the significance of a hypothesis, such as scoring with respect to inactives and QSAR model development.

Scoring pharmacophores with respect to inactives

Oftentimes, pharmacophore models are developed from a set of actives that are built on a common framework. As a result, any number of high-scoring hypotheses may emerge with pharmacophore features from this framework, and it may not be easy to identify the spurious models using only information from the actives. Accordingly, PHASE provides a means for penalizing hypotheses that fail to discriminate actives from inactives, thus effectively elevating pharmacophore models composed only of features that are essential for high-affinity binding.

Before detailing this procedure, it is important to recognize that a molecule may be inactive for any number of reasons, including steric clashes with the receptor, a large desolvation penalty, or excessive entropy loss upon binding. Unfortunately, molecules that fail to bind for these reasons may not be helpful for purposes of identifying spurious hypotheses, because they may in fact be capable of adopting a structure that places all the critical pharmacophore features in the correct spatial arrangement. Therefore, the process of scoring with respect to inactives *assumes* that a failure to bind is due entirely to a pharmacophoric deficiency in the associated molecule. While it is usually not possible to ascertain the cause of inactivity in every case, by choosing the inactives from among weak binders (e.g., low micromolar compounds) one can increase the chances of including molecules that contain some, but not all, of the essential pharmacophore features in the correct spatial arrangement.

A k-point hypothesis is scored with respect to inactives by searching the pharmacophore space of those inactives and finding all m-point matches to the hypothesis, where $3 \leq m \leq k$. The best match i provided by each inactive is determined by way of a fitness score:

$$\begin{aligned}\mathrm{Fitness}_i = {} & w_{\mathrm{site}}\mathrm{Site\_Score}_i + w_{\mathrm{vector}}\mathrm{Vector\_Score}_i \\ & + w_{\mathrm{volume}}\mathrm{Volume\_Score}_i\end{aligned} \tag{8}$$

The quantities in this formula are directly analogous to those in eqs. 3, 4, and 6, except for a modification of RMSD_i that corrects for matching m out of k sites:

$$\mathrm{RMSD}_i = \left[\frac{m}{k}\mathrm{RMSD}_{i,m}^2 + \frac{k-m}{k}\mathrm{cutoff}_{\mathrm{RMSD}}^2\right]^{1/2} \tag{9}$$

Observe that a volume term is included in Eq. 8, so that even if steric clashes are the root cause of inactivity for certain molecules, the potential exists for incorporating those effects into the fitness score. More specifically, if an inactive molecule contains bulky groups that are absent from the reference ligand for a given hypothesis, the volume score for that inactive will be correspondingly reduced.

For a valid hypothesis, all inactives should ideally exhibit relatively low fitness, so the overall score is reduced by the average fitness observed across a set of N inactives, multiplied by a user-adjustable weight w_{inactive} (1.0 by default):

$$\text{Adjusted\_Score} = \text{Active\_Score} - w_{\text{inactive}} \frac{1}{N} \sum_{i=1}^{N} \text{Fitness}_i \quad (10)$$

It should be apparent that greater penalties will be assessed against hypotheses with high fitness scores, i.e., hypotheses that are readily matched by inactives.

Building 3D QSAR models

If a sufficient number of molecules of varying activity are available, a 3D QSAR model can be developed for each hypothesis using training set structures that match the pharmacophore on three or more sites, and employing fitness (Eq. 8) as the criterion to select the best alignment for each molecule. A QSAR model may be atom-based or pharmacophore-based, the difference being whether all atoms are taken into account, or merely the pharmacophore sites that can be matched to the hypothesis. The choice of which type of model to create depends largely on whether or not the training set molecules are sufficiently rigid and congeneric. If the structures contain a relatively small number of rotatable bonds and some common structural framework, then an atom-based model may work quite well. However, if the molecules are highly flexible or if they exhibit significant chemical diversity, a pharmacophore-based model may be more appropriate.

In atom-based QSAR, a molecule is treated as a set of overlapping van der Waals spheres. To encode the basic characteristics of local chemical structure, each atom (and hence each sphere) is placed into one of six categories according to a simple set of rules: hydrogens attached to polar atoms are classified as hydrogen bond donors (D); carbons, halogens, and C–H hydrogens are classified as hydrophobic/non-polar (H); atoms with an explicit negative ionic charge are classified as negative ionic (N); atoms with an explicit positive ionic charge are classified as positive ionic (P); non-ionic nitrogens and oxygens are classified as electron-withdrawing (W); and all other types of atoms are classified as miscellaneous (X).

Alternate atom-typing schemes were investigated, including the reassignment of electron-withdrawing atoms into separate categories to account for different hybridization, hydrogen bond acceptor tendency, π-electron donation, etc. However, we found that the creation of additional categories generally led to poorer predictions on compounds outside the training set for a variety of systems. This is probably due to the tendency to over-fit the training set data when using a larger number of independent variables with lower statistical significance per variable.

The QSAR atom typing rules are generally consistent with the default pharmacophore feature definitions, but there are some important differences. For example, the pharmacophore feature definitions use fairly complex rules to identify hydrophobic regions, whereas atom-based QSAR does not. Pharmacophore feature definitions are also able to treat a given atom as part of two different pharmacophore sites, but atom-based QSAR requires that each atom be assigned to only one category.

In sharp contrast to the atom-based approach, pharmacophore-based QSAR models are concerned only with the sites on a molecule that can be matched to the hypothesis. Each such site is treated as a sphere with a user-adjustable radius, and categories are assigned according to the feature type (A, D, H, N, P, R, X, Y, Z). Pharmacophore-based models offer the advantage of not being as sensitive to the consistency of the overall molecular superpositions, so the models generated have greater applicability to molecules from diverse chemical families. However, a pharmacophore-based QSAR cannot account for factors beyond the pharmacophore model itself, such as possible steric clashes with the receptor. This requires consideration of the entire molecular structure, i.e., an atom-based QSAR.

Prior to constructing a QSAR model, a rectangular grid is defined to encompass the space occupied by the aligned training set molecules. This grid divides space into uniformly sized cubes, typically 1 Å on each side, which are occupied by the atoms or pharmacophore sites that define each molecule. A given atom or site is deemed to occupy a cube if the center of that cube falls within the radius of the corresponding sphere. A single cube may be occupied by more than one atom or site, and that occupation may come from the same molecule or from different molecules. Each occupied cube gives rise to one or more *volume bits*, where a separate bit is allocated for each different category of atom/site that occupies the cube. The total number of volume bits assigned to a given cube is based on occupations from all training set molecules. A molecule may thus be represented by a string of zeros and ones, according to the cubes it occupies, and the different types of atoms/sites that reside in those cubes. Figure 7 illustrates how the occupation pattern of a molecule is mapped to a bit string. Note that while different molecules will exhibit different patterns of zeros and ones, the total number of bits and the meaning of each bit are the same for every molecule.

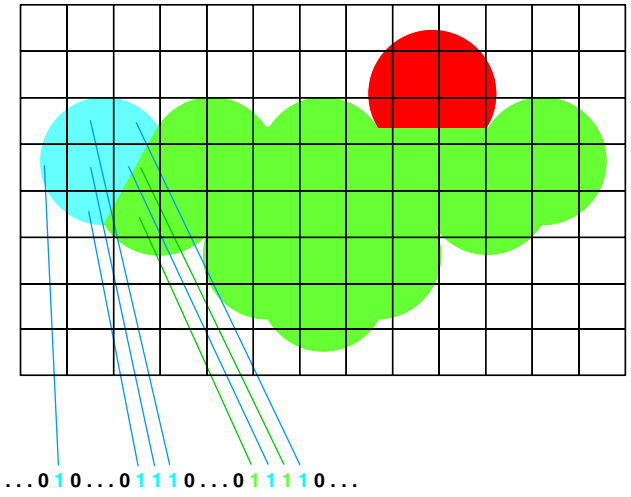

Fig. 7 The mapping of a molecule to a volume bit pattern that provides the independent variables for a PHASE 3D QSAR model. For simplicity, only two dimensions are represented

Since the bit string is simply a collection of binary-valued 3D descriptors, it is possible to treat the bits as a pool of independent variables for purposes of QSAR model development. Because the number of bits is typically much larger than the number of training set molecules, PHASE QSAR models are created by applying partial least squares (PLS) regression to this pool of binary-valued independent variables. The PLS procedure ultimately results in the assignment of a regression coefficient to each bit, which facilitates the identification of specific chemical features that tend to increase or decrease the estimated activity. The reader is referred to Appendix B for further details about the PLS algorithm used in PHASE.

As with most regression-based QSAR techniques, it is assumed that a linear relationship exists between the independent variables and the free energy of binding, so the dependent activity variable should be directly proportional to free energy, for example, $-\log K_i$, or $-\log IC_{50}$. For a given QSAR, PHASE creates a series of regression models, incorporating progressively more PLS factors, with the maximum number of factors being no larger than 1/5 the number of training set molecules. Although it has become rather customary to use *leave-n-out* predictions as a means of determining an appropriate number of PLS factors, it should be recognized that internal cross-validation techniques such as this cannot provide a reliable assessment of how a particular QSAR model will perform on molecules outside the training set. Therefore, PHASE supports only external validation, using a true test set whose structures and activities are not considered when QSAR models are developed.

Figure 8 provides a visual representation of an atom-based PHASE QSAR model in the context of a single high-affinity ligand. Blue cubes correspond to volume bits that are set (i.e., occupied) by the ligand, and for which the associated regression coefficients are above a particular positive, user-defined threshold. These regions represent characteristics of the ligand structure that have a moderate to strong positive effect on the calculated activity. By contrast, the red cubes represent moderate to strong negative effects on calculated activity.

