*Journal of Computer-Aided Molecular Design*, **14:** 787–803, 2000.
*KLUWER/ESCOM*
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

# Simple knowledge-based descriptors to predict protein-ligand interactions. Methodology and validation

J. Willem M. Nissink[a], Marcel L. Verdonk[b] & Gerhard Klebe[a,*]
[a]*Department of Pharmaceutical Chemistry, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany;* [b]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.*

Received 29 December 1999; Accepted 5 June 2000

*Key words:* composite crystal field environments, drug design, hydrogen bonding, hydrophobic effect, IsoStar database, parametrisation of interaction geometries, predicting protein ligand interactions, SuperStar

## Summary

A new type of shape descriptor is proposed to describe the spatial orientation for non-covalent interactions. It is built from simple, anisotropic Gaussian contributions that are parameterised by 10 adjustable values. The descriptors have been used to fit propensity distributions derived from scatter data stored in the IsoStar database. This database holds composite pictures of possible interaction geometries between a common central group and various interacting moieties, as extracted from small-molecule crystal structures. These distributions can be related to probabilities for the occurrence of certain interaction geometries among different functional groups. A fitting procedure is described that generates the descriptors in a fully automated way. For this purpose, we apply a similarity index that is tailored to the problem, the Split Hodgkin Index. It accounts for the similarity in regions of either high or low propensity in a separate way. Although dependent on the division into these two subregions, the index is robust and performs better than the regular Hodgkin index. The reliability and coverage of the fitted descriptors was assessed using SuperStar. SuperStar usually operates on the raw IsoStar data to calculate propensity distributions, e.g., for a binding site in a protein. For our purpose we modified the code to have it operate on our descriptors instead. This resulted in a substantial reduction in calculation time (factor of five to eight) compared to the original implementation. A validation procedure was performed on a set of 130 protein-ligand complexes, using four representative interacting probes to map the properties of the various binding sites: ammonium nitrogen, alcohol oxygen, carbonyl oxygen, and methyl carbon. The predicted ‘hot spots’ for the binding of these probes were compared to the actual arrangement of ligand atoms in experimentally determined protein-ligand complexes. Results indicate that the version of SuperStar that applies to our descriptors is capable to predict the above-mentioned atom types in ligands correctly with success rates of 59% and 74%, respectively, for all ligand atoms (regardless of their solvent accessibility), and a subset of solvent-inaccessible ones. If not only exact atom-type matches are counted, but also those that identify ligand atoms of similar physicochemical properties, the prediction rates rise to 75% and 89%. These rates are close to those obtained by the original SuperStar method (being 67% and 82%, respectively, for the prediction of exact matching atom types, and 81% and 91% in the case of predicting similar atom types).

## Introduction

A number of tools have been published in the field of molecular modeling to assess the recognition of potential drugs by their target receptors, and yet, it has become increasingly obvious that the intricate interaction between a ligand and its surrounding protein binding site cannot be described by simple models. For instance, the contribution of hydrogen bonds to the binding energy of a ligand to a receptor is usually large, but in many cases it has been recognised that neglect of the more subtle hydrophobic contributions leads to an incorrect prediction of the binding geom-

[*] To whom correspondence should be addressed. E-mail: klebe@mailer.uni-marburg.de

etry; in some cases, they even represent the largest contribution to the binding energy [1].

Attempts have therefore been made to survey possible intermolecular interactions [2, 3] and rules for their prediction have been established in many of the commonly applied methods for drug design, e.g. hydrogen-bond prediction [4–6], docking, molecular superposition [7–9], and QSAR [10]. The recently introduced IsoStar database [11], a library that contains a large number of non-bonded interaction geometries, is a powerful tool to validate these models. Basically, it is a useful guide to recognise possible interaction geometries while designing ligands. All entries in IsoStar have been derived from crystal structures stored in the Cambridge Structural Database [12] and Brookhaven Protein Databank [13]. IsoStar is therefore based solely on experimental data.

A particularly interesting feature of IsoStar is that also interactions between hydrophobic groups have been analysed. As recently pointed out by Davis and Teague [1], the relative contributions of ion-ion interactions, hydrogen bonding, dipole-dipole interactions and lipophilicity in the binding process of a ligand are still poorly understood, and the importance of hydrophobic contacts may well be underestimated in current applications. Therefore, the IsoStar database is a valuable starting point for the development of modeling tools, as it offers a wealth of information on a wide range of interactions.

SuperStar, a recently published approach that utilises information from IsoStar, clearly demonstrates its value for the prediction of favorable binding interactions [14]. This method assembles and scales the relevant distribution data from IsoStar to calculate a composite map reflecting the probability (propensity) of a certain probe (an atom in a contact group, e.g., the oxygen atom of a carbonyl group) of forming an interaction with a template at a certain position in space (the template, e.g., being an amino acid residue in the binding site of a protein). The results obtained using SuperStar are comparable with those of the widely applied GRID program [15]. However, instead of using a force-field approach as in the latter program, SuperStar is fully knowledge-based.

The IsoStar database contains a large number of experimentally observed distributions of contact groups around central groups (the so-called scatterplots). Currently, it contains about 10 000 distributions that reflect combinations of 300 central and 45 contact groups. Although covering a wide range of possible interactions, and being very instructive when used interactively (as in SuperStar), the presently stored representation of the empirical information in terms of scatterplots hampers a fast processing of the data and limits its direct implementation in other modeling tools. For a convenient (i.e., fast and accurate) implementation in, e.g., a docking or superpositioning tool, (a) the information on interaction geometries present in IsoStar should be converted to a format that reflects the orientation of the probe and the central group in a transparent and physically meaningful way, and (b) the derived representation should permit fast calculation of the interaction geometry probabilities.

Accordingly, we present in this paper anisotropic Gaussian-type shape descriptors to approximate the propensity distributions derived from IsoStar scatterplots. We present a data-fitting procedure to convert IsoStar distributions into a set of parameterised descriptors in a fully automated and flexible way. A new type of similarity index is proposed for the fitting of the propensity distributions. To validate the quality of the obtained descriptors, the resulting Gaussian shapes are implemented into the SuperStar program. A comparison is made between the results generated using the original IsoStar data, and those obtained while applying the set of Gaussian descriptors during the calculation of the SuperStar fields.

## Methodology

### *Non-spherical Gaussian distributions*

IsoStar scatterplots represent the distribution of contact groups around a central group. These are assembled from a large set of crystal structures that have this central group in common. These empirically derived distributions can be converted to propensity-plots[1], that reflect the probability of finding a contact group in a particular region in space. An example is shown in Figure 1 for hydroxyl groups surrounding an ether fragment. A visual inspection reveals that a large number of IsoStar distributions are extended along curved

[1] The propensity is a measure that reflects the non-randomness of a contact group being found at a certain point in the volume that encloses the central group. It is obtained by scaling the density-of-occurrence (expressed in terms of Å$^{-3}$). A propensity of $p = 1.0$ at a certain point X indicates that the chance of observing the contact group at this particular position equals the probability that would be found for a random distribution of the contact group; likewise, $p = 2.0$ means that the chance of finding a group at this position is twice the random probability, etc. For a more detailed description of the calculation of the propensity, we refer to Bruno et al. [11].

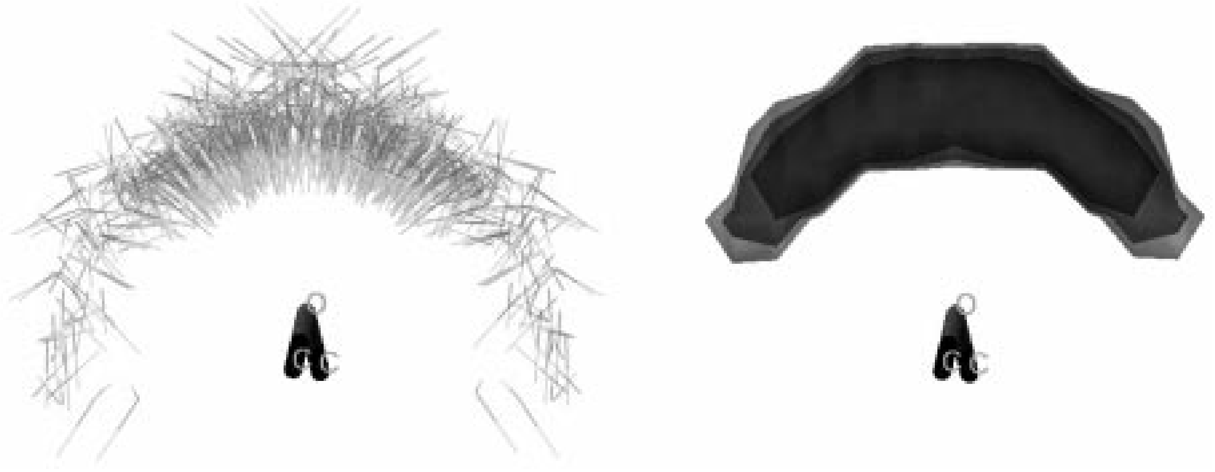


*Figure 1.* Example of a scatterplot from IsoStar (left) and the corresponding propensity plot (right) for a dialkyl C–O–C ether fragment as central group and alcohol COH moieties as contact groups. The gray-scaled contours correspond to propensity levels for the oxygen atom of 1.0, 2.0, and 4.0 (light to dark gray). All contour plots were generated using the program Sybyl v6.5, by Tripos, Inc.

surfaces. We therefore chose a mathematical description for the propensities of such distributions based on a spherical coordinate system. The modeled property falls off with the distance from a predefined centre. Furthermore, we assume that the decay is anisotropic, and accordingly the rate of decay will be different along different directions. The mathematical form is given in Equation 1

$$\mathrm{p}^b(v) = H \cdot f_1 \cdot f_2 \cdot f_3, \tag{1}$$

where the propensity value *p* at a certain point $v$ is given by the height *H* at centre position *b* multiplied by the three decay functions $f_n(n = 1, 2, 3)$. They account for the decay of the propensity p along three directions, one along direction *O*b, and two being along the circular arcs *p* and *q*, as depicted in Figure 2. Vector *O*b is called the base vector. It points to the centre of the distribution (see Figure 2a).

