ARTICLE

OBC
www.rsc.org/obc

# Combined use of physicochemical data and small-molecule crystallographic contact propensities to predict interactions in protein binding sites†

**J. Willem M. Nissink* and Robin Taylor**
*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK CB2 1EZ.*
*E-mail: nissink@ccdc.cam.ac.uk*

***Received 6th April 2004, Accepted 6th July 2004***
***First published as an Advance Article on the web 27th August 2004***

Knowledge-based methods are a good alternative to force-field-based methods for the analysis of sites of interaction in protein binding cavities. Both the Protein Data Bank (PDB) and the Cambridge Structural Database (CSD) offer a good amount of data on non-covalent interactions. Although different from protein-derived data, small-molecule crystal data from the CSD are worth looking at as they provide a much more abundant and diverse set of intermolecular contacts. CSD data, when properly corrected by use of octanol–water $\pi$ values, can be used to predict the type of ligand chemical group most likely to occupy a given position within a protein binding site. Comparison with observed positions of ligand groups shows that the success rates of these predictions vary from 23% to 84%. Often, the group predicted to be most preferred at a given position is similar but not identical to the observed ligand group; if these are considered successes, prediction success rates range from 71% to 94%. Using PDB data, the corresponding rates are 16% to 79%, and 61% to 96%. Specificity of prediction of NH groups is somewhat better when using PDB interaction data, but results of prediction of hydrophobic groups seem worse than those obtained with CSD data.

We have analysed the importance of data selection by applying different filters to eliminate unwanted interactions from our knowledge-base. The presence of certain types of interactions can be undesirable if they are unrepresentative of biological situations (contact to solvent molecules in small-molecule crystal structures, secondary crystallographic contacts) or if they are likely to add noise to the data without conveying much new information (long-distance contacts, sparsely-populated data sets). The elimination of solvent contacts was found to have no effect on the prediction of ligand groups in our test set. Both secondary-contact filtering and noise filtering were found to have a clear beneficial effect on predictive ability.

## 1 Introduction

Understanding the principles of molecular recognition in protein–ligand complexes is a key issue in drug design, and crystal structures are our prime source of information in this field. Phillips and coworkers[1] elucidated the first enzyme structure, lysozyme, in 1965, and the crystallography provided strong leads for its mechanism of catalysis. An early example of structure-based design was reported by the group of Beddell, Goodford *et al.* at Wellcome Laboratories in the United Kingdom in the early 1970s.[2] Their research focussed on haemoglobin, which was one of the few pharmacologically relevant targets for which a crystal structure was available at that time. More recently, the development of non-peptidic human immunodeficiency virus (HIV) protease inhibitors has convincingly demonstrated the importance of crystallographic data in structure-based drug design.[3,4]

Currently, there are two major repositories of crystallographic data that are relevant to drug design. The Protein Data Bank[5,6] (PDB) is a source of protein structures, containing both crystallographic and NMR entries. The Cambridge Structural Database[7] (CSD) contains crystal structures of small molecules. Both databases provide a rich collection of information on non-bonded contacts. Tintelnot and Andrews[8] were among the first to suggest that atomic environments of small functional groups in binding sites taken from the PDB could be used to predict non-bonded protein–ligand interactions. Klebe[9] used a similar approach to analyse non-bonded environments of functional groups in the CSD, using the spatial distributions ('composite crystal fields') of probe groups to map putative interaction sites in protein binding cavities. Such distributions are compiled in the IsoStar database, a knowledge-base of nonbonded interactions[10] (*vide infra*). Other applications that use crystallographic data directly for the assessment of molecular interactions are: the *de novo* design tool LUDI;[11,12] HSITE;[13] XSITE by Laskowski *et al.*;[14] SuperStar[15–18] (*vide infra*); and AQUARIUS.[19] Watson *et al.*[20] use the IsoStar crystallographic knowledge-base of non-bonded interactions[10] and report a method for finding small-fragment candidates for bioisosteric replacement. Nissink *et al.*,[18] Labute[21] and Rantanen *et al.*[22] report approaches for parameterising the geometries of interaction of non-bonded contacts taken from crystal structures. Crystal-structure data are also typically used in knowledge-based scoring functions for use in docking programs. Of these methods, we only mention DrugScore[23] here, as this scoring function was also explicitly applied to the prediction of preferential binding spots in protein cavities.

Knowledge-based methods typically use dedicated databases, *i.e.* data are compiled specifically for the application. Alternative, non-knowledge-based methods for analysis of non-bonded interactions usually apply energy force-fields that are fitted to represent certain observations. Examples of such methods are: HINT;[24] GRID;[25] and MCSS.[26,27] A related approach, computational solvent mapping, has been proposed recently by Vajda *et al.* to characterise protein binding sites using small solvent molecules.[28,29] Fragment-based docking approaches[30,31] are reminiscent of these methods, although they are not explicitly used to map binding sites, but attempt to find appropriate small lead structures that fit a binding site directly.

In this paper we discuss the combined use of small-molecule crystallographic contact propensities from IsoStar[10] and hydrophobicity data from octanol–water partitioning coefficients for the prediction of protein–ligand interactions. Using hydrophobicity data, we propose a method for relating the preferences of polar and hydrophobic interactions in the CSD to a solvated reference state such that small-molecule data can be used reliably for predicting interactions in protein binding sites.

We further investigate the effects of data significance and address the influence of data selection on performance by assessing results for pruned interaction knowledge-bases. We apply filters that

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b405205f

eliminate 'noisy' interactions, solvent-specific interactions, and secondary interactions in crystals, and report on their effect.

## 2 Computational details

### 2.1 Calculation of propensities of interaction

By superimposing crystallographically-observed contacts between two groups *X* and *Y* so that the *Y* moieties are overlaid, a three-dimensional scatterplot can be produced showing the experimental distribution of *X* (the "contact" or "probe" group) around an average *Y* (the "central" group). IsoStar[10] is a database of such scatterplots for many *X*,*Y* pairs. Most of the scatterplots are based on contacts in CSD structures but a substantial minority are based on protein–ligand interactions in the PDB. Any scatterplot can be converted to a contoured surface showing the density of contact groups around the central group. This surface can be put on a meaningful scale by dividing the raw densities by the uniform density of contacts that would be expected in the scatterplot if the *X*,*Y* groups were distributed at random in the contributing crystal structures. This gives a "propensity" surface.[10] By implication, regions of the plot with propensity >1 (density greater than the expectation of the uniform distribution) correspond to energetically favourable positions for the contact group around the central group, and the greater the propensity, the more favourable the position is likely to be.

SuperStar[15,16] uses IsoStar data to generate knowledge-based propensity maps that indicate the likelihood of occurrence of certain probe groups in protein binding sites (or around small molecules), not unlike the well-known GRID program.[25] The program works as follows. *1*) A target molecule (binding site, small molecule) is dissected into its constituent fragments in such a way that each fragment corresponds to a central group in the IsoStar database. For binding sites, a cavity detection algorithm can be used to narrow down the relevant part of the protein prior to analysis. Typical fragments are small and comprise, *e.g.*, terminal groups like carboxylate, methyl groups, charged amines, or links like methylenes and peptides. *2*) For a selected probe, scatterplots are retrieved from the IsoStar database for the constituent fragments. Probe groups can be hydrogen-bonding, like alcohol groups or carbonyl groups, but may also be apolar, *e.g.* aliphatic or aromatic CH. Probes correspond to the contact groups found in IsoStar (*e.g.* an alcohol probe relies on IsoStar's OH contact data). *3*) Each scatterplot is overlaid on all parts of the binding site that it matches and each overlaid scatterplot converted to a propensity surface. *4*) The separate surfaces ("sub-maps") are combined to produce an overall map for the complete binding site or molecule. In those regions where a sub-map overlaps with an adjoining sub-map for another nearby fragment, propensities at coinciding points are multiplied. *5*) For protein maps that are calculated using small-molecule crystallographic data, a correction is applied (hydrophobicity correction) that adjusts the propensities to levels that correspond to those observed in macromolecules (*vide infra*).

The resulting map can be viewed after contouring at suitable levels of propensity. SuperStar uses either CSD or PDB interaction data for map calculation. For further details of this procedure, we refer the reader to Verdonk *et al.*[15] and Bruno *et al.*[10]

### 2.2 Reference state and hydrophobicity correction

Ideally, the propensity *P* we calculate should be a measure of the likelihood that a probe group prefers a certain protein environment over the solvent (assumed to be water), and should relate to an equilibrium constant *K* [eqn. (1)]

$$P \propto K \equiv \frac{[c_{protein}]}{[c_{water}]} \qquad (1)$$

where $[c_{protein}]$ is the concentration of a probe in a specific protein environment, and $[c_{water}]$ is the concentration of the probe in the solvent. Concentrations here are hypothetical as a probe is an average of fragments observed in many similar chemical environments. Unfortunately, interaction propensities that are derived from crystal structures cannot reflect this equilibrium with a solvent; they are usually normalised by application of a reference state other than a solvent (see Bruno *et al.*[10]).