Creating excluded volumes

Once a pharmacophore hypothesis has been developed, PHASE provides a variety of ways to define excluded volumes, i.e., regions of space that cannot be occupied by any molecule which is aligned to that hypothesis. In all cases, excluded volumes are represented by hard spheres, and a violation occurs if the van der Waals surface of a molecule penetrates the surface of an excluded volume sphere. PHASE supports the use of van der Waals models that include only the heavy atoms in a molecule and models that include all atoms. Currently, excluded volumes are applied as a filter when searching a 3D database or other source of structures, but they are not incorporated into QSAR models.

The most basic method for defining excluded volumes is a simple ''point-and-click'' approach that creates a sphere of user-definable radius at any desired location. Manual placement allows precise control over the excluded volume model, and it is generally practical when a small number of spheres will suffice. It

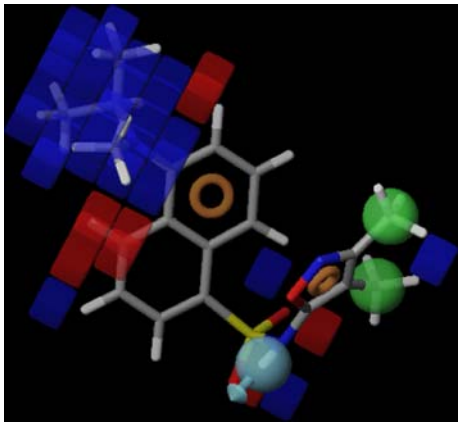

Fig. 8 Visual representation of an atom-based PHASE QSAR model in the context of a single high-affinity ligand, where favorable (blue) characteristics dominate unfavorable (red) characteristics

may become somewhat unwieldy, however, if large regions of space must be covered and/or complex shapes must be represented. Therefore, PHASE provides various automated means of creating excluded volumes that require minimal effort on the part of the user.

For example, given a set of aligned actives and inactives, PHASE can systematically identify locations where the placement of a sphere would cause excluded volume violations only for the inactives. This is done by defining a rectangular grid around the aligned molecules, placing a probe sphere at each point on the grid, and retaining only spheres that intersect the van der Waals surface of one or more inactives, while remaining clear of the actives. To provide a measure of tolerance when matching new, potentially active molecules, a buffer construct can be invoked to require that there be a gap of user-defined width between the excluded volume surface and the active van der Waals surface. Grid points that do not preserve this gap are then skipped.

It is also possible to create excluded volumes to define an overall shape constraint for molecules that are oriented to the hypothesis. In this case, a rectangular grid is constructed to surround one or more aligned actives, and spheres are positioned to form an excluded volume shell, with or without a buffer. Figure 9 contains an example created using a single active, a grid spacing of 1.0 Å, and a buffer distance of 2.0 Å. Only half of the shell is shown so that the view of the ligand is not obscured. This technique is rather simplistic compared to Van Drie's ''Shrink-Wrap'' method [30], but the granularity of the grid can be adjusted to achieve any desired level of precision in the shape.

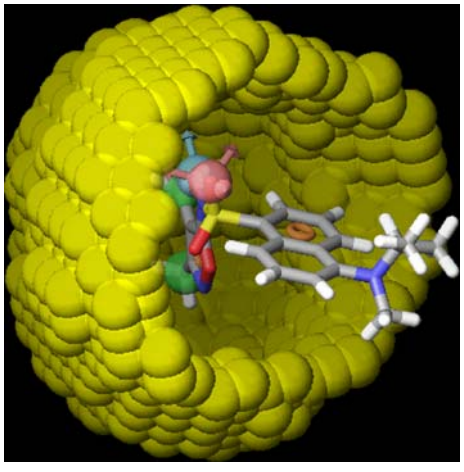

Fig. 9 Illustration of a molecular shape constraint comprised of excluded volume spheres. For clarity, only half of the shell is shown

Finally, if a hypothesis has been constructed from a docked or co-crystallized ligand, then PHASE can create excluded volume spheres whose locations are coincident with any predefined subset of atoms in the receptor, so long as those spheres do not clash with the reference ligand. The radii of the excluded volume spheres can be user-defined or set equal to van der Waals radii of the corresponding receptor atoms. As with the other automated approaches, a buffer can be invoked so that a gap is maintained between the excluded volume surface and the van der Waals surface of the reference ligand.

Searching 3D databases

While the term 3D database formally refers to a specially formatted repository of chemical structure information, the database searching methods used in PHASE extend to other storage formats, such as SD files. Whether searching a 3D database or a flat-file, both standard and *on-the-fly* searching modes are supported. In the standard mode, a pharmacophore hypothesis is matched against a set of pre-computed conformers for each molecule in the database or flat-file. When searching on-the-fly, conformers are generated in memory as needed, using the rapid torsion search method described in the Preparing Ligands section.

The primary method of matching applies a single user-defined tolerance to each intersite distance in a hypothesis. This approach is consistent with the common pharmacophore perception algorithm because the terminal box size reflects an allowed range in each intersite distance. Note, however, that if the terminal box size is ε, the corresponding tolerance on matching a hypothesis should be 2ε, because the direct neighbors of a box are considered when determining whether or not that box represents a common pharmacophore. Although a tighter tolerance can certainly be applied when searching a database, it should be recognized that if the actives used to develop the associated hypothesis were included in the database, it's possible that some of them would not be matched.

A user may require that all sites in a hypothesis be matched, or merely a subset of m out of k sites. In the latter case, it is possible to restrict subsets to include or exclude specific sites. Matching a particular subset of m sites requires that the user-defined distance tolerance be satisfied for all $(m\cdot\ (m–1))/2$ intersite distances in that subset.

When a molecule produces matches, the applicable site points are aligned to the hypothesis using a standard least-squares procedure. If desired, positional

tolerances may be applied at this stage, so that a match will be eliminated if any single site in the matching molecule deviates from the corresponding hypothesis site point by more than the associated positional tolerance. Note that positional tolerances are defined independently from the intersite distance tolerance, so if matching is to be governed purely by the former, the distance tolerance must be at least twice the largest positional tolerance.

All the matches for a particular molecule are sorted by decreasing fitness (Eq. 8), and the user has the option of retaining any number of high-ranking matches for each molecule. As matches are found, the aligned conformers are placed in a *hit list*, which is also sorted by decreasing fitness. When excluded volumes have been defined for a hypothesis, a match will be checked for violations before it is added to the hit list. If it fails the test, the next match in the sorted list is considered, and this process is repeated until the requested number of matches for that molecule have been stored in the hit list, or until all matches have been considered. The total size of the hit list can be capped, so that when the limit is reached, hits with the lowest fitness are ejected from the end list as new hits are added.

In addition to applying positional tolerances and excluded volume filters, users may require that the vector score or volume score of a match exceed a particular threshold. The volume score threshold may be considered to be a filter that forces hits to resemble the reference ligand to any desired degree in overall size and shape. When combined with an excluded volume shell as described in the previous section, a great deal of control may be exercised over the shape of molecules that make it into the hit list.

PHASE application: identification of target pharmacophores

Target pharmacophores: background

The protocol for this investigation was first described by Patel et al. [31], who compared the programs Catalyst/HipHop [16], DISCO [17], and GASP [18] for their ability to generate pharmacophores in accordance with ligand features that were overlaid in X-ray complexes from the Protein Data Bank. For each of five proteins, a number of complexes were visually inspected to identify a *target pharmacophore*, which was used as a standard to establish the accuracy of all programs. More specifically, the ligands for a given protein were run through the automated pharmacophore perception workflow of each program, and the pharmacophore models produced that contained the correct features were aligned to the target pharmacophore to obtain an RMSD in the corresponding site point positions. For a given program, the pharmacophore with the lowest RMSD was judged to be the best, irrespective of how it was ranked by the program's own scoring function.

Patel et al. investigated both rigid and flexible approaches, where rigid refers to pharmacophore perception using only the X-ray ligand structures, and flexible refers to pharmacophore perception based on multi-conformer models generated by the applicable program. Since the rigid approach is a fairly trivial test of whether a method correctly identifies the features in the X-ray structures that can be overlaid, we have carried out tests using only the flexible approach. Further, because Catalyst performed consistently better than the other methods, we report here only head-to-comparisons between PHASE and Catalyst.

Target pharmacophores: methods

Figures 10–14 contain the ligands and target pharmacophores for each of the five systems studied: thrombin, cyclin dependent kinase 2 (CDK2), dihydrofolate reductase (DHFR), HIV reverse transcriptase (HIV-RT), and thermolysin. Note that in the original investigation [31], the HIV-RT target pharmacophore consisted of only a single hydrophobic feature (H), because it was the only characteristic that was overlaid in all 10 ligands. It was not readily apparent to us how a non-zero RMSD can result when aligning a pair of one-point pharmacophores, so we included two additional features (A and D) in the target pharmacophore, which Patel et al. identified as being common to six of the 10 ligands. Similarly, the reported CDK2 target pharmacophore consisted of only two features (A and H), but we included a third feature (D), which was common to five of the six ligands. In both cases we chose a minimum of three features so that when the Phase pharmacophore was aligned to the target pharmacophore, an unambiguous alignment of the calculated reference ligand structure and the X-ray ligand structure would result. Inclusion of additional features in the target pharmacophore does not lend any particular advantage to PHASE since RMSD values tend to be higher as the number of points to be fit increases.

Prior to creating conformers for the flexible analysis, a single low-energy 3D structure was generated for each ligand with the aid of LigPrep [21]. This included hydrogen addition and energy minimization using the MMFFs force field [23] with a constant dielectric of 1.0.

