Åbo Akademi
University

Riku Hakulinen

Probabilistic contact preferences in protein-ligand and protein-protein complexes

PROBABILISTIC CONTACT PREFERENCES IN PROTEIN-LIGAND AND PROTEIN-PROTEIN COMPLEXES

Riku Hakulinen

Department of Natural Sciences
Åbo Akademi University
Åbo, Finland, 2013

Department of Natural Sciences
Åbo Akademi University
Åbo, Finland

Supervised by

Professor Jukka Corander,
Åbo Akademi University,
Åbo

Pre-examiners

Docent Olli Pentikäinen
Nanoscience center,
Jyväskylä University,
Jyväskylä

and

PhD Gerd Wohlfahrt,
Computer-Aided Drug Design,
Orion Pharma,
Espoo

ISBN 978-952-12-2954-1 (Hard copy)
ISBN 978-952-12-2955-8 (Electronic)
Painosalama Oy – Åbo, Finland 2013

Contents

Acknowledgements

The research for this thesis was conducted at the mathematics department of Åbo Akademi University between January 2008 and June 2013. Financial support from the Academy of Finland and Sigrid Jusélius Foundation is gratefully acknowledged. I wish to thank the math department for all their durable support and I am grateful to thesis supervisor Jukka Corander for patient guidance. I would also like to thank pre-examiners Olli Pentikäinen and Gerd Wohlfahrt for their thorough work and useful comments. The idea for the research in this thesis originated in the Department of Biochemistry and Pharmacy of Åbo Akademi University and I am thankful for that starting point and for the helpful discussions with professor Mark Johnson, Santeri Puranen and Jukka Lehtonen, just to mention a few. Finally, I would like to acknowledge my family for family life, which is fun and gives a lot of refreshment.

Introduction

Structural data of biomolecules and biomolecular complexes has been accumulating in databases since the 1970's. Today, both quality and diversity of the data is starting to allow reliable 3D models to be built based on information contained in the coordinate files. These models are then used for various tasks like identifying binding sites, predicting preferred binders and, in general, as tools for quantifying inter- and intramolecular interactions.

Most of the structural data has been collected with X-ray diffraction method, and to a lesser, but possibly increasing extent using nuclear magnetic resonance (NMR), electron microscopy and neutron diffraction. The experimental methods and processing of the raw measurement data generate error in the structure files [1]. Statistical modeling allows taking into account the uncertainty in the data. Chemical and physical knowledge is used for guidance to avoid unrealistic schemes and to reduce computational load in applying the model.

In this work, a Bayesian statistical framework was utilised to build a probabilistic model that is used as a new type of knowledge-based scoring function in assessing preference and relative strength of molecular contacts. This scoring function is novel in the sense that it takes into account also directional data, not only distance data, like knowledge-based scoring functions traditionally have done. Recent reviews of the subject can be found in [2], [3], [4].

The physico-chemical factors that create the directional nature of intermolecular interactions and the fragmentation of molecules needed to capture all relevant structural information in the model are reviewed in the first chapter of this thesis. The second chapter provides an overview of the statistical modeling approach utilised in the development of the scoring function. The third chapter demonstrates use of the model by applying it in a biological context, and the fourth chapter corresponds to the research article presenting a proof of concept for the model, which is further developed and tested in other chapters of this thesis.

Chapter 1

The 3D nature of interaction preferences

1.1 Types of interactions between molecules

Intermolecular interactions are ultimately formed from electrostatic attraction and repulsion between electrons and nucleii of the interacting molecules. Dynamical aspects like oscillations of the electron cloud produce a time averaged electrostatic field. Two standard references on the subject are [5] and [6]. An overview is given in the textbook [7], where also the types of molecular complexes and functional groups relevant for this work can be found. In this section, short definitions of what creates intermolecular forces are listed to provide background details for the subsequent discussion on molecular fragments.

1.1.1 Polar

Different electronegativities of elements produce permanent electrical dipoles in molecules. Interactions between permanent dipoles form strong non-covalent bonds, like the hydrogen bond. A special case of hydrogen bonding is a water bridge, water molecule mediated double hydrogen bond between, for example, two amino acid residues.

1.1.2 Hydrophobicity and van der Waals contacts

Hydrophobic effect is the agglomeration of nonpolar molecules, or parts of molecules, in aqueous solution to minimize the nonpolar structure contact area with water. A van der Waals interaction (vdWi) is the attraction between transient electrical dipoles created by synchronized fluctuations in the electron clouds of the molecules. A vdWi is in principle always present and for individual fragments it is much weaker than a polar interaction. It can be quantified as binding energy per unit surface area.

1.1.3 Ionic

Starting from the electrically neutral form of a functional group, protonated and deptrotonated forms have a net electric charge. Electrostatic attraction and repulsion between these forms, together with interactions involving ionized metal atoms like divalent magnesium (Mg^{2+}), represent a generic class called ionic interactions. The strongest non-covalent bonds are ionic.

1.1.4 Polar-neutral and charged-neutral

Permanent electrical dipoles can induce polarization in electrically neutral structures, which leads to attraction between the permanent and induced dipole.

1.2 Fragmentation of molecules

Modeling the interactions of a molecule with its environment, is in this approach segmented into interactions relating to individual fragments. The fragments are typically either a part of, or contain a functional group. Definition of a fragment type, with an orientation in three-dimensional space, requires specifying at least three atoms for the fragment. The fragmentation is utilised such that contact data for each fragment type is collected and the implied contact patterns are coded in three-dimensional coordinate probability densities and distributions of *a priori* information. Contact patterns of a larger molecular fragment, e.g. the entire molecule, are built combining the patterns for individual fragments. An atom alleged to be in contact with the fragment is here called a target atom.

1.2.1 Fragment class specific prior information

Chemical properties of a molecular fragment are the main form of *a priori* information used for model guidance. These properties include, for example, electronegativity of the fragment atoms and whether or not the fragment is part of an aromatic structure. Also, theoretically known aspects like directional preferences are important, especially when the amount and quality of contact atom position data are limited. The fragment classification used in this work is presented in article [17], where the developed probabilistic model is published, and also in Table 1.1. The target atom classes are given in Table 1.2.

Choosing the classification requires balancing between coverage of chemical space and sufficiency of the alleged amount of contact data for each fragment class. The former determines, not only the range of applicability for the classification, but also the starting point for possible merging of classes based on equivalent contact patterns. In this work, the classification contains most of the frequently encountered fragment types, in training data set, as defined with some broad definition of a molecular setting, like an aromatic structure. This allows testing the method to verify its usefulness, and also lays the basis for further development.