**Influence of reference state and solvent.** The propensity of interaction for a given probe group (contact group) with a given counter-group (central group) as derived from an IsoStar scatterplot based on PDB or CSD data equals a partitioning coefficient $P_E$:

$$P_E \equiv \frac{[c_{crystal}]}{[c_{reference,E}]}, \quad E = PDB \text{ or } CSD \qquad (2)$$

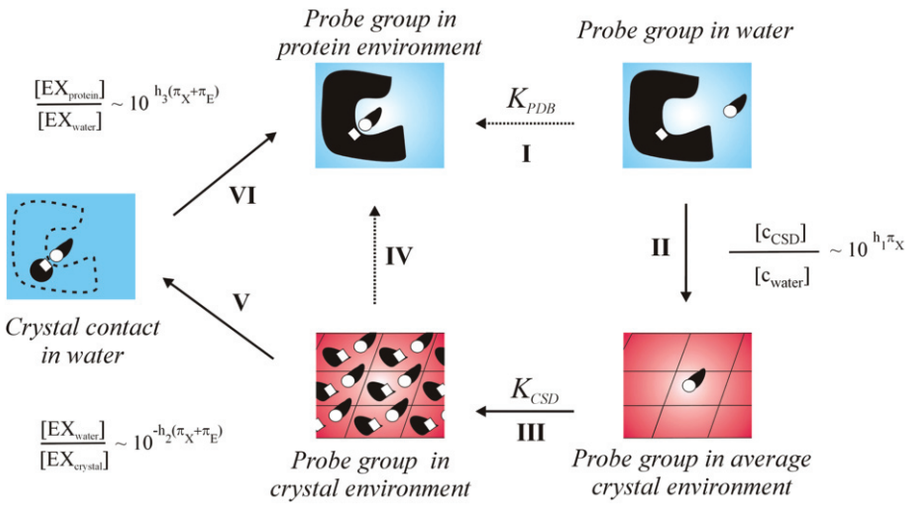
Here, $[c_{crystal}]$ is the "concentration of" contact groups (number of groups per volume) forming interactions with the central group in PDB protein binding sites or CSD crystal structures, and $[c_{reference}]$ is the concentration of contact groups in the reference state, *i.e.* whether or not within interaction distance of the central group. Thus, the propensity that is calculated can be regarded as indicative of the likelihood of a contact group–central group interaction in a protein or small-molecule crystal environment *E*.

There are different ways to determine a reference state, but usually they rely on the definition of a state that is random and where the interaction that is being investigated is not prevalent. For CSD contact data in IsoStar, the reference concentration $[c_{reference}]$ is determined from a set of crystals that contain both groups involved in the interaction we are looking at, either interacting or non-interacting.

In order to use small-molecule crystal structure data for the purpose of predicting interactions in binding sites, this situation has to be corrected. We want to predict interactions taking place in an equilibrium state relative to an aqueous solvent, as found in a biological situation. In proteins, formation of hydrophobic contacts usually causes water to be expelled from the relatively apolar binding site into the water environment, and the driving force for hydrophobic contacts is entropic to a large extent; on the other hand, formation of polar protein–ligand interactions usually involves desolvation of protein and ligand, and may be somewhat less favourable than in small molecule crystals. Given that small-molecule crystals are usually grown from relatively apolar solvents, we speculate that, although interaction enthalpies are expected to be equal for contacts formed in CSD and PDB environments, the net entropic contribution of water expulsion from protein cavities causes free energies of interaction to be different. As a result, CSD structures favour hydrophobic contacts to a lesser degree than do protein–ligand complexes, and the predominant driving force for formation of a small-molecule crystal lattice will be the formation of hydrophilic interactions. In practice, we observe that the relative occurrence of polar and hydrophobic interactions does differ in protein and small-molecule crystals.[15]

**Calculation of correction factor.** Early versions of SuperStar correct this skew in importance of interactions by multiplying all propensities around hydrophobic residues by a factor of 10.0 when CSD-derived maps are calculated using hydrophobic probes. This gave acceptable results.[15] The factor was defined purely empirically, and is applied only to interactions between a predefined set of apolar protein side chains and apolar probes. Though reasonably effective, this approach is essentially flawed as it does not take into account the hydrophobicity of the probe for other probe-to-protein combinations. The single correction factor is a compromise that attempts to account for all influences.

The hydrophobicity correction that relates the crystallographic reference state to a water reference state, implicitly incorporating the solvent-expulsion effect (see above), can be derived from a hypothetical cycle as depicted in Scheme 1. Step **I** indicates the equilibrium between contact groups in the solvent state and those interacting with a central group in the protein environment. It is this step we want to quantify. We use octanol–water partitioning coefficients[32] (log *P* contributions) for fragments as an estimate of the hydrophobicity of a reference state with respect to the water-




**Scheme 1**

solvent state. The log *P* value of a molecule can be estimated as a sum of individual $\pi$ contributions from its component chemical groups.[33–35] We assume that these so-called $\pi$-values are constant (*i.e.* independent of the nature of the molecule). $\pi$-values as found in the literature have been estimated usually for a large number of functional groups by regression against a training set of measured log *P* values.

The fragments involved in the hypothetical cycle of Scheme 1 are the probe group *X*, with its associated log *P* contribution $\pi_X$, and its environment *E*, which is either a solvent, the contacting group(s) in the small-molecule crystal, or the contacting group(s) in the protein. The hydrophobicity of this environment is measured by log *P* contribution $\pi_E$. We first quantify the influence of the hydrophobicity of the small-molecule crystal reference state relative to water (step **II**), by treating it as an octanol–water partitioning step that is attenuated by a factor $h_1$; effectively, $h_1$ is the lipophilicity of the CSD reference state on a scale of 0 (water) to 1 (octanol). $K_{CSD}$ is estimated by the propensity derived from IsoStar CSD data (step **III**). The difference in hydrophobicity between the environment of the $X\cdots E$ pair in the crystal and in the protein (step **IV**) is approximated similarly, but in two steps **V**–**VI** using water as an intermediary solvent. The term $h_3 - h_2$ reflects the difference in hydrophobicity between the crystal and protein environments of the contact, where $h_2$ and $h_3$ are again on a scale of 0 (water) to 1 (octanol). Combining the factors involved in steps **II**, **III**, **V**, and **VI** we derive eqn. (3)

$$K_{PDB} = 10^{h_1\pi_X + (h_3 - h_2)(\pi_X + \pi_E)} K_{CSD} \quad (3)$$

Defining attenuation coefficient $\Delta a = h_1$ for step **I** and $\Delta b = h_3 - h_2$ for step **IV**, eqn. (3) can be rewritten with a correction factor *C* as

$$K_{PDB} = CP_{CSD} \quad (4)$$

with

$$C = 10^{(\Delta a + \Delta b)\pi_X + \Delta b\pi_E} \quad (5)$$

The separation of $\pi_X$ and $\pi_E$ terms in eqn. (5) allows precalculation of the probe- and protein (environment)-dependent parts of the correction factor. Factor *C* is applied to the propensity at each grid point in a raw propensity map.

Hydrophobicity of probe and environment are estimated by contributions $\pi_X$ and $\pi_E$. Probe contributions are approximated by the contribution of the group alone, *e.g.* for an aliphatic hydroxyl probe, the contribution $\pi_{OH}$ is taken. The environment term can be estimated by taking the contribution of the contacting (nearest) group only, or by averaging the $\pi$ terms of *n* nearest groups up to a certain distance. Although the approach depicted in Scheme 1 is of an approximate nature due to the estimations used (averaged apolarity of reference states, approximated hydrophobicity of probe and environment fragments), it does give us an appropriate functional form for the correction to be applied.

### 2.3 Improvement of the knowledge-base

The approaches outlined below focus on selection and improvement of the source data. We describe the use of filters to eliminate noisy data, and pruning of interactions that are of doubtful relevance to biological situations (secondary contacts; solvent contacts).

**Data significance and restriction of noisy regions.** Having compiled a set of crystallographic interactions between groups *X* and *Y*, we can construct from this a map that depicts the propensity of *X* to form an interaction with *Y* at a given position in space. This set of interactions $X\cdots Y$ can be regarded as a sample from a more comprehensive data set that comprises all crystals harbouring central group *Y* and contact *X*, regardless of an interaction between the two. We can estimate the significance of any observation in the interaction map from the distribution of such observations in the comprehensive set. Assuming a Poisson distribution for the observed number of contact groups per cubic Ångstrom in this reference set, the chance of finding more or less than *n* contact groups in a small volume at a given position in the map (*e.g.* a grid cube) can be estimated numerically as

$$p(n \le a) = \sum_{n=0}^{a} \frac{n_{\text{expected}}^{x}}{n!} e^{-n_{\text{expected}}} \quad (6a)$$

$$p(n \ge a) = 1 - p(n < a) \quad (6b)$$

Here, $n_{\text{expected}}$ is the expected number of contact groups per volume unit, which is known from the comprehensive set.[10] The assumption of a Poisson distribution comes naturally as the random observations in the reference set are spread over a large volume, so that the chance of an observation occurring in a given grid cube is reasonably small. eqn. (6a) is used for observations smaller than $n_{\text{expected}}$, (6b) for those larger than $n_{\text{expected}}$. An observation will be considered significant if its chance of occurrence *p* is smaller than a certain threshold *S*. If not, the expected value (corresponding to a propensity of 1) will be substituted.

When regarding regions at large distances from the central group, it is expected that the value of *n*, the observed number of observations per unit volume, levels off towards the $n_{\text{expected}}$ value. This is not always observed. One of the reasons may be a lack of data, which causes remote regions to be sparsely populated. Another reason of a more systemic nature may be that these regions feature a low incidence of contacts because packing in small-molecule crystals is extremely efficient; as a result of this compressive lattice effect, long-distance contacts will hardly ever occur.