The three functions $f_n$ decay along the three above-mentioned directions. To derive these, the position of point $v$ is expressed in a local spherical coordinate system, the first coordinate being radial and the latter two being angular (see Figure 2b). Three deviations $(\delta_1, \delta_2, \delta_3)$ are expressed in Ångstrom and can be derived easily from the spherical coordinates. The decay functions are modeled by Gaussians (Equation 2)

$$f_n(\delta_n) = \exp(-c_n \cdot \delta_n^2) \tag{2}$$

with coefficients $\mathrm{c}_n$ given by Equation 3

$$\mathrm{c}_n = \frac{1}{2\sigma_n^2} \tag{3}$$

The widths $\sigma_n$ are determined by the shape of the distribution centred at b and stretched along the directions *p* and *q* on a spherical surface of radius $|O\mathrm{b}|$, and along the normal to this surface (see Figure 2). A distribution that is described in such a way can be modeled by a total of 10 parameters, being (number of parameters given in parentheses): the origin *O*(3); base vector *O*b(3); Gaussian width coefficients $\sigma_n$ (3); maximum height *H* (1). The curvature of the distribution is that of the surface of a sphere with radius *O*b. In order to adjust the curvature in a case where *O* is fixed (and thus, the length of $|O\mathrm{b}|$), an additional parameter, *k*, may be added that modifies the curvature according to Equation 4.

$$R_{\mathrm{curv}} = (1 - k) \cdot |O\mathrm{b}|. \tag{4}$$

For a simple interaction between two groups A and B, the selected parameters have the following geometrical meaning: given the central fragment A, with an atom $\mathrm{X}^{\mathrm{A}}$ placed at the origin *O*, an interacting atom $\mathrm{Y}^{\mathrm{B}}$ of contact group B is preferentially found at the position indicated by b, at a (bonding) distance |*O*b|. The parameters $c_n$ designate the width of the Gaussian functions and thus they model the decay of the propensity distribution along three above-defined directions.

Since interacting groups often feature more than one preferred binding position b, we model the overall propensity distribution *P* more generally by a set of *m* fit functions (Equation 5)

$$P(v) = \sum_m p^b(v) \tag{5}$$

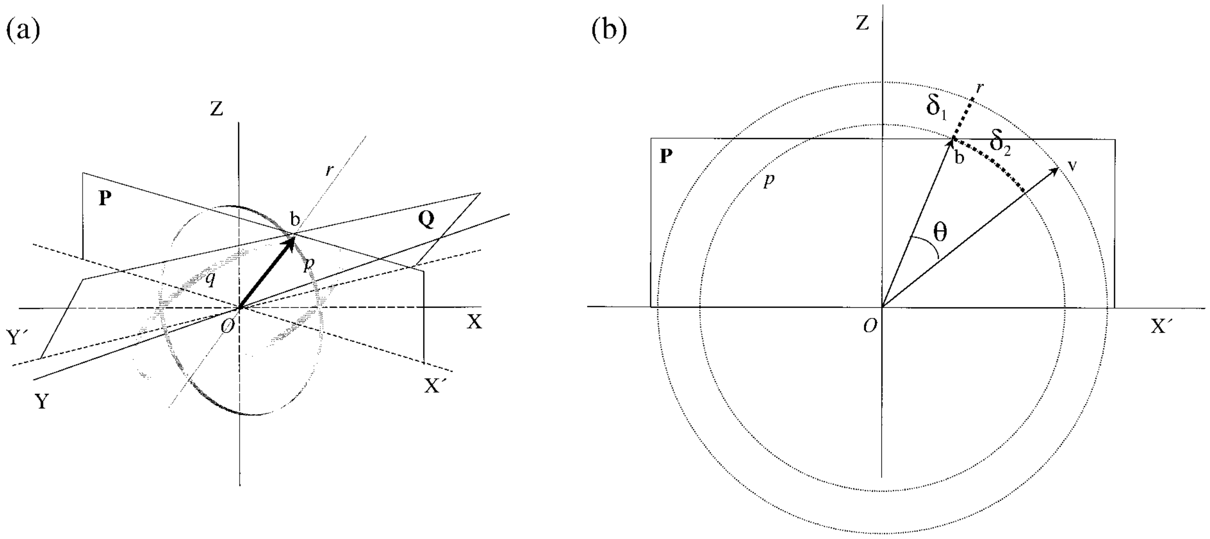


*Figure 2.* The coordinate system used to describe a non-spherical Gaussian-type distribution is displayed in (a) and the **P**-plane is shown in detail in (b). The base vector *Ob* (displayed as an arrow) points towards the centre of the distribution and starts from origin *O*. The deviations $(\delta_1, \delta_2, \delta_3)$ used for the evaluation of the decay functions are measured along three directions *p*, *q*, and *r.* For a given point v in space, $\delta_1$ measures the distance along the radial direction r. It is calculated as the difference between the length of vector *Ov* and base vector $|Ob|$ (see b); $\delta_2$ and $\delta_3$ measure the distances from *v* to the planes **Q**, and **P**. They are measured along the arcs *p* and *q* ($\delta_2$ is shown in b.). The deviations can be easily calculated from the angles between the projection of vector *Ov* onto **P** (or **Q**) and plane **Q** (or **P**); the angle θ is shown in (b).

rather than using one single Gaussian representation (Equation 1). In Equation 5, the superscript b designates that the proper local origin *O*, base vector *Ob*, height *H*, coefficients $F_Q$, and directions *p*, *q*, and *r* should be applied to calculate the contribution of each fit function $p^b(v)$.

*Fitting protocol*

The shape descriptors of the form introduced in Equation 5 are fitted to a given IsoStar scatter distribution by the following steps:

*Step 1. Generate the propensity distributions from IsoStar data.* The discrete propensity distributions to be fitted are derived from a subset of the IsoStar scatterplots using the approach described by Bruno et al. [11]. In this approach, a suitable grid is superimposed onto the scatterplot and a density of contact groups is calculated for each grid point by counting the number of contact groups contained in the grid cube that surrounds the point. This density is then converted to a propensity by a normalisation procedure.

The spacing of the data grid is selected such that the statistical significance of the obtained propensities is optimal. In our experience, this is usually achieved with a grid spacing in the range from 0.5 to 1.4 Å. Only scatterplots having on average at least 4 contact groups per Å$^3$ in the considered volume around the central group are converted.

*Step 2. Determine the number of fit functions to be used.* In order to determine the number of required Gaussian descriptors *m* (see Equation 5), preferred binding orientations are singled out by determining well-separated clusters of points with high propensity in the neighbourhood of the central fragment. The members of one cluster are found by first finding all neighbours of a given starting grid point with propensity $P > L_{\min}$; the procedure is then repeated recursively for all neighbouring points that have thus been found, until no additional cluster members are discovered. As a starting grid point for a new cluster we choose the point with the highest density that has yet not been accounted for. This procedure is repeated until all grid points have been analysed. The minimum threshold $L_{\min}$ is chosen in an iterative way such that the total number of clusters *m* is maximal.[2]

*Step 3. Estimate the fit parameters.* The clusters of grid points obtained by the procedure in the previ-

[2] Clusters with less than a minimum number of points (currently, 4) are not used to prevent over-fitting of very small clusters. Usually, this additional criterium does not lead to problems when the grid spacing is small (< 1 Å).

ous step are processed separately during the fitting process. Thus, initially only points with $P{>}L_{\min}$ are used. Each cluster of grid points is then fitted by one Gaussian descriptor (Equation 1). The origin $O$ for this descriptor is placed at an appropriate position in space (e.g., the nearest non-hydrogen atom of the central fragment, or the global origin given by the IsoStar coordinate system); the base vector $O$b is then computed by determining the weighted centre of the cluster, and the height H is set to the highest propensity observed in this cluster. An initial fit of the descriptor (Equation 1) to the cluster points is then performed, in which only the Gaussian coefficients $c_n$ are allowed to vary. This procedure yields a parameterised fit function (Equation 1) with a preliminary set of 10 fit parameters for each cluster.

