Journal of Computer-Aided Molecular Design **17:** 435–461, 2003.

A Bayesian molecular interaction library

Ville-Veikko Rantanen[a,b,*], Mats Gyllenberg[a], Timo Koski[c] & Mark S. Johnson[b]
[a]*Department of Mathematics, University of Turku, FIN-20014 Turku, Finland;* [b]*Department of Biochemistry and Pharmacy, Åbo Akademi University, PO Box 66, FIN-20521 Turku, Finland;* [c]*Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden*

Received 13 February 2003; accepted 17 July 2003

Key words: combining posterior probabilities, expectation-maximization algorithm, *maximum a posteriori* estimates, mixture model, protein-ligand recognition

Summary

We describe a library of molecular fragments designed to model and predict non-bonded interactions between atoms. We apply the Bayesian approach, whereby prior knowledge and uncertainty of the mathematical model are incorporated into the estimated model and its parameters. The molecular interaction data are strengthened by narrowing the atom classification to 14 atom types, focusing on independent molecular contacts that lie within a short cutoff distance, and symmetrizing the interaction data for the molecular fragments. Furthermore, the location of atoms in contact with a molecular fragment are modeled by Gaussian mixture densities whose *maximum a posteriori* estimates are obtained by applying a version of the expectation-maximization algorithm that incorporates *hyperparameters* for the components of the Gaussian mixtures. A routine is introduced providing the hyperparameters and the initial values of the parameters of the Gaussian mixture densities. A model selection criterion, based on the concept of a 'minimum message length' is used to automatically select the optimal complexity of a mixture model and the most suitable orientation of a reference frame for a fragment in a coordinate system. The type of atom interacting with a molecular fragment is predicted by values of the posterior probability function and the accuracy of these predictions is evaluated by comparing the predicted atom type with the actual atom type seen in crystal structures. The fact that an atom will simultaneously interact with several molecular fragments forming a cohesive network of interactions is exploited by introducing two strategies that combine the predictions of atom types given by multiple fragments. The accuracy of these combined predictions is compared with those based on an individual fragment. Exhaustive validation analyses and qualitative examples (e.g., the ligand-binding domain of glutamate receptors) demonstrate that these improvements lead to effective modeling and prediction of molecular interactions.

Introduction

Proteins function through the specific recognition and binding of other molecules, their ligands. Consequently, understanding protein-ligand interactions is a crucial step in understanding the molecular roles that proteins play. The construction of models that predict protein-ligand interactions from experimental data (including ligand binding studies and results from site-directed mutagenesis) is therefore of importance, as such models may suggest new experiments and have obvious application to knowledge-based drug discovery. Hence, a large variety of computer programs have been developed to exploit structural information in the prediction of protein-ligand binding. These programs include, *inter alia*, GRID [1–3], HINT [4], HS-ITE [5, 6], X-SITE [7], AQUARIUS [8], LUDI [9–11] and IsoStar [12] along with the related works [13–16]. Most often, such programs are based on the analysis of molecular interactions observed in the crystal structures of small molecules (e.g., Klebe [17]) and of protein-ligand complexes [18–23].

*To whom correspondence should be addressed, E-mail: vira@utu.fi; vrantane@abo.fi

Recently, we have accumulated statistics on protein-ligand and protein-protein non-bonded atom interactions, obtained from the Protein Data Bank (PDB) [24], and used standard (or frequentist) concepts of probability calculus to model the spatial locations of atoms surrounding molecular fragments (Rantanen et al. [25]). The estimation methods presented in Rantanen et al. [25] rely on a view of statistical modeling that ignores the uncertainty in the estimated model and its parameters.

In the present work, we incorporate the uncertainty of the model in the prediction of molecular interactions. Prior knowledge, experience and prejudices play a part in the estimation procedure according to coherent rules of the Bayesian theory of probability (Bernardo and Smith [26], pp. 13–104).

Here, the archived molecular interaction data is strengthened considerably by narrowing the original atom classification from 25 to 14 atom types, by excluding non-independent molecular interactions and instead focusing on independent molecular contacts that lie within a short cutoff distance, and by symmetrizing the interaction data for the molecular fragments.

Moreover, we model molecular interaction data by introducing *a priori* assumptions to the learning process leading to the Gaussian mixture densities that describe the location of atoms in contact with a molecular fragment. We apply a version of the expectation-maximization (EM) algorithm [27, 28], incorporating *hyperparameters* for the components of the Gaussian mixtures in order to obtain the *maximum a posteriori* (MAP; e.g., Durbin et al. [29]) estimates of the Gaussian mixture densities. A routine is introduced that provides the hyperparameters and initializes the parameters of the Gaussian mixture densities. A model selection criterion, based on the concept of a 'minimum message length' (MML; see, e.g., Lanterman [30] for a review) and tailored for the requirements of the Gaussian mixture models, is used to automatically choose the optimal complexity of a mixture model and the most suitable orientation of a reference frame for a fragment in a coordinate system.

Values of the posterior probability function (e.g., Ref. [29]) are used to predict the type of atom interacting with a molecular fragment. The accuracy of these predictions is evaluated by comparing the predicted atom type with the actual atom type seen in crystal structures. We have exploited the fact that an atom will simultaneously interact with several molecular fragments forming a cohesive network of interactions. In this study, two strategies are introduced to combine the predictions of atom types given by multiple fragments and the accuracy of these combined predictions is compared with those based on an individual fragment.

The methodology, applied in the case of the ligand-binding domain of the glutamate receptor GluR2, can be used to characterize the structure of a ligand bound to a target protein, as well as the water molecules that interact with both the protein and the ligand.

Theory

Modeling of molecular interactions

In this paper, non-bonded interactions between atoms found at interfaces of protein-ligand and protein-protein complexes are modeled. For each complex, the atoms of the first molecule (i.e., the protein) are categorized into different classes, while the second molecule (i.e., the ligand or another protein) is divided into molecular fragments such that each fragment contains a *main atom* and at least two other atoms. (We refer to Rantanen et al. [25] (Figure 1) for a description of the topology of the molecular fragments.) Partitioning of the second molecule into fragments is needed in order to capture the exact spatial positions of atoms across the interface surrounding the main atom of a molecular fragment. A definition for independent atom contacts is introduced and only atoms participating in independent contacts are considered in this analysis. Each molecular fragment is placed on a coordinate system so that spherical coordinates can be used to locate the contact atoms that surround the molecular fragment; and the most suitable orientation of the molecular fragment is chosen automatically. Distributions of contact atoms are modeled with mixtures of Gaussian distributions using *a priori* information on the data.

Atom classification

The analysis of Rantanen et al. [25] resulted in a strong indication that the classification of protein atoms into 24 classes (Li and Nussinov [31]) is too large for our purposes and that a reduction in the number of classes would lead to improved predictions. Thus, Rantanen et al. [32] studied similarities amongst these 24 atom classes. Taken together with the visual inspection and availability of the interaction data, the results were used to reduce the atom classification to 14 atom classes denoted by C_k (Table 1). This atom

Table 1. Contact atom classes.

Class	Atom	Definition
1	>***C***H-	Side-chain tertiary carbons of isoleucine, leucine, threonine and valine (Ile, Leu, Thr and Val) and main-chain α-carbons (except the α-carbon of glycine (Gly))
2	-***C***H_2-	Side-chain methylene (secondary) carbons (Cγ of Arg, Ile, Met, Lys and Pro; Cδ of Lys and Pro; Cβ of Arg, Asn, Asp, Cys, Gln, Glu, His, Leu, Lys, Met, Phe, Pro, Ser, Trp and Tyr; Cδ of Arg, Cγ of Glu and Cϵ of Lys) and main-chain α-carbon of glycine (Gly)
3	-***C***H_3	Side-chain methyl (primary) carbons (Ala, Ile, Leu, Thr, Val and Met)
4	>$\boldsymbol{C}_{ar}$H, >$\boldsymbol{C}_{ar}$R	Aromatic carbons with hydrogen (carbon atoms on the rings of Phe, Trp and Tyr), substituted aromatic carbons (Cγ of Phe, Cγ and Cϵ2 of Trp, Cγ Tyr) and side-chain imidazole carbons (Cδ, Cϵ and Cγ) of histidine (His)
5	-***C***$R_1(R_2)$	Side-chain carbonyl carbon of asparagine and glutamine (Asn and Gln), side-chain carboxylate carbon of aspartic and glutamic acid (Asp and Glu), guanido carbon of arginine (Arg) and main-chain carbonyl carbons
6	-***S***H	Side-chain thiol sulphur of cysteine (Cys)
7	-***S***-	Side-chain sulfide sulphur of methionine (Met)
8	>***N***H, -***N***H_2	Main-chain amide nitrogens and side-chain amide nitrogens of asparagine and glutamine (Asn and Gln)
9	>***N***H, -***N***H_2	Side-chain indole nitrogen of tryptophan (Trp), side-chain imidazole nitrogens of histidine (His) and side-chain guanido nitrogens of arginine (Arg)
10	-***N***H_3^+	Side-chain amino nitrogen of lysine (Lys)
11	-C***O***(NH_2)	Side-chain carbonyl oxygens of asparagine and glutamine (Asn and Gln)
12	-C***OO***H, >C=***O***	Side-chain carboxylate oxygens of aspartic and glutamic acid (Asp and Glu) and main-chain carbonyl oxygens
13	-***O***H	Side-chain hydroxyl oxygens of serine, threonine and tyrosine (Ser, Thr and Tyr)
14	H_2***O***	Water oxygen

classification takes into account the classical, chemical classification of atom types while also including environment-dependent features of the molecular interactions.

Carbon atoms are classified into five groups. The first three classes (1, 2 and 3; Table 1) differ according to the number of bound hydrogen atoms (>***C***H- vs. >***C***H_2 vs. -***C***H_3), whereas classes 4 and 5 represent different covalent bonding environments (>$\boldsymbol{C}_{ar}$H, >$\boldsymbol{C}_{ar}$R (ar=aromatic) vs. -***C***$R_1(R_2)$). A large amount of data exists for both methyl carbon (class 3; -***C***H_3) and aromatic carbon (class 4; >$\boldsymbol{C}_{ar}$H, >$\boldsymbol{C}_{ar}$R) atoms and, therefore, these classes were not joined together. The two sulphur atom classes, thiols (class 6; -***S***H) and sulfides (class 7; -***S***-), are considered separately. Nitrogen atoms are represented by three classes: amide nitrogen atoms (class 8; >***N***H, -***N***H_2), resonating or planar nitrogen atoms (class 9; >***N***H, -***N***H_2) and the amino nitrogen atom group (class 10; -***N***H_3^+). Oxygen atoms are divided into four classes based on their bonding environments: double-bonded carbonyl oxygen atoms (class 11; -C***O***(NH_2)), partially double-bonded oxygen atoms (class 12; -C***OO***H, main-chain >C=***O***), single-bonded hydroxyl oxygen atoms (class 13; -***O***H) and water oxygen atoms (class 14; H_2***O***). The atom classification is presented in Table 1.

Selection of independent contact atoms

Consider the crystal structure of the *Chlorella* virus DNA ligase (PDB code 1fvi) in complex with the adenosine monophosphate (AMP). The interactions between the adenine ring of the ligand and protein atoms in the binding site are shown in Figures 7 and 8 of Rantanen et al. [25]. A molecular fragment of the adenine ring having an amide nitrogen (-***N***H_2) as the main atom type, interacts with the side-chain hydroxyl oxygen atom (-***O***H) of Thr25 (the left-most oxygen atom in Figure 7(b) of Ref. [25]). If we would collect all atoms that lie within a predefined cutoff distance, e.g., 4.0 Å, from the amide nitrogen atom of the ad-

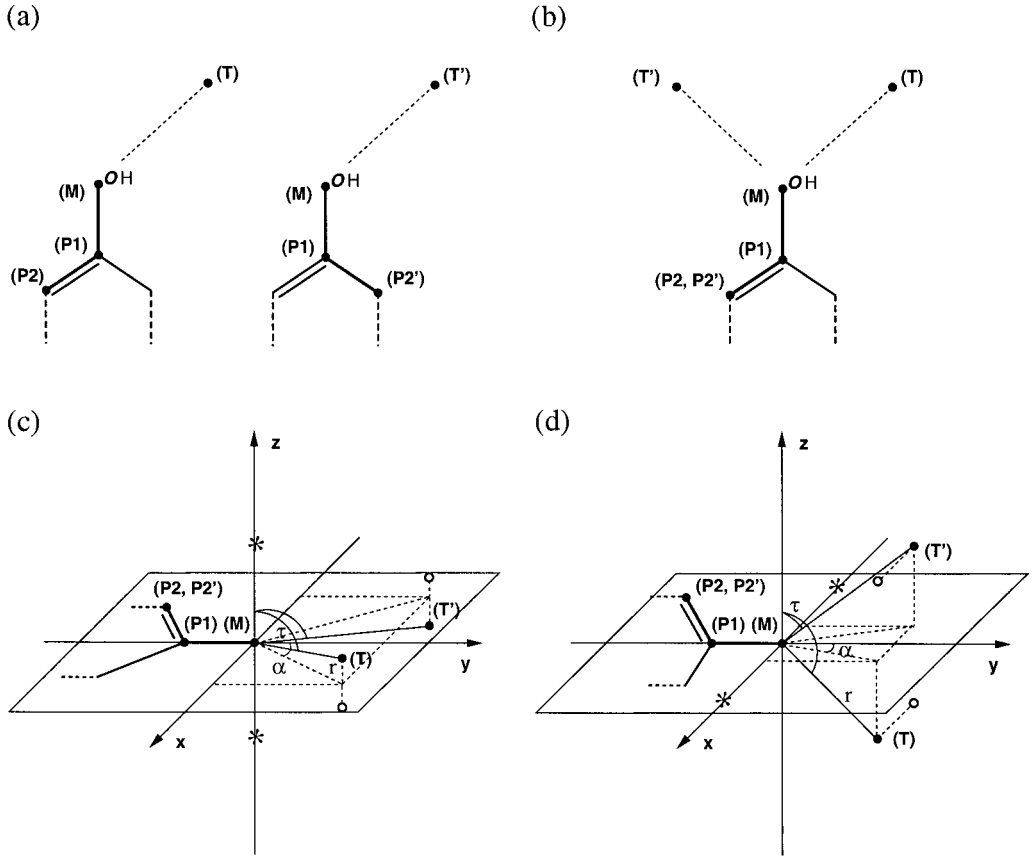

Figure 1. Schematic drawing describing the collection of data and placement of the reference frame for molecular fragments. (a) Alternative referencing of the points M, P1 and P2 for a fragment of a phenol group. The third reference point (P2) can be placed at the position of the aromatic carbon atom on either the left-hand side, P2, or on the right-hand side, P2′. Consequently, two different measurements for the relative position of a contact atom are obtained. (b) The reference points presented in (a) are superimposed on top of each other and the 3D-position of the contact atom is represented by both measurements, T and T′. (c) The phenol group fragment and positions T and T′ (•) of the contact atom are placed on the coordinate system such that the reference points lie in the *xy*-plane. Mirror-images of the positions (T and T′) lying across the *xy*-plane are indicated with two open circles (○). (d) The phenol fragment is placed in the *yz*-plane and the mirror-images are reflected across the *yz*-plane.

enine ring, then the Cβ-carbon atom (>***C***H-) of the Thr25 side-chain would be included in the analysis, too. However, the Cβ-carbon atom represents a non-independent interaction between the amide nitrogen atom of the adenine ring and the Thr25 side-chain of the protein. Such non-independent interactions are omitted in our current molecular interaction data, since their inclusion will increase the level of background noise in the data (non-independent effects are included in the analysis of Rantanen et al. [25]). Nevertheless, the amide nitrogen atom (-***N***H_2) of the adenine ring interacts with other independent functional groups (the main-chain carbonyl group (>C=***O***) of Pro26 (the right-most oxygen atom in Figure 7(a) of Ref. [25]) and a side-chain methyl group (-***C***H_3) of Leu137 (cf., Figure 7(c) of Ref. [25])) whose contribution we would want to include in our study.