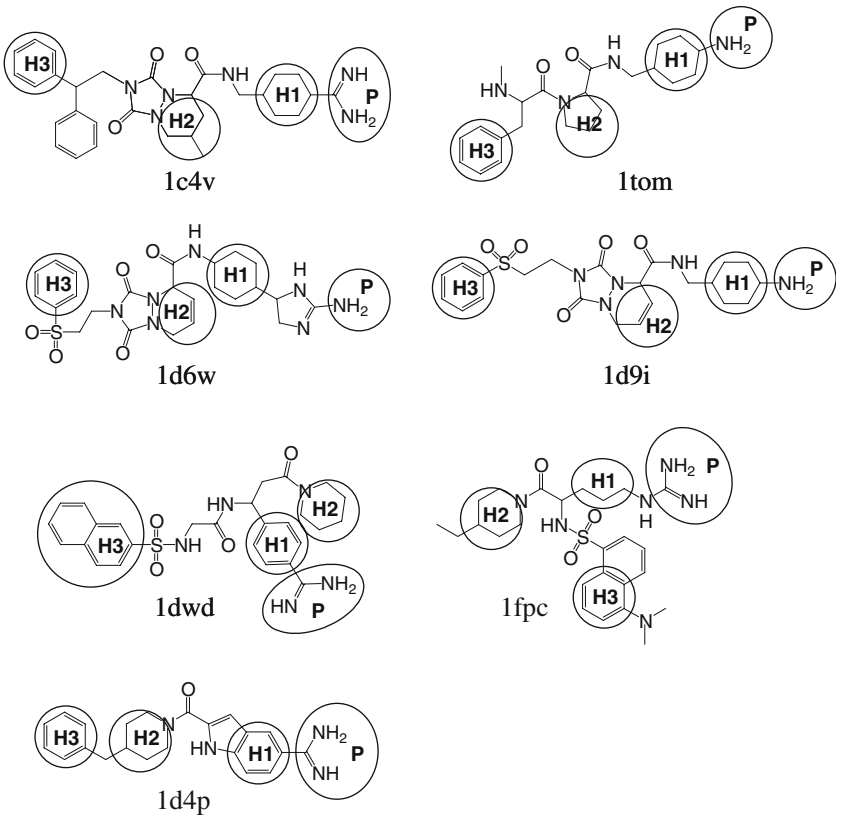

Fig. 10 Thrombin ligands used for target pharmacophore identification. The reference ligand from which the target pharmacophore was derived is 1c4v

The stereochemistries observed for chiral carbons within the applicable X-ray structures were preserved throughout this process. In the case of thrombin inhibitors, a formal + 1 ionic charge was assigned to amidine, piperidine, and primary amine moieties.

Starting with these structures, conformers were generated using MCMM sampling, MMFFs force field, implicit GB/SA solvent model, a maximum of 500 iterations of post-minimization, a relative conformational energy window of 50 kJ/mol, and a redundancy check of 2 Å in the heavy atom positions. The number of Monte–Carlo steps was set to between 1000 and 5000, depending upon the flexibility of the ligands. For example, 5000 steps were used for the highly flexible thermolysin inhibitors, leading to an average of about 500 unique conformers per ligand. By contrast, only 1000 steps were used for the relatively rigid CDK2 inhibitors, which yielded an average of about 50 unique conformers per ligand. It should be noted that the numbers of conformations generated by PHASE were higher on average than Catalyst or DISCO, which have upper limits of 255 and 80, respectively. The numbers of conformations generated by GASP was not reported in the original investigation [31], because these were created internally on-the-fly.

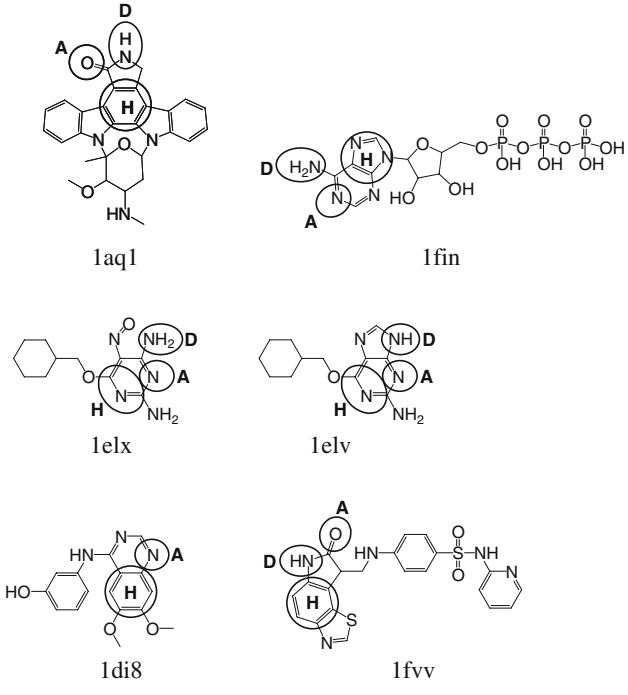

Fig. 11 CDK2 ligands used for target pharmacophore identification. The reference ligand from which the target pharmacophore was derived is 1aq1

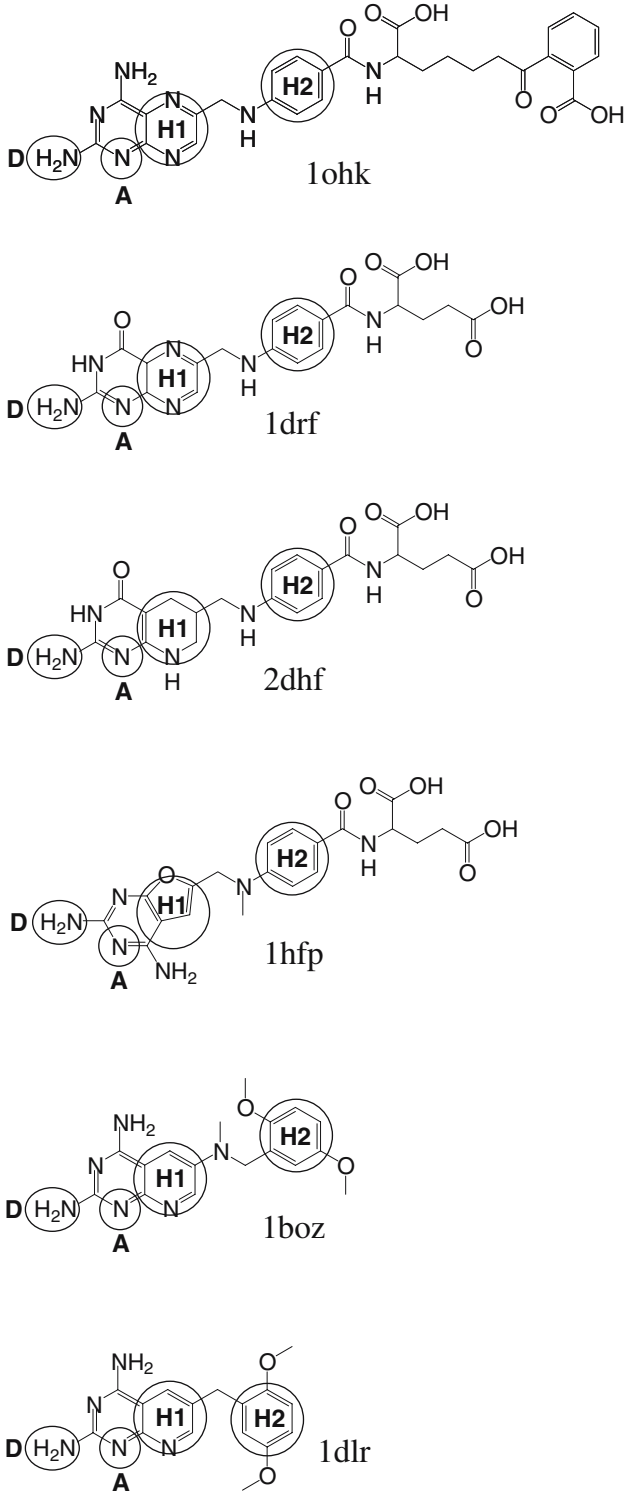

Fig. 12 DHFR ligands used for target pharmacophore identification. The reference ligand from which the target pharmacophore was derived is 1ohk

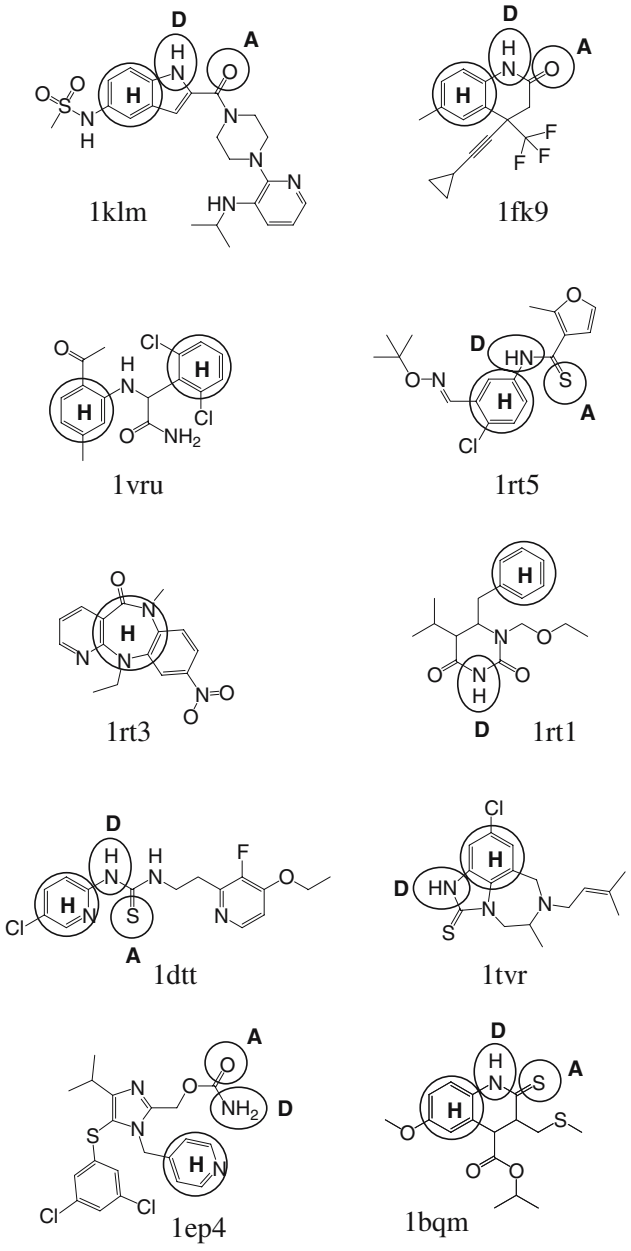

Fig. 13 HIV-RT ligands used for target pharmacophore identification. The reference ligand from which the target pharmacophore was derived is 1klm

Common pharmacophore models containing three, four and five sites were generated using a terminal box size of 1.0 Å and relaxing the requirement, as necessary, that all actives match the pharmacophore. Scoring with respect to actives was done using default parameters values, with incorporation of a binary property value (Eq. 5) to promote the selection of reference ligands in accordance with the ligand used to define each target pharmacophore. Relative conformational energy was not incorporated into the scoring process, nor was there any scoring with respect to inactives.