*Step 4. Expanding the clusters.* The propensity at a particular grid point is described by a sum over all fit functions according to Equation 5. Each grid point in the input distribution that has not been assigned yet is now added to one of the present clusters by finding which of the fit functions contributes most to the propensity at this point. Once the data points are assigned, the original IsoStar propensity value at each data point in a cluster is modified by subtracting the calculated contributions of the fit functions of *other* clusters. Following this concept, we account for the overlap of neighbouring functions and thus we prevent points from being multiply fitted. The net result of this step is that the entire distribution is split into partial distributions (the clusters) that can be fitted separately.

*Step 5. Refine the fit solution.* The uniquely assigned data points in each cluster are fitted using a single Gaussian descriptor (Equation 1). Minimisation of the fit residual (*vide infra*) is performed using a Simplex minimiser [16]. Steps 4 and 5 are iterated and after each cycle a goodness-of-fit (see next section) is calculated for the fitted and experimental distributions. The iteration is stopped if the goodness-of-fit does not improve in consecutive cycles.

As a final result the fitting procedure yields a set of parameters for each Gaussian function. To assess the quality of the derived descriptors compared to the original data, a figure-of-merit is computed as described in the following section.

### *Similarity index: Split Hodgkin Index*

The fitting procedure described above requires a reliable goodness-of-fit indicator. Several similarity indices have been proposed in literature; for a recent review, we refer to Willett et al. [17]. The simplest (non-scaled) indicator for the similarity between an actual and a fitted distribution is the well-established Root Mean Squared deviation (RMSd) of fitted and actual data values (not shown). A normalised form of this index is the Hodgkin index [18] $S_{AB}$, that yields the similarity between two distributions A and B, one corresponding to the experimental, the other to the fitted property $P$ (Equation 6)

$$S_{AB} = \frac{2 \sum\limits_{\text{all points } i} P_i^A P_i^B}{\sum\limits_{\text{all points } i} (P_i^A)^2 + \sum\limits_{\text{all points } i} (P_i^B)^2} \tag{6}$$

A fundamental disadvantage of both the RMSd and Hodgkin-type indicator is that they do not consider the similarity in the region of low values appropriately, since the resemblance in regions of high $P$ values tends to dominate the final similarity index. We felt that in this particular application, a satisfactory similarity for the low-propensity areas is equally desirable as for the high-propensity ones. We therefore introduce a *Split Hodgkin Index* $S_{AB}^{\text{split}}$ (Equation 7)[3]

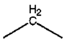
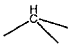
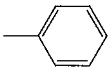
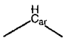
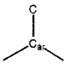
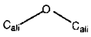
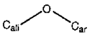
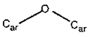
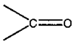
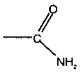
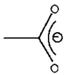
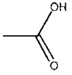
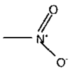
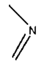
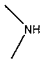
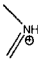
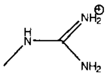
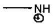
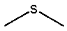
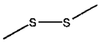
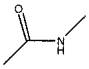
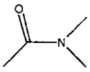
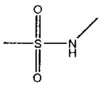
$$S_{AB}^{\text{split}} = \sqrt{S_{AB}^{\text{high}} \cdot S_{AB}^{\text{low}}} \tag{7}$$

In Equation 7, the Hodgkin sub-indices $S_{AB}^{\text{high}}$ and $S_{AB}^{\text{low}}$ are calculated each for a subset, rather than for the complete set of data points. The two subsets are obtained by splitting the total set into one containing all data points with $P > P_{\text{boundary}}$, and another with all points $P \leq P_{\text{boundary}}$. The Split Hodgkin Index reflects the goodness-of-fit in these two regions separately. A superior ranking of the quality of the fit is expected compared to the RMSd or plain Hodgkin index, since it takes both the high and low-propensity regions into account with equal weights.

In order to obtain two subsets, a reasonable level for the boundary $P_{\text{boundary}}$ must be defined. This can be done by taking $P_{\text{boundary}}$ as the average propensity of all non-zero grid points in the cluster. Thus, it falls between the minimum and maximum propensity val-

[3] The same concept can of course be applied to generate a Split Carbó Index. However, we decided to use the Hodgkin Index, as it usually performs better when comparing distributions of different intensities [18].

*Table 1.* List of the central fragments from IsoStar for which propensity plots have been generated

| Central groups[a] | | |
|---|---|---|
| 1 | methyl | $—CH_3$ |
| 2 | methylene |  |
| 3 | tertiary CH |  |
| 4 | phenyl |  |
| 5 | aromatic CH |  |
| 6 | aromatic CC |  |
| 7 | aliphatic-aliphatic ether |  |
| 8 | aliphatic-aromatic ether |  |
| 9 | aromatic-aromatic ether |  |
| 10 | carbonyl |  |
| 11 | carbamoyl |  |
| 12 | carboxylate |  |
| 13 | carboxylic acid (cis,trans) |  |
| 14 | nitro |  |
| 15 | aliphatic NH, | $C_{ali}—NH_2$ |
| 16 | aromatic NH, | $C_{ar}—NH_2$ |
| 17 | planar ring N |  |
| 18 | planar ring NH, uncharged |  |
| 19 | planar ring NH, charged |  |
| 20 | guanidinio |  |
| 21,22 | charged amino (flexible, rigid) |  |
| 23,24 | aliphatic OH (flexible, rigid) | $C_{ali}—OH$ |
| 25,26 | aromatic OH (flexible, rigid) | $C_{ar}—OH$ |
| 27 | thioether |  |
| 28 | thiol | —SH |
| 29 | disulfide |  |
| 30 | peptide |  |
| 31 | N,N-disubstituted peptide |  |
| 32 | sulfonamide, uncharged |  |
| 33 | water | $OH_2$ |

*Abbreviations: $C_{ar}$ = aromatic carbon atom ($sp^2$); $C_{ali}$ = aliphatic carbon atom ($sp^3$).

*Table 2.* The following contact groups from IsoStar have been used to generate propensity plots. The validation was performed using a methyl carbon, ammonium nitrogen, alcohol oxygen, and carbonyl oxygen probe

| | Contact groups |
|---|---|
| 1 | amino |
| 2 | ammonium |
| 3 | pyridin-type nitrogen ($CN^2_{sp}C$) |
| 4 | charged amino ($RNH_3^+$) |
| 5 | pyramidal amino ($R_3N$) |
| 6 | alcohol |
| 7 | water |
| 8 | carbonyl |
| 9 | aliphatic ether |
| 10 | aromatic ether |
| 11 | oxygen (general probe)[a] |
| 12 | carboxylate |
| 13 | aromatic CH |
| 14 | aliphatic CH |
| 15 | methyl |
| 16 | nitro |
| 17 | fluoro |
| 18 | chloro |
| 19 | cyano |
| 20 | sulfur[b] |
| 21 | chloride |
| 22 | iodide |

[a] Terminal, accepting oxygen atoms, e.g., from carbonyl groups, phosphates, etc. (corresponds to the 'Terminal oxygen' entry in IsoStar).

[b] All types of sulfur (corresponds to the 'Any sulfur' entry in IsoStar).

ues of the distribution, and its actual value depends on the data distribution itself.

To assure that the region of low propensity values contains points with a propensity below a certain level only, we introduce an additional parameter $P^{\mathrm{max}}_{\mathrm{boundary}}$. It places an upper limit on $P_{\mathrm{boundary}}$. In this way, a subset of low-propensity points can be selected, even in cases where the observed maximal propensity is very high.