Class	**Description**	**Class**	**Description**
f2	Hydroxyl O / aliphatic	**f18**	Fluorine / non-arom.
f3	Hydroxyl O / aromatic	**f20**	Chlorine / aromatic
f5	Carbonyl O (\ **f9,f10**)	**f21**	Chlorine / non-arom.
f6	Carboxyl O	**f22**	N in aromatic (w/o subst.)
f7	Carbamoyl O	**f18**	Fluorine / non-arom.
f8	Phosphate group O	**f23**	N in non-arom. planar
f9	Amide O / non-arom.	**f26**	Amino (prim.) N / non-arom.
f10	Amide O / aromatic	**f27**	Amino (prim.) N / aromatic
f11	Secondary C in aromatic	**f29**	Amino (prim.) N / planar
f12	Secondary C in non-arom.	**f34**	Bromine / aromatic
f13	Primary carbon	**f35**	Bromine / aliphatic
f17	Fluorine / aromatic	**f36**	Iodine / aromatic

Table 1.1: Fragment classes used in this study. Main forms of intermolecular interaction for these fragment types are hydrogen bonding, dispersion, charged group based electrostatic and halogen bonding. The fragment classification was partly adopted from the previous work of Rantanen et al. [48] (see chapter Article). Here common slash ('/') means bonded to and backslash means excluding. Symbols are used for oxygen, carbon and nitrogen.

Class	**Description**
C3	Carbon of a methyl group
C4	Alpha carbon
C5	Carbon in an aromatic structure
C6	Sulfur of a thioether group
C7	Sulfur of a thiol group
C8	Nitrogen of an amide group
C9	Nitrogen of indole, imidazole and guanido groups
C10	Nitrogen of an amino group
C11	Oxygen of a carboxamide group
C12	Oxygen of a carboxyl group
C13	Oxygen of a hydroxyl group
C14	Main chain carbonyl oxygen
C15	Main chain amide nitrogen

Table 1.2: Classification of contact atoms, or targets. The target classification was partly adopted from the previous work of Rantanen et al. [48] (see chapter Article).

1.3 Specificity and energetics in intermolecular contacts

Examples of chemically compatible structures are molecular fragments that have deformable electronic distributions, and two functional groups with opposite net charge. Non-covalent and covalent bonds are formed between compatible structures. When two molecules form a complex with largely compatible contacts, they can be called structurally and chemically complementary.

1.3.1 Complementarity

In biochemical reactions like enzymatic catalysis, the catalytic site and the reactant are highly complementary, especially in the transition state [7]. Relating to this, a type of therapeutic molecule candidate is the so called transition state analogue that binds tightly to the active site of an enzyme, preventing it from performing catalysis. Finding a suitable candidate molecule based on complementarity, is called molecular docking [4], [8]. Evaluation of the preference of the docked molecule, which reflects its binding strength and is useful for assessing relative affinity, is the task of a scoring function.

1.3.2 Distance and direction dependent preferences

Distance between interacting atoms, or more precisely, between the average positions of two nucleii, can be taken as determining the strength of the interaction. This assumption is used as the basis for traditional knowledge-based scoring functions that estimate binding strength in molecular docking [2]. Docking is part of the computerized stage of drug discovery, and an efficient scoring function is a potent factor in reducing the work load of the wet laboratories testing binding affinities of candidate molecules [4].

Directional data of the contact atom distributions contain relevant information with respect to, not only the plain contact preferences, but also to strength of the binding, as discussed in the next section. When probabilistic preferences are combined with molecular ensemble properties, also relative affinity could be evaluated.

1.3.3 Correspondence with quantum mechanics

In terms of molecular orbital theory, the electron cloud of a molecule is described with bonding and nonbonding electron orbitals. These are in one-to-one correspondence with electron densities, and when molecules are in close proximity, the densities contribute through their 3D structure to the total energy of the contact. The densities reform in the process, so that the system of nucleii and electrons approach the minimum energy conformation continuously. In other words, electrons mediate interactions between nucleii of separate molecules as part of long distance forces, discussed briefly in the first section of this chapter, for a detailed treatment see [5].

The positions of hydrogen atoms in X-ray diffraction based structures are not usually known, and even if they were known, the positions would not be as confined with respect to the rest of the structure, as they are for the heavy atoms. This is because hydrogens are an order of magnitude lighter, so their covalent bonds reorient readily with changing enviroment. Therefore, hydrogens bonded to heavy atoms are included in the model implicitly. Protons are considered to be mediators like electrons, though heavier and carrying opposite charge.

Taken that the fragment classification is non-redundant, the distance and direction dependent distribution of relative positions of the nucleii, together with molecular ensemble properties influencing physical quantities like temperature, is interpreted to contain all relevant information to quantify the strength of the contact. In other words, it reveals the basic form of the potential energy function governing the relative motion of nucleii during the intermolecular contact. This reasoning takes advantage of the well-known fact that intermolecular interactions can be considered energetically to be on the border of quantum and classical mechanics, so that concepts from both sides can be applied. Namely, quantum mechanics is needed for the description of electronic structure of the molecule, and as already discussed, it is the electron distribution that dictates the directional preferences of the intermolecular contacts, through lone electron pairs, p-orbitals and other basic features. A molecular complex is not a static structure though, because thermal energy brings about nuclear motion, and this kinetic energy is confined in the multidimensional potential energy landscape of the complex. The motion of the nucleii allows a classical description, in the potential created by electrons and single protons. In conclusion, assuming a nonredundant molecular fragment classification, the three-dimensional probability densities that capture relative positions of nucleii correspond to potential energy wells. The quantum mechanical formation of an effective potential in molecules is next reviewed.

The adiabatic or Born-Oppenheimer approximation

The fundamental assumption behind the approach described in this section, that the relative positions of nucleii suffice for describing the form of potential energy of the interactions between molecules, is based on the so called adiabatic approximation, see e.g. [9] for details. It is also known as Born-Oppenheimer approximation and in case of molecules it states that the electronic motion can be separated from nuclear motion. This is justified by that the electrons are much lighter than nucleii and therefore move in the mutual electric field much faster, so much faster that their distribution can be taken to adapt to the nuclear motion without delay. Many textbooks on quantum mechanics have an exposition of this approximation included [10], [11]. The separation of motions is formalized by factoring the joint state of electrons and nucleii to a product of nuclear and electronic states, also parametrising the electronic states with nuclear coordinates of a fixed molecular conformation. The outcome is a group of equations for nucleii, moving in the effective potential created by electrons. This standard procedure traditionally models nuclear motion as part of a molecule, and is in this work used in the framework of relative motion between molecules in a complex.

Intermolecular motion

Deformations of the electronic clouds during an intermolecular contact depend on the molecular orbitals of the free molecules, and on the types of contacting molecular surfaces, and on the distance between contacting atoms. Also the hydrogen, or single proton positions are taken here to continuously adjust between the molecules in relative motion. These changes in the electron and proton arrangements generate the potential of interaction, time-average of which is used in intermolecular contact modeling. Typical forms of these potentials can be found for example in [12].