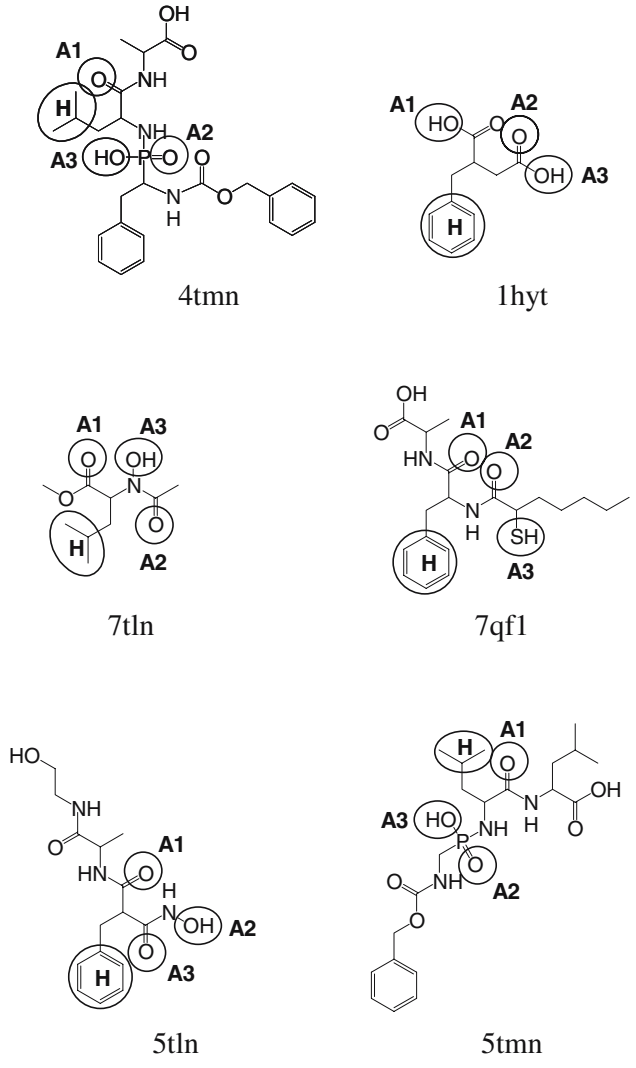

Fig. 14 Thermolysin ligands used for target pharmacophore identification. The reference ligand from which the target pharmacophore was derived is 4tmn

Target pharmacophores were constructed by adding hydrogens to the applicable crystallographic ligand structures (1c4v, 1aq1, 1ohk, 1klm, and 4tmn) via LigPrep, and manually selecting the appropriate pharmacophore sites from those structures. Each hypothesis that emerged from the flexible analysis and which contained the correct features was aligned to its associated target pharmacophore using a standard least-squares technique, and an RMSD in the matching site point positions was computed. When the pharmacophore contained more than one occurrence of a particular feature type, each possible mapping to the target pharmacophore was considered. After processing all hypotheses in this manner, the one yielding the lowest RMSD was selected, and its overall ranking according to the PHASE scoring function was recorded.

Target pharmacophores: results and discussion

Table 1 summarizes the results of target pharmacophore identification using PHASE and Catalyst, where data for the latter were taken directly from the flexible analysis reported by Patel et al. [31]. In four out of five cases, the PHASE RMSD values were significantly lower than those of Catalyst, with an average reduction of 1.18 Å. While Catalyst provided a better match to the HIV-RT target pharmacophore, both programs did quite well for this system, and the difference in the RMSD values was small compared to the other cases.

The Scoring Rank column indicates that for thrombin, CDK2 and DHFR, the PHASE hypothesis that yielded the lowest RMSD was also the hypothesis that received the highest active score as computed by Eq. 7 (with $w_{\text{conf}} = 0$). Catalyst performed almost as

Table 1 Identification of target pharmacophores: PHASE vs. Catalyst

Target	PHASE		Catalyst	
	Lowest RMSD[a] (Å)	Scoring Rank[b]	Lowest RMSD[a] (Å)	Scoring Rank[b]
Thrombin	1.27	1	2.27	2[c,d]
CDK2	0.02	1	1.40	2
DHFR	0.21	1	1.06	2
HIV-RT	0.21	4	0.065	18
Thermolysin	0.48	2	1.96	8

[a] Lowest root-mean-squared deviation from the target pharmacophore among all hypotheses generated by the program

[b] The rank assigned to the lowest RMSD hypothesis based on the program's own scoring function

[c] Hypothesis contained a hydrogen bond donor (D) in place of the positive ionizable center (P) that was indicated in the target pharmacophore

[d] A single target pharmacophore feature in each of two ligands (1d9i and 1fpc) was missed

well in these three cases, with the minimum RMSD hypothesis being ranked second by the Catalyst scoring function. However, the Catalyst thrombin hypothesis contained a hydrogen bond donor (D) in place of the positive ionizable center (P) that was indicated in the target pharmacophore, so the reported RMSD is based on an alignment that overlays these mismatched features. Further, Catalyst did not correctly map all features in thrombin ligands 1d9i and 1fpc, so there were two ''misses'' overall. By contrast, PHASE did not exhibit any misses across the five datasets.

Patel et al. [31] reported results only for the hypotheses that were most similar to the corresponding target pharmacophores. The question naturally arises, then, as to how other high scoring hypotheses compare to the target pharmacophore. Accordingly, we report in Table 2 the RMSD for each of the top five hypotheses as ranked by the PHASE scoring function. Results for CDK2 are omitted because the highly rigid nature of the reference ligand (1aq1) leads to solutions with essentially no variation in the feature positions. Although Table 2 indicates no obvious correlation between the PHASE rankings and the RMSD, in each case there are multiple solutions among the first five that exhibit low RMSD values. These results illustrate that it is indeed possible to uncover some of the most crystallographically relevant pharmacophore models by focusing on a just a few of the highest scoring hypotheses.

While a minimum goal of pharmacophore analysis is to identify chemical features that are essential for binding, a more elusive goal is to infer the structures of bound ligands. Even when a given technique affords a crystallographically accurate pharmacophore model, it may or may not produce a ligand conformation that closely resembles the associated bioactive structure. To illustrate how PHASE performs in this regard, we provide the superposition of each reference ligand conformation and the corresponding crystallographic structure, Figs. 15–19. These overlays were generated by subjecting each PHASE reference ligand to the same transformation that was used to align the minimum RMSD hypothesis to the target pharmacophore.

Table 2 Root-mean-squared deviations (Å) from the target pharmacophore for the five highest scoring PHASE hypotheses

Target[a]	#1	#2	#3	#4	#5
Thrombin	1.27	2.28	2.15	1.51	1.92
DHFR	0.21	0.66	1.68	1.77	0.66
HIV-RT	2.26	1.92	0.65	0.21	1.89
Thermolysin	0.85	0.48	0.55	0.77	0.79

[a] Results for CDK2 are omitted because the highly rigid nature of the reference ligand (1aq1) leads to essentially no variation among solutions that contain the correct features

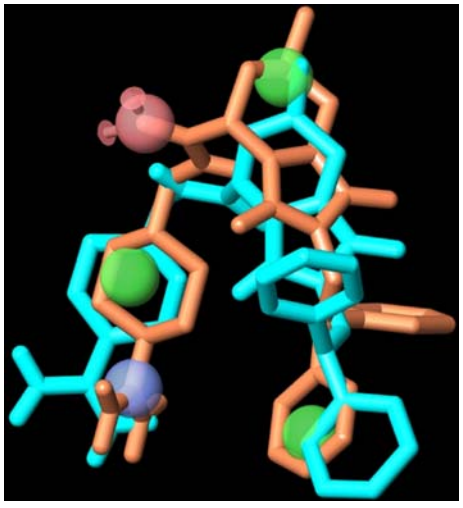

Fig. 15 Overlay of the crystallographic (cyan) and PHASE (brown) structure for the thrombin reference ligand 1c4v

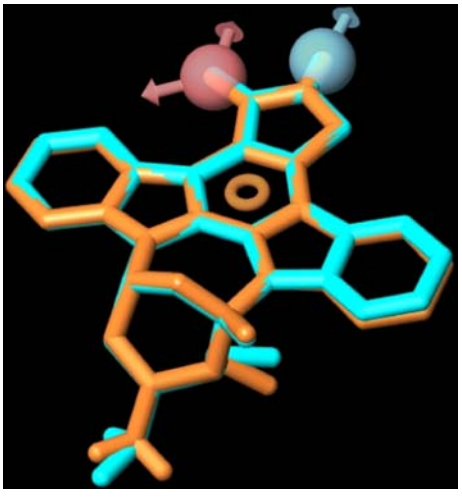

Fig. 16 Overlay of the crystallographic (cyan) and PHASE (brown) structure for the CDK2 reference ligand 1aq1

The thrombin ligand superposition in Fig. 15 indicates reasonably good correspondence in the structures except in the region of the biphenyl. This moiety can adopt a range of rotational states that still allow satisfactory superposition of one phenyl ring and the corresponding hydrophobic sites on the other thrombin ligands. Thus in the absence of a receptor structure, there is probably no overriding force to pin down the positions of both phenyl rings. Excluding the biphenyl moiety, the RMSD between the heavy atom positions in the two structures is 2.4 Å.