**Table 1** Success rates (%) for different hydrophobicity correction protocols. *f* percentage of ligand groups predicted by correct probe; *f'* percentage of predictions by probe with appropriate physicochemical properties. $\Delta_{shell}$ = 0.0 Å for contacts to apolar groups, $\Delta_{shell}$ = 0.5 Å for polar groups. The error was derived using a bootstrapping procedure: for 100 'bootstrapped' validations of entries that were picked randomly with replacement from the original set of entries, success rates were analysed. Only one set of error results is shown as errors proved to be similar for different validation runs

| | Aliphatic CH | | Aromatic CH | | C=O | | OH | | $RNH_2$ | | RR'NH | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *n groups* | *468* | | *224* | | *176* | | *96* | | *29* | | *74* | |
| *Error estimate* | *2* | | *3* | | *4* | | *4* | | *5* | | *4* | |
| | *f* | *f'* | *f* | *f'* | *f* | *f'* | *f* | *f'* | *f* | *f'* | *f* | *f'* |
| PDB | 60 | *66* | 16 | *72* | 61 | *61* | 65 | *96* | 79 | *86* | 76 | *87* |
| CSD, no correction | 4 | *17* | 7 | *9* | 88 | *88* | 81 | *100* | 7 | *100* | 38 | *88* |
| CSD, original | 10 | *71* | 81 | *86* | 78 | *78* | 81 | *96* | 7 | *86* | 18 | *71* |
| CSD, $\pi$-based, single | 18 | *81* | 75 | *80* | 85 | *85* | 70 | *97* | 17 | *72* | 41 | *71* |
| CSD, $\pi$-based, extended | 18 | *80* | 75 | *80* | 85 | *85* | 69 | *97* | 17 | *72* | 41 | *71* |
| CSD, $\pi$-based/$\Delta_{shell}$ | 23 | *84* | 78 | *83* | 84 | *84* | 64 | *94* | 35 | *76* | 53 | *71* |

This may cause problems in practice for two reasons: first, noise may be introduced in such underpopulated areas, and second, the presence of very low propensities in the above mentioned regions may lead to cancellation of high propensities in areas where they overlap with other sub-maps during binding site analysis. We therefore limit contributions from $X\cdots Y$ interaction data to a core region within a radius of $R < r_X + r_Y + \Delta_{shell}$ Å of the interacting group; here, $r_X$ and $r_Y$ denote the van der Waals radii of the contacting atoms of the contact and central group, respectively, and $\Delta_{shell}$ the thickness of the region beyond the sum of the van der Waals radii. We extend this core region only to areas beyond $R$ when $n$ is significantly larger (*i.e.* $p < S$) than $n_{expected}$. For other areas outside the core region, we assume that there are $n_{expected}$ observations (*i.e.* corresponding to a propensity value of 1), although in the actual data there may be less. The maximum extension distance is currently that of the underlying IsoStar data, $R_{max} = r_X + r_Y + 0.5$ Å.

**Secondary contacts.** An interaction of a contact group $X$ with a given group $A$ in the database does not preclude a second contact of the group with another moiety $B$. An example is shown in Fig. 1. In the CSD crystal structure shown in Fig. 1, the contact of the OH to the phenyl ring is considered to be secondary to the main contact, a hydrogen bond to a carbonyl group. If the interaction with $B$ is much stronger than the one with $A$, we should not use the latter for modelling $XA$-type interactions. When trying to predict primary, one-on-one interactions between $X$ and $A$ only, including such information about an incidental secondary contact is expected to be detrimental to predictions. The effect is expected to be more important for CSD data than for PDB data, as small-molecule crystals are closely packed. In order to remove such contacts from the analysis, we filtered IsoStar CSD data, regarding the contact group-central group distance as a measure of its strength, and eliminated a contact from $XA$ interaction data if

$$d_{XB} < d_{XA} + \Delta_{filter} \quad (7)$$

where $d_{XA}$ is the $X\cdots A$ distance minus the sum of the van der Waals radii of $X$ and $A$, as determined in the original crystal structure, and $d_{XB}$ has an analogous definition. The parameter $\Delta_{filter}$ sets the strength of the filter, *i.e.* it determines the threshold distance at which we deem contacts to be secondary and hence omit them. Filters with negative $\Delta_{filter}$ values are less stringent than those with positive values for $\Delta_{filter}$.

**Solvent contacts.** Inclusion of solvents in crystal structures may be induced by more than just a primary intermolecular interaction. Solvent molecules in crystals typically have at least two of the following three different functions: participation in hydrogen-bonding networks; space-filling, with no pronounced interactions between solvent and other molecules; and as ligands completing the coordination around metal ions.[36]

The particular role solvents play in crystal formation may bias their interactions. It has been reported that solvent inclusion in crystals can result in stronger H-bonds and an increase in favourable CH–π interactions.[37] We investigated this influence by eliminating contact data due to solvent molecules from our knowledge-base, and assessing predictive performance of SuperStar calculations.

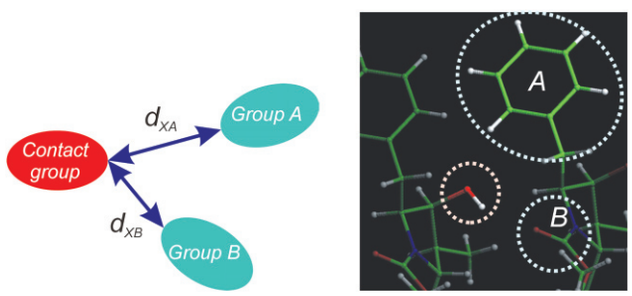



**Fig. 1** Schematic depiction of a secondary contact (left), and an example of a secondary contact in crystal structure RUWMAX (right) (M. Bolte, *Acta Crystallogr., Sect. C*, 1997, **53**, 9700028), where the primary contact is a hydrogen bond between the OH contact group and the carbonyl.

Another reason for excluding solvents from interaction data is the bias they may introduce in terms of diversity. An area where this is known to play a role is, *e.g.*, that of small-molecule chlorine contacts. A majority of these contacts is found to be due to interactions with common chlorinated solvents like chloroform, dichloromethane and trichloroethane that are used abundantly in crystallisation liquors because of their physicochemical properties. Such contacts may not be representative of ligand chlorine contacts in protein environments. Here, we investigate whether this is also the case for common solvent inclusions like acetone or ethanol.

## 3 Results and discussion

### 3.1 Evaluation of π-based hydrophobicity correction mechanism

Success rates of prediction for a set of map probes are given in Table 1 (see section 5.1 for validation procedure). Rates are given for prediction by the *correct* probe (*f*) or by *appropriate* probes with similar properties (*f'*). Results are shown for SuperStar using PDB-based interaction data (*PDB*), applying raw small-molecule crystallographic data (*CSD*, *no correction*), CSD data in combination with the original single-factor hydrophobicity correction mechanism[15] (*CSD original*), and CSD data with the π-based protocol (*π-based*). Results for the latter are shown for two estimates of the $\pi_E$ term: by regarding the closest protein–ligand contact within van der Waals distance + 0.5 Å only, *i.e.* distance < $r_{probe} + r_{proteincontact} + 0.5$ Å ($r_{probe}$ and $r_{proteincontact}$ are van der Waals radii of the contacting atoms) (*CSD π-based single*), and by averaging the $\pi_E$ contributions of up to three nearest protein groups within distance $r_{probe} + r_{proteincontact} + 0.5$ Å (*CSD π-based extended*).

Errors were estimated using a bootstrapping approach: success rates of prediction were calculated for 100 'random sets' equal in size to the original set of entries. These sets were compiled randomly

**Table 2** $\pi$ contributions assumed for SuperStar map probes

| *Probe name* | $\pi$ |
|---|---|
| Aliphatic CH carbon | +0.5 |
| Aromatic CH carbon | +0.5 |
| Carbonyl oxygen | −0.4 |
| Alcohol oxygen | −0.9 |
| Uncharged nitrogen | −0.3 |
| Methyl carbon | 0.5 |
| Organic chlorine | 0.9 |
| Organic fluorine | 0.5 |
| Charged $RNH_3$ nitrogen | −1.5 |
| Carboxylate oxygen | −2.0 |
| Nitro oxygen | 0.2 |
| Sulfur | 0.5 |

from entries of the original set (allowing replacement). The variance of the bootstrapped results is indicative of the uncertainty in the success rates. Only one set of errors is reported, as results were similar for validation runs that employ different settings.

**Ligand group prediction, no hydrophobicity correction.** Using raw CSD data without applying the hydrophobicity correction clearly has a detrimental effect on prediction of hydrophobic ligand fragments like aromatic CH and aliphatic CH (Table 1, *CSD, no correction*). This is the result of the difference in importance of specific types of interaction observed for contacts in the CSD and PDB. Assessment of hydrophilic ligand groups is more successful with CSD interaction data than with PDB-based data, but in this uncorrected case this may just be the result of over-emphasising these types of interaction. The amino ligand fragments are generally predicted to be OH rather than NH. This does not mean that NH-containing groups do not occur in positions predicted to be favourable for the uncharged-NH probe; it does indicate that propensities found for the OH maps at these ligand positions are higher than those for the NH probe.

**Ligand group prediction, original scheme vs. $\pi$-based protocol.** Application of the single correction factor protocol (Table 1, *CSD original*) improves prediction of hydrophobic aliphatic and aromatic CH moieties by CSD interaction data considerably, while losing some of the performance for hydrophilic groups. Prediction rates for CO and OH ligand groups are better than those obtained from PDB-based interaction data; prediction of the NH-containing ligand fragments ($RNH_2$, RR′NH) is less good, although most are predicted when taking into account prediction by the OH probe, as indicated by the high $f'$ rates. This is acceptable since the OH probe has donor properties, as has uncharged NH, and the former is generally the better donor of the two.

For the $\pi$-based correction protocol, values for $\pi$ contributions were estimated from log $P$ contributions from the literature.[33,34] These were checked for consistency against experimental values of octanol–water partitioning coefficients for small compounds where available, and adjusted to match protein fragments reasonably (Tables 2 and 3). Parameters were not optimised to yield the best validation results possible, with the exception of the value for $\pi_{NH}$; slightly better results were obtained with a value of −0.3 rather than −0.5. A value of −0.5 would apply to strong NH donor groups (*c.f.* $\pi_{OH} = -0.9$), whereas NH-containing ligand groups often correspond to somewhat weaker donors. Setting the $\pi_{NH}$ contribution to −0.3 yields satisfactory results and this value was used throughout.