The frequency of the intermolecular vibrations are commonly found in the far infrared region, located between 300 GHz and 3 THz, whereas intramolecular vibrations are typically in the near infrared region, i.e. from 30 to 300 THz [13], [14]. Teraherz (THz) corresponds to 10^{12} Herz. In other words, vibrational motion between molecules is on average slower than internal vibrations, which is tentatively taken here to allow also protons to adjust their positions to intermolecular motion instantly. These arguments are the basis for applying the adiabatic approximation to intermolecular motion, including protons in addition to electrons as mediators forming the potential.

Additivity of the individual fragment contributions

A relevant issue relating to fragment based modeling is whether the contributions of individual fragments can be straightforwardly added to give a meaningful estimate for a larger portion of a molecule, for example, an entire ligand or an area of protein surface. A recent study of this subject with respect to fragment specific contributions of energy is published in [15] and there it is concluded that contributions can not automatically be taken as additive. The reason for this is that biochemistry takes place in water based solution, and the hydration state of a binding site, or contact site, has a role in the binding process. The working assumption here is that effects of the hydration state manifest themselves through the probabilistic contact preferences described in this work, when water molecules involved in binding are modeled explicitly, but it is acknowledged that this needs to be studied further to assess if a more complicated than strictly additive approach should be preferred.

Formalism

This section is concluded by a formal presentation of intermolecular adiabatic approximation as applied to intermolecular interactions. The quantum mechanical derivation of the equation for intramolecular nuclear motion in an average potential can be found in many textbooks on quantum mechanics, for example [9], [10], [16].

Our exposition starts from the time-independent Schrödinger equation:

$$\hat{H}(\{\bar{r}\},\{\bar{R}\})\Psi(\{\bar{r}\},\{\bar{R}\}) = E\Psi(\{\bar{r}\},\{\bar{R}\}). \tag{1.1}$$

The variables $\{\bar{r}\}$ and $\{\bar{R}\}$ in (1.1) represent electronic and nuclear positions, respectively. On the right hand side of the equation, E is the total energy of the contact. The total energy operator, Hamiltonian $\hat{H}$, is composed of several terms:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN}, \tag{1.2}$$

where $\hat{T}$ is a kinetic energy operator and $\hat{V}$ a potential energy operator. Subscript e refers to electrons and subricpt N to nucleii. According to the central dogma of non-relativistic quantum mechanics, the wave function contains all information about the system described by the Hamiltonian operator. Part of the adiabatic approximation is factoring the wave function Ψ to a product of nuclear Θ and electronic Φ wave function:

$$\Psi(\{\bar{r}\},\{\bar{R}\}) = \Theta(\{\bar{R}\})\Phi(\{\bar{r}\};\{\bar{R}\}), \tag{1.3}$$

where Φ is parametrized with relative nuclear positions $\{\bar{R}\}$. When function (1.3) is used in equation (1.1), taking into account that the kinetic energy operators depend only on either $\{\bar{r}\}$ or $\{\bar{R}\}$, formally solving the electronic equation $(\hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee})\Phi = E_e(\{\bar{R}\})\Phi$, and finally integrating over the degrees of freedom of the mediators (electrons and protons), the equation for nuclear motion in the average potential $(\hat{E}_{e^-p^+})$ of the contact is obtained:

$$\hat{T}_N\Theta(\{\bar{R}\}) + \hat{E}_{e^-p^+}(\{\bar{R}\})\Theta(\{\bar{R}\}) = E\Theta(\{\bar{R}\}). \tag{1.4}$$

Reminding that a quantum mechanical wave function is in a one-to-one correspondence with a probability density, the wave function for relative, intermolecular nuclear motion $\Theta(\{\bar{R}\})$ is considered to have a functional connection

$$f \equiv |\Theta|^2. \tag{1.5}$$

In other words, the squared norm of the wave function for nuclear motion coincides with the three-dimensional probability density f obtained by modeling the structural data of the molecular contacts. Interpreting (1.5) together with (1.4) from the side of f, the atomic positions defined by distance and direction, that attain the highest probability density values correspond to those atomic configurations that are located deepest in the potential well of the contact. This concludes the part of the work describing conceptual foundations for using structural coordinate data modeling probability densities as knowledge-based scoring functions. The discussion is next moved on to the statistical model used for capturing information in the structural coordinates.

Chapter 2

Statistical modeling

This chapter gives first a review of the basic statistical tools and concepts used in the model developed in this work and published in [17]. Contents of the publication are also given in chapter Article in this book. Second, a possible future development step, a study of a predictive version for the model is presented. These are followed by exemplifying results, which are compared with the corresponding results from the presently in use parametric version of the model [17].

2.1 Bayesian data analysis

The product of probability density point values, determined for a set of observed data points, is called a likelihood function. A statistical model can be guided using theoretical and empirical facts or modeler's beliefs that do not depend on the observed raw data. In the framework of Bayesian data analysis, this can be done, for example, through the so called conjugate prior densities by forming a product with the likelihood. Parameters of the prior densities, the hyperparameters, hold the external information and are used to calculate regularized estimates for the model parameters. The level of regularization can be adjusted by one or more parameters of the prior density. The update procedure creates a probability density of the parameters, called the posterior density, which can then be used to determine the regularized estimates and also different measures of the uncertainty in the estimate. These work as a test, for correctness of the modeler's beliefs, or for the support that a theoretical *a priori* description of the modeled system attains from the data. The posterior density has the same functional form as the conjugate prior density and it may be used as a prior for subsequent rounds of update with new data.

This interplay between external information and data is useful, and many times necessary, when the datasets are not exhaustive but limited, which they often are. Useful references on the subject include [18], [19].

2.1.1 Chemical *a priori* information

Modeling of molecular processes allows chemical information about the alleged types of interactions between molecules to be included in the form of experimental observations and theoretical results. An example of theoretical information is a value for the angle between a reference direction of a fragment and an average contact direction, of which the latter can be estimated from known directional aspects of molecular orbitals. Another level of guidance, in addition to the prior distributions, emerges when the model contains submodels based on a classification that requires weights for the classes. The weights can represent plausibility of a contact type for a molecular fragment class.

2.2 Choosing the densities

The raw data on which this probabilistic model was trained, is atom position coordinates in three-dimensional space, mainly collected from the Protein Data Bank [20]. Starting point for choosing the present structure of the model, was that the three-dimensional probability density to be used, should be flexible enough to capture the details of the contact atom cloud and still be readily fitted to data. The rationale behind this was to allow fast training of the model, several times, in order to investigate the properties of coordinate data sets. This way, a workable model is obtained, a model that also works as a platform for future development, one example of a possible development is presented in section Predictive model, later in this chapter. The most straightforward way to achieve the above mentioned goal was to use interconnected one-dimensional parametric densities. They are straightforward to fit to the data of the corresponding variable and their parameters can be updated with simple formula. Some complexity comes from that, in principle, each individual density requires $n \times m \times l$ terms, where (n, m, l) are numbers of modes for the distributions, corresponding to each of the three spherical polar coordinates. Numbers n, m and l are typically less than or equal to three, so that a reference maximum number of terms in the modeling probability density is 27. This number can be reduced by relying on the characteristic features in the 3D contact data collected for molecular fragments that have a uniquely defined spatial orientation.