In principle, another possibility to determine the goodness-of-fit would be to use a chi-square statistic to asses whether the experimental and fitted distributions differ significantly. This would eliminate a possible bias of the Hodgkin index towards the points of high propensity in the separate regions. In practice, however, this dependency was not observed to have a very large influence on the final result (given a reasonable

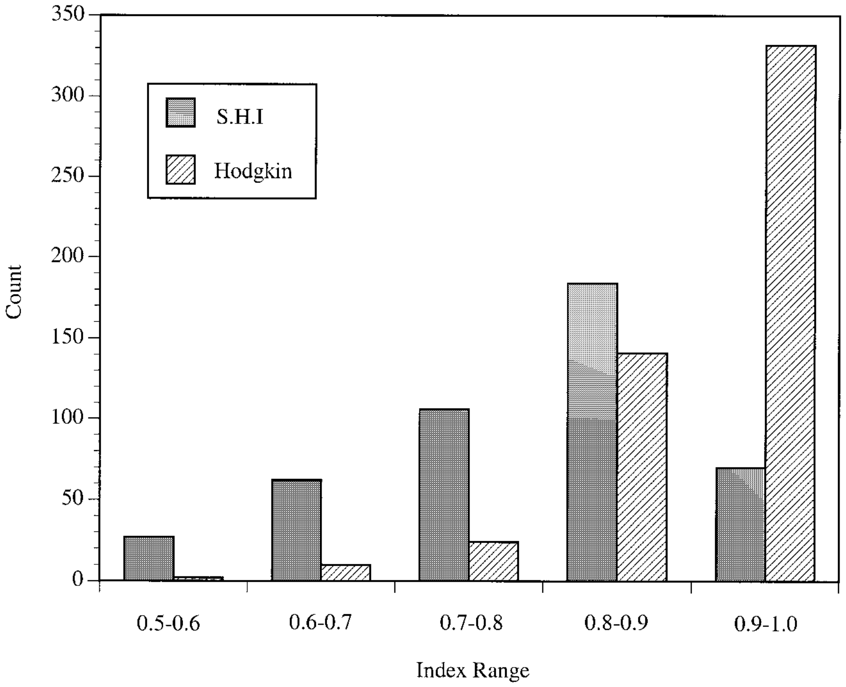


*Figure 3.* Distribution of the regular Hodgkin index and Split Hodgkin Index over the index range between 0.5 and 1.0 for a set of 583 fitted distributions. All distributions were fitted using the root mean squared deviation between experimental and fitted distributions as a residual, and subsequently evaluated using the above-mentioned indices. For both indices, a value of 1.0 indicates a perfect match between experimental and fitted data.

estimate of the boundary) and we have applied the Hodgkin indices for computational convenience.

*Validation procedure*

The predictive power and accuracy of the fitted functional forms is evaluated in a procedure that is analogous to the one applied to validate SuperStar. For this purpose, the SuperStar program was modified and the fitted Gaussian descriptors were used instead of the actual IsoStar propensity distributions for the generation of the maps.

For a detailed description of the validation procedure we refer to Verdonk et al. [14]. To summarise, during the validation procedure, propensity maps are generated for all receptor cavities of a set of 130 protein-ligand complexes using different probes. The ligands were removed from all complexes. In order to map hydrogen-bond donor sites, acceptor sites, and non-polar regions in the receptor cavities, a carbonyl oxygen, an alcohol oxygen, an ammonium nitrogen, and a methyl carbon were selected as probes. In the validation process, the percentage of cases is determined for which an atom type is suggested correctly by the SuperStar propensity. In order to determine the percentage of correct predictions of e.g. OH oxygen atoms in ligands, we count those cases where the propensity for an alcohol oxygen probe at the crystallographically determined position of an actual ligand OH group is higher than the propensities calculated for any of the other probes.

To assess the quality of the set of fitted descriptors, the results thus obtained are compared to those generated by the original SuperStar method. The validation procedure is applied to both a set of all matching ligand atoms, and a subset of the solvent-inaccessible atoms only (solvent accessibility less than 2%[4]). This subdivision is made as the solvent-accessible ligand atoms are likely to bind to bulk or surface water, interactions of which SuperStar has no knowledge. Solvent-inaccessible atoms partake in protein-ligand interactions only and are therefore expected to be predicted more reliably.

[4] Solvent accessibility is defined as the percentage of the surface of a ligand atom where a neighbouring water molecule would not overlap with protein or other parts of the ligand. This surface portion was approximated by a cube algorithm.

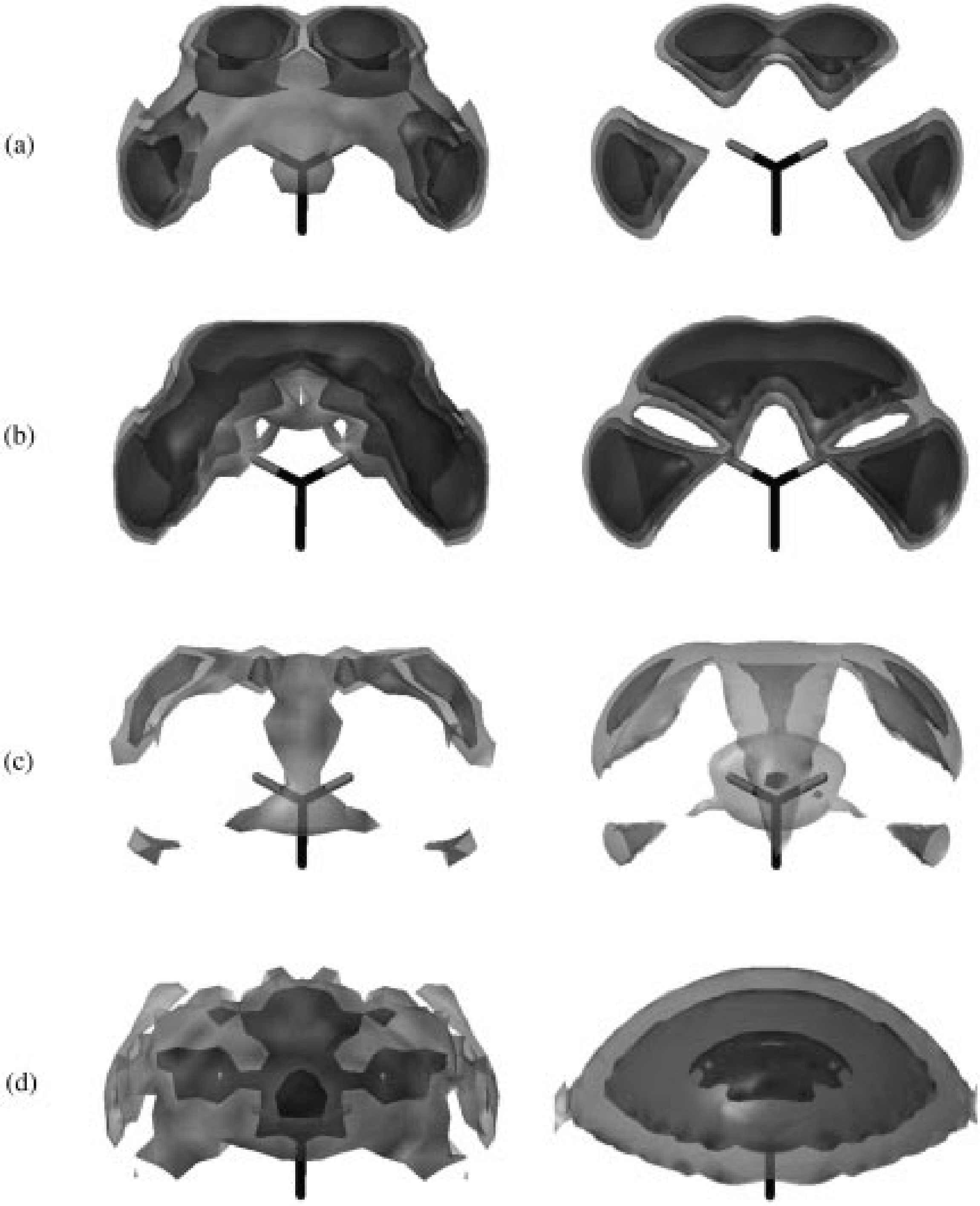


*Figure 4.* Original propensity distribution from IsoStar (left) and optimised Gaussian descriptors (right) for a carboxylate as central group. The descriptors for the distributions of the following probes are shown: (a) alcohol oxygen, (b) ammonium nitrogen, (c) carbonyl oxygen, and (d) methyl carbon. Pairs on the left: frontal view, pairs on the right: corresponding view along the C–C bond (view from top); contour levels (from light to dark gray) 1.0, 2.0, and 4.0. For further details, see text.