With this set of $\pi$-contribution parameters for the probes (Table 2), the $\pi$-based correction protocol (Scheme 1) yields validation results as shown in Table 1. Optimal results were obtained for $\Delta a = 0.4$ and $\Delta b = 0.2$ [eqn. (5)]. Being on a scale from 0 (water) to 1 (octanol) these values are quite reasonable, with $\Delta a$ indicating that the average crystal environment has an apolarity of 0.4 times that of octanol; the $\Delta b$ value indicates that protein environments surrounding ligands are about 0.2 units more apolar than small-molecule crystal environments. The values $\Delta a = 0.4$ and $\Delta b = 0.2$ have been used throughout. Validation results for the $\pi$-based scheme indicate a slight improvement in the prediction of aliphatic CH ligand groups, and a definite improvement in prediction of carbonyl groups and NH moieties by their corresponding map probes. Although the correct prediction of hydroxyl groups by the OH probe is down, its prediction by appropriate probes (*i.e.* OH, CO, or NH) is similar to that found for the original protocol.

Prediction of ligand hydroxyl groups is complicated because a given hydroxyl might be a donor only (in which case its prediction as an NH is a reasonable result), an acceptor only (when prediction as C=O would be reasonable) or both a donor and an acceptor (in which case its prediction as either NH or C=O would be unsatisfactory). Looking at this in more detail, maps calculated with the original hydrophobicity correction predict both donating-only, and donating-and-accepting ligand hydroxyl groups as OH nearly exclusively (Table 4). Using the $\pi$-based hydrophobicity correction, stronger competition is introduced between NH, CO, and OH map probes, which is not unexpected. Ligand hydroxyl groups that both donate and accept seem to be predicted wrongly as NH slightly less often than groups that donate only. From a subset of 33 buried OH groups in positions that favour a donor group only, 27 are predicted as OH, 1 as carbonyl, and 5 as uncharged NH (15%); for a set of 49 hydroxyl groups in positions with both donors and acceptors within contact distance, 37 are predicted correctly as OH, another 7 are predicted to be carbonyls, and 5 to be NH groups (10%). All OH groups from a set of five buried ligand hydroxyls in positions that would favour an acceptor are predicted to be carbonyls.

Fig. 2 shows example SuperStar maps for lytic transglycosylase binding bulgecin A (PDB code 1D0L). This type of protein catalyses the cleavage of a glycosidic bond in peptidoglycan, but its precise functions in peptidoglycan metabolism is unknown.[38] The binding site is quite polar and a comparison of maps calculated with uncharged NH and aromatic CH map probes shows that, in particular, the aromatic CH maps are much more according to expectation for the $\pi$-based protocol than for the original hydrophobicity correction method. The sparser NH maps seem to compare better to the PDB-based maps than the maps calculated using the original correction scheme. The additional hot-spot in the top left-hand corner in the plots based on CSD data is the result of small-molecule data showing an interaction preference for an NH probe to both lone pairs of the Gln98 terminal carbamoyl CO group, whereas PDB data do not display this preference. It is unclear whether the absence of this spot in a PDB-data-based map is genuine, or caused by a lack of or bias in protein–ligand interaction data.

**Map descriptors, original *vs.* $\pi$-based correction.** Apart from prediction of ligand groups by their corresponding map probes, other factors of interest are the extent to which a map contains points with high propensity values, and whether the ligand groups examined in the validation actually fall within regions of high propensity ('hot spots'). Table 5 lists the map descriptors $f_{P>1}$ and $f_{P<P'}$. The former, $f_{P>1}$, represents the percentage of ligand groups observed to lie at positions for which the map indicates a propensity larger than 1.0, calculated per map probe for the appropriate ligand groups. Propensities larger than 1.0 indicate a likelihood of occurrence that exceeds random expectation. The latter, $f_{P<P'}$, is the (probe-accessible) fraction of a map with propensities smaller than the propensity found at the ligand group's position (averaged over all ligand groups, per map probe). High values for this descriptor indicate that the ligand groups occur in 'hot-spots' in the map. Ideally, both descriptors should have high values, but it is not known how high these values should be. Although ligand placement in binding sites usually can be considered optimal, ligands exhibit a broad range of binding affinities, and the local positioning of a functional group in the ligand with respect to its environment may be sub-optimal.

A comparison of values for the original and $\pi$-based hydrophobicity correction methods shows that the latter yields comparable values for $f_{P<P'}$, but lower ones for $f_{P>1}$. This is an indication that maps for the $\pi$-based mechanism feature less and/or smaller regions

**Table 3** Estimated log *P* contributions for protein fragments. Fragment contributions as stated by Klopman ([a]) and Suzuki ([b]) are shown for comparison. Experimental values for related compounds were used for checking whether values were appropriate and consistent. Contributions here are for generic fragments, and generally follow the trend of the values given by Klopman

| *Protein fragment* | $\pi$ | *Source* |
|---|---|---|
| Carbamoyl | −0.7 | $NH_2$-CO −0.24[b], −0.795[a]; *N*-methylacetamide −1.05; *N*-methylformamide −0.97 |
| Methyl | 0.7 | $CH_3$-Cal 0.764[b], 0.661[a]; $CH_3$-Car0.614 (*S*); ethane 1.81; 2-methylpropane 2.36; 2,2-dimethylpropane 3.11 |
| Methylene | 0.5 | $CH_2$-C2 0.897[b], 0.415[a]; $CH_2$-Car-C 0.369 (*S*); propane 2.36; diethyl ether 0.89; diethylamine 0.58 |
| Tertiary CH | 0.4 | CH-(C)3 0.233[b], 0.104[a]; 2-methyl-propane 2.76 |
| Phenyl | 0.7 | CarH-(Car)2 0.367[b]; CarH-(Nar)20.863[b]; benzene 2.13; *p*-cresol 1.94 (assume that contacts 'see' one-third of phenyl ring) |
| Aromatic CH | 0.6 | CarH-(Car)2 0.367[b] 0.380[a]; CarH-(Nar)2 0.863[b]; CarH-(Car)-(Nar) |
| Histidine Aromatic CH | 0.0 | 0.367[b]; benzene 2.13; *p*-cresol 1.94 (assume that contacts 'see' one-third of a ring) |
| Aromatic CC or CX | 0.5 | =Car < 0.129[a]; set as methylene |
| al-al ether | 0.0 | O-(C)2 −1.093[b]; −0.402[a] O-(C)-(CO) −0.093[b]; diethyl ether 0.89; methylethylether 0.56 |
| al-ar ether | −0.1 | |
| Ester | −0.5 | O-(C)-(CO) −0.062[b], −0.414[a]; methylformate 0.03; benzyl methyl ester 2.12; methyl acetate 0.18; ethyl acetate 0.73 |
| Ketone | −0.2 | CO-(C)2 −1.747[b], −0.493[a]; acetone −0.24; 2 butanone 0.29 |
| Carboxylate | −1.0 | CO-(C)-O −1.357[b]; charged fragment. |
| Carboxylic acid | −0.4 | acetic acid −0.17[b] −0.263[a] (aliphatic) 0.467[a] (aromatic); propanoic acid 0.33; benzoic acid 1.87 |
| Planar ring NH (uncharged) | −0.3 | NH-Car-N −0.615[b]; NH (Car)2 −0.720[b] −0.160[a]; pyridine 0.65; pyrrole 0.75 |
| Planar ring NH (charged) | −0.5 | |
| Guanidinio | −0.8 | charged group, estimated |
| Charged amino | −1.0 | charged group, estimated |
| Aliphatic OH | −0.7 | OH-Cal −1.287[b], −0.681[a] (primary) −0.575[a] (secondary); methanol −0.77; ethanol −0.31 |
| Aromatic OH | −0.4 | OH-Car −1.102[b], 0.135[a]; phenol 1.46 |
| Thiol | 0.9 | SH-C 0.052[b] 0.875[a] |
| Amide | −0.8 | NH-C-CO −0.060[b] −1.006[a]; *N*-methylformamide −0.97; *N*-methyl acetamide −1.05 |
| Disulfide link | 0.5 | S-X 0.079[b]; -S- 0.485[a]; dimethyldisulfide 1.77; diethyldisulfide 1.95 |
| Water | −1.4 | water 1.38[b] |

[a] Ref. 40. [b] Ref. 33. Cal aliphatic CH; Car aromatic CH. Experimental octanol–water partitioning coefficients were obtained from: http://www.syrres.com/esc/kowdemo.htm.

**Table 4** Competitive prediction of buried ligand hydroxyl groups that only donate, donate and accept, or only accept (numbers given in brackets). Maps were calculated for all probes as before (ARCH aromatic CH probe; ALCH aliphatic CH probe)

| Predicted as: | OH | CO | NH | ARCH | ALCH |
|---|---|---|---|---|---|
| *Donating ligand OH groups* (*33*) | | | | | |
| original hydrophobicity correction | 32 | 0 | 1 | 0 | 0 |
| $\pi$-based hydrophobicity correction | 27 | 1 | 5 | 0 | 0 |
| $\pi$-based/shell | 21 | 1 | 8 | 3 | 0 |
| $\pi$-based/filter | 24 | 2 | 6 | 1 | 0 |
| $\pi$-based/filter/shell | 23 | 2 | 6 | 2 | 0 |
| *Donating & accepting ligand OH groups* (*49*) | | | | | |
| original hydrophobicity correction | 42 | 5 | 2 | 0 | 0 |
| $\pi$-based hydrophobicity correction | 37 | 7 | 5 | 0 | 0 |
| $\pi$-based/shell | 37 | 7 | 5 | 0 | 0 |
| $\pi$-based/filter | 41 | 6 | 2 | 0 | 0 |
| $\pi$-based/filter/shell | 41 | 6 | 2 | 0 | 0 |
| *Accepting ligand OH groups* (*5*) | | | | | |
| original hydrophobicity correction | 0 | 5 | 0 | 0 | 0 |
| $\pi$-based hydrophobicity correction | 0 | 5 | 0 | 0 | 0 |
| $\pi$-based/shell | 0 | 5 | 0 | 0 | 0 |
| $\pi$-based/filter | 0 | 5 | 0 | 0 | 0 |
| $\pi$-based/filter/shell | 0 | 5 | 0 | 0 | 0 |

of high propensity. Figs. 2 and 3 both show example SuperStar maps for a selection of map probes. A comparison of corresponding maps for *CSD original* and *π-based* protocols clearly shows that the latter are much more sparse. In Fig. 3 (1AOE binding site of quinazoline derivative) it is interesting to note that although PDB maps point out the ligand quinazoline NH and $NH_2$ groups that bind Glu32, CSD original and $\pi$-corrected maps for the uncharged NH probe do not. Maps for OH do point out the appropriate regions in all cases.