In Figure 8 of Rantanen et al. [25], the adenine ring system is sandwiched between an aromatic side-chain (Phe98) on one side of the ring and a non-polar side-chain (Met164) on the other side, forming a cohesive network of interactions. Although the aromatic carbon and nitrogen atoms of the adenine ring probably contribute most to interactions with atoms of the Phe98 and Met194 side-chains, the closest atoms of these side-chains also interact with the amide nitrogen atom (-***N***H_2) of the adenine ring and thus these atoms cannot be ignored. In general, an atom may interact simultaneously with several adjacent molecular fragments and therefore each molecular fragment should

define its own interactions without knowledge of the interactions of the surrounding molecular fragments. In the end, the modeling of partially overlapping data provides added support to atom type predictions based on combined predictions from multiple fragments.

In order to consider only relevant molecular interactions, a definition for independent atom contacts is introduced. An atom is designated as an independent contact with a molecular fragment if both of the following conditions, (a) and (b), hold.

(a) *Independence.* Atoms of each amino acid are divided into standard side-chain and main-chain units. For the main atom of each molecular fragment the closest protein atom from each unit throughout the entire protein is identified – neither distance nor direction criterion is involved. Thus, one amino acid yields at most two independent atoms, one from the side-chain and one from the main-chain. If any of these atoms are covalently bonded to each other (i.e., main-chain (Cα) to side-chain (Cβ) bond of a residue, or the backbone peptide bond of adjacent residues), then only the closest of these two bonded atoms is recognized. In addition, if one or more water molecules surrounds the main atom of a fragment, then the oxygen atom of the closest water molecule is included. This procedure results in a sufficiently large set of data while excluding local non-independent interactions that primarily contribute noise. However, each amino acid contributes atoms and many remote atoms are included, too. Such irrelevant atoms are eliminated using condition (b).

(b) *Contact.* Since non-independent atoms are removed by applying condition (a), we are able to identify dense and clear data clusters in the 3D-scatterplots of atoms surrounding the main atom of a fragment (cf., Figure 3(a) of Rantanen et al. [25]). In order to choose the essential atoms involved in a contact, we assign both the van der Waals atom radius (w_i) and a tolerance value ϵ_i for each atom type i involved in a potential contact: $w_{\mathrm{i}} = 1.52$ Å for oxygen, $w_{\mathrm{i}} = 1.55$ Å for nitrogen, $w_{\mathrm{i}} = 1.7$ Å for carbon and $w_{\mathrm{i}} = 1.8$ Å for sulphur atoms (Bondi [33]). If the distance between an independent atom of type i and the main atom j of a molecular fragment is d_{ij}, then these atoms are taken to be in contact only if

$$d_{ij} \leq w_i + \epsilon_i + w_j + \epsilon_j. \quad (1)$$

The tolerance values ϵ_i were manually determined from the experimental 3D-distributions of interactions so that the observed, short-range, data clusters are included in the analysis ($\epsilon_i = 0.125$ Å for oxygen, nitrogen and sulphur atoms, $\epsilon_i = 0.5$ Å for carbon atoms). The distance criterion (1) includes strong non-covalent interactions such as electrostatic interactions between oppositely charged atoms (ionic interactions), electrostatic interactions between partial charges (e.g., aromatic-aromatic interactions), hydrogen bonding (including amino-aromatic interactions) and hydrophobic interactions, while it excludes *weak* long-range attractive electrostatic effects between atoms that are not directly connected to each other.

Distance criteria similar to (1) have been used in other approaches, too; e.g., IsoStar [12] applies a threshold value that is set to the sum of the van der Waals radii of the atoms plus 0.5 Å. In addition, non-independent atom interactions are eliminated, e.g., in the work on probabilistic receptor potentials using a 'line of sight' test (Labute, P., 1997–2003, Chemical Computing Group Inc., `http://www.chemcomp.com/feature/cstat.htm`).

Automated selection of reference frames for molecular fragments

In the present approach, fragments of molecular structures are divided into 30 different types, each of which includes one main atom (M) and at least two additional covalently bonded atoms. This is required to generate reference points (P1, P2) that assist in the definition of the 3D-position of the contact atom surrounding the molecular fragment. In most cases, the reference point P2 can be put into more than one appropriate place. Consider, e.g., the case of a symmetrical fragment of a hydroxyl oxygen (-***O***H) bonded to an aromatic ring (Figure 1(a)). The reference point can be placed on the position of the aromatic carbon atom either at the left-hand side, P2, or at the right-hand side, P2′.

Here, we strengthen the statistics for the fragments by sampling the data using all possible locations of the reference point P2. For example, reconsider the case of the hydroxyl oxygen atom (Figure 1(a)) where there were two different ways to arrange the reference points for this fragment, (M, P1, P2) and (M, P1, P2′). This leads to two different measurements for the relative position of a contact atom. If the reference points are superimposed on top of each other and the 3D-position of the contact atom is regenerated using both of the measurements, then the scatterplot contains two images (T and T′) of the contact atom (Figure 1(b)). Furthermore, the mirror-image of the contact atom lying across the plane given by the reference points is included in the data, too. Thus, in Figure 1(c), the open circle (○) below the *xy*-plane corresponds to the complementary of data-point T (●),

while the open circle above the plane corresponds to the counterpart of point T′. In Figure 1(d), the datapoints are reflected across the *yz*-plane. Consequently, in both cases the collected interaction data of the fragments are symmetrized. (Symmetrization of dataplots is also featured in IsoStar [12].)

In our previous study, each molecular fragment, together with its reference points (M, P1, P2), was placed on a coordinate system manually (see, e.g., Figure 2 of Rantanen et al. [25]). In the current study, we automatically select the best orientation of a molecular fragment in the coordinate system for each contact atom type C_k. Thus, each molecular fragment f and the associated 3D-distribution of contact atoms C_k are placed on the coordinate system in two orientations: reference points P1 and P2 are positioned in (i) the *xy*-plane (Figure 1(c)) and (ii) the *yz*-plane (Figure 1(d)). For both cases, the 3D-distribution of contact atoms C_k is modeled with Gaussian mixture models, which are fitted to the data using an expectation-maximization algorithm. The optimal Gaussian mixture model per orientation is selected with a model selection criterion (see below). In addition, the lowest value of the model selection criterion identifies the best reference frame given the candidate orientations (i) and (ii). The optimal mixture model obtained using the best reference frame contributes to the final interaction library.

Mathematical modeling of molecular interactions

The *class-conditional* probability density function $\mathbf{p}_f(\mathbf{x}|C_k)$ specifies the probability densities that the variables $\mathbf{x} = \{r, \alpha, \tau\}$ (i.e., spherical coordinates, see Figure 1(c)-(d)) have for a molecular fragment f interacting with a contact atom from class C_k. In order to accurately model the unknown distributional shapes of interaction data, which are clustered to several local regions, it is necessary to use a mixture model. The following mixture model was introduced (Rantanen et al. [25]):

$$\mathbf{p}_f(\mathbf{x}|C_k, \Theta) = \sum_{j=1}^{M_{f_k}} \beta_j \mathbf{p}_f(\mathbf{x}|C_k, \theta_j), \tag{2}$$

in which individual component densities are given by the Gaussian distribution functions

$$\mathbf{p}_f(\mathbf{x}|C_k, \theta_j) = \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \mu_j)^t \; T_j \; (\mathbf{x} - \mu_j)\}}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}, \tag{3}$$

where $\beta_j \geq 0$, $\sum_j \beta_j = 1$ and d is the dimension of the vector $\mathbf{x}$. The parameter vector Θ has mixing proportions β_j, mean vectors μ_j, and covariance matrices Σ_j as its components. For convenience, the *precision matrix* T_j is defined to be the inverse of the covariance matrix Σ_j, that is, $T_j = \Sigma_j^{-1}$. The term $(\mathbf{x} - \mu_j)^t$ denotes the transpose of the column vector $\mathbf{x} - \mu_j$ and $|\Sigma_j|$ is the determinant of the covariance matrix Σ_j.

The standard EM-algorithm [27, 28] obtains *maximum likelihood* (ML) estimates of the Gaussian mixture model (2) by maximizing a log-likelihood function (see equation (7) of Rantanen et al. [25]). This approach has led to good results in several applications, although serious problems have been reported in the literature. For example, Böhning et al. [34] have shown that the EM-algorithm is highly sensitive to initialization, and different starting values of $\Theta^{(0)}$ can lead to different final estimates. The EM-algorithm is also relatively slow computationally and, thus, improper initialization can considerably slow down the rate of convergence. Furthermore, the EM-algorithm is prone to converge to the boundary of the parameter space, i.e., during the estimation process one of the mixing proportions β_j may approach zero and the corresponding covariance matrix Σ_j may tend towards a singular matrix; such parameter estimates are not useful (see, e.g, Ref. [27], pp. 33–34, summarizing the properties of the EM-algorithm). In order to avoid the above-mentioned problems, we replace here our previous use of the standard EM-algorithm and ML-estimates (Rantanen et al. [25]), estimating the parameters of the Gaussian mixture using the *maximum a posteriori* (MAP) framework. In this concept, some probability information $p(\Theta)$ about the parameters Θ of the Gaussian mixture is given *a priori* and the modified log-likelihood function

$$\begin{aligned} \mathcal{L}_f(C_k, \Theta) &= \log \prod_{t=1}^{N_{f_k}} \mathbf{p}_f(\mathbf{x}^{(t)}|C_k, \Theta) + \log p(\Theta) \\ &= \sum_{t=1}^{N_{f_k}} \log \mathbf{p}_f(\mathbf{x}^{(t)}|C_k, \Theta) + \log p(\Theta) \end{aligned} \tag{4}$$

is maximized given M_{f_k} mixture components and N_{f_k} independent, identically distributed samples $\{\mathbf{x}^{(t)}\}_{t=1}^{N_{f_k}}$. The major advantage of the MAP-approach is that it allows user-dependent information to be incorporated into the parameter estimation process, thus giving

better control on the final values of the estimates. However, the computational burden of the parameter estimation is increased: the choice of the prior distribution, the specification of the parameters of the prior distribution and the actual process for evaluation of the MAP-estimates have to be addressed.

The following *a priori* assumptions about the unknown parameter vector Θ of the Gaussian mixture are made. The mixing probabilities $\{\beta_1, ..., \beta_{M_{f_k}}\}$ are assumed to be *Dirichlet* distributed (Ref. [26], p. 134) with the parametric vector $\{\gamma_1, ..., \gamma_{M_{f_k}}\}$, $(\gamma_i > 0;\ j = 1, ..., M_{f_k})$:

$$\mathbf{D}(\beta_1, ..., \beta_{M_{f_k}}|\gamma_1, ..., \gamma_{M_{f_k}}) = \frac{\Gamma(\gamma_1 + \cdots + \gamma_{M_{f_k}})}{\Gamma(\gamma_1)\cdots\Gamma(\gamma_{M_{f_k}})} \prod_{j=1}^{M_{f_k}} \beta_j^{(\gamma_j - 1)}, \quad (5)$$

where $\Gamma(\cdot)$ is the Euler *gamma function* (Ref. [35], pp. 31–32 and 441–442). The quantities $t_{jik} = t_{jki}$ of the precision matrix T_j of the Gaussian component densities (3) are *Wishart* distributed (Ref. [26], p. 138) with parameters δ_j ($\delta_j > d - 1$) and S_j ($d \times d$ precision matrix), if the probability density function satisfies the following equation:

$$\mathbf{W}(T_j|\delta_j, S_j) = c\,|S_j|^{\delta_j/2}\,|T_j|^{(\delta_j - d - 1)/2} \exp\{-\frac{1}{2}\mathrm{tr}(S_j T_j)\}, \quad (6)$$

where

$$c = \left[2^{(\delta_j d)/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left(\frac{2\delta_j + 1 - i}{2}\right)\right]^{-1}, \quad (7)$$

and $\mathrm{tr}(\cdot)$ denotes the *trace* of the matrix argument. On the other hand, the components of the random vector μ_j are *normally* distributed with parameters λ_j (d-dimensional mean vector) and U_j ($d \times d$ precision matrix), if

$$\mathbf{N}(\mu_j|\lambda_j, U_j) = \frac{1}{(2\pi)^{d/2}|U_j^{-1}|^{1/2}} \exp\{-\frac{1}{2}(\mu_j - \lambda_j)^t\, U_j\, (\mu_j - \lambda_j)\}, \quad (8)$$

where λ_j is the d-dimensional vector, U_j^{-1} is the $d \times d$ covariance matrix and $|U_j^{-1}|$ is the determinant of the covariance matrix U_j^{-1}. Using assumptions (6) and (8), the parameter vector μ_j and the precision matrix T_j of the Gaussian component densities (3) are assumed to have a joint *Normal-Wishart* distribution **NW** of dimension d (Ref. [26], p. 140):

$$\begin{aligned}&\mathbf{NW}(\mu_j, T_j|\lambda_j, \eta, \delta_j, S_j) = \mathbf{N}(\mu_j|\lambda_j, \eta T_j)\\ &\mathbf{W}(T_j|\delta_j, S_j) \propto |T_j|^{(\delta_j - d)/2}\\ &\exp\{-\tfrac{\eta}{2}(\mu_j - \lambda_j)^t\, T_j\, (\mu_j - \lambda_j)\}\\ &\exp\{-\tfrac{1}{2}\mathrm{tr}(S_j T_j)\},\end{aligned} \quad (9)$$

where the multivariate normal, $\mathbf{N}(\cdot)$, and the Wishart, $\mathbf{W}(\cdot)$, densities are defined above. If the mixing probabilities $\{\beta_1, ..., \beta_{M_{f_k}}\}$ and the parameters $(\mu_j,\ T_j)$ of the component densities are assumed to be independent, then the prior probability distribution about parameters of the mixture model is the product of equations (5) and (9), i.e.,

$$p(\Theta) = \mathbf{D}(\beta_1, ..., \beta_{M_{f_k}}|\gamma_1, ..., \gamma_{M_{f_k}}) \prod_{j=1}^{M_{f_k}} \mathbf{NW}(\mu_j, T_j|\lambda_j, \eta, \delta_j, S_j) \quad (10)$$

with *hyperparameters* $\gamma_1, ..., \gamma_{M_{f_k}}, \lambda_j, \eta, \delta_j, S_j$.

The Dirichlet and Normal-Wishart prior densities are, of course, chosen for pragmatical reasons and/or for reasons of computational tractability. With these choices of the prior densities, the posterior densities of the parameters are of a known form and we obtain an explicit EM-algorithm for the Gaussian mixture densities (2). The selection of priors is regarded by some researchers as a problem of *meta-statistics*, which cannot be resolved by reference to data, but only by judgement. It is, however, reasonable to be able to express these judgements in some explicit mathematical terms. In this vein, Geiger and Heckerman [36] have shown that if the parameters of certain marginal variables in a statistical model are judged to be independent and to have a positive density, then this density must be the Dirichlet density. Another argument for the use of the Dirichlet prior, technically known as *sufficientness*, is found in Gyllenberg and Koski [37] and in its references. Furthermore, Heckerman and Geiger [38] have shown that the Normal-Wishart distribution is characterized by *global parameter independence* with regard to a Directed Acyclic Graph (DAG) model. (Note that a finite mixture model, such as we have used in the present work, can be seen as a special case of a DAG.) The influence of these choices of the prior probability densities in the overall prediction system is probably a minor one, but this can be strictly assessed only by extensive simulations using, e.g., the method of Markov Chain Monte Carlo.