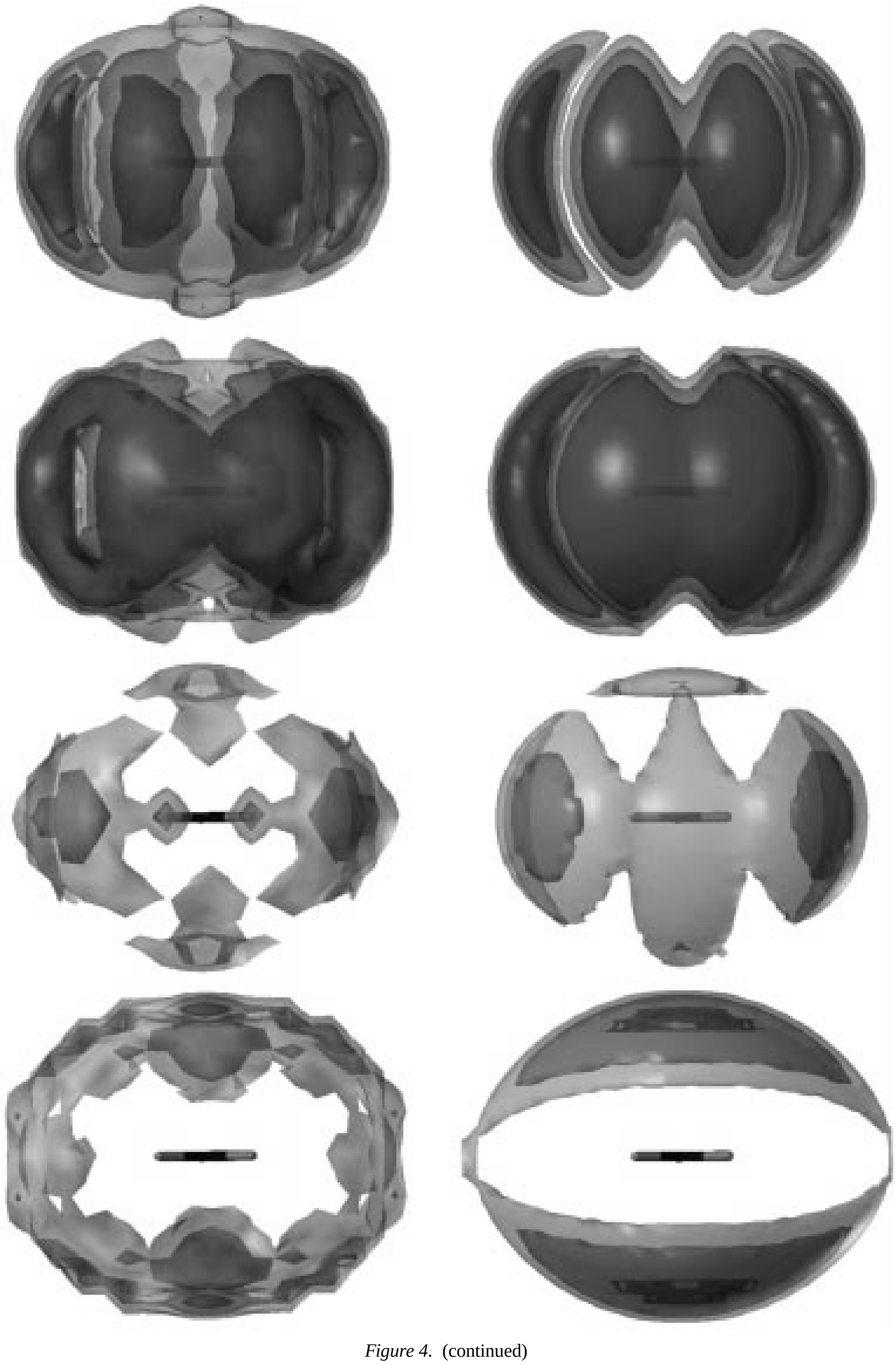


*Figure 4.* (continued)

## Results and discussion

### *Split Hodgkin Index*

As an input for the fitting algorithm, a subset of sufficiently populated scatterplots from IsoStar was selected, covering in total the combinations of 33 central and 22 contact groups. The same set is currently accounted for in SuperStar. In the validation procedure (*vide infra*), a subset was used, covering all central groups present in amino acids, and interacting with one of the above-mentioned probes. All groups are listed in Table 1 (central groups) and Table 2 (contact groups). The selection of the full set of probes is based on a statistical analysis of the similarities between all contact groups covered by IsoStar [19].

All suitable scatterplots were converted to propensity distributions and fitted using a simple RMSd residual. In order to determine the performance of the regular versus the Split Hodgkin Index (Equations 6 and 7) as a goodness-of-fit indicator, both similarity indices were subsequently calculated for the fit between the generated descriptors and the experimental distributions. A histogram for the distribution of the indices for this set of descriptors is shown in Figure 3 (indices $>0.5$).

The observation that both Hodgkin and Split Hodgkin indices yield high values of similarity in most cases suggests that the Gaussian descriptors (Equation 5) approximate the observed distributions satisfactorily. Figure 3 clearly shows that the regular Hodgkin index is not very discriminative in the high-similarity region, since most of the fits have been ranked with a Hodgkin similarity higher than 0.9 (1.0 indicating identity of the descriptions on this scale). The Split Hodgkin Index (S.H.I.), however, is a much more sensitive measure of similarity compared to the regular Hodgkin index, as a result of the explicit consideration of the low-propensity region. We decided therefore to use the S.H.I. also during the optimisation process, in order to derive a set of fit parameters that describes both the low- and the high-propensity region appropriately.

### *Fitting results*

IsoStar propensity plots were generated for the combinations of central and contact groups displayed in Tables 1 and 2. They were fitted in a fully automated way using the Split Hodgkin Index. In this section, some selected examples will be discussed, to give an impression of the results obtained.

Figure 4 displays four different propensity distributions around a carboxylate group together with the fitted descriptors. Part (a) shows the original and computed map for an alcohol-type oxygen atom as a probe. The distribution is fitted nicely by four ‘banana-shaped’ Gaussian functions (Equation 1), which describe the experimental propensity with a S.H.I of 0.95 (1.0 indicating identity of the descriptions on this scale). The distribution of $NH_3^+$ nitrogen atoms around carboxylate groups is also described well by four Gaussian functions (part (b)). In this case, the S.H.I. is 0.91. The distribution of carbonyl oxygen atoms as a probe is shown in part (c) and is fitted by 6 functions. Here, a more complex asymmetric arrangement of fitted shapes had to be generated; however, a S.H.I. of 0.88 indicates convincing similarity between original and fitted distribution.

The distribution of a methyl carbon around a carboxylate central group (Figure 4d) is rather fuzzy. The generated descriptor comprises two expanded functions, that fit the density around the central group with a S.H.I. of 0.84. Though this solution has the lowest index of the four cases described here, it still can be considered acceptable.

The distribution of carbonyl oxygen atoms around a charged carboxylic acid group (Figure 4c) highlights another application of our shape descriptors: they can be used to remove artefacts and undesired effects from the raw IsoStar data. In this distribution, part of the density in the plane of the carboxylate group is caused by an incorrect coding of carboxylic acid oxygen atoms in crystal structures as carbonyl oxygen atoms. Clusters caused by the latter are the result of protonated carboxylates forming acid dimers (e.g., the two small clusters ahead of the oxygens of the carboxylate). If these effects result in separate clusters in a distribution, the fitted representation can be easily modified to omit the corresponding Gaussian functions.

In Figure 5, the distribution of carbonyl oxygen atoms around a phenyl ring as central group is shown. This is a rather diffuse distribution with low propensity values featuring several overlapping maxima. The descriptor generated by the fitting algorithm consists of 5 Gaussian functions that describe the full experimental distribution with a similarity of 0.94.