2.2.1 Simplifying the functional form through regularities in data

The type of probability density described in this section, can be made simpler by using observed or alleged regular features in the data.This means, e.g., that it may be possible to combine modes of two variables, like distance and polar angle. Combining modes of these variables starts from the assumption that when a contact atom cloud genuinely has a polar angle distribution with more than one peak, or mode, the distribution with respect to distance follows the same pattern, because both reflect the strength of the bond. This assumption

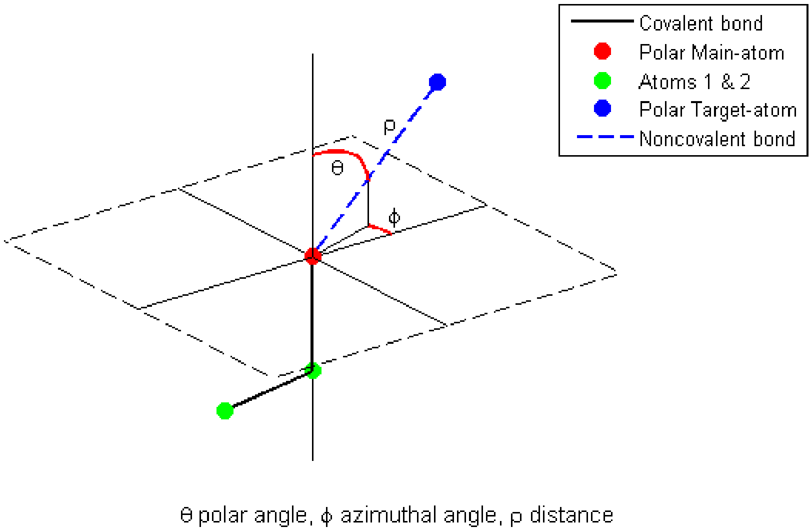


Figure 2.1: A possible polar contact geometry showing the use of spherical polar coordinates to describe the position of the target atom with respect to the fragment.

is based on chemistry like the double polarization of halogens [21], and can be observed in the data.

Distributions of the third spherical polar coordinate variable, azimuthal angle, were modeled separately for each combined polar angle and distance mode, because no similar argument was found for connecting the azimuthal angle modes with those of either of the other two variables.

The number of modes for either polar angle or distance can be determined first and then used for the other variable, because possible multimodality of distance distribution will imply interaction with electrons on different molecular orbital. Though, it was found that when the count of contact atom positions for a fragment is relatively low, but large enough to be informative, starting form the distance distribution gives more robust parameter estimates.

The model

Building the model starts from presenting the distribution of contact atom positions in spherical polar coordinates, i.e. distance ρ, polar angle θ and azimuthal angle ϕ. A graphical presentation of the geometry of an exemplifying fragment-target contact is given in Figure 2.1.

A one-dimensional probability density was fitted to data distributed along each coordinate separately. The piecewise defined densities obtained: f, g and h, can now be used to put together a three-dimensional density describing the cloud of observed contact atom positions. Generic functional form of the density p, is

$$p(\rho,\theta,\phi \mid \overline{\Theta}) = \sum_{i=1}^{N_i} \frac{f_i(\rho \mid \overline{\Theta}_i)}{\rho^2} \times \left[\sum_{j=1}^{N_{ij}} \frac{g_{ij}(\theta \mid \overline{\Theta}_{ij})}{\sin(\theta)} \times \left[\sum_{k=1}^{N_{ijk}} h_{ijk}(\phi \mid \overline{\Theta}_{ijk}) \right] \right] . \tag{2.1}$$

In equation (2.1), vectors $\overline{\Theta}$, $\overline{\Theta}_i$, $\overline{\Theta}_{ij}$ and $\overline{\Theta}_{ijk}$ symbolize the parameters of the corresponding density, and N_i, N_{ij} and N_{ijk} are numbers of density function modes. The number of modes for the polar angle density in (2.1) is distance density mode specific and the number of modes for azimuthal angle density is polar angle density mode specific. The order in which variables appear in the chain of dependences is exchangeable in this generic form. Correction terms ρ^2 and $\sin(\theta)$ in (2.1) are required in compiling the three-dimensional density, following from that the atom location distributions along each variable, i.e. spherical polar coordinate, are modeled as one dimensional variables. Another equivalent choice would be to use the correction terms to the data points before fitting probability densities. Modeling data was in this work done using the approach corresponding to equation (2.1).

The form of the density (2.1) can be simplified when, for example, the rationale discussed earlier in this section holds, namely that there are prevalent regularities allowing a preferred order of dependensies to be defined, together with possibly removing a dependence of two variables.

2.3 Updating parameters with new data

This section demonstrates the use of conjugate prior densities for a model of the form given in equation (2.1).

2.3.1 Conjugate prior densities

A conjugate prior probability density, or a probability mass function in case of a discrete prior, is such that the parameter update procedure leads to a posterior probability density, or a probability mass function in case of a discrete prior, that has the same functional form as the prior density. The prior is then said to be closed under sampling.

The parameter update procedure is here demonstrated formally using one of the one-dimensional densities in the model, the von Mises distribution:

$$
\begin{aligned}
&f(\lambda, \kappa|\lambda_1\left(\{x_i\}_{i=1,..,n}, \lambda_0\right), \kappa_1\left(\{x_i\}_{i=1,..,n}, \kappa_0\right)) = \\
&= \frac{1}{K_1 * I_0^{n+c}(\kappa)} \exp(\kappa_1 * \cos(\lambda - \lambda_1)) = \\
&= \left[\Pi_{i=1}^{n} \frac{1}{K * I_0(\kappa)} * \exp(\kappa * \cos(x_i - \lambda))\right] * \frac{1}{K_0 * I_0^{c}(\kappa)} \exp(\kappa_0 * \cos(\lambda - \lambda_0)),
\end{aligned}
\tag{2.2}
$$

where the left hand side (LHS) is the posterior density, K_x are normalizing constants and $(c, \kappa_0 = R_{0*}\kappa, \lambda_0)$ are the hyperparameters. I_0 is the modified Bessel function of zeroth order. This may be written in the form:

$$
\begin{aligned}
&f(\lambda, \kappa|\lambda_1, R_1 = \frac{\kappa_1}{\kappa}) \propto \frac{1}{I_0^{n+c}(\kappa)} * \qquad (2.3) \\
&* \exp\left(\kappa\left\{\cos(\lambda)[\Sigma_{i=1}^{n} \cos(x_i) + \frac{\kappa_0}{\kappa}\cos(\lambda_0)] + \sin(\lambda)[\Sigma_{i=1}^{n} \sin(x_i) + \frac{\kappa_0}{\kappa}\sin(\lambda_0)]\right\}\right).
\end{aligned}
$$

The definitions for the updated parameters can be found from several sources in the literature: e.g. [22], [23], but suitable values for the hyperparameters, like the concentration κ_0 and therefore the ratio $R_0 = \frac{\kappa_0}{\kappa}$, are still required. Parameter R_0 fixes the chosen level for guidance by the external information stored in λ_0 - the larger R_0, the more λ_0 is emphasized. The third hyperparameter c represents uncertainty in the given value of κ_0, the larger it is, the more certain κ_0 is taken to be.