Gauvain et al. [39] revised the standard EM re-estimation formulas [27, 28] to the maximum a posterior (MAP) approach. If the number of mixture components M_{f_k} is fixed and N_{f_k} independent, identically distributed samples $\{\mathbf{x}^{(t)}\}_{t=1}^{N_{f_k}}$ are given, then the MAP-estimates for parameters of the Gaussian mixture, i.e., $\hat{\Theta}_{MAP} = \arg\max_{\Theta}\{\mathcal{L}_f(C_k, \Theta)\}$, where the log-likelihood function $\mathcal{L}_f(C_k, \Theta)$ defined in (4) includes the prior distribution $p(\Theta)$ of parameters Θ as described above (10), can be obtained by using the following iterative algorithm:

$$\beta_j^{(q+1)} = \frac{(\gamma_j - 1) + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t)}{\sum_{j=1}^{M_{f_k}}(\gamma_j - 1) + N_{f_k}}, \tag{11}$$

$$\mu_j^{(q+1)} = \frac{\eta_j \lambda_j + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t) x^{(t)}}{\eta_j + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t)}, \tag{12}$$

$$\Sigma_j^{(q+1)} = \frac{S_j + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t)[x^{(t)} - \mu_j^{(q+1)}][x^{(t)} - \mu_j^{(q+1)}]^t}{(\delta_j - d) + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t)} + \frac{\eta[\lambda_j - \mu_j^{(q+1)}][\lambda_j - \mu_j^{(q+1)}]^t}{(\delta_j - d) + \sum_{t=1}^{N_{f_k}} h_j^{(q)}(t)}, \tag{13}$$

where the posterior probabilities $h_j^{(q)}$ are defined by

$$h_j^{(q)}(t) = \frac{\beta_j^{(q)} \mathbf{p}_f(\mathbf{x}^{(t)}|C_k, \theta_j^{(q)})}{\sum_{i=1}^{M_{f_k}} \beta_i^{(q)} \mathbf{p}_f(\mathbf{x}^{(t)}|C_k, \theta_i^{(q)})}. \tag{14}$$

The re-estimation formulas (11)–(13) reduce to the standard EM-algorithm by selecting hyperparameters $\gamma_j = 0$, $\eta_j = 0$, $S_j = 0$ and $\delta_j = d$ for $j = 1, ..., M_{f_k}$. In other words, if no subjective information is added to the parameter estimation process then the *a priori* distribution (10) does not have any influence on the parameter estimates of the Gaussian mixture model. In this case, the standard maximum likelihood estimates are obtained.

The molecular fragments possess known symmetries that can be automatically exploited in order to symmetrize the interaction data (see the data collection phase above). However, the observed symmetry of the interactions ultimately depends on the type of the molecular fragment and the class of the contact atom. In order to explicitly include the symmetries in the mathematical models, one would need to derive several submodels from the mixture model (2). In practice, the handling of different models is laborious and an additional automatic procedure is needed to select the suitable submodel for each input. In addition, a further manipulation of the mixture functions or their variables could wreck the EM-algorithm; extra attention would be needed to keep the densitity functions as coherent probability functions, too. In the end, the complexity of the molecular library would be increased. Moreover, the general mixture model (2) is capable of modeling all kinds of interaction data, whether the given data are symmetrical or asymmetrical. For these reasons, we have refrained from including the observed symmetries of the interaction data directly in the mathematical models.

Choosing the hyperparameters and the initial values for the EM-algorithm

A discrete set of 3D-points (i.e., a 3D-grid) $\{(x_i, y_j, z_k) = (\pm 0.1n \text{ Å}, \pm 0.1m \text{ Å}, \pm 0.1l \text{ Å}), \forall m, n, l \in \mathcal{N}\}$ is placed around a molecular fragment. Each sample point $\mathbf{x}^{(t)}$, originally presented in a spherical coordinate system, is transformed to an equivalent location in a rectangular coordinate system, $\mathbf{x}_{rect}^{(t)}$. The closest grid point is assigned to a location $\mathbf{x}_{rect}^{(t)}$ and the corresponding count at that grid point is increased. Statistics for each grid point are collected and the top M_{f_k} populated points $\{(x_{i_j}^*, y_{j_j}^*, z_{k_j}^*)\}_{j=1}^{M_{f_k}}$ selected. These grid points $\{(x_{i_j}^*, y_{j_j}^*, z_{k_j}^*)\}_{j=1}^{M_{f_k}}$ are mapped back to the spherical coordinate system $\{\mathbf{x}_j^*\}_{j=1}^{M_{f_k}}$ and they provide an initial value of the mean vector $\mu_j^{(0)} = \mathbf{x}_j^*$ for the Gaussian component density j. This information is also used to determine the mean value vector λ_j for the mean vector μ_j of the Gaussian component density j, i.e., $\lambda_j = \mathbf{x}_j^*$. Therefore, the hyperparameter vector λ_j of the prior distribution is data-dependent. In other words, we (i) search for the M_{f_k} most populated locations in the interaction data using the independent 3D-grid method, (ii) start the EM-algorithm from those locations, and moreover, (iii) expect that the optimal mean values of the final parameter estimates are near these high-density regions. However, our initialization strategy does not partition the data for the EM-algorithm. (We refer to Hastie and Tibshirani [40] for an account of using the k-means clustering algorithm in determining the starting values of the EM-algorithm; see also Rantanen et al. [25] for a comparison of the present method with the method of Nissink et al. [14]). Consequently, we do not have a straightforward procedure to calculate

either the initial covariance matrix Σ_j or the mixing proportion β_j for each Gaussian contributing to the mixture model. Fortunately, according to our experiences, the EM-algorithm is not particularly sensitive to initialization of matrices Σ_j and proportions β_j. Starting with equal mixing proportions $\beta_j = 1/M_{f_k}$ and small values of the covariances, $1/10$ of the *identity* matrix (i.e., $\Sigma_j = 0.1I$), the EM-algorithm results in a fair fit of the component densities.

The rest of the hyperparameters (γ_j, η_j, δ_j, S_j) of the *a priori* distribution are determined subjectively. We have influence on the parameter estimation process and thus can inhibit the mixing proportions β_j from approaching zero: by setting $\gamma_j = 10$, we steer the mixing proportions $\beta_1, ..., \beta_{M_{f_k}}$ towards a value of $1/M_{f_k}$. Each covariance matrix Σ_j is constrained not to turn into a singular matrix by setting the hyperparameter $S_j = I$ for $j = 1, ..., M_{f_k}$. In addition, the simplest hyperparameter values $\eta_j = 1$ and $\delta_j = 4$ are applied.

Optimal model complexity of the Gaussian mixture

The principle of the 'minimum description length' (MDL) [41, 42] is often used as an objective criterion to adjust the complexity of a Gaussian mixture model (e.g., Rantanen et al. [25]). During the derivation of the MDL criterion the first order terms of the function are ignored. This approximation is too rough for our application, given that the mixture model has first order terms as mixing proportions. In addition, we foresee that the inclusion of the prior distribution of parameters requires a more sensitive model selection than the standard MDL criterion can provide.

Figueiredo et al. [43] proposed a model order criterion for mixture probabilities that is based on the 'minimum message length' (MML) [44, 45]; the model order criterion adopts the *Jeffreys' priors* (Bernardo and Smith [26], pp. 356–362) for the mixture component densities (3) and mixing probabilities β_j (2) in order to express the lack of knowledge about the parameters. If we assume that the mixture probability density function (2) does not contain any zero-probability components (i.e., $\beta_j \neq 0$, for all $j = 1, ..., M_{f_k}$), then the optimal model dimension is $\hat{M}_{f_k} = \arg\min_{M_{f_k}} \mathbf{MML}(M_{f_k})$, where

$$\mathbf{MML}(M_{f_k}) = -2 \sum_{t=1}^{N_{f_k}} \log \mathbf{p}_f(\mathbf{x}^{(t)}|C_k, \hat{\Theta}_{M_{f_k}}) + \kappa \sum_{j=1}^{M_{f_k}} \log\left(\frac{N_{f_k}\beta_j}{12}\right) + M_{f_k}\left(1 + \kappa + \log\frac{N_{f_k}}{12}\right). \quad (15)$$

The term $\kappa = (\frac{d(d+1)}{2} + d)$ specifies the number of parameters of an individual mixture component. This special MML criterion exploits knowledge of the mixing proportions β_j in assessing the optimal number of component densities, and thus we expect that it leads to better model selection than the standard MDL criterion. We refer to Lanterman [30] for an extensive survey of MDL, MML and related model selection principles.

Prediction of molecular interactions

An atom may contact simultaneously several molecular fragments. In the first case, we will describe the situation where we treat each interacting molecular fragment as an independent self-reliant unit. In the second case, the interactions made by all interacting fragments are considered as a cohesive network. Finally, predictive maps are generated for the ligand-binding domain of the glutamate receptor GluR2.

Predicting atom types based on a single fragment

The suitability of a contact atom from class C_k for a given molecular fragment f can be expressed without knowledge of the directional information on the molecular interaction by introducing the prior probability function $\mathbf{P}_f(C_k)$. The prior probability function $\mathbf{P}_f(C_k)$ is estimated directly from the collected data; for N_{f_k} contact atom observations from class C_k the prior probability function is $\mathbf{P}_f(C_k) = \frac{N_{f_k}}{\sum_{i=1}^{14} N_{f_i}}$. On the basis of the values of this function, we can rank the contact atom types for a molecular fragment without further knowledge of the variables $\mathbf{x}$.

However, the class-conditional probability density function $\mathbf{p}_f(\mathbf{x}|C_k)$ is modeled with the mixture model (2) and it specifies the 3D probability density that the variables $\mathbf{x}$ have for a molecular fragment f interacting with a contact atom C_k, giving further information on which to base the atom type prediction

upon. Namely, we have the required functions to obtain the posterior probability function $\mathbf{P}_f(C_k|\mathbf{x})$ using Bayes' theorem (see, e.g., Durbin et al. [29], p. 6):

$$\mathbf{P}_f(C_k|\mathbf{x}) = \frac{\mathbf{p}_f(\mathbf{x}|C_k)\mathbf{P}_f(C_k)}{\mathbf{p}_f(\mathbf{x})}, \tag{16}$$

where the *unconditional* probability density $\mathbf{p}_f(\mathbf{x})$ is given by

$$\mathbf{p}_f(\mathbf{x}) = \sum_{k=1}^{14} \mathbf{p}_f(\mathbf{x}|C_k)\mathbf{P}_f(C_k). \tag{17}$$

The value of the posterior probability function tells us how likely it is that a contact atom at a given position $\mathbf{x}$ surrounding a molecular fragment f belongs to class C_k. Thus, values of the posterior probability function can be used to predict the most suitable contact atom class for a molecular fragment in the situation where the only information available is the location $\mathbf{x}$ for the unidentified atom in contact. Given the value of vector $\mathbf{x}$, the posterior probabilities are estimated and arranged in decreasing numerical order: $\mathbf{P}_f(C_{k_1}|\mathbf{x})$, $\mathbf{P}_f(C_{k_2}|\mathbf{x})$, ..., $\mathbf{P}_f(C_{k_{14}}|\mathbf{x})$. The largest posterior probability value $\mathbf{P}_f(C_{k_1}|\mathbf{x})$ gives greatest support for the corresponding class C_{k_1} and thus, this is the first prediction for the type of unknown atom at position $\mathbf{x}$. The second largest posterior probability value $\mathbf{P}_f(C_{k_2}|\mathbf{x})$ provides the next best choice, C_{k_2}, for the unknown atom type, and so on.

In a similar context, Samudrala and Mould [46] used the conditional probability formalism to model the statistics of interatomic distances in proteins, which led to a Bayesian criterion for predicting the structure of a protein. Furthermore, we must point out in this connection that the prior probabilities as well as the class-conditional probabilities are based on inter-molecular atom contacts obtained from protein-ligand and protein-protein interfaces rather than intramolecular atom contacts obtained from proteins (cf., Ref. [7]). In addition, we note that the methods presented here rest on the coherent mathematical concepts that are widely used, e.g., in many statistical pattern recognition problems, the field of which is well established and has a long history.

In contrast, the initial 'probe densities' presented in X-SITE [7] and SuperStar [13] correspond to raw counts of atom contacts in scatterplots. In order to be able to compare these densities (number of probes per unit volume), heuristical normalization procedures based on 'average densities' are performed. This leads to 'propensities' in the spirit of Chou and Fasman [47]. As far as we understand it, this normalization of densities would correspond to a situation in which we had somehow 'lost' the normalization constants $\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}$ of the mixture models (3) and we would then attempt to recalculate them. In other words, the normalization processes in Refs. [7, 13] are part of the estimation procedure of 'propensities' ultimately attempting to scale the probe densities to comparable quantities, whereas the mixture densities presented here are automatically comparable since they are well-established probability densities and the step 'normalization to the same scale' is purposeless. Consequently, the normalization procedures of Refs. [7, 13] have nothing to do with the Bayesian formalism, in which prior expectations and likelihood functions are converted to posterior odds that are quantities of central interest in mathematical prediction.

Combining predictions of atom types using multiple fragments

An atom may contact simultaneously several molecular fragments $\mathbf{f} = \{f_1, f_2, ..., f_n\}$. Since each fragment f_i has its own unique interpretation of the vector $\mathbf{x}_i$, in addition to unique parameters of the class-conditional, $\mathbf{p}_{f_i}(\mathbf{x}_i|C_k)$, and prior, $\mathbf{P}_{f_i}(C_k)$, probability functions, each fragment contributes to the pool of posterior probabilities $\{\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)\}$, $k = 1, ..., 14$, $i = 1, ..., n$ for the prediction of a suitable contact atom type. The combination of these fragment-conditional posterior probabilities potentially leads to more effective prediction of a suitable contact atom type. Unfortunately, there is no straightforward way to suggest an optimal combined posterior probability function $\mathbf{P}_\mathbf{f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ for multiple fragments using Bayes' theorem. However, several practical schemes have been proposed where posterior probabilities are combined in order to approximate the probability $\mathbf{P}_\mathbf{f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ as well as possible. These schemes include, e.g., product, sum, weighted average, minimum, maximum, median and majority voting combination rules (see, e.g., Kittler et al. [48]).

In our case, we have applied two extremes of the commonly used aggregation rules. In (i) the sum combination strategy, the posterior probabilities $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ are summed together and the estimate of

$\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ is obtained by

$$\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n) = \frac{\sum_{i=1}^{n} \mathbf{P}_{f_i}(C_k|\mathbf{x}_i)}{\sum_{k=1}^{14} \sum_{i=1}^{n} \mathbf{P}_{f_i}(C_k|\mathbf{x}_i)} = \frac{\sum_{i=1}^{n} \mathbf{P}_{f_i}(C_k|\mathbf{x}_i)}{n}. \quad (18)$$

Whereas in (ii) the product combination strategy, the posterior probabilities $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ are multiplied together and the estimate of $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ is

$$\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n) = \frac{\prod_{i=1}^{n} \mathbf{P}_{f_i}(C_k|\mathbf{x}_i)}{\sum_{k=1}^{14} \prod_{i=1}^{n} \mathbf{P}_{f_i}(C_k|\mathbf{x}_i)}. \quad (19)$$

In all cases, the function $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ can be seen as the overall *responsibility* of the contact atom class C_k for the observations $\mathbf{x}_1, ..., \mathbf{x}_n$ of the molecular fragments.

Weighting of the components $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ has been omitted in the sum combination strategy (18) and the rule is expected to be accurate in the case of well-correlated probability estimates $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ that result in independent prediction errors. Errors in the posterior probability estimates of a single fragment f_i will be reflected only weakly in the overall posterior probability estimates $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$. In contrast, the product combination strategy (19) heavily penalizes the posterior probability estimate $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ of class C_k if the contact atom class C_k does not seem to be responsible for the data observed for a fragment f_i (e.g., Tax et al. [49]). Nevertheless, the values of the combined posterior probability function $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ approximated with the sum or product combination strategy can be used to predict the most suitable atom for a network of molecular fragments, $\mathbf{f} = \{f_1, f_2, ..., f_n\}$, when each of the values of vector $\mathbf{x}_i$ is measured for the corresponding fragment f_i.