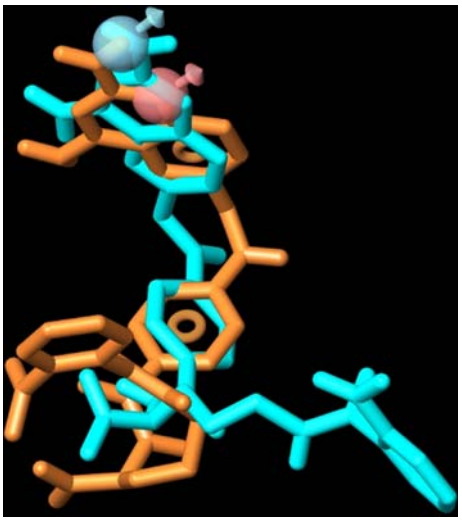

Fig. 17 Overlay of the crystallographic (cyan) and PHASE (brown) structure for the DHFR reference ligand 1ohk

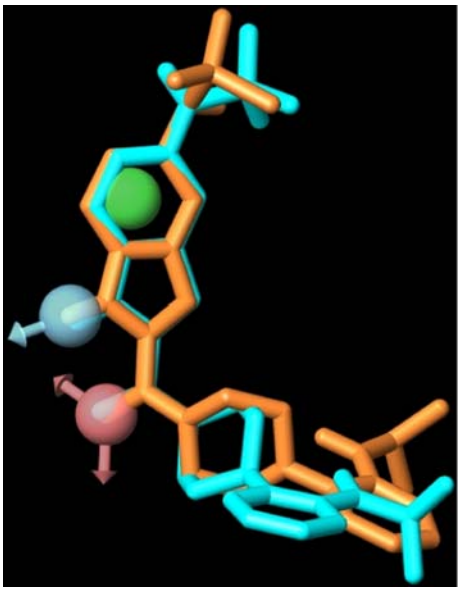

Fig. 18 Overlay of the crystallographic (cyan) and PHASE (brown) structure for the HIV-RT reference ligand 1klm

Due to the high rigidity of the 1aq1 ligand, the CDK2 structures in Figure 16 overlay almost perfectly. This is of course an idealized case, but it emphasizes the advantages of incorporating ligands with conformational constraints into the pharmacophore model development process. By considering both rigid and flexible structures, the chances of identifying a crystallographically accurate pharmacophore model with the correct features are greatly enhanced.

The DHFR pharmacophore spans only about half the structure of the 1ohk ligand, leading to satisfactory superposition throughout that portion of the molecule, Fig. 17. However, beyond the pharmacophore region, there is significant fraying between the PHASE conformation and the X-ray structure, although it should be noted that this portion of the ligand is solvent exposed. Considering only the heavy atoms of the ligand that lie within the binding pocket, the RMSD between the two structures is 2.2 Å.

As shown in Fig. 18, some degree of fraying away from the pharmacophore also occurs for the HIV-RT ligand 1klm, although it is far less pronounced than in the case of DHFR. The RMSD between all heavy atoms in the superimposed HIV-RT ligand structures is 2.9 Å, but it drops to 0.3 Å when the substituted piperazine moiety is ignored.

The final example, thermolysin, demonstrates the worst-case scenario, where the geometry of the hypothesis is essentially correct, but the PHASE reference ligand conformation is very much at odds with the bound structure, Fig. 19. This result illustrates the difficulties associated with deriving models from ligands with significant peptide character. The flexibility and feature-rich nature of these molecules allow them to match a given target pharmacophore in a number of different ways, making it especially challenging to identify which conformation is most likely to be preferred in the binding site.

PHASE application: 3D QSAR of human DHFR inhibitors

hDHFR QSAR: background

Suling et al. [32, 33] reported inhibition data against Mycobacterium avium complex dihydrofolate reductase (MAC DHFR) and human dihydrofolate reductase (hDHFR) for a large number of 2,4-diamino-5-deazapteridines. Debnath [34] utilized the two sets of published IC_{50} values to develop Catalyst/HypoGen QSAR models for each type of inhibition. In this application we focus on just the hDHFR data, and we develop PHASE QSAR models following a protocol that is consistent with the one used to create the Catalyst/HypoGen models.

In the Debnath study, a training set of 20 molecules was selected from a pool of 77, and the remaining 57 molecules were held out as an external test set. Figure 20 summarizes the basic scaffolds investigated, but the reader is referred to reference [34] for complete specification of the chemical structures. Debnath notes that the training set was chosen with an

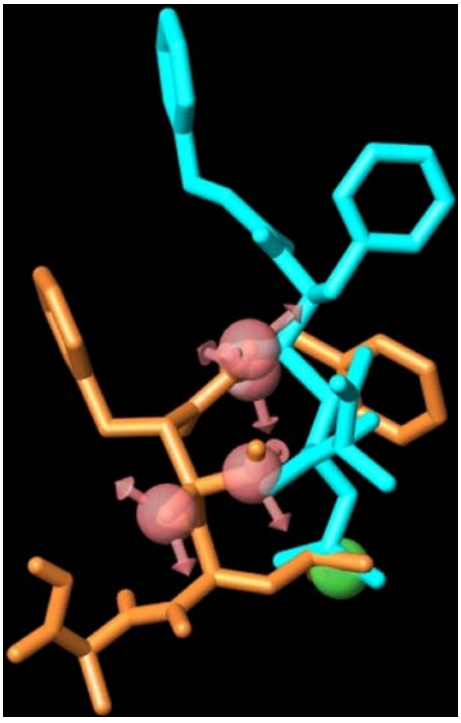

Fig. 19 Overlay of the crystallographic (cyan) and PHASE (brown) structure for the thermolysin reference ligand 4tmn

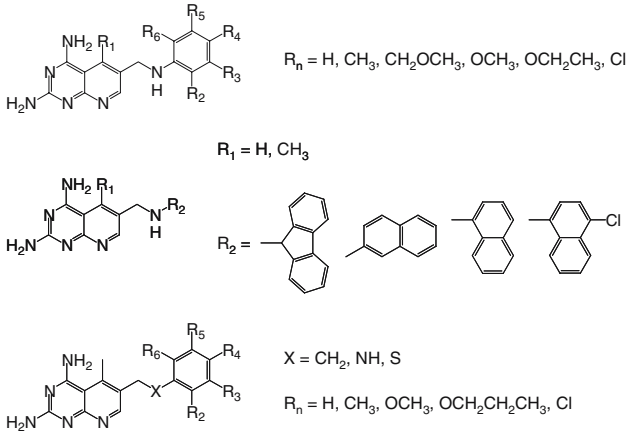

Fig. 20 Basic scaffolds summarizing the structures of the ligands used in the hDHFR QSAR study

effort toward minimizing structural redundancy while maximizing coverage of the experimental activity coordinate. This is the recommended Catalyst/HypoGen protocol, which is aimed at ensuring that each training set molecule provides as much new and independent information as possible. PHASE pharmacophore models and QSAR models were developed from this same training set.

hDHFR QSAR: methods

The Maestro graphical interface [35] was used to build initial 3D models of the 77 hDHFR ligands from the Debnath study [34]. Each structure was subsequently minimized using the MMFFs force field with a constant dielectric of 1.0. The 3-nitrogen in each deazapteridine system was maintained in its neutral form throughout this process.

Conformers were generated using a maximum of 2000 steps of MCMM sampling, followed by up to 500 iterations of truncated Newton conjugate gradient minimization. Potentials were computed using the MMFFs force field with implicit GB/SA solvent. Each minimized conformer was filtered through a relative energy window of 50 kJ/mol and a redundancy check of 2 Å in the heavy atom positions.

Active and inactive thresholds of 100 nM and 800 nM, respectively, were applied to the training set IC_{50} values to yield 10 actives and seven inactives. After applying default feature definitions to each ligand, common pharmacophores containing five and six sites were generated, using a terminal box size of 1 Å, and with the requirement that all 10 actives match.

Scoring with respect to actives was conducted using default parameters for site, vector, and volume terms. Ligand activity, expressed as $-\log_{10}(IC_{50})$, was incorporated into the score with a weight of 1.0, and relative conformational energy (kJ/mol) was included with a weight of 0.01. Hypotheses that emerged from this process were subsequently scored with respect to the seven inactives, using a weight of 1.0.