Following from equations (2.2) and (2.3) it is deduced that the updated parameters, i.e. the posterior density parameters, are found in the form:

$$
\kappa_1 = \kappa\left[\left(\Sigma_{i=1}^{n} \cos(x_i) + \frac{\kappa_0}{\kappa}\cos(\lambda_0)\right)^2 + \left(\Sigma_{i=1}^{n} \sin(x_i) + \frac{\kappa_0}{\kappa}\sin(\lambda_0)\right)^2\right]^{\frac{1}{2}},
\tag{2.4}
$$

and

$$
\begin{aligned}
\lambda_1 &= \arctan\left(\frac{\Sigma_{i=1}^{n} \sin(x_i) + \frac{\kappa_0}{\kappa}\sin(\lambda_0)}{\Sigma_{i=1}^{n} \cos(x_i) + \frac{\kappa_0}{\kappa}\cos(\lambda_0)}\right) = \\
&= \arctan\left(\frac{n * \overline{\{\sin(x_i)\}} + R_0 \sin(\lambda_0)}{n * \overline{\{\cos(x_i)\}} + R_0 \cos(\lambda_0)}\right),
\end{aligned}
\tag{2.5}
$$

where n is the amount of data points and, in the latter form for $\lambda_{1,}$ the result is given using arithmetic means of the sum terms.

Equations (2.4) and (2.5) show that choosing ratio $\frac{\kappa_0}{\kappa}$ large enough, $\kappa_1 - > \kappa_0$ and $\lambda_1 - > \lambda_0$, or choosing it small enough, the hyperparameters of the prior density would not have influence on posterior parameter values.

The posterior density is then used for inference about the paramerters, for example, by determining the so called maximum a posteriori, or MAP estimates for the parameters. In this study, a MAP estimate for the direction parameter λ was used and it is straightforwardly defined as λ_1. The MAP estimate of the concentration parameter κ is calculated using condition

$$\left(\frac{\partial f(\lambda, \kappa \mid \lambda_1, R_1)}{\partial \kappa}\right)_{\lambda=\lambda_1} = 0. \tag{2.6}$$

The resulting equation is

$$\frac{I_0(\kappa_M)}{I_1(\kappa_M)} = \frac{c+n}{R_1}, \tag{2.7}$$

where R_1 is the updated ratio of concentration parameters for λ and θ. The solution κ_M is the maximum a posteriori estimate for κ. Parameter sum $c+n$ balances the effect of R_1, as c balanced R_0 in the prior, so that the density of κ is concentrated around a larger than zero value for κ, in a non-symmetrical way. The prior and the posterior have the same functional form in the case of a conjugate prior, and equation determining the peak of the prior density with respect to κ, has the same form as equation (2.7), namely

$$\frac{I_0(\kappa_m)}{I_1(\kappa_m)} = \frac{c}{R_0}. \tag{2.8}$$

Equation (2.8) shows how one can define numerically the hyperparameters c and R_0 so that the value for κ, corresponding to the peak (κ_m), gets a pre-chosen reference value, a suggestion based on modeler's beliefs. When the posterior peak value κ_M stays relatively close to the prior peak value κ_m after one or more updates, it can be concluded that κ_m is a reasonable choice, possibly to be used also in other equivalent modeling cases. The MAP estimates $\lambda_1$and κ_M can be used as parameters of the coordinate density for calculations, in addition to being used as prior parameters in consequtive update procedures.

In this work, a justified way to determine κ_m from chemical information was not yet found, so a formal update for κ was used instead to avoid unnecessary numerical calculations. The update procedure in general starts from the functional form of the joint probability density for the coordinates and the parameters. The joint density p was factorized in the following way:

$$p(\theta, \lambda, \kappa \mid \lambda_0, \kappa_0, c) \propto p_1(\theta \mid \lambda, \kappa) * p_2(\lambda \mid \lambda_0, \kappa_0) * p_3(\kappa \mid \kappa_0, c), \tag{2.9}$$

where von Mises density of the polar angle is

$$p_1 = \frac{e^{\kappa * \cos(\theta-\lambda)}}{2\pi * I_0(\kappa)}, \tag{2.10}$$

and prior density for expected direction λ is

$$p_2 = \frac{e^{\kappa_0 * \cos(\lambda-\lambda_0)}}{2\pi * I_0(\kappa_0)}, \tag{2.11}$$

and prior density for concentration κ around expected direction λ is

$$p_3 = \frac{1}{I_0^c(\kappa - \kappa_0)}. \tag{2.12}$$

So, in this case κ_0, representing concentration of λ around λ_0, is a predefined constant that does not depend on κ, and the corresponding prior p_3 is symmetrical in the interval $[0,2\kappa_0]$ with respect to κ_0. In other words, κ_0 describes how much the *a priori* information on the direction parameter is emphasized, and can be defined for each modeled system separately, e.g., based on the count of data points available for modeling. Hyperparameter c has in this form of implementing external information, the role of being a measure of modeler's belief in the expected concentration κ_0. The posterior density parameter κ_1 is still a function of κ, though the dependence is no longer linear. Approximate linearity appears when the ratio $\frac{\kappa_0}{\kappa}$ can be considered small, as seen from equation (2.4). In practise, estimates for κ used in p_1 for calculations, were determined numerically directly from data, i.e., from the observed positions of the contacting atoms.

The prior density for κ in (2.12) is not normalized, like p_1 and p_2 are, and the constant needed to normalize the joint density (2.9) is found by integrating $\frac{p_3(\kappa)}{I_0(\kappa)}$ from 0 to ∞, and the result will depend on c. After decision is made on how *a priori* information is used, one proceeds to the update (2.2) and also possibly further, for example, to integrating out parameters like it is done when forming a predictive model, explored next.

2.4 Predictive model

A representative three-dimensional structure of the contact atom distribution is an essential component for this type of approach to modeling molecular interactions. It can be emphasized by taking into account simultaneously all contributions with accepted parameter values, with a relative weight for each contribution. This corresponds to calculating a predictive probability density and is accomplished by integrating out the parameters from the joint probability density like in eq. (2.9), possibly first updating the hyperparameters of the model. A versatile treatment of predictive inference can be found in [24].