Visualization of combined predictions in the ligand-binding domain of the glutamate receptor GluR2

The structure of the protein binding site (S1S2; see the Results and discussion section below) is divided into molecular fragments $\mathbf{f} = \{f_1, f_2, ..., f_n\}$ and a dense 3D-grid, composed of points (x_i, y_j, z_k), is placed around the binding site. The location of a grid point (x_i, y_j, z_k) is identified by each of the fragments f_i using the fragment-specific representation of the vector $\mathbf{x}_i$ so that the values of the posterior probability function $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ can be estimated. For each grid point, the predictions $\mathbf{P}_{f_i}(C_k|\mathbf{x}_i)$ of several molecular fragments $\mathbf{f} = \{f_1, f_2, ..., f_n\}$ are combined together using the product combination rule (equation (19)) and the estimates of the overall posterior probability function $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ are obtained. Furthermore, for each grid point, the values of the posterior probability function $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ obtained for *hydrophobic* carbon atoms (classes from 1 to 5) are summed up and the resulting probability value is assigned to the grid point (x_i, y_j, z_k). Consequently, a posterior probability map $\mathbf{P}_\mathbf{f}^h(x_i, y_j, z_k)$ characterizing favorable positions for hydrophobic (h=hydrophobic) ligand atoms is obtained. Similarly, posterior probability maps are calculated for hydrogen bond donor ($\mathbf{P}_\mathbf{f}^d$, d=donor; classes 8, 9, 10 and 13) and hydrogen bond acceptor ($\mathbf{P}_\mathbf{f}^a$, a=acceptor; classes 11 and 12) atoms. A map for water oxygen atoms ($\mathbf{P}_\mathbf{f}^w$, w=water; class 14) is obtained, too. At the high density regions of the maps, the posterior probability values approach a value of 1.0. In order to ease the visualization of the maps, the values of the posterior probabilities are scaled using the logit transform of the probability $\mathbf{P}$: $\text{logit}(\mathbf{P}) = \log \frac{\mathbf{P}}{1-\mathbf{P}}$. Consequently, the values of the final maps vary from negative to positive infinity. Furthermore, atoms whose presence in the binding site is known *a priori* (including, e.g., water molecules and metal ions) may overlap with regions of the predictive maps. In order to avoid interatomic 'bumps', where van der Waals surfaces of two atoms overlap by more than 0.6 Å, the values of the non-accessible grid points are set by default to negative infinity.

Results and discussion

Modeling

An archive of molecular structures was scanned for non-bonded intermolecular interactions (see Materials and methods). Ligand structures in binding sites and protein structures in protein-protein interfaces were divided into 30 different types of molecular fragments. Contact atoms surrounding a molecular fragment f, representing independent molecular interactions, were categorized into 14 atom classes C_k according to the atom type (Table 1). Their 3D-positions relative to molecular fragments were located using spherical coordinates, $\mathbf{x}$ (Figure 1(c)–(d)). Statistics for the fragments were symmetrized by sampling the data using all possible combinations of reference points and by including the complementary data-points lying across the plane given by the reference points (Figure 1).

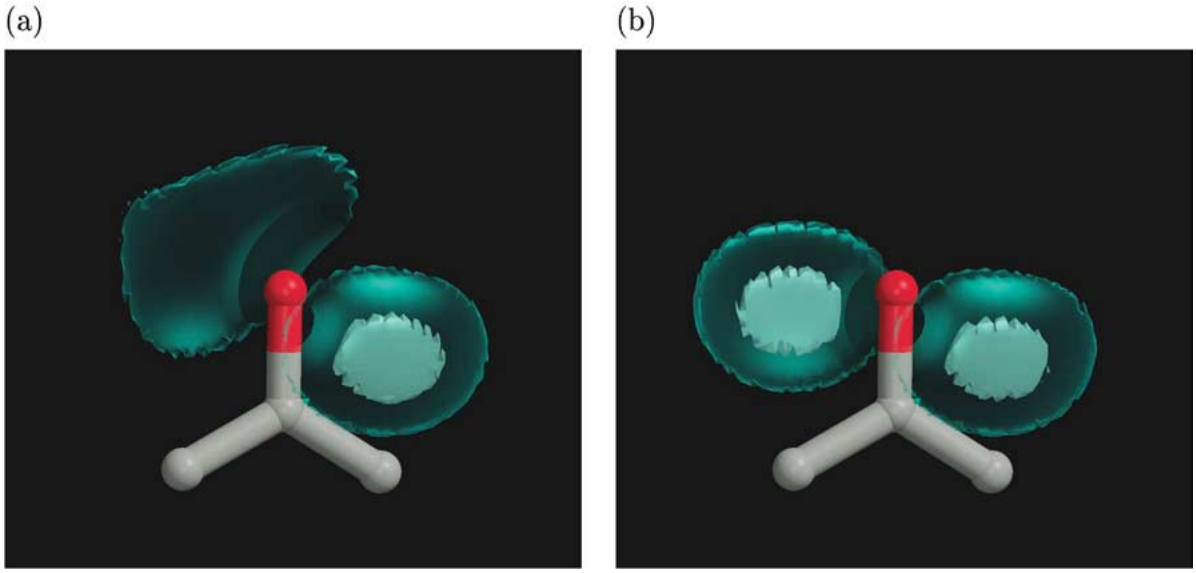

Figure 2. Comparison of two different methods to select the optimal model complexity. A symmetrized distribution of methyl carbon atoms (-***C***H_3) surrounding a hydroxyl oxygen (-***O***H) atom bonded to a planar structure is modeled with Gaussian mixtures. (a) The model selection criterion based on MDL (Rantanen et al. [25]) suggests a mixture model with 7 component densities. The resulting model is an asymmetric (and undesirable) representation of the data. (b) The criterion based on MML (this work) proposes a mixture model with 10 component densities, which does successfully represent the data. (The hydroxyl oxygen fragment system is slightly rotated around the center point; it is indeed symmetric.) Both mixture models are plotted with density levels of 0.5 (cyan) and 0.8 (white).

The 3D-distributions of the positions of the contact atoms C_k were modeled with Gaussian mixture models (equation (2)), which were fitted to the data using the EM-algorithm (equations (11)–(13)) with an increasing number of mixture components $M_{f_k} = 1, ..., 20$ until convergence. However, in all cases considered here, the number of parameters in the Gaussian mixture model was less than the number of items in the training data. The hyperparameters and initial values were determined for the EM-algorithm using a 3D-grid method (see the Theory section above). For each contact atom type C_k, both the optimal number of components, $\hat{M}_{f_k} \in \{1, ..., 20\}$, in the Gaussian mixture model and the most suitable orientation of a molecular fragment on the coordinate system was chosen automatically using the model selection criterion based on the 'minimum message length' (MML; equation (15)). Consequently, the best possible maximum *a posteriori* parameter estimates of the Gaussian mixture model were obtained.

Selection of reference frames

The 3D-positions of contact atoms surrounding the main atom of a molecular fragment were located using spherical coordinates after the fragment was placed on a coordinate system. For each contact atom class, the orientation of a fragment was selected automatically using MML as an objective model selection criterion (see equation (15) above). Primarily, we use this criterion to select the optimal complexity for a mixture model, but we found that this criterion can also be used to choose the most suitable orientation for a molecular fragment.

Consider, e.g., a molecular fragment with a hydroxyl oxygen (-***O***H) attached to a planar ring (cf., hydroxyl oxygen of tyrosine) that is positioned on the reference frames in Figure 1(c) and 1(d). The data for carboxylate oxygen atoms (class 12; -C***OO***H, >C=***O***) are clustered into two clouds that lie in the plane defined by the atoms of the fragment and are centered roughly about points T and T′. Gaussian mixture models were fitted to the data with the EM-algorithm. Both reference frames led to proper mixture models: the optimal mixture model obtained using the frame of reference in Figure 1(c) had five component densities (M_{f_k}=5) and the optimal mixture obtained using the frame of reference in Figure 1(d) had six component densities (M_{f_k}=6). Nevertheless, after visual comparison of the models one readily selects the mixture model obtained using the frame of reference in Figure 1(c) to contribute to the final interaction library. Indeed, the values of the MML function indicated

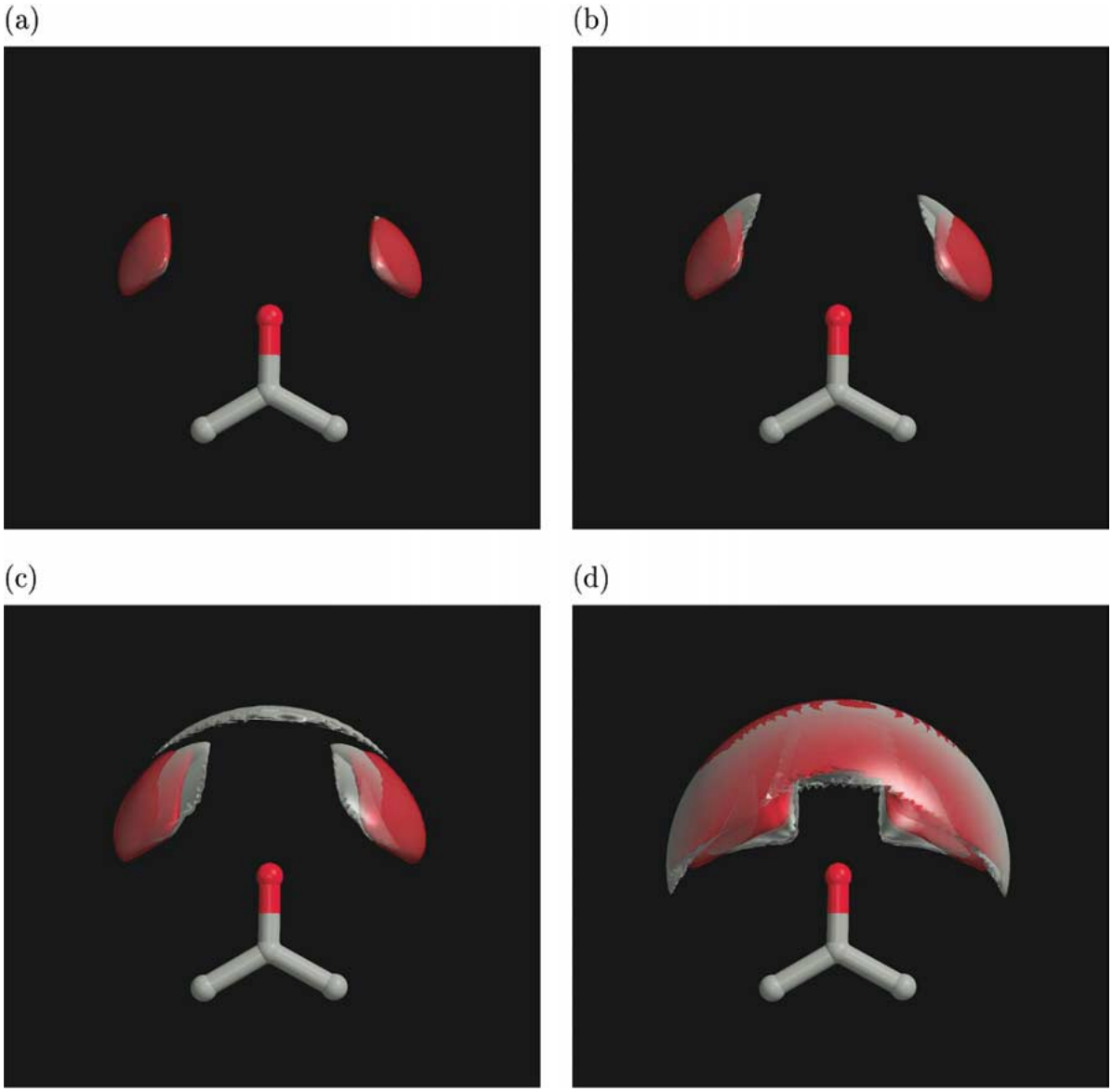

Figure 3. Comparison of two different mixture models (EM-algorithms) composed of an equal number of component densities. A distribution of carboxylate oxygen atoms surrounding a fragment containing a hydroxyl oxygen (-***O***H) is considered. In each case, a mixture model obtained using the standard EM-algorithm (ML-estimates; Rantanen et al. [25]) is depicted with a white isosurface, while a model obtained using the modified EM-algorithm (MAP-estimates; this work) is depicted with a red isosurface. (a) Contour value of 0.3: the 'plump' red surface (MAP) covers the 'thin' white surface (ML) and the white surface comes into view only at the edges of the red density. (b) Contour value of 0.2: the red surface is compact while the white surface broadens and looses its compactness. (c) Contour value of 0.125: again, the red surface is more compact than the white surface. Additional component density, modeling outlying data-points, becomes visible for the white density. (d) Contour value of 0.05: the edges of the red density are smoother than the edges of the white density.

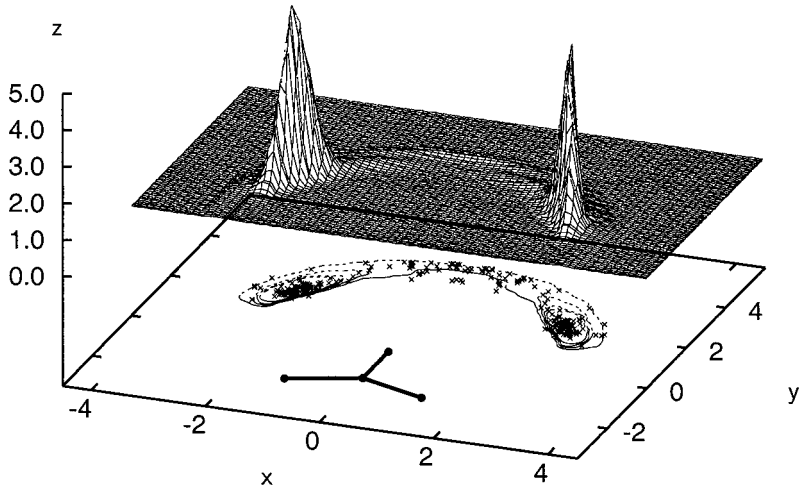

Figure 4. Comparison of a theoretical distribution with empirical data. A molecular fragment having a hydroxyl oxygen atom (-***O***H) as the main atom type is placed on the *xy*-plane (cf., Figures 2 and 3). The positions of data-points (class 12), which locate within a distance of 0.25 Å from the *xy*-plane, are indicated by crosses; other data-points are not displayed. The probability values of the corresponding 3D mixture model (MAP-estimates; the present work) are evaluated at the *xy*-plane and they are represented by isocontours (levels 0.06, 0.125, 0.2 and 0.3). Above the *xy*-plane, the elevations of the isocontours are depicted as a surface. The figure indicates that the mixture model (five component densities) provides a suitable representation for the clustered data.

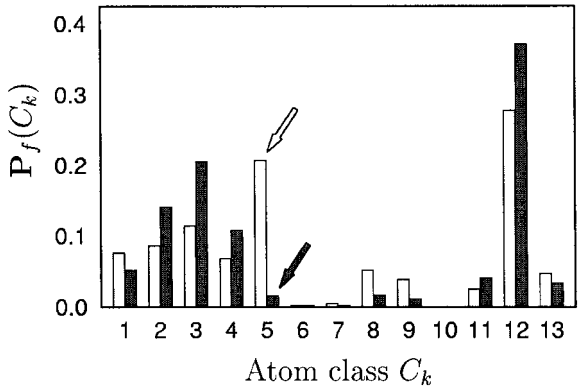

Figure 5. Two variants of the prior probability distribution, $\mathbf{P}_f(C_k)$, which models the suitability of a contact atom type (Table 1) for a molecular fragment without knowledge of the 3D-position of the particular contact atom, are shown for a molecular fragment with a hydroxyl oxygen atom (-***O***H). Dark-gray bars were obtained using the definition for independent contact atoms (this work; see the Theory section), whereas the light-gray bars correspond to the prior probability distribution that was obtained bypassing the definition for independent contact atoms (Rantanen et al. [25]). In this example, the definition of independent contact atoms improves the weighting of independent interactions (i.e., atoms of class 12 (side-chain carboxylate and main-chain carbonyl oxygen atoms), 2 (methylene carbon atoms), 3 (methyl carbon atoms) and 4 (aromatic carbon atoms)), while dramatically reducing the influence of non-independent interactions (i.e, carbon atoms of carbonyl-, carboxylate-, and guanido-groups (class 5) that are directly attached to atoms of class 12), from 21% (dark-gray arrow) to 1% (light-gray arrow).

that the setting in Figure 1(c) is better than that in Figure 1(d).