Atom-based QSAR models were generated for all hypotheses using the 20-member training set and a grid spacing of 1.0 Å. Because models containing three or more PLS factors tended to fit the $-\log_{10}(IC_{50})$ values beyond their experimental uncertainty, only one-factor and two-factor models were considered. Each of these was validated by predicting activities of the 57 test set molecules.

hDHFR QSAR: results and discussion

A total of 264 five-point hypotheses and 135 six-point hypotheses were obtained upon completion of the scoring process. Focusing only on those pharmacophore models whose scores ranked in the top 1%, the most predictive QSAR model was found to be associated with the five-point hypothesis shown in Fig. 21, which contains two hydrogen bond donors and three aromatic ring features. The reported Catalyst hypothesis [34] also contained donor features mapped to N–H bonds from the $-CH_2NH-$ linker and the 2-amino

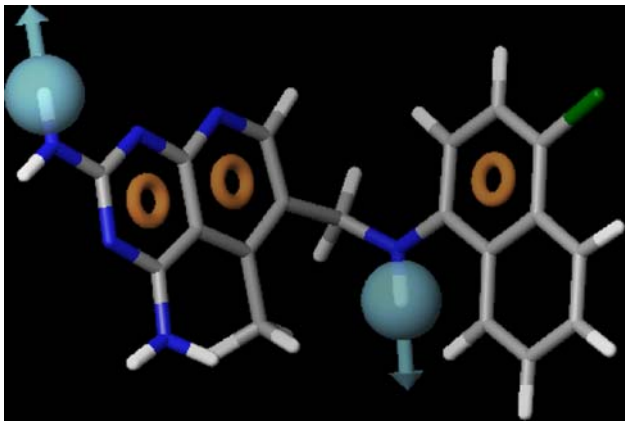

Fig. 21 The PHASE hypothesis (DDRRR) that yielded the most predictive atom-based QSAR model

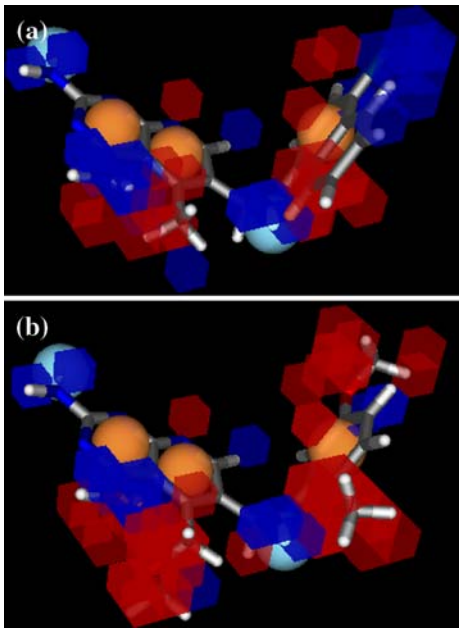

Fig. 22 The hDHFR QSAR model visualized in the context of the most active (**a**) and least active (**b**) molecules in the training set

group, but in the 2-amino case Catalyst chose the alternate N–H mapping compared to PHASE. In addition, the Catalyst hypothesis included a donor feature mapped to the 4-amino, and a hydrophobic feature mapped to the 5-methyl.

It's important to recognize that a superior atom-based QSAR model may or may not come from a hypothesis that is itself fully consistent with the SAR data. This is in contrast to the basic presumptions that apply to pharmacophore-based QSAR, where the presence, absence and locations of pharmacophore sites directly determine the predicted activity. Because the overall alignment of ligands is such a critical aspect of atom-based QSAR, certain pharmacophore features can play a role that is primarily associated with ligand superposition rather than with the direct rationalization of high affinity binding. So, for example, the fact that two aromatic ring features are incorporated from the deazapteridine system does not necessarily imply that both rings are involved in critical interactions with the enzyme.

However, by visualizing the hypothesis and various ligands in the context of the QSAR model, additional insight may be gained about the relevance of each feature in explaining activity. Figure 22a illustrates the most significant favorable and unfavorable interactions that arise when the two-factor QSAR model is applied to the reference ligand, which happens to be the most active compound in the training set (IC_{50} = 2 nM). The blue regions around all three potential hydrogen bond donors suggest that these features are important for high activity, even though the hypothesis contained only two of them. In addition, the model indicates that the presence of the chlorine atom on the naphthyl group tends to increase the predicted activity, though this hydrophobic feature is also missing from the hypothesis.

While some unfavorable regions are indicated for the reference ligand, it is best to examine one or more inactive molecules to identify structural features that may be detrimental to activity. For example, Fig. 22b shows a clear predominance of unfavorable interactions for the most inactive training set molecule (IC_{50} = 31μM). In particular, a methoxymethy group ($–CH_2–O–CH_3$) protrudes significantly out of the plane of the deazapteridine ring system, with a concomitant increase in the unfavorable volume compared to the reference ligand, suggesting the possibility of reduced activity due to steric factors.

The performance of the two-factor QSAR model on the training and test set molecules is illustrated in Figure 23. One fairly striking observation is the complete lack of training set compounds with pIC_{50} values in the range 6.5 to 7.5. Including these midrange molecules in the learning process would almost certainly improve the chances of developing a predictive model because they help differentiate structural modifications that lead to activity changes on the order of one log unit. When training only on compounds at the extremes of the activity scale, it is more difficult to separate the effect of one structural change from another, and whether those changes have a reinforcing or an opposing effect on activity.

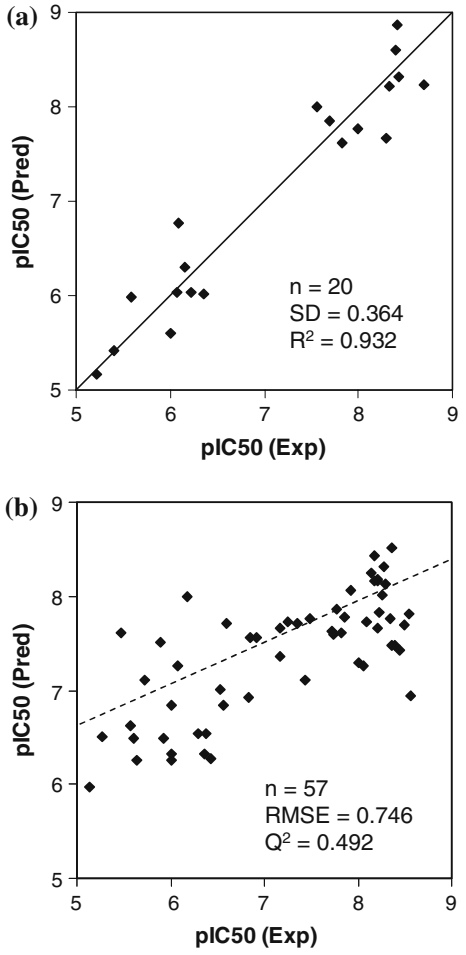

Fig. 23 Scatter plots for the hDHFR QSAR model applied to the training set (**a**) and the test set (**b**). The dashed line in the test set plot indicates the hypothetical ''best fit'' line between the predicted and experimental pIC_{50} values

The scatter plot for the test set (Fig. 23b) indicates a reasonably good correlation between the predicted and experimental activities ($r = 0.73$), but there appears to be a systematic positive shift in the predictions of molecules with low activities, so the slope of the hypothetical ''best-fit'' line is less than one, and it does not pass through the origin. As a result of this shift, the RMSE in the test set predictions is 0.75 log units, which is roughly twice the model standard deviation of regression.

Debnath did not report the actual IC_{50} or pIC_{50} values predicted by the Catalyst model, but instead compared the experimental and predicted activities after mapping each to the following categorical scale:

$$pIC_{50} < 6.0 \rightarrow \text{``+''}$$
$$pIC_{50} \in [6.0, 7.0] \rightarrow \text{``++''}$$
$$pIC_{50} > 7.0 \rightarrow \text{``+++''}$$

In Table 3, we have applied the same scheme to the PHASE test set predictions, and we include the Catalyst predictions from reference [34] for direct comparison. Results for each activity category are summarized in Table 4, and these totals indicate that the PHASE QSAR model is more accurate for ligands with moderate (++) and strong (+++) activities, while Catalyst does better for molecules with low activities (+), although both models perform poorly at this end of the scale. A more judicious choice of training set molecules might overcome this shortcoming for reasons discussed previously. Overall, PHASE predicted the activity category correctly in 42 out of 57 cases (74%), while Catalyst was correct in 38 out of 57 cases (67%).

It is interesting to note that even though the training set contained no compounds with pIC_{50} values in the range 6.5 to 7.5, PHASE still did unusually well on midrange activity ligands in the test set, as indicated by the distinct ''dumbbell'' shape of the scatter plot, rather than the more conventional ''cigar'' shape. There were a total of 15 ligands in this activity range, and PHASE predicted the activity to within 0.5 log units in all but four cases, with a maximum error of 1.18 and an RMSE of 0.58 across the set of 15. Figure 23b indicates several predictions in this range that lie almost perfectly on the hypothetical best fit line, suggesting that the structural variations among these molecules are very well accounted for, in a relative sense, by the PHASE QSAR model.

Conclusions

Although much of computer-aided drug discovery is focused heavily on the use of structure-based methods, there continues to be a steady stream of pharmacologically important problems whose solution by these means is impossible, impractical, or unreliable. Consequently, ligand-based design is the natural alternative when the structure of the biological target is unsolved and/or poorly modeled, when a rapid in silico screen of a large corporate database is needed, or when a quantitative, predictive model of biological activity is sought.