2.4.1 Integrating out the parameters

The predictive method utilizes a marginal density, which is obtained by integrating the fragment class (f) - target class class (C) -specific joint probability density over domains of the parameters. In the model used in this work, the random variable parameters were $\{\mu_\rho, \kappa_\theta, \lambda_\theta, \mu_\phi\}$, and the joint density, without correction terms μ_ρ^{-2} and $\sin(\lambda_\theta)^{-1}$, has the form:

$$
\begin{aligned}
&J_{fC}(\rho,\theta,\phi,\mu_\rho,\kappa_\theta,\lambda_\theta,\mu_\phi|\overline{\Theta}) \propto \\
&\propto \sum_{i=1}^{N_i} \exp(-\frac{1}{2\sigma^2_{\rho,i}}(\rho-\mu_\rho)^2) \times \exp(-\frac{1}{2\sigma^2_{\rho,post,i}}(\mu_\rho-\mu_{\rho,MAP,i})^2)\times \\
&\times \frac{\exp(\kappa_\theta\cos(\theta-\lambda_\theta))}{I_0(\kappa_\theta)} \times \frac{\exp(\kappa_{\theta,post,i}\cdot\cos(\lambda_\theta-\lambda_{\theta,MAP,i}))}{I_0^{n_i}(\kappa_\theta-\kappa_{\theta,post,i})}\times \\
&\times \left(\sum_{j=1}^{N_{ij}} \exp(-\frac{1}{2\sigma^2_{\phi,ij}}(\phi-\mu_\phi)^2) \times \exp(-\frac{1}{2\sigma^2_{\phi,post,ij}}(\mu_\phi-\mu_{\phi,MAP,ij})^2)\right),
\end{aligned}
\tag{2.13}
$$

where $I_0^{n_i}$ is the modified bessel function of the first kind, raised to (n_i):th power and n_i is a measure of the modeler's belief in the correctness of the concentration parameter estimate $\kappa_{\theta,i}$. It is seen from the definition of the function (2.13) that κ_θ is formally a random variable.

All parameters that are not treated as random variables, are described with one symbol $\overline{\Theta}$. In the hyperparameter subscripts, it is shown which are maximum a posteriori (subscript MAP) estimates and which are posterior (subscript *post*) parameters without being MAP estimates. The rest of the parameters, i.e. standard deviations for distance and azimuthal angle, were defined numerically directly from data. The deviations of the modeling, or coordinate, density were chosen not to be handled as random variables, because no meaningful way has not yet been found to determine *a priori* values for them from external data. The same applies, as discussed in previous section, to concentration for polar angle κ_θ.

The aimed at marginal density is evaluated as the integral (2.14) and has the functional form (2.15):

$$
p(\rho,\theta,\phi) = \int_0^\infty \int_0^\pi \int_0^{2\pi} \int_0^{\rho_{\max,fC}} J_{fC}(\rho,\theta,\phi,\mu_\rho,\kappa_\theta,\lambda_\theta,\mu_\phi|\overline{\Theta}) d\mu_\rho d\mu_\phi d\lambda_\theta d\kappa_\theta
\tag{2.14}
$$

$$
\begin{aligned}
p(\rho,\theta,\phi) \propto \sum_{i=1}^{N_i} \exp(-\frac{1}{2\sigma^2_{\rho,pred,i}}(\rho-\mu_{\rho,pred,i})^2)\times \\
\times\{\pi \cdot I_0(\kappa_{\theta,post,i}\cdot\sqrt{(c^2_{\kappa,i}+1)+2\cdot c_{\kappa,i}\cdot\cos(\theta-\lambda_{\theta,MAP,i})})+ \\
2\cdot\sum_{k=1}^{\infty} I_k(\kappa_{\theta,post,i}\cdot\sqrt{(c^2_{\kappa,i}+1)+2\cdot c_{\kappa,i}\cdot\cos(\theta-\lambda_{\theta,MAP,i})})\cdot \\
\cdot\frac{(-1)^{k+1}+1}{k}\cdot\sin(k\cdot a\tan(\frac{\mid\sin\theta\mid+\mid\sin\lambda_{\theta,MAP,i}\mid}{\mid\cos\theta\mid+\mid\cos\lambda_{\theta,MAP,i}\mid}))\}\times \\
\times\{\sum_{j=1}^{N_{ij}}\exp(-\frac{1}{2\sigma^2_{\phi,pred,ij}}(\rho-\mu_{\phi,pred,ij})^2)\}.
\end{aligned}
\tag{2.15}
$$

In derivation of the function (2.15) were needed equation 9.6.34 from [25] and the additivity of the normal density, through which $\mu_{\rho,pred,i}=\mu_{\rho,map,i}$, $\sigma^2_{\rho,pred,i} = 2\cdot\sigma^2_{\rho,post,i}$, $\mu_{\phi,pred,ij}=\mu_{\phi,map,ij}$ and $\sigma^2_{\phi,pred,ij}=2\cdot\sigma^2_{\phi,post,ij}$, see e.g. [26]. The approximate additivity of the vonMises density, described in [26], could not be used, following from the definition of the interval of the polar angles ($\theta\in[0,\pi]$) in spherical polar coordinates. Outside this interval, in practice when $\theta<0$, polar angle does not obtain values, and consequently the integral (2.14) is not a product of full convolutions giving symmetric results with respect to the expectation values $\{\lambda_{\theta,MAP,i}\}$. The same restriction of the variable values to an interval is true for the distance ρ, though the situation is different, because the data is strongly concentrated on the third third of the interval $\rho\in[0,r_{cutoff}]\cdot\mathring{A}$ and the restriction has an effect only for values of ρ beyond the cutoff distance, not below zero. On the other hand, the cutoff is an artificial limit, see Table 2.1, used in collecting the data, and therefore possible influence of any beyond cutoff values in the form of the predictive density can be considered reasonable, or meaningful, and the additivity is expected to hold. So, the predictive distance density has the form of a normal mixture and is normalized to the interval $[0,r_{cutoff}]\cdot\mathring{A}$. In case of azimuthal angle then, the support is cyclic in the interval $[0,2\pi]$ and the densities are defined without any cutting, from which it follows that the additivity is naturally there.

A noteworthy part in the evaluation of integral (2.14) is the double integral with respect to the parameters of the polar angle θ, namely κ_θ and λ_θ, where the following step was taken:

$$\int_0^\infty \frac{d\kappa_\theta}{I_0(\kappa_\theta)\cdot I_0^{n_i}(\kappa_\theta-\kappa_{\theta,post,i})}\cdot$$
$$\cdot\{\pi\cdot I_0(\sqrt{\kappa_\theta^2+\kappa_{\theta,post,i}^2+2\cdot\kappa_\theta\cdot\kappa_{\theta,post,i}\cdot\cos(\theta-\lambda_{\theta,MAP,i})})+$$
$$+2\cdot\sum_{j=1}^{\infty}I_j(\sqrt{\kappa_\theta^2+\kappa_{\theta,post,i}^2+2\cdot\kappa_\theta\cdot\kappa_{\theta,post,i}\cdot\cos(\theta-\lambda_{\theta,MAP,i})})\cdot$$
$$\cdot\frac{(-1)^{j+1}+1}{j}\cdot\sin(j\cdot a\tan(\frac{\mid\sin\theta\mid+\mid\sin\lambda_{\theta,MAP,i}\mid}{\mid\cos\theta\mid+\mid\cos\lambda_{\theta,MAP,i}\mid}))\}\approx$$
$$\approx a\cdot\{\pi\cdot I_0(\sqrt{(c_{\kappa,i}^2+1)\cdot\kappa_{\theta,post,i}^2+2\cdot c_{\kappa,i}\cdot\kappa_{\theta,post,i}^2\cdot\cos(\theta-\lambda_{\theta,MAP,i})})+$$
$$+2\cdot\sum_{j=1}^{\infty}I_j(\sqrt{(c_{\kappa,i}^2+1)\cdot\kappa_{\theta,post,i}^2+2\cdot c_{\kappa,i}\cdot\kappa_{\theta,post,i}^2\cdot\cos(\theta-\lambda_{\theta,MAP,i})})\cdot$$
$$\cdot\frac{(-1)^{j+1}+1}{j}\cdot\sin(j\cdot a\tan(\frac{\mid\sin\theta\mid+\mid\sin\lambda_{\theta,MAP,i}\mid}{\mid\cos\theta\mid+\mid\cos\lambda_{\theta,MAP,i}\mid}))\}. \tag{2.16}$$

The new parameters in equation (2.16), i.e. a and $c_{\kappa,i}$, are determined numerically for each polar angle density mode i and are based on the trapezoidal approximation of the integral (2.16): a is half the base of the area-approximating triangle and $c_{\kappa,i}$ is the relation of the peak with respect to $\kappa_{\theta,post,i}$, see Figure 2.2. The specific value of the parameter a is not very useful, because the result in (2.16) is only proportional to the predictive density, but a is kept here to illustrate how the integral was evaluated. The absolute values in the last term of eq. (2.16) are for crossing the point $\theta=\frac{\pi}{2}$ so, that the values of the integral stay positive. The curves of Figure 2.2 are calculated using values $\theta=0$ and $\lambda_{\theta,MAP,i}=0$. Integrand's asymmetric form that prevails for small values of n_i, like for the red curve with $n_i=5$, follows from that the term $\frac{1}{I_0^{n_i}(\kappa_\theta-\kappa_{\theta,post,i})}$ concentrates the integrand, around $\kappa_{\theta,post,i}$, only after the number of observations supporting $\kappa_{\theta,post,i}$, namely n_i, is high enough, as seen in Figure 2.2.

Also function $\frac{1}{I_0(\kappa_\theta)}$ in the integrand of (2.16) influences the position of the peak, which is here described through condition

$$\frac{1}{n_i}\times\frac{I_1(\kappa_{\theta,peak})}{I_0(\kappa_{\theta,peak})}=\frac{I_1(\kappa_{\theta,peak}-\kappa_{\theta,post,i})}{I_0(\kappa_{\theta,peak}-\kappa_{\theta,post,i})}, \tag{2.17}$$

solution of which ($\kappa_{\theta,peak}$) gives the location of the peak for function

$$\frac{1}{I_0(\kappa_\theta)*I_0^{n_i}(\kappa_\theta-\kappa_{\theta,post,i})}. \tag{2.18}$$

I_1 is the modified Bessel function of the first kind and of order one. Equation (2.17) shows that already for modest values for n_i, $\kappa_{\theta,post,i}$ will be an approximation for the location of the peak, because the LHS ratio of modified Bessel functions gets values between zero and one, and the right hand side (RHS) ratio is close to zero when $\kappa_{\theta,peak}$ is close to $\kappa_{\theta,post,i}$.

The sum over orders of the modified Bessel function in eq. (2.16), though it extends to infinity, converges with a few more terms than $j \approx \kappa_{\theta,post,i}$, and $j_{\max} = \kappa_{\theta,post,i} + 5$ is used. The convergence is depicted in Figure 2.3.

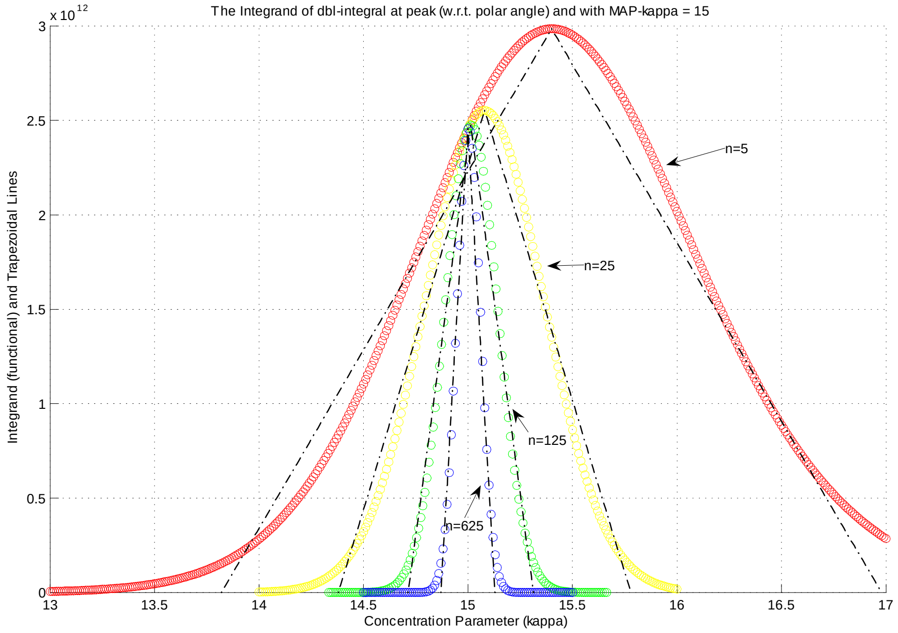


Figure 2.2: Form of the integrand for concentration κ, with four values of n, i.e n_i: 5, 25, 125 and 625. Shown also the straight lines (dot-slash) of the trapezoidal method, defining triangles that approximate the value of the definite integral. The asymmetric form of the integrand is seen for the smallest concentration ($n = 5$).

The purpose of the exercise in this section was to demonstrate a possible way to have an approximate fuctional form for the predictive density and to show how the choices for hyperparameters affect the procedure. A functional form for the predictive density is useful for determining and comparing the characteristics, like numbers of genuine modes, of the three-dimensional shapes of the contact patterns, given in the predictive form by the marginal densities for the coordinates.