In contrast, the data for methylene carbon atoms (class 2; -***C***H_2) are accumulated on both sides of the hydroxyl oxygen so that the high density regions locate perpendicular to the plane of the planar fragment. These regions are indicated with asterisks ($*$) on the z-axis (Figure 1(c)) and the x-axis (Figure 1(d)). With the frame of reference in Figure 1(c), data-points populate the border of the variable space such that the values of the angle $\tau \in [0, \pi]$ are spread either towards the boundary value of 0 or π. These data are not *normally* distributed and 11 component densities (M_{f_k}=11) were needed to obtain a good Gaussian mixture model. Instead, if the reference frame of Figure 1(d) is used, then the values of vector **x** do not populate the border of variable space and the optimal mixture model is obtained using only eight components (M_{f_k}=8). This is a much better choice for the final model and the fitness of this mixture model is indicated with the lower value of the MML model selection criterion. Consequently, different reference frames are selected for the molecular fragment containing a hydroxyl oxygen (-***O***H) when its interactions with either the methylene carbon or the carboxylate oxygen atoms are considered.

As a result, we conclude that it is difficult to define only one orientation of a molecular fragment in a coordinate system that would be simultaneously optimal

for each of the different contact atom classes. Instead, the orientation of the molecular fragment should be selected relying on the input data so that each contact atom class has its own optimal reference frame for a given molecular fragment. Here, this dynamic and adaptive selection of the optimal frame of reference for a molecular fragment and a particular atom class is fully automated.

Model selection and the quality of the mixtures

The optimal model complexity ($\hat{M}_{f_k}$) of the Gaussian mixture was determined by choosing the mixture model that minimizes the MML criterion (equation (15)). This model selection criterion includes information from the mixing proportions β_j of the Gaussian mixture and leads to more sensitive model selection than the MDL criterion used in Rantanen et al. [25]. When we compared the results of the MML model selection criterion with the results of the standard MDL criterion we found that in 40% of the cases these criteria led to an equal number of component densities ($\hat{M}_{f_k}$) for the Gaussian mixture. In the remaining cases, the MML criterion selected Gaussian mixture models that were more complex than the choices made using the MDL criterion: e.g., in 20% of the cases, the mixture model obtained with the MML criterion had one component density more than the model obtained with the MDL criterion, whereas in 15% of the cases the MML citerion led to models with two additional component densities.

However, the risk that the selected mixture models are overdetermined for large numbers of component densities and small numbers of data is low. In the majority of cases, there were relatively high numbers of data available for fitting the mixture densities. Only in $\sim$9% of the cases, the number of data-points turned out to be less than four times the total number of adjustable parameters. In any event, the mixture densities are *semiparametric* models that are designed to handle situations, in which the structures of data need to be estimated using only a limited number of data-points.

Although the differences in terms of model complexity appear nominal, the mixture model chosen by the MML criterion may be more suitable than the one selected by the MDL criterion. Indeed, visual inspection of the mixture models revealed that there were cases where the MDL criterion resulted in an underestimation of the data whereas the MML criterion gave a suitable model. For example, if the symmetrized distribution of methyl carbon atoms (-$\boldsymbol{C}$H$_3$) surrounding a hydroxyl oxygen (-$\boldsymbol{O}$H) atom bonded to a planar structure is modeled with 7 component densities, as the MDL criterion suggests, then we will obtain a final Gaussian mixture model that is clearly asymmetrical (see Figure 2(a)). Whereas, the MML criterion proposes a mixture model with 10 Gaussian component densities and the molecular interactions are successfully represented (see Figure 2(b)). However, in our previous study (Rantanen et al. [25]) we found that the MDL criterion leads to good modeling if the standard EM-algorithm and ML-estimates are used and no *a priori* information is considered. Here, the inclusion of *a priori* information in the parameter estimation process has an important influence on the final estimates of the mixture model. Nonetheless, we can make the most of the MAP-estimates when the mixture models are selected using the MML criterion rather than the MDL criterion.

We can also compare the quality of the class-conditional probability distributions presented by Gaussian mixture models, $\mathbf{p}_f(\mathbf{x}|C_k, \Theta)$, and obtained using the standard EM-algorithm (ML-estimates; Rantanen et al. [25]) with those obtained using the modified EM-algorithm (MAP-estimates; this work). (We refer to Rantanen et al. [25] for comparison of the quality of the Gaussian mixture models with the experimentally determined 3D-distributions of contact atoms.) Consider, for example, the distribution of carboxylate oxygen atoms (class 12; Table 1) surrounding a hydroxyl oxygen (-$\boldsymbol{O}$H) atom bonded to a planar structure (Figure 3). In each section of Figure 3, the probability distribution of the mixture model corresponding to the MAP-estimates is plotted as a red contour, and the model corresponding to the ML-estimates is plotted as a white contour. Both distributions have an equal number of mixture components (M_{f_k}=5) and comparable threshold values for the contours are applied: 0.3 in (a), 0.2 in (b), 0.125 in (c) and 0.05 in (d).

In the core of the density, Figure 3(a), the red surface corresponding to the MAP-estimates is thicker than the white surface due to the influence of the *a priori* assumptions over the parameter estimates (MAP). The red density covers the white density and the white surface pierces the red surface only at the edges of the red density. The use of a smaller threshold value, 0.2 in Figure 3(b), reveals that the red density (MAP) is compact and it effectively highlights the favorable spatial position for the carboxylate oxygens, while the white density obtained without *a priori* assumptions (the standard ML-estimates) loses its compactness and is broader than the red density. This trend can be seen

also in Figure 3(c), where the threshold value 0.125 is used. In the case of the ML-estimates, an additional white component density that is attached to the outermost data-points is visible, in contrast to the red distribution (MAP-estimates), in which the effect of the equivalent component density is damped *a priori* and thus is not noticed. If an exceedingly small threshold value is applied, 0.05 in Figure 3(d), then the sharp edges of the outermost white component density (ML) can be detected. At those regions, the red component density is smoother than the white one, since its covariance matrix is prevented from approaching a singularity matrix.

Consequently, in using the MAP-approach we obtain Gaussian mixture densities that are compact and smoothly surround the cores of the data. The component densities that are attached to the peripheral data-points are well controlled and, thus, stable overall densities are obtained.

Furthermore, referring to Figures 2 and 3, we point out that in cases where the clustering of the data is the primary task of the analysis it is relevant to have one-to-one correspondence between the mixture components and the actual groups of data. However, our primary task here is to model accurately the class-conditional probability functions for molecular interactions and the actual clustering of the data is of secondary importance. Consequently, if one component density is not enough to model the unknown shape of an empirically observed data 'cloud', then the cloud is modeled using a mixture formulation meaning that the data-points of the cloud are further divided into smaller subgroups and modeled with several component densities (Ref. [28]). In the end, the combined effect of all component densities provides a suitable representation for the data. For example, in Figure 4, the molecular fragment shown in Figures 2 and 3 is placed on the *xy*-plane so that the hydroxyl oxygen atom of the fragment lies at the origin. The positions of data-points that belong to contact atom class 12 and which are closer than 0.25 Å from the *xy*-plane are indicated with crosses; other data-points, which locate further away from the plane, are not considered here. Furthermore, the probability values of the corresponding mixture model (MAP-estimates; the present work) at the *xy*-plane are depicted using isocontours obtained at levels of 0.06, 0.125, 0.2 and 0.3. (Note that the mixture density is modeled using 3D-data, but here only the probability values at the 2D-plane are considered.) In addition, above the *xy*-plane, the elevations of the isocontours are described with a surface obtained for the probability values of the mixture model. Clearly, the data-points of the contact atoms forming two distinct groups on the *xy*-plane are in good agreement with the mixture model that is comprised of five component densities.

The prior probability function of independent contact atoms

In this work, we have adapted the Bayesian way of thinking in the modeling of molecular interactions. Namely, both (i) the parameter estimates of the Gaussian mixtures that describe the 3D-distributions of interacting atoms and (ii) the EM-algorithm, which is used to fit the mixture models to the data, incorporate subjective information or judgement given *a priori*. Compatible with the Bayesian framework, we have also taken into account *a priori* knowledge of the relevant molecular interactions in our data collection procedure and considered only the atoms that are in independent contact with a molecular fragment (see the Theory section above).

In Figure 5, we illustrate the effectiveness of the definition for independent contact atoms by plotting the values of the prior probability distribution $\mathbf{P}_f(C_k)$ for each contact atom class C_k in the case of the molecular fragment containing a hydroxyl oxygen (-***O***H). The dark-gray bars in Figure 5 correspond to the prior probability distribution, which was obtained using the definition for independent contact atoms, whereas the light-gray bars (Figure 5) correspond to the prior probability distribution that was obtained collecting all possible atoms lying within a cutoff distance of 4 Å from the hydroxyl oxygen atom of the fragment, so that the definition for independent contacts was bypassed.

The carbon atoms of carbonyl-, carboxylate- and guanido-groups (class 5; -***C***$R_1(R_2)$) are located in the core of their respective functional groups and they are not very exposed and available for interactions. Indeed, the definition for independent contact atoms significantly reduces the proportion of this atom class, from $\mathbf{P}_f(C_k{=}5)$ $\approx$21% (light-gray arrow) to $\mathbf{P}_f(C_k{=}5)$ $\approx$1% (dark-gray arrow). In contrast, the proportions of the methylene (class 2; -***C***H_2-), methyl (class 3; -***C***H_3) and aromatic (class 4; >$\boldsymbol{C}_{ar}$H, >$\boldsymbol{C}_{ar}$R) carbon atoms are increased in the prior probability distribution, since these atoms are more accessible to forming independent interactions. Furthermore, the prior probability value for the oxygen atoms in the carboxylic acid and main-chain carbonyl groups (class 12; -C***OO***H, >C=***O***) is increased, emphasizing their

capability to participate in hydrogen bonding with the hydroxyl oxygen of the fragment. Changes for other atom classes are rather nominal and consequently proportions corresponding to independent contact atoms are emphasized in the prior probability distribution.

Prediction

The type of atom in contact with one or more molecular fragments can be predicted using the values of the posterior probability function $\mathbf{P}_f(C_k|\mathbf{x})$ or $\mathbf{P_f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$. Namely, given the relative location of contact atom in terms of the vector $\mathbf{x}$ (a single molecular fragment) or a set of vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ (a network of molecular fragments), the arranged numerical order of the posterior probability values ranks the suitable atom types for an unidentified contact atom (see the Theory section above).

We have evaluated the degree of error in our atom type predictions by comparing the actual atom class C_k of a contact atom for a fragment (observed in crystal structures) with the predicted ranked position. If the best-ranked atom type is the correct one, then the prediction resulted in an error value of zero (error=0). If the best prediction was incorrect, but the second best prediction matched with reality, then the error value of one (error=1) was assigned, and so on. Thus, we divided each ligand in the data set into molecular fragments and predicted the atom types of contacting protein atoms considering separately the results for fragments containing either a nitrogen, a carbon, an oxygen, or a sulphur atom as the main atom (a total of 63 000, 339 000, 132 000 and 6 400 independent predictions, respectively). The proportions of the prediction errors, P(error), are presented in Figure 6.

The continuous line in each plot (Figure 6(a)–(d)) corresponds to the prediction errors that were obtained treating each fragment as an independent self-reliant unit and was constructed by joining the consecutive error values P(error) with lines. For example, all ligand fragments containing a nitrogen atom as the main atom are considered in Figure 6(a). In 37.3% of the cases (i.e., P(error=0)=37.3%), the correct protein atom type was predicted immediately, while in 26.1% of the cases the correct atom type was predicted after one mistake (P(error=1)=26.1%). The prediction curve decreases very rapidly, indicating clean predictions. Similarly, the degree of success in the predictions is also high for ligand fragments containing a carbon (Figure 6(b)), an oxygen (Figure 6(c)) or a sulphur (Figure 6(d)) atom as the main atom. The error curves decrease rapidly and plateau only at their tails (cf., Figure 9(c) of Rantanen et al. [25], where the too widely divided atom classification is reflected in the central 'hump' for the predictions for the carbon containing fragments).

The graphs of Figure 6 also include two broken lines that lie nearly on top of each other. These lines correspond to the prediction errors that were obtained using either the sum (equation (18)) or the product (equation (19)) strategies in order to give a combined prediction of atom types. It is striking that, although two entirely different combination strategies were used, almost identical changes in the prediction curves were obtained. For example, in Figure 6(a) both combination rules increase the proportion of successful predictions for the nitrogen containing fragments from P(0)=37.3% to P(0)=42.0% (sum-rule) and P(0)=42.3% (product-rule), whereas in Figure 6(b) the predictions for the carbon atom containing fragments sharpen from P(0)=30.9% to P(0)=35.9% (sum) and P(0)=36.6% (product). Furthermore, the use of multiple molecular fragments improves predictions for fragments containing either an oxygen (Figure 6(c)) or a sulphur (Figure 6(d)) atom as the main atom, too (oxygen, from P(0)=38.3% to P(0)=40.5% (sum) and P(0)=40.9% (product); sulphur, from P(0)=34.6% to P(0)=35.9% (sum) and P(0)=36.3% (product)). These results indicate that the combination of posterior probabilities given by individual fragments leads to more effective atom type predictions and that the product combination strategy is slightly better than the sum combination strategy.

We have also extended the prediction analysis to an opposite problem, in which we first divided each binding site of a protein into molecular fragments and then predicted the atom types of the interacting ligand and water molecules. (We remind the reader that nonbonded interactions between ligand atoms and protein fragments are not included in the training data set. These interactions are preserved intentionally as an independent validation data.)