But despite decades of ligand-based methodology development, most practitioners would agree that the field still suffers from a lack of robust, flexible, and easy-to-use tools. In response to these needs, we have developed PHASE, a highly functional system for pharmacophore perception, 3D QSAR model development, and database screening. To promote a

Table 3 Comparison of PHASE and Catalyst hDHFR QSAR test set predictions

Number[a]	Experimental Activity		PHASE		Catalyst
	pIC_{50}[b]	Category[c]	pIC_{50}	Category	Category
78	5.14	+	5.97	+	+++
54	5.28	+	6.51	++	+++
56	5.47	+	7.60	+++	+++
57	5.47	+	8.06	+++	+++
76	5.57	+	6.62	++	++
73	5.60	+	6.49	++	++
1	5.64	+	6.26	++	+++
5	5.72	+	7.11	+++	+
50	5.89	+	7.51	+++	+
70	5.92	+	6.49	++	++
2	6.00	++	6.25	++	+++
3	6.00	++	6.31	++	++
6	6.00	++	6.83	++	++
10	6.07	++	7.26	+++	+++
32	6.17	++	7.99	+++	+
71	6.30	++	6.53	++	++
69	6.36	++	6.32	++	++
74	6.38	++	6.54	++	++
12	6.43	++	6.27	++	++
7	6.52	++	7.01	+++	+++
8	6.56	++	6.84	++	+++
9	6.60	++	7.70	+++	+++
4	6.82	++	6.92	++	+++
53	6.85	++	7.57	+++	+++
11	6.92	++	7.57	+++	+++
13	7.16	+++	7.66	+++	+++
14	7.17	+++	7.35	+++	+++
15	7.24	+++	7.73	+++	+++
16	7.36	+++	7.71	+++	+++
19	7.44	+++	7.10	+++	++
17	7.48	+++	7.76	+++	+++
21	7.72	+++	7.63	+++	+++
24	7.74	+++	7.59	+++	+++
22	7.77	+++	7.85	+++	+++
25	7.82	+++	7.61	+++	+++
27	7.85	+++	7.78	+++	+++
31	8.00	+++	7.28	+++	+++
30	8.06	+++	7.26	+++	++
33	8.09	+++	7.72	+++	+++
60	8.15	+++	8.25	+++	+++
35	8.17	+++	8.16	+++	+++
59	8.17	+++	8.42	+++	+++
34	8.21	+++	7.67	+++	+++
37	8.21	+++	8.17	+++	+++
61	8.21	+++	8.18	+++	+++
38	8.23	+++	7.83	+++	+++
42	8.26	+++	8.01	+++	+++
40	8.28	+++	8.31	+++	+++
44	8.30	+++	8.13	+++	+++
39	8.34	+++	7.76	+++	+++
62	8.37	+++	7.47	+++	+++
65	8.37	+++	8.50	+++	+++
46	8.39	+++	7.48	+++	+++
48	8.44	+++	7.43	+++	++
49	8.50	+++	7.69	+++	+++
28	8.55	+++	7.81	+++	+++
47	8.57	+++	6.94	++	+++

[a] Test set compound number as assigned by Debnath. [34]

[b] $-\log_{10}(IC_{50})$, where IC_{50} is expressed in moles/liter

[c] Categorical activity: $pIC_{50} < 6.0 \rightarrow$ "+"; $pIC_{50} \in [6.0, 7.0] \rightarrow$ "++"; $pIC_{50} > 7.0 \rightarrow$ "+++"

Table 4 Summary of categorical hDHFR QSAR test set predictions

Category	Frequency	Correct Predictions	
		PHASE	Catalyst
+	9	1	2
++	15	9	6
+++	33	32	30
All	57	42	38

thorough understanding of the full range of capabilities, we have provided a comprehensive description of the major workflows and their underlying scientific methodologies.

Because ligand-based design is frequently applied to problems with a high degree of technical difficulty, it is important to be aware of both the practical limitations of a given method and the strategies that are most likely to yield meaningful results. Accordingly, we have addressed questions regarding how the user's data, needs, and expectations can impact the appropriate course of action throughout various PHASE workflows.

Finally, any new methodology should be validated against experimental findings, and it should be compared directly to existing approaches that are in widespread use. In a series of exercises involving target pharmacophore identification using a flexible ligand treatment, PHASE was shown to provide solutions that reproduced crystallographic pharmacophores more accurately than Catalyst/HipHop, GASP, or DISCO. Further, in a QSAR study of inhibition of human dihidrofolate reductase by 2,4-diamino-5-deazapteridines, PHASE was more accurate than Catalyst/HipHop at classifying a large test set of inhibitors into three activity categories.

Appendix A: selectivity estimation

In PHASE, the selectivity of a pharmacophore hypothesis H is defined as follows:

$$\text{Selectivity}(\mathrm{H}) = -\log_{10}[p(\mathrm{H})], \tag{A1}$$

where $p(\mathrm{H})$ is the probability that a random drug-like molecule will match the hypothesis, irrespective of any activity exhibited by that molecule toward the biological target in question. Given a database of drug-like molecules, it is straightforward to search that database for matches to a hypothesis, and thereby arrive at an estimate of selectivity based on that particular sample population of molecules. However, application of such a procedure is far too time-consuming to be practical when scoring a large number of hypotheses, so a rapid means of estimating selectivity based on the physical characteristics of a hypothesis is sought.

Van Drie [12] has shown that selectivities of two-point pharmacophores can be reliably estimated with respect to a given database using pre-tabulated probabilities that cover discrete distance ranges. He went on to show that highly selective three-point pharmacophores can be constructed by combining two-point pharmacophores with the highest selectivities. This is a natural consequence of the fact that the probability of matching a k-point pharmacophore $\mathrm{H}^{\langle k\rangle}$ is less than or equal to the probability of matching all $(k\cdot(k–1))/2$ two-point pharmacophores embedded within $\mathrm{H}^{\langle k\rangle}$:

$$p(\mathrm{H}^{\langle k\rangle}) \le p\left(\bigcap_{i<j\le k} \mathrm{H}_{ij}^{\langle k\rangle}\right) \tag{A2}$$

Strict equality is not preserved because a given molecule may match each of the two-point pharmacophores even if it fails to contain a single arrangement of k features that matches $\mathrm{H}^{\langle k\rangle}$. Nevertheless, since matching the two-point pharmacophores is a necessary condition for matching $\mathrm{H}^{\langle k\rangle}$, the right-hand-side of Eq. A2 is of interest for purposes of estimating selectivity.

If the two-point probabilities are independent, then the following relation holds:

$$p\left(\bigcap_{i<j\le k} \mathrm{H}_{ij}^{\langle k\rangle}\right) = \prod_{i<j\le k} p(\mathrm{H}_{ij}^{\langle k\rangle}) \tag{A3}$$

Further, if sites i and j are separated by a distance of d_{ij}, and their pharmacophore feature types are $\alpha(i)$ and $\alpha(j)$, respectively, then Eq. A3 can be rewritten in terms of probabilities of matching specific inter-feature distances to within a tolerance Δd:

$$p\left(\bigcap_{i<j\le k} \mathrm{H}_{ij}^{\langle k\rangle}\right) = \prod_{i<j\le k} p\left(d_{\alpha(i)\alpha(j)} \in [d_{ij}-\Delta d,\, d_{ij}+\Delta d]\right) \tag{A4}$$

Given a population of drug like molecules and a pair of feature types x and y, there is a *probability density* $p^*(d_{xy})$ that describes the distribution of xy pharmacophores within that population. While $p^*(d_{xy})$ may be complex and possibly discontinuous, for purposes of estimating selectivity a simple Gaussian dependence is assumed, so that the probability density may be written as:

$$p^*(d_{xy}) = \frac{1}{\sigma_{xy}\sqrt{2\pi}} \exp\left[-\frac{(d_{xy}-\mu_{xy})^2}{2\sigma_{xy}^2}\right] \tag{A5}$$

For small values of Δd, the following approximation can be made:

$$p\left(d_{xy} \in [d-\Delta d,\, d+\Delta d]\right) \approx \Delta d \cdot p^{*}(d_{xy})\big|_{d_{xy}=d} \qquad \text{(A6)}$$

Substituting A5 and A6 into A4 yields

$$p\left(\bigcap_{i<j\le k} \mathrm{H}_{ij}^{\langle k\rangle}\right) \approx \prod_{i<j\le k} \frac{\Delta d}{\sigma_{\alpha(i)\alpha(i)}\sqrt{2\pi}} \exp\left[-\frac{(d_{ij}-\mu_{\alpha(i)\alpha(j)})^2}{2\sigma^2_{\alpha(i)\alpha(j)}}\right] \qquad \text{(A7)}$$

Taking logarithms,

$$-\log_{10}\left[p\left(\bigcap_{i<j\le k} \mathrm{H}_{ij}^{\langle k\rangle}\right)\right] \approx \sum_{i<j\le k}\left\{-\log_{10}\left[\frac{\Delta d}{\sigma_{\alpha(i)\alpha(i)}\sqrt{2\pi}}\right] + \frac{(d_{ij}-\mu_{\alpha(i)\alpha(j)})^2}{\log(10)\cdot 2\sigma^2_{\alpha(i)\alpha(j)}}\right\} \qquad \text{(A8)}$$

Although it is certainly possible to estimate the univariate parameters $\sigma_{\alpha\ (i)\alpha\ (j)}$ and $\mu_{\alpha\ (i)\alpha\ (j)}$ for each possible pair of feature types, it is advantageous to treat the right-hand-side of Eq. A8 as a general polynomial in d_{ij}, and fit the associated coefficients to observed probabilities for a large number and variety of pharmacophores:

$$-\log_{10}\left[p(\mathrm{H}^{\langle k\rangle})\right] \approx \sum_{i<j\le k}\left(A_{\alpha(i)\alpha(j)} + B_{\alpha(i)\alpha(j)} d_{ij} + C_{\alpha(i)\alpha(j)} d_{ij}^2\right) \qquad \text{(A9)}$$

This treatment can help overcome certain deficiencies in the model, such as the assumption that the two-point probabilities are independent of each other (Eq. A3). In practice, the second-order terms in Eq. A9 do not add much statistically independent information to the model, and we have found a first-order approximation to be satisfactory:

$$-\log_{10}\left[p(\mathrm{H}^{\langle k\rangle})\right] \approx \sum_{i<j\le k}\left(A_{\alpha(i)\alpha(j)} + B_{\alpha(i)\alpha(j)} d_{ij}\right) \qquad \text{(A10)}$$

To determine appropriate values for the A and B parameters, a training set was assembled by randomly selecting 1000 minimized structures from a conformational database of the World Drug Index [36], then randomly choosing between two and seven pharmacophore sites from each structure. This yielded a training set of 1000 pharmacophores containing varying numbers of sites and different combinations of the features A, D, H, N, P, and R. A sample probability was computed for each pharmacophore H_λ by determining the number of structures M_λ out of the original 1000 that matched the pharmacophore to within a tolerance of 2.0 Å in all intersite distances:

$$p(\mathrm{H}_\lambda) \equiv \frac{M_\lambda}{1000} \qquad \text{(A11)}$$

Since there were six types of features in the sampled pharmacophores, the number of unique feature pairs was 21, requiring a total of 42 adjustable parameters. No attempt was made to optimize all of these independently because of the possibility of only limited information for certain pairs of features. For example, pharmacophores that contain both negative and positive ionizable features tend to be very rare among drug-like structures, so they cannot be expected to be well-represented in a relatively small population sample. Therefore, parameter values were determined by applying a partial least-squares (PLS) procedure to fit the $-\log_{10}(\mathrm{H}_\lambda)$ values in terms of latent factors constructed from the pool of 42 variables. Details of the PLS algorithm used in PHASE are provided in Appendix B.