Descriptor values for hydrogen-bonding OH and CO ligand moieties are somewhat higher than those for hydrophobic ligand groups; CSD data tend to yield more favourable $f_{P>1}$ values for aromatic CH than PDB data do. $f_{P<P'}$ values are relatively low for hydrophobic probes due to the nature of the maps: these are usually quite disperse and devoid of strong features. Maps for hydrogen-bonding probes like OH or CO tend to display localised areas of high favourability ('hot-spots'), and this is expressed in the $f_{P<P'}$ values of around 80%, indicating that hydrogen-bonding ligand groups are, on average, found in quite specific areas of high likelihood that make up only about 20% of the map.

**Map descriptors, distance-limited π-based correction.** Introduction of a distance-limitation $\Delta_{shell}$ = 0.0 Å for contacts to apolar central groups increases $f_{P>1}$, which points out that more ligand groups are now in favoured regions with propensities larger than 1; values for $f_{P<P'}$ are very similar, indicating that maps are similar in shape to those that do not include the correction (Table 5). Introduction of a distance limitation for contacts to polar central groups was found to have a less beneficial effect, so this was not applied. Table 1 (*CSD*, *π-based*/$\Delta_{shell}$) shows application of the distance restriction to be favourable for the prediction of apolar

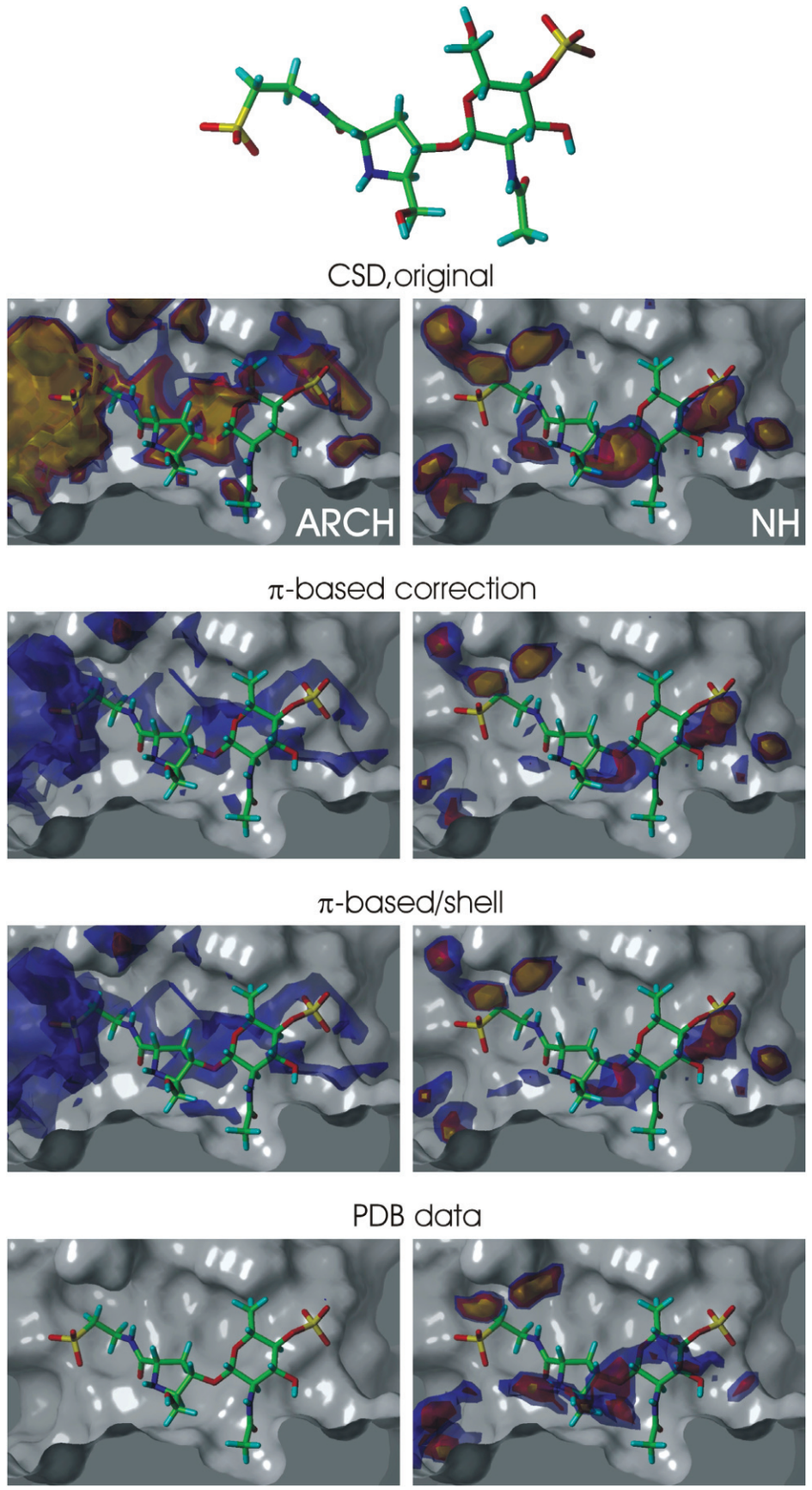



**Fig. 2** SuperStar maps for PDB entry 1D0L (E. J. van Asselt, K. H. Kalk and B. W. Dijkstra, *Biochemistry*, 2000, **39**, 1924–1934); waters have been omitted from map calculation. Maps are shown for probes uncharged NH (right) and aromatic CH (left), corrected with the original hydrophobicity adjustment, the $\pi$-based protocol, and the $\pi$-based protocol with shell-correction. Maps calculated from PDB data are shown for comparison at the bottom. Propensities contoured at levels: 2 (blue); 4 (red); 8 (yellow).

ligand groups by either correct or appropriate probes. A strong improvement is observed for prediction of ligand NH groups by the correct uncharged NH map probe. This unexpectedly large influence of $\Delta_{shell}$ on the prediction of NH groups is likely to be the result of these groups being planar, and therefore often sitting in tight cavities. It is in such cavities that the influence of distant, noisy regions is most noted, as the likelihood of sub-map regions overlapping is high in such narrow binding sites. This can also be

**Table 5** Map descriptors $f_{P>1}$ and $f_{P<P'}$ describing propensities at ligand group positions. H.C. hydrophobicity correction method. $f_{P>1}$ average percentage of ligand groups observed to lie at positions with propensity >1.0; $f_{P<P'}$ percentage of map with propensities smaller than the propensity found at the ligand group position

| | | | *Ligand groups* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Aliphatic CH* | | *Aromatic CH* | | *C=O* | | *OH* | | *RR′NH* | | *$RNH_2$* | |
| *Source* | *H.C.* | *filter* | $f_{P>1}$ | $f_{P<P'}$ | $f_{P>1}$ | $f_{P<P'}$ | $f_{P>1}$ | $f_{P<P'}$ | $f_{P>1}$ | $f_{P<P'}$ | $f_{P>1}$ | $f_{P<P'}$ | $f_{P>1}$ | $f_{P<P'}$ |
| PDB | n.a. | — | 62 | 71 | 26 | 19 | 71 | 85 | 79 | 82 | 89 | 74 | 90 | 86 |
| CSD | original | — | 52 | 57 | 86 | 63 | 84 | 86 | 82 | 82 | 53 | 65 | 35 | 48 |
| CSD | π-based | — | 31 | 53 | 71 | 55 | 78 | 87 | 53 | 80 | 35 | 63 | 10 | 42 |
| CSD | π-based/shell | — | 64 | 56 | 77 | 55 | 77 | 86 | 51 | 80 | 41 | 64 | 41 | 57 |
| CSD | π-based | Δ = −0.3 | 37 | 52 | 79 | 53 | 84 | 87 | 68 | 84 | 59 | 71 | 52 | 63 |
| CSD | π-based | Δ = −0.1 | 40 | 48 | 80 | 53 | 85 | 87 | 66 | 83 | 59 | 70 | 59 | 67 |
| CSD | π-based | Δ = +0.1 | 53 | 50 | 82 | 50 | 82 | 86 | 65 | 82 | 47 | 58 | 52 | 56 |
| CSD | π-based/shell | Δ = −0.3 | 69 | 55 | 82 | 54 | 84 | 87 | 68 | 84 | 59 | 71 | 55 | 64 |
| CSD | π-based/shell | Δ = −0.1 | 71 | 51 | 83 | 54 | 85 | 87 | 66 | 83 | 59 | 70 | 59 | 67 |
| CSD | π-based/shell | Δ = +0.1 | 78 | 51 | 83 | 50 | 82 | 86 | 65 | 82 | 47 | 58 | 52 | 58 |

observed in Fig. 3. A comparison of NH maps for *CSD*, *π-corrected/shell* and *CSD*, *π-corrected* calculations shows that the former now does point out the quinazoline ring NH and $NH_2$ groups that bind to protein residue Glu32. This protein cavity (PDB code 1AOE) is a typical example of a narrow binding site that accommodates a planar nitrogen heterocycle compound.