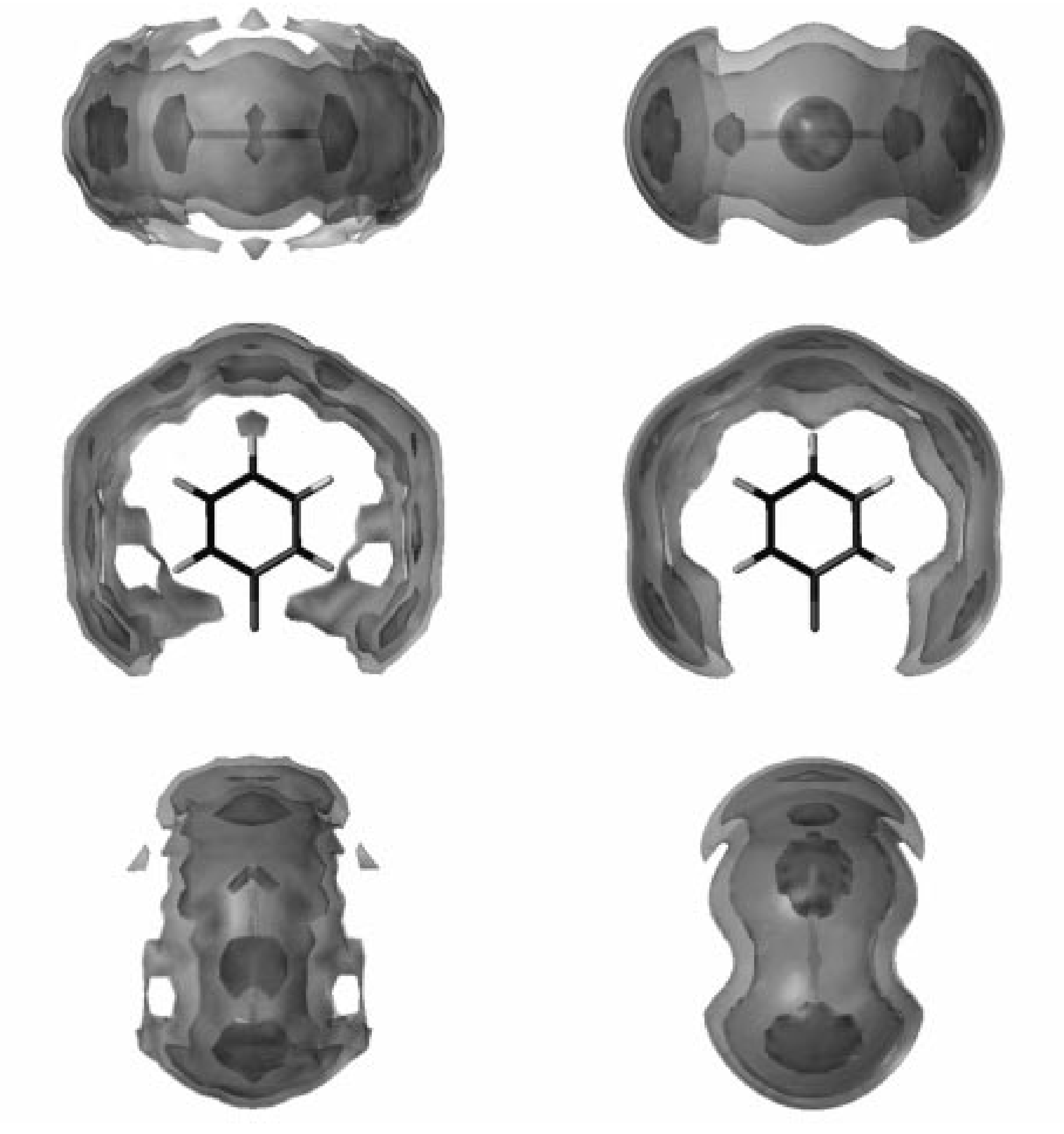


*Figure 5.* Example of optimised Gaussian descriptors attributed to a diffuse propensity distribution of carbonyl oxygen atoms around a phenyl group (left: Isostar data; right: fitted Gaussian descriptors; top: edge-on view along the upper CH bond of the phenyl ring, middle: view perpendicular to the ring plane, bottom: edge-on view from the side of the ring towards the center). In this case, five Gaussian functions were used as a descriptor, and the S.H.I. value amounts to 0.94. Contours were generated for propensity levels of 0.5, 1.0, and 2.0.

## *Validation*

### *Sets of descriptors optimised with the S.H.I.*

It was found that the incorporation of the descriptors in SuperStar yielded an improvement in calculational speed by a factor of five or more. A typical SuperStar calculation for a binding site of approximately 78 Å$^3$ (in a structure of Proteinase K, PDB entry 2prk, map for one probe) amounts to about 85 s. On application of the Gaussian descriptors, calculation times were reduced to approximately 14 s (calculations were performed on a Silicon Graphics O2 machine with a 225 MHz MIPS R10000 processor).

The validation results obtained for the different sets of Gaussian descriptors are displayed in Table 3, together with the data found for the original evaluation of SuperStar. Various similarity measures used in the optimisation of the fit are recorded.

The validation results are given in terms of two percentages indicating the success rates of prediction for all matching ligand atoms ($f_{\mathrm{correct}}^{\mathrm{all}}$) and for the solvent-inaccessible ligand atoms only ($f_{\mathrm{correct}}^{\mathrm{solv.-inacc.}}$). Furthermore, the values for $f_{P>1}$ and $f_{P'<P}$ are given, averaged over all ligands. $f_{P>1}$ lists the percentage of matching ligand atoms at positions that have propensity values larger than the expectation value of 1.0. $f_{P'<P}$ represents the percentage of the total volume of the distribution that covers propensity values smaller than the predicted propensity at a ligand atom position. $f_{P>1}$ indicates whether the descriptors assign propensity values larger than random to the ligand atom positions. $f_{P'<P}$ shows whether the matching ligand atom positions coincide with regions of high propensity (i.e. within or near ‘hot spots’ in the map); e.g., a value $f_{P'<P} = 75\%$ would indicate that, on average, the matching ligand atoms fall within the 25% of the volume of the distribution that accounts for the highest densities. All results in Table 3 have been averaged for the four probes.

It can be seen that the descriptors perform well (Table 3). Successful prediction rates $f_{\mathrm{correct}}$ for SuperStar are 67% and 82%, respectively, for all matching ligand atoms or the solvent-inaccessible ones. For the same data, the set of Gaussian descriptors derived using the S.H.I. yields 58% and 71%.

A comparison of the results for fits derived using the RMSd, Hodgkin index, and the Split Hodgkin Index demonstrates that descriptors optimised using the latter perform best (Table 3). The prediction rates and values for $f_{P'<P}$ are comparable, indicating that the extent and localisation of the high-density regions is described similarly. Yet, the higher value for $f_{P>1}$ obtained for the S.H.I. set (64% vs. 58% for all atoms, 70% vs. 67% for solvent-inaccessible ones, Table 3) shows that the S.H.I. allows for a better approximation of the low-propensity regions. Thereby, this set of descriptors approximates best the results observed for the original data.

High values are observed for $f_{P'<P}$. This indicates that ligand atoms frequently fall next to regions of high propensity, so-called ‘hot spots’ of interaction. The lower values obtained for $f_{P>1}$ for the Gaussian descriptors in comparison to the original data evaluated in SuperStar indicate that apparently, some information is lost during the fitting process. The Gaussian shapes represent a smoothed representation of the original data and supposedly this smoothing is responsible for the slightly lower prediction rate. Most likely, differences are observed in regions of low propensities. They account for less favourable interaction geometries, and accordingly, they are predicted less reliably. Considering the resolution and accuracy to which most protein structures are solved, this observation is not unexpected.

A beneficial influence, however, of the above-mentioned smoothing effect is its influence on the $f_{P'<P}$ values. They are higher for our descriptors than for SuperStar. The value of $f_{P'<P}^{\mathrm{all}} = 69\%$ and $f_{P'<P}^{\mathrm{solv.-inacc.}} = 79\%$ for SuperStar, indicates that the matching ligand atoms, on average, coincide with regions that account for 31% or 21% of the highest density in the map. The values obtained with the descriptor set are 75% and 83%, respectively. They are larger than the ones observed for SuperStar, and indicate that the ligand atoms are on average found in the densest 25% and 17% of the maps. Apparently, the descriptors highlight better the favourable positions, presumably as a result of focusing on the regions of high propensity.

Extended validation data for SuperStar and the sets of descriptors are shown in Table 4. This table lists the frequency that a matching ligand atom is predicted correctly. The successful predictions can be found on the diagonal of each 4×4 sub-table. The prediction rates $f_{\mathrm{sim}}$, summarising correct predictions of chemically similar probes[5], are almost equal to the success

[5] To calculate the $f_{\mathrm{sim}}$ values, a prediction is considered to be successful if a probe with similar hydrogen-bonding properties is predicted. This means that OH is also considered a successful prediction for an $NH_3^+$ group, and vice versa; C=O is allowed to be predicted by OH, and vice versa, assuming the hydroxyl acts as an acceptor.

*Table 3.* Results for the validation of sets of descriptors optimised with respect to different residuals. The best overall unmodified parametrisation is indicated in bold

| Calculation | SHI settings | All ligand atoms[a] | | | Solvent-inaccessible ligand atoms[bc] | | |
|---|---|---|---|---|---|---|---|
| | $P^{\mathrm{max}}_{\mathrm{boundary}}$ | $f^{\mathrm{all}}_{\mathrm{correct}}$ (%) | $f_{P>1}$ (%) | $f_{P'<P}$ (%) | $f^{\mathrm{solv.in}}_{\mathrm{correct}}$ (%) | $f_{P>1}$ (%) | $f_{P'<P}$ (%) |
| Nonmodified sets of descriptors | | | | | | | |
| SuperStar | – | *67* | 78 | 69 | *82* | 89 | 79 |
| RMSd | – | *55* | 58 | 75 | *65* | 66 | 81 |
| Hodgkin | – | *57* | 58 | 74 | *70* | 67 | 81 |
| SHI | $\infty$ | *58* | 64 | 75 | *71* | 70 | 83 |
| ***SHI*** | **1** | ***59*** | **62** | **75** | ***74*** | **70** | **82** |
| *SHI* | 3 | *59* | 61 | 75 | *70* | 66 | 81 |
| *SHI* | 10 | *59* | 62 | 75 | *71* | 68 | 82 |
| Modified sets of descriptors ($P^{\mathrm{max}}_{\mathrm{boundary}} = \infty$) | | | | | | | |
| | $x_\sigma$ | $f^{\mathrm{all}}_{\mathrm{correct}}$ (%) | $f_{P>1}$ (%) | $f_{P'<P}$ (%) | $f^{\mathrm{solv.in}}_{\mathrm{correct}}$ (%) | $f_{P>1}$ (%) | $f_{P'<P}$ (%) |
| *SHI* | 1.25 | *60* | 78 | 75 | *69* | 88 | 85 |
| *SHI* | 1.50 | *58* | 82 | 75 | *66* | 93 | 85 |

[a] This set contains 457 matching ligand atoms in a total of 130 protein-ligand complexes.