To arrive at an appropriate number of PLS factors to include in the model, predictions were made for a test set of 500 pharmacophores drawn from the same sample population of 1000 WDI structures. As successively more PLS factors were incorporated into the model, test set errors trended downward until reaching a minimum at 23 factors. At this point, the test set RMSE was 0.372 log units and Q^2 was 0.786. This compared to a training set RMSE of 0.343 and R^2 of 0.826. This model has been integrated into PHASE for computation of the Selectivity_Score term that appears in Eq. 7.

It is worth noting that training sets containing as many as 5000 structures were also investigated, and no significant improvement in the test set predictions was observed. The protocol of using 1000 structures was adopted because it is far less computationally demanding, and therefore represents a practical approach for users who wish to calibrate selectivity models based on a different set of structures.

Appendix B: partial least-squares regression

PHASE utilizes a standard recursive procedure for extracting orthogonal latent factors from a data matrix in a predetermined number of steps. It is distinguished from the NIPALS algorithm [37, 38], which is an iterative approach with a user-defined stopping criterion, but no absolute control over the total number of steps.

Let $\mathbf{X}\in\mathbf{R}^{n\times m}$ represent the independent variable data matrix for a training set of n observations and a pool of m variables. Let $\mathbf{y}\in\mathbf{R}^{n\times 1}$ represent the training set dependent data, which will be estimated using latent factors extracted from $\mathbf{X}$. Creation of the PLS regression model proceeds as follows:
Center each column of $\mathbf{X}$:

for $i = 1,\ldots,m$
  $\mu_i^x = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}(k,i)$
  for $k = 1,\ldots,n$
    $\mathbf{X}(k,i) \rightarrow \mathbf{X}(k,i) - \mu_i^x$
  next k
next i

Center $\mathbf{y}$:

$\mu^y = \frac{1}{n}\sum_{k=1}^{n}\mathbf{y}(k)$
for $k = 1,\ldots,n$
  $\mathbf{y}(k) \rightarrow \mathbf{y}(k) - \mu^y$
next k

Determine PLS factors and regression coefficients for up to M PLS factors ($M \le m$):

$\mathbf{X}_1 = \mathbf{X}$
for $i = 1,\ldots,M$
  Compute the vector of weights that define PLS factor i :
  $\mathbf{w}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{y}/|\mathbf{X}_i^{\mathrm{T}}\mathbf{y}| \quad (\mathbf{w}_i \in \mathbf{R}^{m\times 1})$
  Project the rows of $\mathbf{X}_i$ onto factor i :
  $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i \qquad (\mathbf{t}_i \in \mathbf{R}^{n\times 1})$
  Project $\mathbf{t}_i$ onto each column of $\mathbf{X}_i$:
  $\mathbf{p}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{t}_i/|\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i| \quad (\mathbf{p}_i \in \mathbf{R}^{m\times 1})$
  Compute the i$^{\mathrm{th}}$ PLS regression coefficient by projecting $\mathbf{t}_i$ onto $\mathbf{y}$:
  $\mathbf{b}(i) = \mathbf{t}_i^{\mathrm{T}}\mathbf{y}/|\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i| \quad (\mathbf{b} \in \mathbf{R}^{M\times 1})$
  Orthogonalize $\mathbf{X}_i$ w.r.t. PLS factor i :
  $\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^{\mathrm{T}}$
next i

For a regression with M PLS factors, the estimates $\hat{\mathbf{y}}$ are then given by:

$$\hat{\mathbf{y}}(k) = \mu^y + \sum_{i=1}^{M}\mathbf{b}(i)\mathbf{t}_i(k) \quad k = 1,\ldots,n$$

To apply the M-factor PLS model to a new set of $\tilde{n}$ observations with data matrix $\tilde{\mathbf{X}}$, the regression coefficients $\mathbf{b}$ must first be translated back to the space of the original $\mathbf{X}$ variables:

Define

$$\mathbf{W} \equiv [\mathbf{w}_1 \ldots \mathbf{w}_M] \qquad (\mathbf{W} \in \mathbf{R}^{m\times M})$$
$$\mathbf{P} \equiv [\mathbf{p}_1 \ldots \mathbf{p}_M] \qquad (\mathbf{P} \in \mathbf{R}^{m\times M})$$
$$\mathbf{b}^x \equiv \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{b} \quad (\mathbf{b}^x \in \mathbf{R}^{m\times 1})$$

The coefficients $\mathbf{b}^x$ may then be used to make estimates for the new observations as follows:

$$\tilde{\mathbf{y}}(k) = \mu^y + \sum_{i=1}^{M}\left[\tilde{\mathbf{X}}(k,i) - \mu_i^x\right]\mathbf{b}^x(i) \quad k = 1,\ldots,\tilde{n}$$

References

1. Guner OF (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, CA
2. Van Drie JH (2003) Curr Pharm Design 9:1649
3. Topliss JG (1983) Quantitative structure-activity relationships of drugs, vol 19. Academic Press, New York
4. Martin YC (1978) Quantitative drug design: a critical introduction. Marcel Dekker, New York
5. Hansch C, Fujita T (1964) J Am Chem Soc 86:1616
6. Gund P, Wipke WT, Langridge R (1974) Computer searching of a molecular structure file for pharmacophoric patterns, vol 3. Elsevier, Amsterdam, pp 33–39

7. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, London
8. Hancsh C, Leo A (1979) Substituent constants for correlation analysis in chemistry and biology. Wiley, New York
9. Hopfinger AJ (1980) J Am Chem Soc 102:7196
10. Van Drie JH, Weininger D, Martin YC (1989) J Comput-Aided Mol Design 3:225
11. Lauri G, Bartlett PA (1994) J Comput-Aided Mol Design 8:51
12. Van Drie JH (1997) J Comput-Aided Mol Design 11:39
13. Chen X, Rusinko A, III Young SS (1998) J Chem Inf Comput Sci 38:1054
14. Chen X, Rusinki A, III Tropsha A, Young SS (1999) J Chem Inf Comput Sci 39:887
15. Greene J, Kahn S, Savoj H, Sprague P, Teig S (1994) J Chem Inf Comput Sci 34:1297
16. Barnum D, Greene J, Smellie A, Sprague P (1996) J Chem Inf Comput Sci 36:563
17. Martin YC, In Hansch C, Fujita T (eds) (1995) Classical and 3D QSAR in agrochemistry. American Chemical Society, Washington, DC, pp 318–329
18. Jones G, Willett P, Glen RC (1995) J Comput-Aided Mol Design 9:532
19. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
20. Van Drie JH, In Guner OF (ed) (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, CA, pp 517–530
21. Ligprep 2.0 (2006) Schrodinger, LLC, New York, NY
22. MacroModel 9.1 (2006) Schrodinger, LLC, New York, NY
23. Halgren TA (1996) J Comput Chem 17:520
24. MacroModel 2.0 (2006) User Manual, Schrodinger LLC, New York, NY
25. Chang G, Guida W, Still WC (1989) J Am Chem Soc 111:4379
26. Kolossvary I, Guida WC (1996) J Am Chem Soc 118:5011
27. SMARTS – Language for Describing Molecular Patterns, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA
28. Marshall GR, Barry CD, Bosshard HE, Dammkoehler RA, Dunn DA, In Olson EC, Christoffersen RE (eds) (1979) Computer-assisted drug design. American Chemical Society, Washington, DC, pp 205–226
29. Beusen DD, Marshall GR, In Guner OF (ed) (2000) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, CA, pp 23–45
30. Van Drie JH (1997) J Chem Inf Comput Sci 37:38
31. Patel Y, Gillet VJ, Bravi G, Leach AR (2002) J Comput-Aided Mol Design 16:653
32. Suling WJ, Reynolds RC, Barrow EW, Wilson LN, Piper JR, Barrow WW (1998) J Antimicrob Chemother 42:811
33. Suling WJ, Seitz LE, Pathak V, Westbrook L, Barrow EW, Zywno-Van-Ginkel S, Reynolds RC, Piper JR, Barrow W (2000) Antimicrob Agents Chemoth 44:2784
34. Debnath AK (2002) J Med Chem 45:41
35. Maestro 7.5 (2006) Schrodinger, LLC, New York, NY
36. World Drug Index (2001) Thomson Scientific
37. Wold H, In Gani J (ed) (1975) Perspectives in probability and statistics, Papers in Honour of Bartlett MS on the Occasion of His Sixty-Fifth Birthday, Academic Press, London, pp 117–142
38. Wold S, Ruhe H, Wold H, Dunn WJI (1984) SIAM J Scientific Stat Comput 5:735