For the main atom of each protein fragment both the closest atom from the ligand molecule and the closest oxygen atom from any nearby water molecules were treated as independent atoms. If the distance between an independent atom and the main atom of a protein fragment was within the cutoff distance (see the Theory section), then the independent atom was determined to be in contact with the main atom of the protein fragment. However, the oxygen atom of a water molecule was included only if it was also found

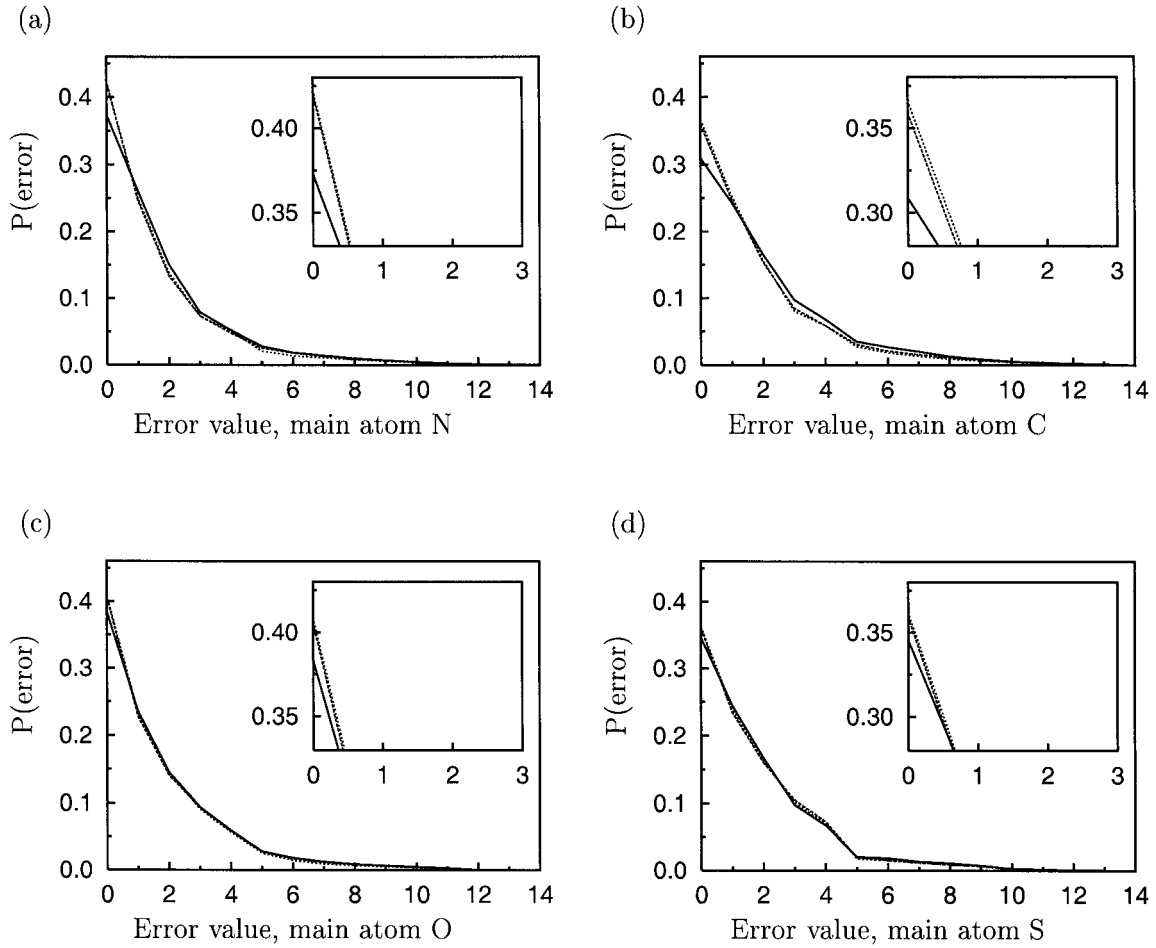

Figure 6. Proportions of prediction errors, P(error), obtained for *ligand fragments* containing (a) a nitrogen, (b) a carbon, (c) an oxygen, or (d) a sulphur atom as the main atom. Each ligand fragment was treated as an independent unit (continuous line) when the atom types of the contacting protein atoms were predicted. Predictions based on multiple ligand fragments were combined using the product rule and the sum rule (two overlapping broken lines; magnified view shown in the inset). These results demonstrate that a consideration of multiple molecular fragments can lead to improved predictions.

to be in an independent contact with one of the ligand fragments. Consequently, water molecules that were not involved in the binding of the ligand molecules were excluded. In addition, each contact atom obtained from the ligand or from water molecules was represented with an eligible atom type from Table 1. If we were not able to determine the applicable atom type for the contact atom (e.g., a triple-bonded carbon atom does not have a corresponding atom type in Table 1), then the contact atom was not included in the analysis.

The prediction results for the protein fragments containing a nitrogen, a carbon, an oxygen, or a sulphur atom as the main atom are presented in Figure 7 (a total of 29 000, 146 000, 42 000 and 1 500 predictions, respectively). As above, the continuous line corresponds to the prediction errors that were obtained considering protein fragments as independent units.

The best predictions are obtained for the protein fragments containing an oxygen atom as the main atom (Figure 7(c)). The correct ligand atom is predicted immediately in 34.8% of the cases (i.e., P(0)=34.8%). The error curve decreases rapidly, indicating good prediction accuracy. Good predictions are obtained also for the protein fragments containing a nitrogen atom as the main atom, since the error-free prediction is obtained in 28.3% of the cases (P(0)=28.3%) and the curve decreases fairly rapidly (Figure 7(a)). However, the prediction curves do not decrease sharply from the beginning, when the protein fragments contain either a carbon or a sulphur atom as

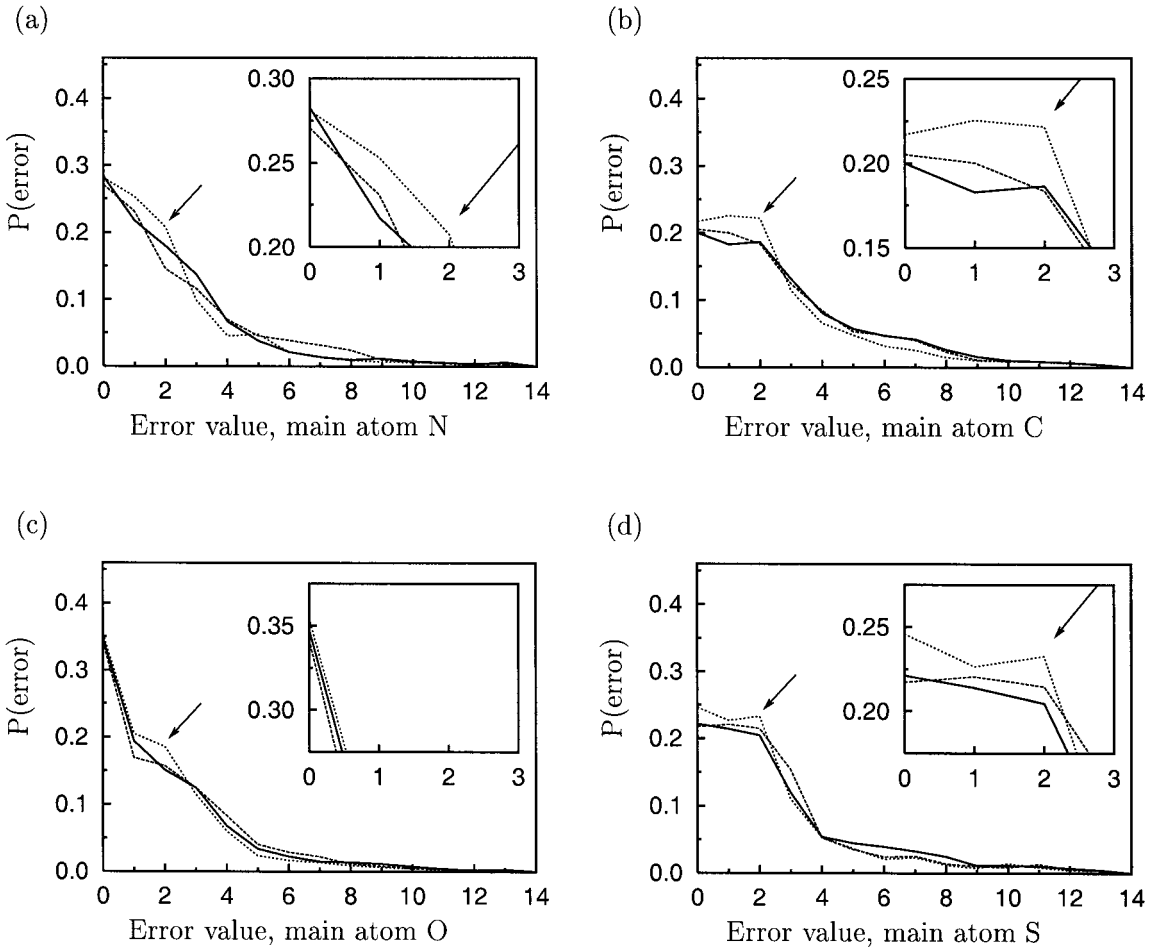

Figure 7. Proportions of prediction errors, P(error), obtained for *protein fragments* containing (a) a nitrogen, (b) a carbon, (c) an oxygen, or (d) a sulphur atom as the main atom. Each protein fragment was treated as an independent unit (continuous line) when the atom types of the contacting ligand atoms were predicted. Predictions based on multiple protein fragments were combined using the product rule (arrow; broken line) and the sum rule (second broken line). In all cases, the product combination rule leads to better predictions than the sum rule.

the main atom (Figures 7(b) and 7(d)). In both cases, the three left-most proportions P(0), P(1) and P(2) in the curves are approximately equal, after which the decrease of the curves is comparable to the previous graphs. Nonetheless, in nearly 60% (Figure 7(b)) or 65% (Figure 7(d)) of the predictions the correct atom class type was within the three best-ranked atom types. These results indicate that the present approach can also be used to predict suitable contact atom types surrounding protein fragments.

As above, the predictions of atom types given by multiple fragments (broken lines), this time for protein fragments, were combined using either the product (arrow) or the sum combination strategy. In all cases, the product combination rule generally leads to better results than the sum rule. Consider, e.g., the protein fragments that contain a carbon atom as the main atom (Figure 7(b)). The values of the three left-most proportions obtained using the product rule (P(0)=21.7%, P(1)=22.6% and P(2)=22.2%) are greater than the values obtained using the sum rule (P(0)=20.5%, P(1)=20.0% and P(2)=18.4%). Similarly, the product combination rule gives somewhat better results than the sum rule in Figure 7(d), in which the predictions for protein fragments containing a sulphur atom as the main atom are considered. Furthermore, in Figure 7(a) the product combination rule results in a better outcome than the sum combination rule.

Above, we have evaluated the degree of error in the predicted ranked order of atom types. Here, we consider the values of the posterior probability

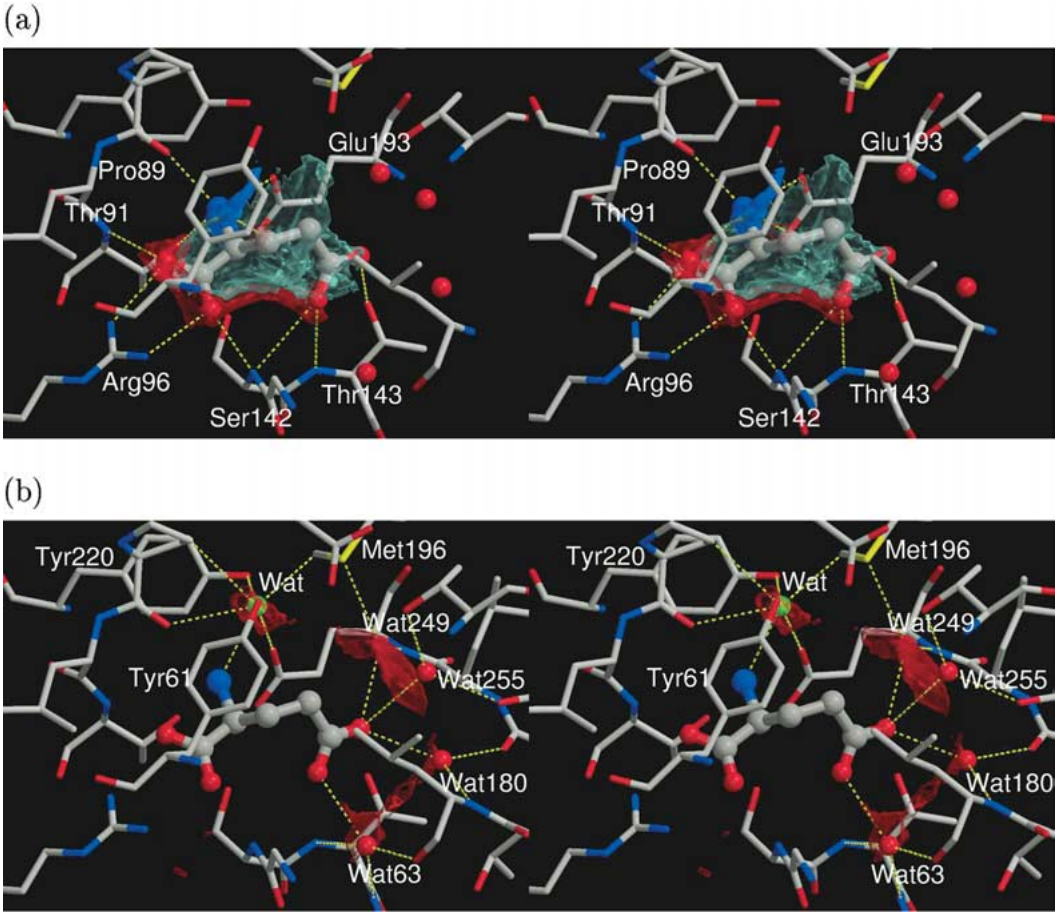

Figure 8. Stereo view of the ligand-binding domain S1S2 of the glutamate receptor GluR2 with bound glutamate (PDB code 1ftj). Only residues interacting with the ligand are shown in (a); additional details near the water molecules are shown in (b) (the same orientation is depicted and labels can be compared across both figures). (a) Molecular fragments obtained from S1S2 were used to construct predictive maps for ligand atoms functioning as hydrogen bond acceptors (red density), hydrogen bond donors (blue density) and as hydrophobic contacts (cyan density). Maps are depicted using density contour levels of 10.0, 5.0 and 15.0, respectively. The actual positions for the two α-carboxylate oxygen atoms and the one side-chain carboxylate oxygen atom of the ligand glutamate overlap with the red density; the actual position for the α-nitrogen atom of glutamate is covered by the blue density; and the five carbon atoms of glutamate are within the cyan density. (b) Molecular fragments obtained from S1S2 and the ligand glutamate were jointly used to predict favorable positions for water oxygen atoms. The red density is plotted using a contour level of 5.0. The actual positions of the water oxygen atoms Wat63, Wat180 and Wat249 overlap with the predicted regions, while the position of Wat255 locates near the red density. Furthermore, a green sphere pinpoints an unoccupied region of the red density, where there is sufficient space for one water molecule, although none is reported in the structure file 1ftj. The labeling scheme follows that present in the crystal structure file 1ftj.

function that are obtained for the true atom type (observed in the crystal structure) without involving the predicted order of atom types. Given the relative location of a contact atom in terms of the vector $\mathbf{x}$ (a single molecular fragment) or a set of vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ (a network of molecular fragments), the posterior probability value $\mathbf{P}_f(C_k|\mathbf{x})$ or $\mathbf{P}_\mathbf{f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ corresponding to the actual atom type C_k was calculated. (For simplicity, we let $\mathbf{P}$ denote either $\mathbf{P}_f(C_k|\mathbf{x})$ or $\mathbf{P}_\mathbf{f}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$ depending on the underlying prediction approach.) All molecular fragments in the data set were examined and the proportions of the posterior probability values $\mathbf{P}$ falling into four ranges (first quarter, $\mathbf{P} \in [0.0, 0.25]$; second quarter, $\mathbf{P} \in (0.25, 0.50]$; third quarter, $\mathbf{P} \in (0.50, 0.75]$; and last quarter, $\mathbf{P} \in (0.75, 1.0]$) are shown in Table 2.