[b] This subset contains 136 solvent-inaccessible matching ligand atoms with solvent inaccessibilities in the range of 0.0–0.02.

[c] Solvent accessibility is defined as the percentage of grid points around a ligand atom where a water molecule would not overlap with protein or ligand. Solvent-inaccessible atoms in the set as used here have accessibilities up to 2%.

rates obtained in the case of SuperStar; values of 75% and 89% are found for the best descriptor set versus 81% and 91% for SuperStar (for all ligand atoms and solvent-inaccessible ones, respectively). The fact that the $f_{\mathrm{sim}}$ values are closer to the corresponding SuperStar values than the $f_{\mathrm{correct}}$ ones points out that the approximative nature of the Gaussian functions has an additional effect: the calculated propensities for one central group can be slightly off in the centre regions of a Gaussian descriptor since the experimental and Gaussian profiles do not match exactly. The error due to the misfit will propagate since SuperStar combines several maps of different central groups in a multiplicative fashion. Propensities in the final map therefore may differ from the ones that would be predicted by SuperStar.

Atom-type predictions for the $NH_3^+$ probe are apparently less satisfactory. A possible explanation for this observation might be that the fitting algorithm is less efficient due to the scarce data in IsoStar for the $NH_3^+$ probe. Since we suppose that the low-density regions of our fits are represented relatively inaccurately due to the inherent smoothing of the Gaussian descriptors, we investigated whether the introduction of a low $P^{\mathrm{max}}_{\mathrm{boundary}}$ would improve the overall results. Table 3 shows that the introduction of a maximum boundary level indeed improves the prediction rate of atom types whereas the other properties ($f_{P>1}, f_{P'<P}$) are not or only slightly modified. The best correct prediction rates increase to 59% and 74%, for the full set and the solvent-inaccessible set, respectively. Detailed results are displayed in Table 4 and show that especially the $NH_3^+$ and $CH_3$ probes benefit from the application of an upper limit ($P^{\mathrm{max}}_{\mathrm{boundary}}$) during the subset-determining stage of the Split Hodgkin Index calculation.

### *Modified descriptors*

Usually, crystal structures of proteins are not determined to atomic resolution. The average error in atomic positions depends on the diffracting power of the crystals. As a rule of thumb, the positional error amounts to about 1/6 of the resolution. This means that a structure resolved to 2.4 Å is affected by positional errors of 0.4 Å. The accuracy of the data in the CSD, usually resolved better than 0.8 Å, is much higher.

*Table 4.* Atom type predictions (counts) for matching ligand atoms. Results are shown for SuperStar, and for two sets of Gaussian fits optimised using the RMSd, Hodgkin, and best S.H.I. residuals

| Probe | All matching ligand atoms<br>Predicted probe | | | | $f_{\text{correct}}$<br>(%) | $f_{\text{sim}}$<br>(%) | Solvent-inaccessible ligand atoms[a]<br>Predicted probe | | | | $f_{\text{correct}}$<br>(%) | $f_{\text{sim}}$<br>(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C=O | OH | $NH_3^+$ | $CH_3$ | | | C=O | OH | $NH_3^+$ | $CH_3$ | | |
| SuperStar | | | | | | | | | | | | |
| C=O | **144** | 18 | 1 | 40 | *71* | *80* | **56** | 1 | 0 | 5 | *90* | *92* |
| OH | 29 | **37** | 13 | 27 | *35* | *75* | 7 | **20** | 3 | 4 | *59* | *88* |
| $NH_3^+$ | 0 | 4 | **18** | 0 | *82* | *100* | 0 | 2 | **4** | 0 | *67* | *100* |
| $CH_3$ | 12 | 2 | 4 | **108** | *86* | *86* | 2 | 1 | 0 | **31** | *91* | *91* |
| Overall | | | | | *67* | *81* | | | | | *82* | *91* |
| SuperStar applying Gaussian descriptors optimised with the RMSd residual | | | | | | | | | | | | |
| C=O | **119** | 41 | 6 | 37 | *59* | *79* | **44** | 13 | 1 | 4 | *71* | *92* |
| OH | 30 | **37** | 9 | 30 | *35* | *72* | 11 | **17** | 3 | 3 | *50* | *91* |
| $NH_3^+$ | 1 | 6 | **11** | 4 | *50* | *77* | 0 | 3 | **3** | 0 | *50* | *100* |
| $CH_3$ | 34 | 8 | 1 | **83** | *66* | *66* | 5 | 4 | 1 | **24** | *71* | *71* |
| Overall | | | | | *55* | *74* | | | | | *65* | *87* |
| SuperStar applying Gaussian descriptors optimised with the regular Hodgkin residual | | | | | | | | | | | | |
| C=O | **133** | 32 | 2 | 36 | *66* | *81* | **46** | 11 | 1 | 4 | *62* | *92* |
| OH | 29 | **44** | 4 | 29 | *42* | *73* | 9 | **21** | 2 | 2 | *74* | *94* |
| $NH_3^+$ | 0 | 9 | **7** | 6 | *32* | *73* | 0 | 4 | **2** | 0 | *33* | *100* |
| $CH_3$ | 34 | 12 | 2 | **78** | *62* | *62* | 3 | 5 | 0 | **26** | *77* | *77* |
| Overall | | | | | *57* | *74* | | | | | *70* | *89* |
| SuperStar applying Gaussian descriptors optimised with the S.H.I. ($P^{\text{max}}_{\text{boundary}} = \infty$) | | | | | | | | | | | | |
| C=O | **127** | 35 | 1 | 40 | *63* | *80* | **44** | 14 | 0 | 4 | *71* | *94* |
| OH | 23 | **48** | 4 | 31 | *45* | *71* | 5 | **25** | 1 | 3 | *74* | *91* |
| $NH_3^+$ | 1 | 13 | **5** | 3 | *23* | *82* | 0 | 5 | **1** | 0 | *17* | *100* |
| $CH_3$ | 33 | 6 | 0 | **87** | *69* | *69* | 5 | 3 | 0 | **26** | *77* | *77* |
| Overall | | | | | *58* | *75* | | | | | *71* | *89* |
| SuperStar applying Gaussian descriptors optimised with the S.H.I. ($P^{\text{max}}_{\text{boundary}} = 1.0$) | | | | | | | | | | | | |
| C=O | **128** | 31 | 1 | 42 | *63* | *79* | **46** | 12 | 0 | 4 | *74* | *94* |
| OH | 26 | **48** | 4 | 27 | *46* | *74* | 5 | **25** | 1 | 3 | *74* | *91* |
| $NH_3^+$ | 1 | 11 | **7** | 3 | *32* | *82* | 0 | 3 | **3** | 0 | *50* | *100* |
| $CH_3$ | 34 | 7 | 0 | **85** | *68* | *68* | 6 | 2 | 0 | **26** | *77* | *77* |
| Overall | | | | | *59* | *75* | | | | | *74* | *89* |

[a] Solvent accessibility of atoms in this set ranges from 0 to 2%. For a definition, see Table 3.

In order to assess the influence of this positional uncertainty on our predictions, we investigated the effect of ‘smearing’ the representation of our descriptors. A ‘blurring’ effect is achieved easily by applying a spatial expansion to the fitted descriptors via their coefficients $c_n$, i.e. (Equation 10)

$$c_n' = \frac{1}{2(x_\sigma \sigma_n)^2} \tag{8}$$

Replacing the actual coefficients by $c_n'$ extends the widths $\sigma_n$ of the distribution by a factor $x_\sigma$. Accordingly, the propensity levels are smeared out, while approximately maintaining the actual position and maximum intensity of the shapes.[6]

Increasing the descriptor widths by 25% ($x_\sigma = 1.25$) does not significantly alter the values for $f_{\text{correct}}$ and $f_{P'<P}$ compared to the unmodified descriptors (see Table 3). The values for $f_{P>1}$ are improved, from 64% to 78% for the full set of matching ligand atoms, and from 70% to 88% for the solvent-inaccessible

[6] As a result of the overlap at the tails of the Gaussian shapes, the positions and heights of the maxima may shift slightly for $x_\sigma$ values larger than 1. This effect is small as long as $(x_\sigma - 1)$ is small.