Focusing on competition between NH and OH map probes when predicting ligand OH groups (Table 4), we see that a larger proportion of donating-only ligand OH groups is predicted by NH probe with $\Delta_{shell}$ correction than for the π-based protocol only: only 5 out of 49 (10%) of donating-and-accepting OH groups are predicted as NH (another 7 are predicted by the carbonyl probe), whereas 8 out of 33 donating-only OH groups (24%) are predicted as NH (another 1 is predicted as CO, and 3 are predicted as aliphatic CH). In such cases a pure donor may indeed be more appropriate.

### 3.2 Elimination of secondary contacts

Fig. 4 shows original and filtered IsoStar scatterplots of hydroxyl groups around phenyl for two different $\Delta_{filter}$ settings. Filtering removes those contacts that have no strong interaction with the ring. What remains is a strip of contact groups in the plane of the ring for the highest settings of $\Delta_{filter}$. At first glance, one might expect that removal of contacts deteriorates statistics for the resulting plots but a large number of the discarded contacts are incidental occurrences and probably just add noise. Filtering tends to remove distant contacts that add density at the very fringe of favourable regions in the contour plots, decreasing ambiguity in those cases where sub-maps overlap.

Filtering of secondary contacts is not expected to be beneficial for all cases where we observe differences between CSD and PDB data. Although the multitude of interaction geometries found in CSD and PDB are the same, differences are found in the ratio of occurrence for a small number of contact pairs.[17] As an example, Fig. 5 shows OH contact data for amide linkages. PDB data favour formation of amide NH hydrogen bonds over amide CO ones, whereas CSD interactions prefer contacts to either acceptor or donor. It is not clear why this difference arises; it may be a genuine effect, but might equally be the result of a much higher diversity of contacts observed in CSD data. Fig. 5 (extreme right) shows the effect of secondary contact filtering in this case: non-relevant, remote contact groups are removed.

SuperStar validation results for different types of filters (Table 6) indicate that a combination of π-based hydrophobicity correction and removal of secondary contacts benefits accuracy of donor group prediction. Both prediction of groups by correct and by appropriate probes is seen to increase for $RNH_2$ groups. Smaller improvements are observed for the prediction of RR′NH ligand groups. The same trend is observed both with and without application of the $\Delta_{shell}$ correction. Overall, results for the application of a secondary contact filter are not dissimilar from those for application of a distance limit $\Delta_{shell}$. This is not unexpected, since both the distance limitation correction and the secondary contact filter tend to eliminate contact groups that form long-distance interactions and do not contribute useful information. Fig. 3 shows that both secondary contact filtering and $\Delta_{shell}$ correction have a similar influence on NH maps. Small differences are observed for competitive prediction of OH groups (Table 4).

Application of a secondary-contact filter in addition to the $\Delta_{shell}$ restriction has minor effects. The $f_{P>1}$ and $f_{P<P'}$ fractions are seen to increase for donor map probes (OH, $RNH_2$, RR′NH, Table 5), but success rates are only affected marginally (Table 6). The effect is observed for both polar and apolar ligand groups. Map descriptors and success rates suggest that a combination of filtering and distance-limitation correction yields optimal results for a $\Delta_{filter}$ value of −0.1.

### 3.3 Solvent bias filtering

Solvent interactions were eliminated and results for such a filter are shown in Table 6 (*SBF*). It is clear from these data that solvent effects have no major influence on accuracy of prediction of ligand groups. We therefore conclude that solvent interactions in small-molecule crystal structures are not distinguishable from interactions between other groups with similar functionality when assessing strongly interacting groups like OH, CO, and NH.

### 3.4 Examples of prediction of ligand groups

Aromatic CH probe maps shown for the relatively polar bulgecin A binding site (PDB code 1D0L) in Fig. 2 show that the original scheme for hydrophobicity correction of CSD data tends to over-emphasise binding of hydrophobic probes. The reason that this does not have a strongly detrimental effect on success rates of prediction is that, usually, maps for polar and apolar probe groups occupy different regions in space, and hence do not compete when it comes to prediction of ligand groups. The PDB-based map for aromatic CH groups is empty at these contouring levels, indicating a low predicted preference for aromatic interactions. The original scheme can be seen to yield NH-probe maps that have larger regions of high propensity than the maps calculated with the π-based correction or from PDB data, *i.e.* the original scheme does not adjust the strength of polar interactions to the level observed in protein data.

Fig. 3 (dihydrofolate reductase binding site, PDB code 1AOE) allows a comparison of maps for different probes calculated with different protocols (*original*, *π-corrected*, *π-corrected*/$\Delta_{shell}$, and *π-corrected*/$\Delta_{filter}$). Contour maps are shown for aromatic CH, carbonyl, hydroxyl, and uncharged NH probes. Comparing PDB-based and original CSD maps, one can see that high-propensity regions in PDB maps are relatively sparse. This may be a result of interaction data derived from protein binding sites being more relevant, but could equally well be a result of a lack of diversity in the PDB data that are available. A comparison of *original* CSD maps and *π-corrected* ones shows that the latter suppress some regions that are featured strongly in the *original* maps. Given that

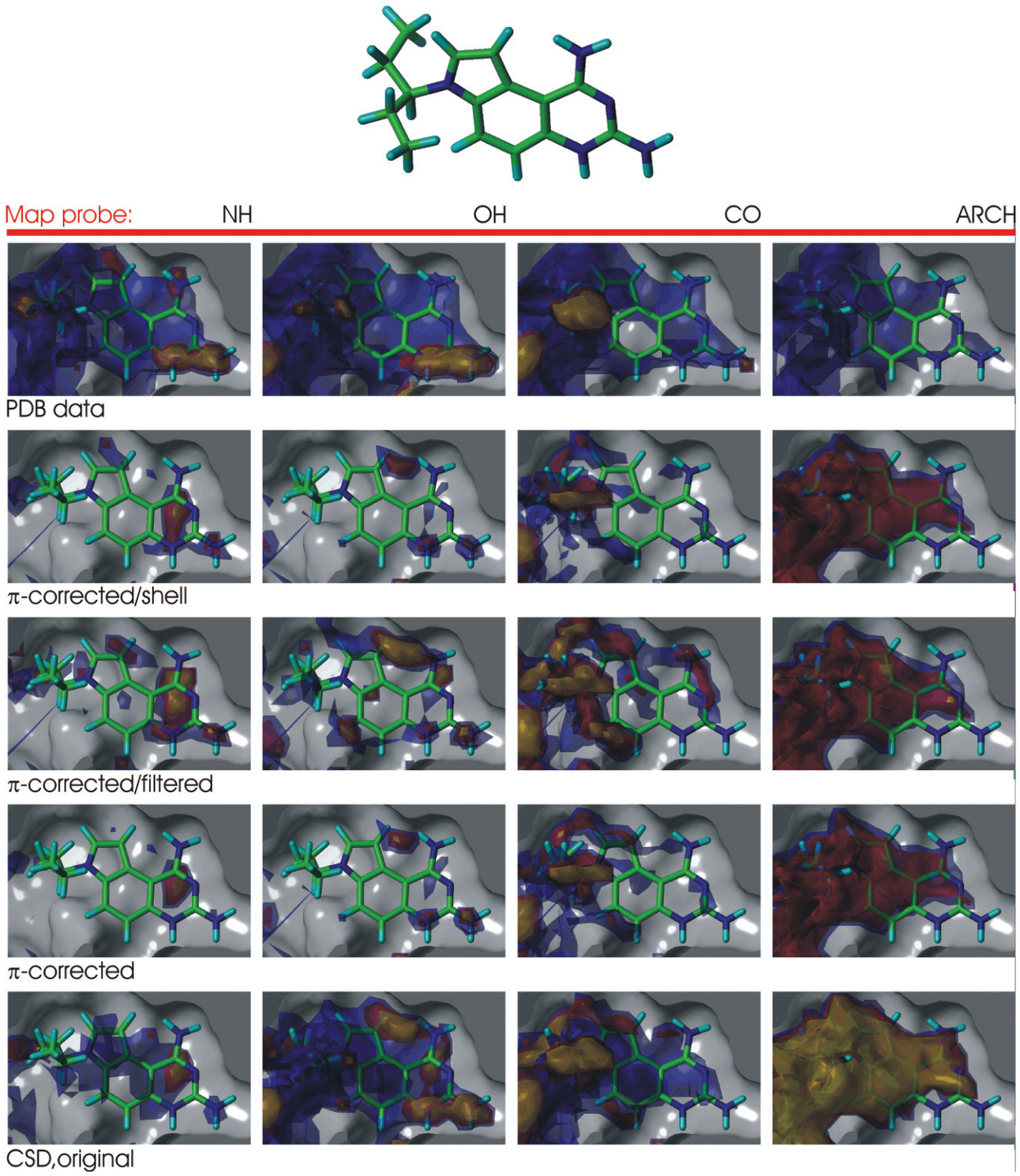



**Fig. 3** Example maps calculated for a dihydrofolate reductase binding site (1AOE; M. Whitlow, A. J. Howard, D. Stewart, K. D. Hardman, L. F. Kuyper, D. P. Baccanari, M. E. Fling and R. L. Tansik, *J. Biol. Chem.*, 1997, **272**, 30289–30298). The quinazoline ligand is shown in green. CSD results are shown for the original hydrophobicity correction scheme, for the $\pi$-based scheme, for the $\pi$-based scheme using data filtered for secondary contacts ($\Delta_{filter} = -0.1$), and for the $\pi$-based scheme using a reduced interaction shell ($\Delta_{shell}$). ARCH aromatic CH probe; CO carbonyl probe; OH hydroxyl probe; NH uncharged NH probe. Propensity contours: 1 (blue); 2 (red); 4 (yellow). Results calculated with PDB interaction data are shown for comparison.

validation results are similar for both, one can assume that the suppressed regions are indeed of lesser interest. Comparing the *π-corrected*-maps to the *π-corrected/filtered* and *π-corrected/shell* contours, changes can be observed for the polar probes, most notably in the maps for the NH probe.