Consider the posterior probability values obtained for ligand fragments. In 55.1% of the cases the posterior probability value $\mathbf{P}$ corresponding to the actual atom type is less than or equal to 0.25 (first quarter). The use of the multiple fragment approach decreases this proportion either to 52.7% (sum rule) or to 49.5%

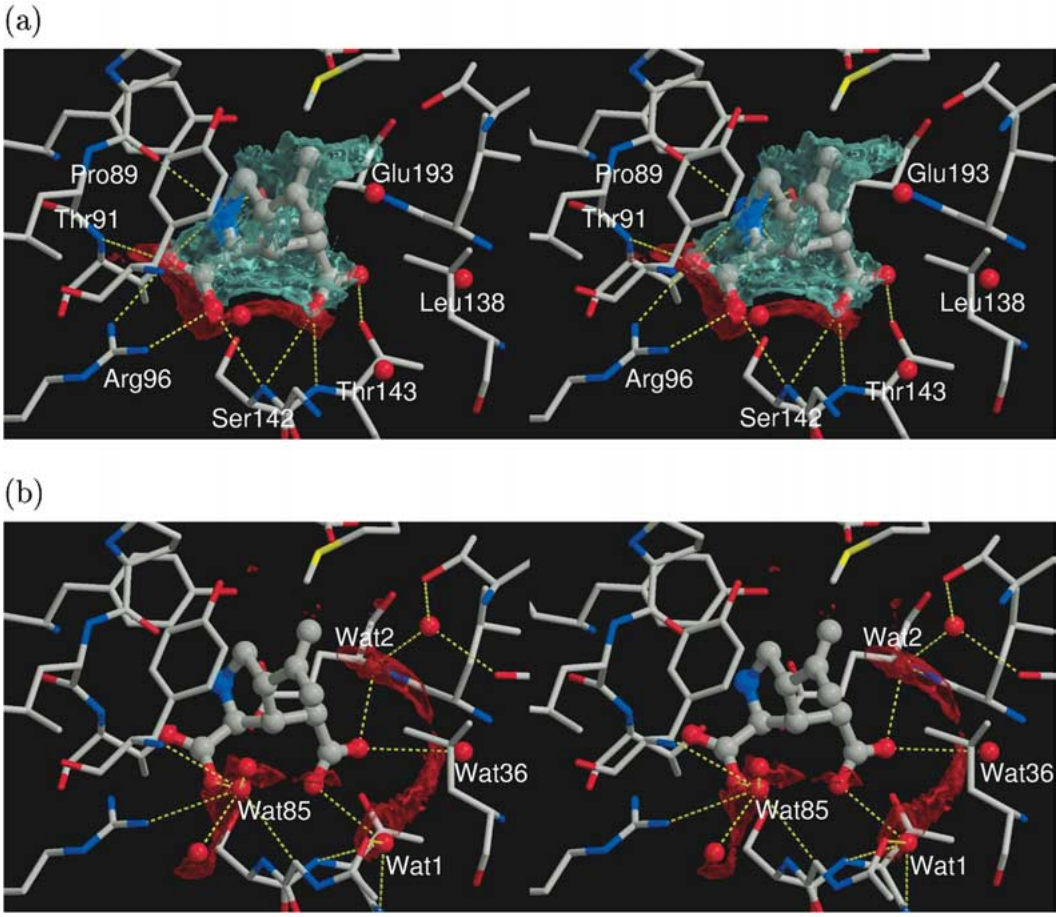

Figure 9. Stereo view of S1S2 with bound kainate (PDB code 1gr2), prepared as described in Figure 8. (a) Favorable positions for the atoms of the ligand kainate were predicted. The red density overlaps with the actual positions for the two oxygen atoms of the α-carboxylate group of kainate. The blue density identifies the actual position for the NH_2^+-nitrogen atom of kainate. Moreover, the altered side-chain conformations of Tyr61 and Leu138 provide extra space for the five additional carbon atoms present in kainate and the actual positions of the carbon atoms of kainate overlap with the cyan density. (b) Favorable positions for water oxygen atoms were predicted. The actual positions of the water oxygen atoms Wat1, Wat2 and Wat85 overlap with the predictive red density; Wat36 locates close to the red density, too. The red density is depicted using a contour value of 6.0; water oxygen labels are those from the structure file 1gr2, otherwise the residue labeling is as in Figure 8.

Table 2. Proportions of the posterior probability values corresponding to the actual atom class that are obtained using different prediction methods.

Method	[a]P ∈ [0.0, 0.25]	P ∈ (0.25, 0.50]	P ∈ (0.50, 0.75]	P ∈ (0.75, 1.0]
Ligand fragments				
Single fragment	55.1%	37.7%	6.2%	0.9%
Multiple frag. (sum rule)	52.7%	42.6%	4.4%	0.2%
Multiple frag. (product rule)	49.5%	23.9%	**14.4%**	**12.3%**
Protein fragments				
Single fragment	70.1%	19.0%	4.2%	6.6%
Multiple frag. (sum rule)	71.8%	16.3%	6.7%	5.2%
Multiple frag. (product rule)	68.5%	12.9%	5.9%	**12.6%**

[a]Depending on the prediction method used, P denotes either $P_f(C_k|\mathbf{x})$ or $P_{\mathbf{f}}(C_k|\mathbf{x}_1, ..., \mathbf{x}_n)$.

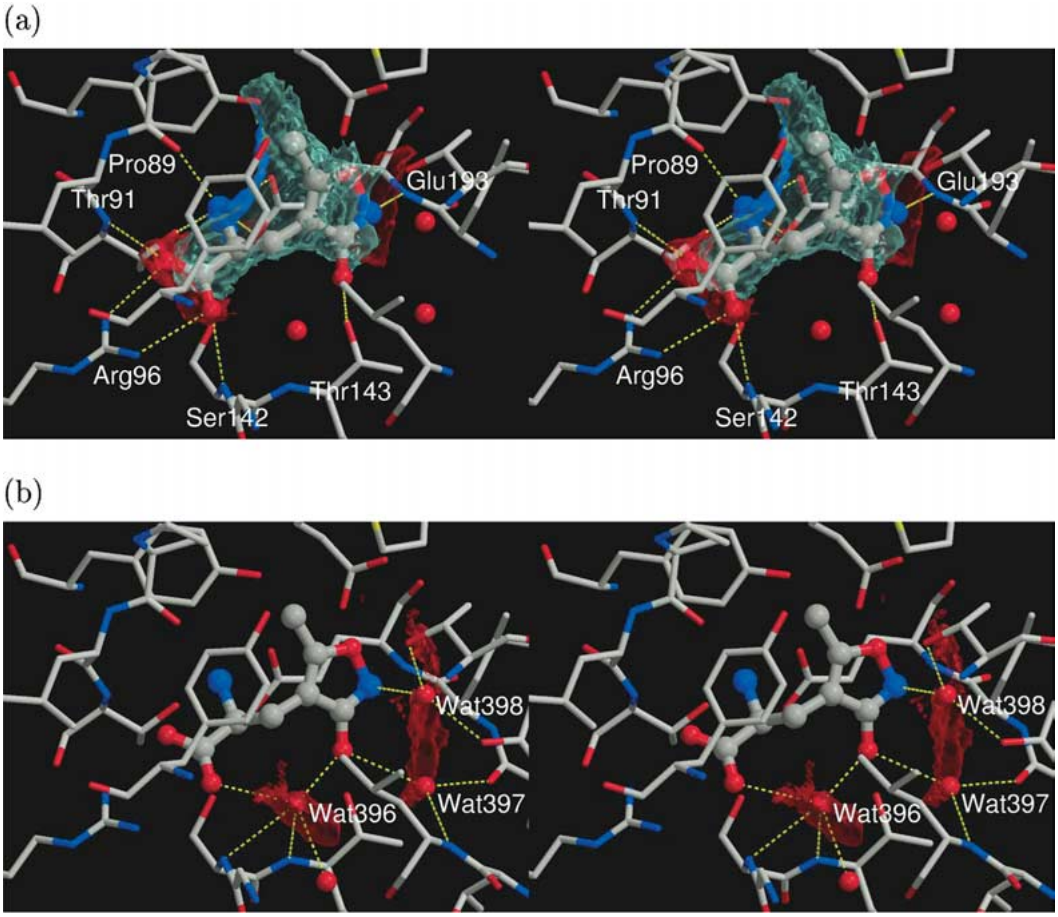

Figure 10. Stereo view of S1S2 with bound AMPA (PDB code 1ftm), prepared as described in Figure 8. (a) Suitable positions for the atoms of AMPA were predicted. The actual positions of the α-carboxylate oxygen atom, the $\mathit{N}H_3^+$-nitrogen atom and the seven carbon atoms of the ligand AMPA fit well with the predictive red, blue and cyan densities. In addition, the actual location for the *negatively-charged* nitrogen atom of the isoxazole ring of the AMPA overlaps with the red density. (b) Favorable positions for water oxygen atoms were predicted. A contour level of 3.0 is used and the actual positions for the water oxygen atoms Wat396, Wat397 and Wat398 are embedded in the red density. The residue labeling scheme in the structure file 1ftm is identical to that in the file 1ftj.

(product rule). Furthermore, the product combination rule efficiently reduces the number of the posterior probability values **P** of the second quarter, from 37.7% (single fragment) to 23.9% (product rule). In contrast, the product rule increases the number of those posterior probability values **P** that fall into the third (from 6.2% (single fragment) to 14.4% (product rule)) and last quarter (from 0.9% (single fragment) to 12.3% (product rule)). In the case of protein fragments, the product combination rule also improves the values of the posterior probabilities **P**. For the first three quarters, the proportions of posterior probabilities **P** have decreased from 70.1%, 19.0% and 4.2% (single protein fragment) to 68.5%, 12.9% and 5.9% (product rule), respectively. Whereas, the proportion of posterior probabilities **P** obtained for the last quarter has increased from 6.6% (single protein fragment) to 12.6% (product rule). In both cases, the product combination rule has a clear positive influence on the posterior probability values obtained for correct atom types, while the sum rule has a relatively small impact.

In summary, our prediction analysis demonstrates that a combination of posterior probabilities given by multiple molecular fragments can be used to strengthen the atom type predictions. Moreover, the product combination rule turned out to be a more effective strategy than the more error tolerant sum rule. This suggests that the molecular fragments provide good posterior probability values, which need not be

corrected with averaging as is the case with the sum combination rule.

Application to the ligand-binding domain of glutamate receptors

Here, the present approach was put into practice by considering the 3D-structures of the ligand-binding domain (engineered S1S2 construct; Kuusinen et al. [50]) of the glutamate receptor subunit GluR2 complexed with the ligand kainate (PDB code 1gr2; Armstrong et al. [51]), AMPA (PDB code 1ftm; Armstrong and Gouaux [52]) and glutamate (PDB code 1ftj; Armstrong and Gouaux [52]). Predictive posterior probability maps were obtained as described in the Theory section.

S1S2 with bound glutamate. The ligand binding site with bound glutamate from the crystal structure 1ftj is shown in Figure 8(a). Favorable positions for the atoms of the ligand molecule were predicted assuming that the locations of water molecules in the binding site were known *a priori*. Consider, e.g., a case in which the positions of the structurally important water molecules are determined from the crystal structure of a protein complex and the aim is to design new ligands. However, knowledge about the ligand glutamate was not used during the prediction.

Multiple protein fragments in the binding site, including fragments of the amino acid arginine (Arg96; note that throughout we are using the amino acid numbering scheme present in the structure file 1ftj, but labels for water molecules follow the numbering scheme of the individual files), predict favorable spatial positions for hydrogen bond acceptor atoms (probability map logit($\mathbf{P}_\mathbf{f}^a$)) as indicated by the red density in Figure 8(a). The red density, plotted with a contour level of 10.0, overlaps with the actual location for the α-carboxylate oxygen atoms of the ligand glutamate, approximately 2.8 Å from the nitrogen atoms of Arg96 (side-chain guanido nitrogen atoms), threonine (Thr91; main-chain amide nitrogen atom) and serine (Ser142; main-chain amide nitrogen atom). Similarly, the red density overlaps with the actual position for one of the carboxylate oxygen atoms from the side-chain of the ligand glutamate, which is 3.2 Å from the main-chain amide nitrogen atoms of Ser142 and Thr143 in the crystal structure. The second side-chain carboxylate oxygen atom of the ligand contacts the side-chain hydroxyl oxygen of Thr143 in the crystal structure and it is not within the red region depicted in Figure 8(a). However, the carboxylate oxygen atom is embedded in the density map (not shown), if we depict it using a lower isocontour value.

The blue density, shown behind the ligand glutamate in Figure 8(a), marks favorable positions for hydrogen bond donor atoms, logit($\mathbf{P}_\mathbf{f}^d$), and was predicted by the oxygen atoms of proline (Pro89; main-chain oxygen atom), glutamic acid (Glu193; side-chain carboxylate oxygen atoms) and Thr91 (side-chain hydroxyl oxygen atom). The use of a contour level of 5.0 outlines a small region, which overlaps with the actual position for the $\boldsymbol{N}H_3^+$-nitrogen (α-nitrogen) atom of the ligand glutamate. In addition, the cyan density (logit($\mathbf{P}_\mathbf{f}^h$); contour level 15.0) predicts favorable regions for hydrophobic ligand atoms and overlaps with the actual positions for the five hydrophobic carbon atoms of the ligand glutamate. (Predicted atom types (hydrogen bond acceptor, hydrogen bond donor or hydrophobic atom) and the atoms of the ligand glutamate are shown in Table 3.)

In Figure 8(b), both the structure of the protein molecule (1ftj) and the position of the ligand molecule (glutamate) in the binding site are assumed to be known *a priori* and favorable positions for water oxygen atoms are subsequently predicted. (Consider, e.g., a case in which the ligand glutamate is fitted into the probability maps of Figure 8(a).)

The red density, logit($\mathbf{P}_\mathbf{f}^w$), is plotted with a contour level of 5.0 and overlaps with the actual positions for water oxygen atoms seen in the crystal structure, labeled as Wat63, Wat180, Wat249 in Figure 8(b). Each of these water oxygen atoms is in contact with one or the other of the side-chain carboxylate oxygen atoms of the ligand glutamate at a distance of approximately 3.0 Å. The position of the water oxygen atom Wat255, which is 3.6 Å away from the right-most oxygen atom of the ligand in Figure 8(b), locates close to a part of the red density.

In addition, the red density highlights a favorable region for a water oxygen atom above the α-nitrogen atom of glutamate. The structure file 1ftj does not report any water oxygens in that region. However, a water oxygen atom (green sphere labeled as Wat) would fit into this density. The hypothetical oxygen atom (green; Wat) is 2.9 Å from the α-nitrogen atom of the ligand glutamate (beneath Wat in Figure 8(b)). It is approximately 3.0 Å from the main-chain oxygen atom of Pro89, the side-chain carboxylate oxygen atom of Glu193 and the phenol $\boldsymbol{O}$H groups of two tyrosines (Tyr61 and Tyr220). The water oxygen atom (Wat) is 3.1 Å from the methyl carbon atom of methionine (Met196) and 3.4 Å from the methylene carbon atom

Table 3. Predicted atom types are related to the experimentally found ligand atoms.

Atom label[a]	Observed atom	Predicted atom type
Glutamate		
OXT	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
O	$-CO\boldsymbol{O}^{\ominus}$	Acceptor
C	$-\boldsymbol{C}OO^{\ominus}$	Hydrophobic
CA	$>\boldsymbol{C}H-$	Hydrophobic
N	$-\boldsymbol{N}H_3^{\oplus}$	Donor
CB	$-\boldsymbol{C}H_2-$	Hydrophobic
CG	$-\boldsymbol{C}H_2-$	Hydrophobic
CD	$-\boldsymbol{C}OO^{\ominus}$	Hydrophobic
OE1	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
OE2	$-CO\boldsymbol{O}^{\ominus}$	Acceptor
Kainate		
OXT	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
O	$-CO\boldsymbol{O}^{\ominus}$	Acceptor
C	$-\boldsymbol{C}OO^{\ominus}$	Hydrophobic
CA	$>\boldsymbol{C}H-$	Hydrophobic
N	$-\boldsymbol{N}H_2^{\oplus}-$	Donor
CD	$-\boldsymbol{C}H_2-$	Hydrophobic
CG	$>\boldsymbol{C}H-$	Hydrophobic
CG2	$>\boldsymbol{C}=$	Hydrophobic
CD1	$=\boldsymbol{C}H_2$	Hydrophobic
CD2	$-\boldsymbol{C}H_3$	Hydrophobic
CB	$>\boldsymbol{C}H-$	Hydrophobic
CB1	$-\boldsymbol{C}H_2-$	Hydrophobic
CG1	$-\boldsymbol{C}OO^{\ominus}$	Hydrophobic
OD1	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
OD2	$-CO\boldsymbol{O}^{\ominus}$	Acceptor
AMPA		
OT1	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
OT2	$-C\boldsymbol{O}O^{\ominus}$	Acceptor
C	$-\boldsymbol{C}OO^{\ominus}$	Hydrophobic
CA	$>\boldsymbol{C}H-$	Hydrophobic
N	$-\boldsymbol{N}H_3^{\oplus}$	Donor
CB	$-\boldsymbol{C}H_2-$	Hydrophobic
CG	$>\boldsymbol{C}=$	Hydrophobic
CD2	$>\boldsymbol{C}=$	Hydrophobic
CE2	$-\boldsymbol{C}H_3$	Hydrophobic
OE2	$-\boldsymbol{O}-$	Hydrophobic
CD1	$>\boldsymbol{C}=$	Hydrophobic
NE1	$-C\overset{\ominus}{O}\boldsymbol{N}-$	Acceptor
OE1	$-C\overset{\ominus}{\boldsymbol{O}}N-$	Acceptor

[a]The ligand atoms are labeled as in the individual coordinate files.

in the pyrrolidine ring of Pro89, too. Consequently, no serious interatomic bumps are detected and there is sufficient space for one water molecule at that location, although none is reported in the crystal structure. Taking into account this predicted water molecule and recalculating the probability map logit($\mathbf{P}_{\mathbf{f}}^{d}$) (the blue density in Figure 8(a)), we were able to further compress the region for the hydrogen bond donor atoms so that the actual position for the α-nitrogen atom of the ligand glutamate was trapped within the predicted density (not shown).