*Table 5.* Atom type predictions (counts) for matching ligand atoms. Results are shown for modified sets of descriptors. See text for details. An upper limit $P^{\max}_{\text{boundary}}$ was not applied

| Probe | All matching ligand atoms<br>Predicted probe | | | | $f_{\text{correct}}$<br>(%) | $f_{\text{sim}}$<br>(%) | Solvent-inaccessible ligand atoms[a]<br>Predicted probe | | | | $f_{\text{correct}}$<br>(%) | $f_{\text{sim}}$<br>(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C=O | OH | $NH_3^+$ | $CH_3$ | | | C=O | OH | $NH_3^+$ | $CH_3$ | | |
| SuperStar applying Gaussian descriptors optimised with the S.H.I. , $x_\sigma = 1.25$ | | | | | | | | | | | | |
| C=O | **132** | 27 | 1 | 43 | *65* | *78* | **46** | 10 | 0 | 6 | *74* | *90* |
| OH | 27 | **44** | 3 | 32 | *42* | *70* | 8 | **21** | 1 | 4 | *62* | *88* |
| $NH_3^+$ | 1 | 12 | **5** | 4 | *23* | *82* | 0 | 5 | **1** | 0 | *17* | *100* |
| $CH_3$ | 28 | 7 | 0 | **91** | *72* | *72* | 4 | 4 | 0 | **26** | *77* | *77* |
| Overall | | | | | *60* | *75* | | | | | *69* | *87* |
| SuperStar applying Gaussian descriptors optimised with the S.H.I. , $x_\sigma = 1.5$ | | | | | | | | | | | | |
| C=O | **135** | 23 | 0 | 45 | *67* | *78* | **48** | 8 | 0 | 6 | *77* | *90* |
| OH | 30 | **33** | 3 | 35 | *33* | *65* | 11 | **16** | 0 | 5 | *50* | *84* |
| $NH_3^+$ | 0 | 14 | **3** | 5 | *14* | *77* | 0 | 6 | **0** | 0 | *0* | *100* |
| $CH_3$ | 28 | 8 | 0 | **90** | *71* | *71* | 5 | 5 | 0 | **24** | *71* | *71* |
| Overall | | | | | *58* | *73* | | | | | *66* | *84* |

[a] Solvent accessibility of atoms in this set ranges from 0 to 2%. For a definition, see Table 3.

set. Similar trends are observed for the case where $x_\sigma = 1.50$ is applied, with slightly lower rates of prediction than for $x_\sigma = 1.25$. Further results are shown in Table 5.

It is rather surprising that $f_{P'<P}$ is still high (75% and 85%, for all matching ligand atoms, and the solvent-inaccessible ones, respectively) for $x_\sigma = 1.25$. In other words, the matching ligand atoms are still within the highest density regions of the map. On the whole, an increase of the descriptor widths seems to be advantageous, especially when trying to predict the atom types disregarding their solvent accessibility.

## Conclusions

We have developed an algorithm to transform crystal-field environments into a set of optimally placed density functions of a Gaussian-type functional form. These descriptors reflect the chance to find an interacting group near a given central group. The fields used were derived from the IsoStar database, and are expressed in terms of propensity distributions of non-bonded contact geometries. The Gaussian-type functional form is parameterised by fitting the original propensity distributions using a similarity index that is tailored to the problem. This index, a Split Hodgkin Index, reflects better the existing similarity in the low-propensity regions of a distribution than a simple RMSd or Hodgkin residual.

The descriptors were validated by assessing their capability to predict favourable interaction sites for different ligand atom types. For this purpose, propensity maps were generated for four probes (OH oxygen, CO oxygen, $NH_3^+$ nitrogen and methyl carbon atoms) in 130 protein binding sites following the same protocol as used in SuperStar. Although the results obtained by use of the original SuperStar application are better than those obtained by the Gaussian descriptors, the loss of prediction rate when using the latter is very acceptable in view of their much lower computational demands.

The descriptor representation can be easily modified. This advantage over the usage of original data might help to filter undesirable bias and background noise due to false data assignment in some of the IsoStar propensity plots. Apart from this, the functional descriptors can easily be adapted to positional uncertainty in the protein data while modeling protein cavities.

A major advantage of using descriptors instead of the original data is the speed of calculation: their application in SuperStar yields an improvement in computing times by a factor of five to eight. This facilitates the application of SuperStar-like fields in a more large-scale analysis of ligands. One disadvantage, however, might be that the fit algorithm sometimes does not find an optimal descriptor, in particular for sparsely populated entries in IsoStar.

As in all knowledge-based approaches, a major drawback of the present method is that its success depends heavily on the presence of information. Nevertheless, a descriptor-based approach can be extended easily to new types if similarities are present to groups that are already parameterised using sufficient data. Since the IsoStar database is still expanding, data will become available for an increasing number of central and contact groups.

## Applications

We intend to further improve and apply the developed descriptors. An obvious application is their implementation into a docking program, to guide the placement of the ligand into its receptor site, or, as a part of a scoring function, to better rank the predicted ligand binding modes. In principle, propensity values can be converted to energy terms (see, e.g., Sippl et al. [20]). In that case, the propensity of finding a certain type of ligand atom at the positions of the respective ligand atoms when docked into a binding site may be used to derive an energy of interaction for that particular binding mode.

Similar to the propensity maps produced by SuperStar, fields generated by our descriptors can be used in QSAR (CoMFA, CoMSIA) or crystal structure predictions. Preliminary QSAR studies based on SuperStar fields have revealed better correlations than usual Lennard-Jones or Coulomb potentials [21]. As the Gaussian descriptors are more versatile and easier to modify than the original IsoStar data, their application in comparative molecular field analysis appears advantageous.

Although less complete, IsoStar also contains data for interactions compiled from the PDB. The conversion of these data would result in descriptors supposedly better suited to predict protein-ligand interactions, since they will implicitly account for binding motifs that are typical for proteins. A major limitation is imposed by the scarce data, the significantly reduced accuracy and the missing information on protons.

Protein flexibility is currently not taken into account during the mapping of a protein binding site. The fast mapping of different frames from a molecular-dynamics run would make it possible to construct a composite view of a binding site that enables one to investigate this aspect.

The directionality of the interactions toward a central group is not yet evaluated in our approach. We only consider the position of an interacting atom rather than exploiting the actual orientation of an interacting group with respect to the central group. Interactions do possess directionality (see, e.g., Figure 1) and contact groups do feature preferred orientations. Indeed, if we consider the descriptors for, e.g., hydroxyl groups around a carboxylate, we observe that both hydrogen and oxygen propensity distributions are represented by four Gaussian functions. Their spatial orientation is clearly related to the lone pair directions at the carboxyl oxygens, indicating that the OH group points towards the carboxylate oxygen.

Finally, the descriptors can provide a strategy to compare similarity in terms of putative molecular interaction properties during a fast screening of large compound databases.

## Acknowledgements

This work was supported by the German Federal Ministry for Education, Science, Research and Technology (BMBF, RELIMO Project, Grant no. 0311619). We gratefully acknowledge an anonymous referee for valuable comments on the approach that we applied to obtain a measure of the goodness-of-fit.

## References

1. Davis, A.M. and Teague, S.J., Angew. Chem. Int. Ed. Engl., 38 (1999) 736.
2. Böhm, H.-J. and Klebe, G., Angew. Chem. Int. Ed. Engl., 35 (1996) 2588.
3. Klebe, G., J. Mol. Biol., 237 (1994) 212.
4. a. Mills, J.E., Perkins, T.D. and Dean, P.M., J. Comput.-Aided Mol. Design, 11 (1997) 229.
   b. Mills, J.E. and Dean, P.M., J. Comput.-Aided Mol. Design, 10 (1996) 607.
5. Laskowski, R.A., Thornton, J.M., Humblet, C. and Singh, J., J. Mol. Biol., 259 (1996) 175.
6. a. Danziger, D.J. and Dean, P.M., Proc. Roy. Soc. Ser. B., 236 (1989) 101.
   b. Danziger, D.J. and Dean, P.M., Proc. Roy. Soc. Ser. B., 236 (1989) 115.
7. Humblett, C. and Dunbar, J.B., Jr, In Venuti, M.C. (Ed.), Annual Reports in Medicinal Chemistry, Vol. 28, Academic Press, London, 1993, p. 275.
8. Klebe, G., In Kubinyi, H. (Ed.), 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM, Leiden, 1993, p. 173.
9. Bures, M.G., In Charifson, P.S. (Ed.), Practical Application of Computer-Aided Drug Design, Marcel Dekker, New York, NY, 1997, p. 39.
10. Klebe, G., Persp. Drug Discov. Design, 12–14 (1998) 87.
11. a. Bruno, I.J., Cole, J.C., Lommerse, J.P., Rowland, R.S., Taylor, R. and Verdonk, M.L., J. Comput.-Aided Mol. Design, 11

(1997) 525.

b. IsoStar Database, Version 1.1, Cambridge Crystallographic Data Centre, Cambridge, 1998.

12. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31 (1991) 187.
13. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
14. Verdonk, M.L., Cole, J.C. and Taylor, R., J. Mol. Biol., 289 (1999) 1093.
15. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
16. Nelder, J.A. and Mead, R., Comput. J., 7 (1965) 308.
17. Willet, P., Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 38 (1998) 983.
18. Hodgkin, E.E. and Richards, W.G., Int. J. Quant. Chem. Quant. Biol. Symp., 14 (1987) 105.
19. Verdonk, M.L., manuscript in preparation (2000).
20. Sippl, M.J., Curr. Opin. Struct. Biol., 5 (1995) 229.
21. Boehm, M. and Klebe, G., unpublished results (1999).