## 4 Conclusions

In this paper we describe the selection and processing of information on intermolecular interactions from crystal structure data for use in knowledge-based methods, and the effect of using such data for the prediction of interactions in protein binding sites. In particular, we explore the use of small-molecule interaction data for the prediction of protein–ligand interactions in protein binding sites. All calculations were performed with SuperStar, an application that generates propensity maps that indicate the likelihood of occurrence of certain probe groups in binding sites.

We implemented a new, improved hydrophobicity correction mechanism that renders CSD data suitable for use in predicting protein–ligand interactions, correcting the reference state to an aqueous-like environment. The protocol uses octanol–water partitioning coefficients to estimate the hydrophobicity of binding

**Table 6** SuperStar success rates for the prediction of ligand groups ($\pi$-based hydrophobicity correction protocol) using filtered IsoStar data, both without and with shell correction ($\Delta_{shell}$ = 0.0 Å for contacts to apolar central groups). Results for unfiltered data have been inserted for comparison. $\Delta_{filter}$: secondary-contact filter with given delta value; SBF: solvent bias filter. *f* percentage of ligand groups predicted by correct probe; *f′* percentage of predictions by probe with appropriate physicochemical properties

| | Aliphatic CH | | Aromatic CH | | C=O | | OH | | $RNH_2$ | | RR′NH | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ngroups* | *468* | | *224* | | *176* | | *96* | | *29* | | *74* | |
| *Filter* | *f* | *f′* | *f* | *f′* | *f* | *f′* | *f* | *f′* | *f* | *f′* | *f* | *f′* |
| *Without $\Delta_{shell}$ correction* | | | | | | | | | | | | |
| *none* | 18 | *81* | 75 | *80* | 85 | *85* | 70 | *97* | 17 | *72* | 41 | *71* |
| *$\Delta_{filter}$ = −0.3* | 17 | *79* | 67 | *73* | 86 | *86* | 75 | *94* | 38 | *90* | 47 | *65* |
| *$\Delta_{filter}$ = −0.1* | 18 | *79* | 66 | *71* | 86 | *86* | 69 | *94* | 45 | *90* | 53 | *65* |
| *$\Delta_{filter}$ = +0.1* | 22 | *84* | 74 | *83* | 84 | *84* | 64 | *92* | 38 | *79* | 47 | *53* |
| *SBF* | 17 | *82* | 76 | *80* | 85 | *85* | 69 | *95* | 17 | *72* | 41 | *65* |
| *With $\Delta_{shell}$ correction (contacts to apolar groups, $\Delta_{shell}$ = 0.0 Å)* | | | | | | | | | | | | |
| *none* | 23 | *84* | 78 | *83* | 84 | *84* | 64 | *94* | 35 | *76* | 53 | *71* |
| *$\Delta_{filter}$ = −0.3* | 23 | *81* | 67 | *77* | 85 | *85* | 73 | *93* | 41 | *90* | 53 | *65* |
| *$\Delta_{filter}$ = −0.1* | 27 | *81* | 64 | *75* | 86 | *86* | 68 | *91* | 45 | *90* | 53 | *65* |
| *$\Delta_{filter}$ = +0.1* | 28 | *86* | 70 | *86* | 84 | *84* | 64 | *91* | 38 | *79* | 47 | *53* |
| *SBF* | 17 | *82* | 76 | *80* | 85 | *85* | 69 | *95* | 17 | *72* | 41 | *65* |

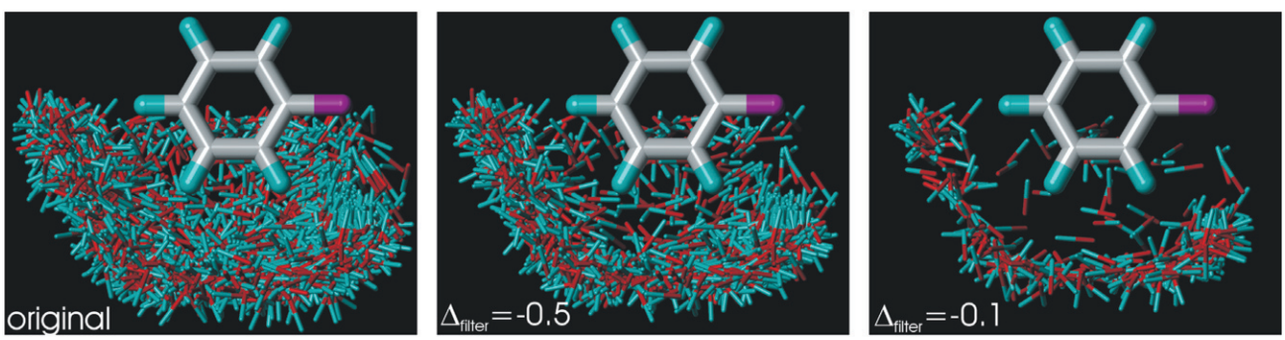



**Fig. 4** Results of secondary-contact filtering of OH contacts about a phenyl ring for different settings of $\Delta_{filter}$. Data are taken from the CSD.

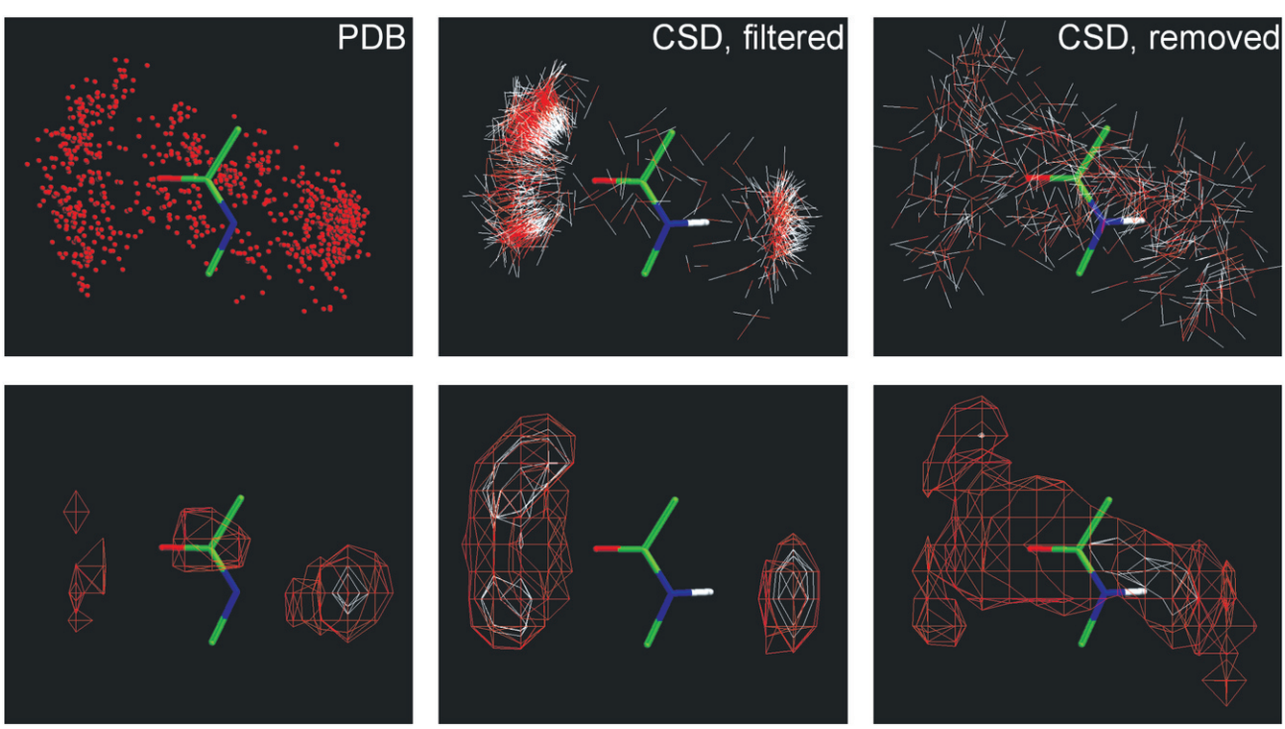



**Fig. 5** Scatterplots and corresponding contour maps for OH contacts surrounding a peptide link. From left to right: contacts from the PDB; contacts from the CSD, filtered for secondary contacts ($\Delta_{filter}$ = −0.1); CSD contacts discarded by the filter. Propensity contours at 1 (red) and 4 (white).

site environments and probes, and uses this information to adjust small-molecule based propensity data to values that apply to protein-type environments.

Improvements were observed for prediction of ligand groups in protein binding sites when irrelevant contacts in 'noisy' regions were eliminated from the knowledge-base. Two approaches were shown to be effective in eliminating contacts that do not convey information. Secondary-contact filtering removes those interactions from the knowledge-base that are incidental to a second, stronger interaction. A second approach eliminates all remote regions beyond a given distance threshold if lacking a number of observations significantly larger than the amount expected at random. Both approaches show favourable influences on predictions of ligand groups by corresponding map probes. Most notably, ligand-NH

group prediction by the uncharged NH map probe is observed to rise from 7% (original map calculation) to approximately 35% ($RNH_2$ groups), and from 18% to up to 53% (RRNH groups, including ring-NH). Such groups typically sit in narrow binding sites and can be difficult to predict correctly. The accuracy of prediction of these groups by either OH or NH probes, which both indicate donor groups, is about 70% whether or not filters are applied, so the filtering approaches increase the specificity of the map probe. This is also observed for ligand hydroxyl groups, where a larger amount of hydroxyls that both donate and accept are predicted correctly by the OH map probe when the filters are applied.