S1S2 with bound kainate. The ligand binding site with bound kainate from the crystal structure 1gr2 is shown in Figure 9(a). Again, the positions of the water oxygen atoms are assumed to be known *a priori* and they are kept at their actual positions seen in the crystal structure. The protein fragments predict favorable positions for ligand atoms functioning as hydrogen bond acceptors (red density), hydrogen bond donors (blue density) and hydrophobic contacts (cyan density); the results are shown in Figure 9(a).

The red density is plotted using a countour level of 10.0 and it overlaps with the positions for the two left-most carboxylate oxygen atoms of the ligand kainate (equivalent to the α-carboxylate group of glutamate). Interatomic contacts from the side-chain guanido nitrogen atoms of Arg96 and main-chain amide nitrogen atoms of Ser142, Thr91 and Thr143 to the carboxylate oxygen atoms of the ligand kainate are indicated with yellow dashed lines, each of which has a length between 2.8 and 3.2 Å. The blue density behind the ligand kainate is predicted by the oxygen atoms of Pro89 (main-chain oxygen atom), Glu193 (carboxylate oxygen atoms) and Thr91 (side-chain hydroxyl oxygen atom). The blue density is plotted with a contour level of 5.0 and it identifies the actual position for the $\boldsymbol{N}H_2^{+}$-nitrogen atom of the five-membered pyrrolidine ring of kainate.

Since the positions for the atoms of the amino acids Pro89, Thr91, Arg96, Ser142, Thr143 and Glu193 match well with the equivalent ones shown in Figure 8, the probability patterns for the hydrogen bond acceptor (red density) and hydrogen bond donor (blue density) atoms are expected to look very alike in Figures 8(a) and 9(a). However, the distance between the phenol oxygen atom of Tyr61 (in front of kainate) and the closest methyl carbon atom of leucine (Leu138) is 8.0 Å, whereas the corresponding distance in the glutamate-bound structure (1ftj) is only 5.3 Å. The altered side-chain conformations of Tyr61

and Leu138 provide extra space for the additional ligand atoms in kainate. As a result, the volume of the cyan density defining probable hydrophobic contacts, which is plotted with the same contour level used in Figure 8(a), has increased. This extra density is occupied by five additional carbon atoms present in the ligand kainate. (Predicted atom types and the atoms of the ligand kainate are related in Table 3.)

If we assume that the structure of the protein binding site together with the position of the ligand kainate is known in advance, it is possible to predict favorable positions for the water oxygen atoms using the molecular fragments obtained from the protein and from kainate. The probability map that predicts water oxygen atoms is illustrated with red density using a contour level of 6.0 (Figure 9(b)).

The water oxygen atoms, which are in potential contact with an atom of the ligand kainate, are considered and labeled as Wat1, Wat2, Wat36 and Wat85 in Figure 9(b). For each of these atoms, the nearest carboxylate oxygen atom of the ligand kainate is identified and indicated with one of the yellow dashed lines; interatomic distances seen in the crystal structure are 3.1, 2.8, 3.5 and 2.7 Å, respectively.

The interatomic distance (3.5 Å) from Wat36 to the nearest carboxylate oxygen atom of kainate exceeds the maximum distance value of 3.3 Å assigned for a pair of two oxygen atoms taken to be in contact (see the Theory section). Consequently, according to our definition, Wat36 is not in contact with atoms of the ligand kainate, nor is the reported position of Wat36 inside the predicted region. Nevertheless, if Wat36 is brought slightly closer to kainate, then all named water oxygen atoms in Figure 9(b), Wat1, Wat2, Wat36 and Wat85, overlap with the predicted red density. The other three water oxygen atoms (unlabeled) in Figure 9(b) are not in direct contact with the atoms of kainate, and are not considered here.

S1S2 with bound AMPA. The ligand binding site with bound AMPA from the crystal structure 1ftm is shown in Figure 10(a). As in Figures 8(a) and 9(a), the positions for the water oxygen atoms are known in advance and the molecular fragments obtained from the protein structure are used to predict the favorable positions for the hydrogen bond acceptor (red density), hydrogen bond donor (blue density) and hydrophobic (cyan density) atoms. The results are shown in Figure 10(a) using threshold values of 10.0, 5.0 and 15.0, respectively.

As in the previous cases, the nitrogen atoms of the residues Arg96, Thr91 and Ser142 participate in the prediction of a part of the red density on the left in Figure 10(a). The actual positions for the oxygen atoms of the carboxylate group (equivalent to the α-carboxylate group of glutamate) of the ligand AMPA match with this density. The hydroxyl oxygen atom of AMPA overlaps with the red density if a lower contour value for the density is used (not shown). A part of the red density on the right of Figure 10(a), located in front of the main-chain amide nitrogen of Glu193, overlaps with the *negatively-charged* nitrogen atom of the isoxazole ring of the ligand AMPA. The blue density overlaps with the actual position of the $\boldsymbol{N}H_3^+$-nitrogen atom of AMPA and the cyan density encloses the actual positions of the seven carbon atoms of AMPA. (Atom type predictions for the ligand AMPA are summarized in Table 3.)

Finally, suitable positions for water oxygen atoms are predicted using the molecular fragments obtained from GluR2 with bound AMPA (1ftm) in Figure 10(b). The isocontours of the probability map are plotted with a density level of 3.0. The water oxygen atoms that are in contact with AMPA are labeled as Wat396, Wat397 and Wat398 and each of them overlaps with the predicted density.

Previously, we have shown that molecular fragments can be used to describe a protein-binding site in cases where the structure of a protein is not available, but the ligands for this protein are known (Rantanen et al. [25]). The above examples show that the same molecular fragments can be used to characterize the structure of unknown ligands when only the structure of the protein-binding site is available.

In summary, we have described the further development of a molecular interaction library that has been built to predict favorable non-bonded interatomic interactions between various biomolecules. During the collection of the molecular interaction data a definition for independent atom contacts and a revised contact atom classification were applied. The collected data have been symmetrized and the most suitable orientation of molecular fragments in the coordinate system has been chosen automatically. The molecular interactions have been modeled using Gaussian mixture models so that subjective information on the interactions can be incorporated in the parameter estimates of the models. The selection of the optimal model complexity was revised, too. Moreover, two commonly used rules have been used in order to combine atom type predictions given by multiple molecular frag-

ments and the accuracy of combined predictions was compared to those given by single fragments. In each case, we have demonstrated, using either exhaustive validation analyses or qualitative examples (e.g., in the case of the ligand-binding domain of glutamate receptors), that these carefully chosen improvements lead to effective prediction of molecular interactions.

A set of pre-calculated posterior probability maps for the most common molecular groups will be made publicly available on the World Wide Web in the near future.

Materials and methods

Data sets of molecular interactions

The known 3D structures of proteins in the Protein Data Bank (PDB) [24] were systematically screened for the non-bonded atom interactions as described in Rantanen et al. [25] Here, the data collection procedure is outlined.

Firstly, our archive of protein-ligand complexes was updated to include PDB entries released between June 2000 and March 2002. The archive includes structures determined by X-ray crystallography with a resolution better than 2.5 Å; files containing DNA/RNA were not considered. For convenience, the coordinates for each ligand and the surrounding protein atoms were placed into separate files, here numbering 21 000. Each of the archived ligand structures was divided into molecular fragments (see Figure 1 in Rantanen et al. [25]) and non-bonded interactions between the ligand fragments and protein atoms were collected into the training data set. Furthermore, an independent validation data set was produced that consists of all protein atom fragments and their non-bonded interactions with ligand atoms. These latter data were not used in any phase of training of the mathematical models presented in this work.

Secondly, in order to increase the amount of interaction data collected for sparsely observed molecular fragments, e.g., fragments with thiol and sulfide groups, the initial training data obtained from archived protein-ligand complexes was augmented by adding the molecular interaction data obtained from protein-protein interfaces, thus forming the final training data set. A representative subset of protein structures deposited in the PDB was obtained from the website: `http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html` (Dunbrack, R.L., Jr and Muquit, M., Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, March 2002 release) by requesting a set of polypeptide chains having a resolution better than 2.5 Å and at most 90% mutual sequence identity. From the resulting set, multi-chain structures were selected. This led to a final set of 178 protein structure files, consisting of 674 pairs of interacting polypeptide chains. Only molecular interactions between non-identical polypeptide chains were considered.

The protein-ligand and protein-protein complexes were analyzed as they are presented in their PDB files. Atom coordinates were not subject to crystallographic symmetry operations; hence, crystal contacts between asymmetric units were not considered.

Figures 2, 3, 8–10 were prepared using the molecular visualization programs Bodil (`http://www.abo.fi/fak/mnf/bkf/research/johnson/bodil.html`; Lehtonen, J.V., Rantanen, V.-V., Still, D.-J., Gyllenberg, M. and Johnson, M.S., unpublished), Molscript [53] and Raster3D [54].

Acknowledgements

We thank Olli T. Pentikäinen (Department of Biochemistry and Pharmacy, Åbo Akademi University, Finland) and Dr Per-Ola Norrby (Department of Organic Chemistry, Technical University of Denmark) for discussion about glutamate receptors. This work has been supported by grants from Tekes (the National Technology Agency of Finland), the Academy of Finland, the Swedish Research Council, the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi) and the National Graduate School of Informational and Structural Biology (ISB).

References

1. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
2. Wade, R.C., Clark, K.J. and Goodford, P.J., J. Med. Chem., 36 (1993) 140.
3. Wade, R.C. and Goodford, P.J., J. Med. Chem., 36 (1993) 148.
4. Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Des., 5 (1991) 545.
5. Danziger, D.J. and Dean, P.M., P. Roy. Soc. Lond. B Biol., 236 (1989) 101.
6. Danziger, D.J. and Dean, P.M., P. Roy. Soc. Lond. B Biol., 236 (1989) 115.
7. Laskowski, R.A., Thornton, J.M., Humblet, C. and Singh, J., J. Mol. Biol., 259 (1996) 175.
8. Pitt, W.R. and Goodfellow, J.M., Protein Eng., 4 (1991) 531.
9. Böhm, H.J., J. Comput.-Aided Mol. Des., 6 (1992) 61.

10. Böhm, H.J., J. Comput.-Aided Mol. Des., 6 (1992) 593.
11. Böhm, H.J., J. Comput.-Aided Mol. Des., 8 (1994) 623.
12. Bruno, I.J., Cole, J.C., Lommerse, J.P., Rowland, R.S., Taylor, R. and Verdonk, M.L., J. Comput.-Aided Mol. Des., 11 (1997) 525.
13. Verdonk, M.L., Cole, J.C. and Taylor R., J. Mol. Biol., 289 (1999) 1093.
14. Nissink, J.W.M., Verdonk, M.L. and Klebe, G., J. Comput.-Aided Mol. Des., 14 (2000) 787.
15. Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V. and Willett, P., J. Mol. Biol., 307 (2001) 841.
16. Boer, D.R., Kroon J., Cole, J.C., Smith, B. and Verdonk, M.L., J. Mol. Biol., 312 (2001) 275.
17. Klebe, G., J. Mol. Biol., 237 (1994) 212.
18. Verkhivker, G., Appelt, K., Freer, S.T. and Villafranca, J.E., Protein Eng., 8 (1995) 677.
19. Mitchell, J.B.O., Laskowki, R.A., Alex, A. and Thornton, J.M., J. Comput. Chem., 20 (1999) 1165.
20. Mitchell, J.B.O., Laskowki, R.A., Alex, A., Forster, M.J. and Thornton, J.M., J. Comput. Chem., 20 (1999) 1177.
21. Muegge, I. and Martin, Y.C., J. Med. Chem., 42 (1999) 791.
22. Gohlke, H., Hendlich, M. and Klebe, G., J. Mol. Biol., 295 (2000) 337.
23. Hendlich, M., Acta Crystallogr. D, 54 (1998) 1178.
24. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
25. Rantanen, V.-V., Denessiouk, K.A., Gyllenberg, M., Koski, T. and Johnson, M.S., J. Mol. Biol., 313 (2001) 197.
26. Bernardo, J.M. and Smith, A.F.M., Bayesian Theory, John Wiley and Sons, Chichester, UK, 1994.
27. McLachlan, G.J. and Krishnan, T., The EM Algorithm and Extensions, John Wiley and Sons, New York, 1997.
28. McLachlan, G.J. and Peel, T., Finite Mixture Models, John Wiley and Sons, New York, 2000.
29. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.J., Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.
30. Lanterman, A.D., Int. Stat. Rev., 69 (2001) 185.
31. Li, A.-J. and Nussinov, R., Proteins, 32 (1998) 111.
32. Rantanen, V.-V., Gyllenberg, M., Koski, T. and Johnson, M.S., Bioinformatics, 18 (2002) 1257.
33. Bondi, A., J. Phys. Chem., 68 (1964) 441.
34. Böhning, D., Schlattman, P. and Lindsay, B.G., Biometrics, 48 (1992) 283.
35. Ewens, W.J. and Grant, G.R., Statistical Methods of Bioinformatics, Springer Verlag, New York, 2001.
36. Geiger, D. and Heckerman, D., Ann. Stat., 25 (1997) 1344.
37. Gyllenberg, M. and Koski, T., Math. Biosci., 177&178 (2002) 161.
38. Geiger, D. and Heckerman, D., Ann. Stat., 30 (2002) 1412.
39. Gauvain, J.-L. and Lee, C.-H., IEEE T. Speech Audi. P., 2 (1994) 291.
40. Hastie, T. and Tibshirani, R., J. Roy. Stat. Soc. B Met., 58 (1996) 158.
41. Rissanen, J., IEEE T. Inform. Theory, 42 (1996) 40.
42. Rissanen, J., J. Comput. Syst. Sci., 55 (1997) 89.
43. Figueiredo, M. and Jain, A.K., IEEE T. Pattern Anal., 24 (2002) 381.
44. Wallace, C.S. and Freeman, P.R., J. Roy. Stat. Soc. B Met., 49 (1987) 241.
45. Wallace, C.S. and Freeman, P.R., J. Roy. Stat. Soc. B Met., 54 (1992) 195.
46. Samudrala, R. and Moult, J., J. Mol. Biol., 275 (1998) 895.
47. Chou, P.Y. and Fasman, G.D., Biochemistry, 13 (1974) 211.
48. Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J., IEEE T. Pattern Anal., 20 (1998) 226.
49. Tax, D.M.J., van Breukelen, M., Duin, R.P.W. and Kittler, J., Pattern Recogn., 33 (2000) 1475.
50. Kuusinen, A., Arvola, M. and Keinänen, K., EMBO J., 14 (1995) 6327.
51. Armstrong, N., Sun, Y., Chen, G.Q. and Gouaux, E., Nature, 395 (1998) 913.
52. Armstrong, N. and Gouaux, E., Neuron, 28 (2000) 165.
53. Kraulis, P.J., J. Appl. Crystallogr., 24 (1991) 946.
54. Merritt, E.A. and Bacon, D.J., Methods Enzymol., 277 (1997) 505.