One might expect that small hydrogen-bonding solvent molecules in crystals have slightly different properties than their main building blocks. This was investigated by filtering out solvent contacts from the knowledge-base and assessing results of prediction; however, an improvement could not be observed for the map probes used.

In general, accuracy of prediction of ligand moieties by interaction maps was observed to be similar or better for corrected and filtered CSD data than for PDB data. The only exception is the prediction of ligand NH groups, where PDB-data excel. This may be a result of the lack of diversity in PDB-data, which may focus NH prediction to those specific types of protein–ligand interaction that are observed in the knowledge-base; NH interaction data from the CSD cover a much broader range of chemical groups.

The contact filters require pre-processing of the knowledge-base. The distance-limitation filter does not require such pre-processing and can easily be applied during calculation of maps. Although the data-selection procedures described here have been applied to small-molecule crystal structure data, we have seen evidence that the distance-limitation filter improves prediction of interactions using PDB data slightly; it may therefore be relevant for other knowledge-based approaches, like scoring functions for protein–ligand docking.

## 5 Methods

### 5.1 SuperStar validation

All calculations were performed using SuperStar v1.5 (release date mid 2004) and IsoStar v1.5 (released November 2003). Validations were carried out using a test set of 224 structures that has been published previously.[39] This set has been checked for errors and diversity. The set contains protein structures and separate ligand structures. All protein structures have had hydrogens added and their protonation states set; they have been inspected visually.

The validation involved comparing the experimental positions of ligand functional groups in binding sites with those predicted by SuperStar. This was done by calculating several SuperStar maps for each binding site, using different probes. A ligand group was deemed to be predicted correctly if it matched the probe which, at the experimentally-observed position of the ligand group, had the highest propensity of all the probes for which maps were generated. Success rates were determined as the percentage of groups that were predicted correctly. Only buried[15] ligand groups were considered, as interactions of groups that are exposed to solvent generally cannot be predicted reliably.

The following map probes were used, with the corresponding ligand groups in brackets: alcohol oxygen (alcohol groups, ROH); carbonyl oxygen (carbonyl groups, RR′CO), uncharged NH nitrogen (disubstituted amino groups RR′NH and amino groups $RNH_2$), aliphatic CH carbon (aliphatic CH, methyl and methylene groups, RR′R″CH), aromatic CH carbon (aromatic CH groups). These map probes cover a broad range of hydrogen bonding and hydrophobic interactions.

Two success rate were determined (*f* and *f*′ in Table 1). *f* is the rate of exactly correct predictions, *i.e.* where the observed ligand group exactly matches the probe giving rise to the highest SuperStar propensity at that point. *f*′ is the success rate if we also count as successful situations where the probe of highest propensity is not an exact match of the ligand group but has similar characteristics (is an "appropriate" probe; *e.g.* if the ligand group were aliphatic CH and the probe of highest propensity were aromatic CH). Specifically, for hydrophobic groups, a prediction would be deemed correct if predicted to be either of the hydrophobic probes; for hydroxyls, prediction as carbonyl or NH is deemed correct; for amino groups, prediction by OH is deemed correct.

### 5.2 Hydrophobicity correction and database filtering

The hydrophobicity correction was implemented in SuperStar, and is applied automatically after calculation of the raw CSD-based SuperStar maps according to eqns. (4) and (5) (*π-based* protocol), or according to the *original* (single correction factor) protocol. For the $\pi$-based protocol, the protein environment of each grid point in the map is looked up, and the $\pi_E$ contribution is based on the $\pi$ values for one or more nearby protein groups. Contribution $\pi_X$ depends on the chosen map probe.

Filtering of contacts was performed using in-house software. IsoStar scatterplot files were analysed and for each contact group the original contact atom-to-central atom pair was retrieved from the relevant CSD crystal structure (CSD version 5.23). Atom coordinates of the crystal structure were expanded by symmetry when needed. For solvent contact filtering, simple heuristics were applied to determine whether the contact atom was part of a solvent molecule. For secondary-contact filtering, the contact group atom and central group were identified and it was then checked whether there were any short distances between the contact group atom and a third moiety in the crystal. The decision whether the contact under investigation should be rejected was made using the distance cut-off value $\Delta_{filter}$ defined in eqn. (7).

The significance filter was implemented in SuperStar, and *S* (see eqn. (7) and explanation) was set to an empirically-optimised value of 0.001; $\Delta_{shell}$ values were set for apolar and polar groups separately.

## Acknowledgements

The authors acknowledge Dr. M. L. Verdonk for suggesting the significance estimation mechanism based on Poisson probabilities.

## References

1 C. C. Blake, D. F. Koenig, G. A. Mair, A. C. North, D. C. Phillips and V. R. Sarma, *Nature*, 1965, **206**, 757–761.
2 C. R. Beddell, P. J. Goodford, F. E. Norrington, S. Wilkinson and R. Wootton, *Br. J. Pharmacol.*, 1976, **57**, 201–209.
3 A. Brik and C. H. Wong, *Org. Biomol. Chem.*, 2003, **1**, 5–14.
4 G. Klebe, *J. Mol. Med.*, 2000, **78**, 269–281.
5 F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, 1977, **112**, 535–542.
6 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
7 F. H. Allen, *Acta Crystallogr., Sect. B*, 2002, **58**, 380–388.
8 M. Tintelnot and P. Andrews, *J. Comput. Aided Mol. Des.*, 1989, **3**, 67–84.
9 G. Klebe, *J. Mol. Biol.*, 1994, **237**, 212–235.
10 I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, *J. Comput. Aided Mol. Des.*, 1997, **11**, 525–537.
11 H. J. Bohm, *J. Comput. Aided Mol. Des.*, 1992, **6**, 593–606.
12 H. J. Bohm, *J. Comput. Aided Mol. Des.*, 1992, **6**, 61–78.
13 D. J. Danziger and P. M. Dean, *Proc. R. Soc. London, Ser. B*, 1989, **236**, 101–113.
14 R. A. Laskowski, J. M. Thornton, C. Humblet and J. Singh, *J. Mol. Biol.*, 1996, **259**, 175–201.
15 M. L. Verdonk, J. C. Cole and R. Taylor, *J. Mol. Biol.*, 1999, **289**, 1093–1108.
16 M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet and P. Willett, *J. Mol. Biol.*, 2001, **307**, 841–859.
17 D. R. Boer, J. Kroon, J. C. Cole, B. Smith and M. L. Verdonk, *J. Mol. Biol.*, 2001, **312**, 275–287.
18 J. W. M. Nissink, M. L. Verdonk and G. Klebe, *J. Comput. Aided Mol. Des.*, 2000, **14**, 787–803.
19 W. R. Pitt and J. M. Goodfellow, *Protein Eng.*, 1991, **4**, 531–537.
20 P. Watson, P. Willett, V. J. Gillet and M. L. Verdonk, *J. Comput. Aided Mol. Des.*, 2001, **15**, 835–857.

21 P. Labute, *Journal of the CCG*, 2000, http://www.chemcomp.com.
22 V. V. Rantanen, K. A. Denessiouk, M. Gyllenberg, T. Koski and M. S. Johnson, *J. Mol. Biol.*, 2001, **313**, 197–214.
23 H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.
24 G. E. Kellogg, S. F. Semus and D. J. Abraham, *J. Comput. Aided Mol. Des.*, 1991, **5**, 545–552.
25 P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.
26 A. Miranker and M. Karplus, *Proteins*, 1991, **11**, 29–34.
27 A. Caflisch, A. Miranker and M. Karplus, *J. Med. Chem.*, 1993, **36**, 2142–2167.
28 T. Kortvelyesi, S. Dennis, M. Silberstein, L. Brown, 3rd and S. Vajda, *Proteins*, 2003, **51**, 340–351.
29 M. Silberstein, S. Dennis, L. Brown, T. Kortvelyesi, K. Clodfelter and S. Vajda, *J. Mol. Biol.*, 2003, **332**, 1095–1113.
30 D. A. Erlanson, A. C. Braisted, D. R. Raphael, M. Randal, R. M. Stroud, E. M. Gordon and J. A. Wells, *Proc Natl Acad Sci U S A*, 2000, **97**, 9367–9372.
31 S. B. Shuker, P. J. Hajduk, R. P. Meadows and S. W. Fesik, *Science*, 1996, **274**, 1531–1534.
32 C. Hansch and A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Wiley & Sons, New York, 1979.
33 T. Suzuki and Y. Kudo, *J. Comput. Aided Mol. Des.*, 1990, **4**, 155–198.
34 G. Klopman, J.-Y. Li, S. Wang and M. Dimayuga, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 752–781.
35 A. J. Leo, *Chem. Rev.*, 1993, **93**, 1281–1306.
36 C. H. Gorbitz and H. P. Hersleth, *Acta Crystallogr., Sect. B*, 2000, **56 (Pt 3)**, 526–534.
37 C. H. Gorbitz and H. P. Hersleth, *Acta Crystallogr., Sect. B*, 2000, **56 (Pt 6)**, 1094–1102.
38 E. J. van Asselt, K. H. Kalk and B. W. Dijkstra, *Biochemistry*, 2000, **39**, 1924–1934.
39 J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins*, 2002, **49**, 457–471.
40 G. Klopman, J.-Y. Li, S. Wang and M. Dimayuga, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 752.