Multimodalities in the predictive density

The apparent number of modes of the one dimensional densities in equation (2.15) , i.e. N_i and N_{ij}, are here defined by first building a kernel estimate [27]

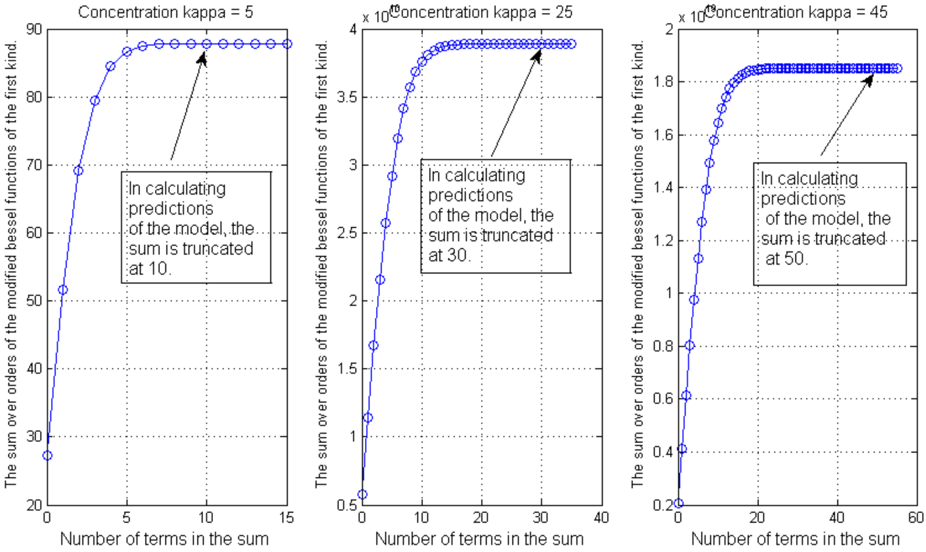


Figure 2.3: Parameter density with respect to concentration κ, convergence of the sum over orders of the modified bessel function. Three exemplifying concentrations: $\kappa = \{5; 25; 45\}$. The sum is considered to have converged when the result is flattened.

Fragment type\Target Class	**C3 - C7**	**C8 - C15**
f2, f3, f5, f8, f22, f23, f26, f27, f29	3.5 Å	3.3 Å
f17, f18, f20, f21, f34, f35, f36, f37	3.9 Å	3.9 Å
f11, f12, f13	3.7 Å	3.7 Å
f6, f7, f9, f10	3.3 Å	3.3 Å

Table 2.1: The cutoff distances used in collecting the training data set. See chapter Article, or Table 1.1, for description of the classification.

of the density, then finding the number of modes in the kernel estimate. After this stage, using the information obtained, fitting von Mises and Normal distributions is done mode by mode to find estimates for the hyperparameters of the conjugate prior distributions used in the update procedure, and for the parameters for which a MAP estimate is not calculated, i.e. $\{\sigma^2_{\rho,i}, \sigma^2_{\phi,ij}\}$. The next step is then to use the update procedure on $\{\mu_\rho, \kappa_\theta, \lambda_\theta, \mu_\phi\}$ to find the parameters of the posterior distributions, $\{\mu_{\rho,MAP,i}, \kappa_{\theta,post,i}, \lambda_{\theta,MAP,i}, \mu_{\phi,MAP,ij}\}$, see eq. (2.13). The joint density is the starting point for calculating the predictive density.

The primary goal of using the kernel estimate in the model construction is not to find a perfect fit, but to deduce the apparent number of modes in the one dimensional densities used in building the actual 3D model. The number of modes was selected by a simple stability criterion in a loop going through a pre-set range of bandwidth values, namely $h = [\pi/200, 2\pi/200, 3\pi/200, ..., 32\pi/200]$ for polar angle and $h = [\pi/50, 2\pi/50, 3\pi/50, ..., 32\pi/50]$ for azimuthal angle. The stability criterion used was that the number of modes and minima in the kernel estimate stays constant during three consecutive rounds of the loop. The third variable, distance, is handled similarly, but with the criterion that relating to certain polar angle mode, the kernel estimate has exactly one mode. The bandwidth is one from $h = [maxdist/100, 2 * maxdist/100, 3 * maxdist/100, ..., 1]$, where $maxdist$ equals the cutoff distance (see Table 2.1), used when collecting the training data set for the model.

The multimodalities in eq. (2.15), i.e. N_i and N_{ij}, are apparent, because some of the mixture components in eq. (2.13) will merge with a neighboring component. This is allowed, because the main point is that genuine multimodality is taken into account. Of course, when the merged components are updated separately, the amount of observations used for each component is smaller than it would be with merged components, but on the other hand, the kernel estimate has in these cases predicted higher multimodality, and as the observed data accumulates, the individual components might well get smaller variances and shifted peaks, so that the higher multimodality arises in the modeling density as well. The procedure is illustrated in Figures 2.4, 2.5, 2.6 and 2.7.

These one-dimensional densities are connected to each other so that first the number of peaks with respect to the polar angle is determined. Then the multimodality of the azimuthal angle density is defined for each polar angle von Mises density separately. For example, in Figure 2.7, the azimuthal angle

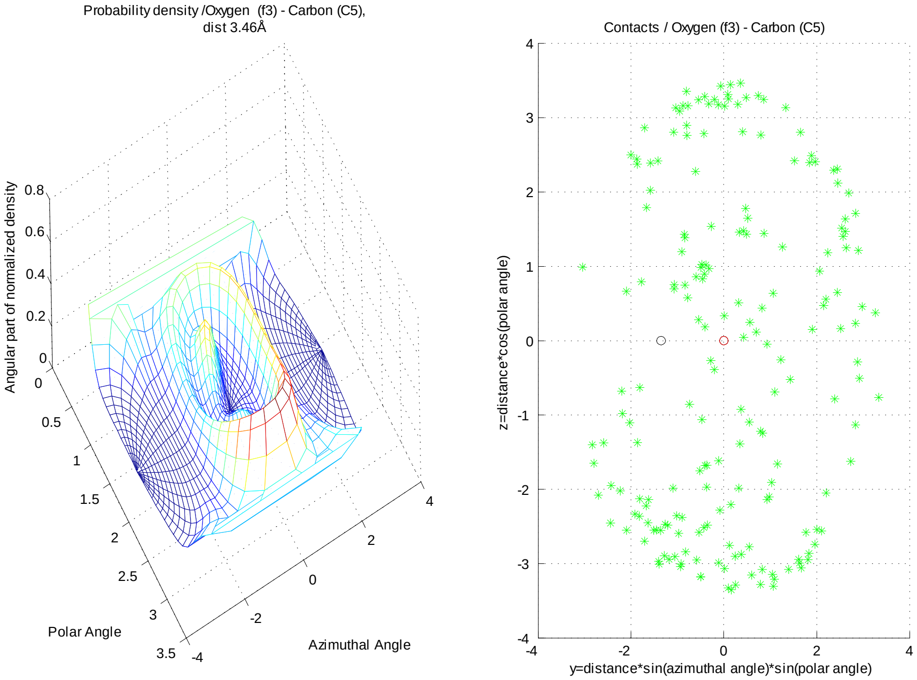


Figure 2.4: Left: The form of the angular part of the modeling density at 3.46 Å. Right: Training data as a scatter plot. The circle in (0,0) is the main atom of the fragment. The left and right figures are oriented so that they can be matched visually by thinking them as superimposed. The fragment class **f3** denotes a hydroxyl oxygen bonded to an aromatic structure and target class **C5** denotes a carbon in an aromatic structure.

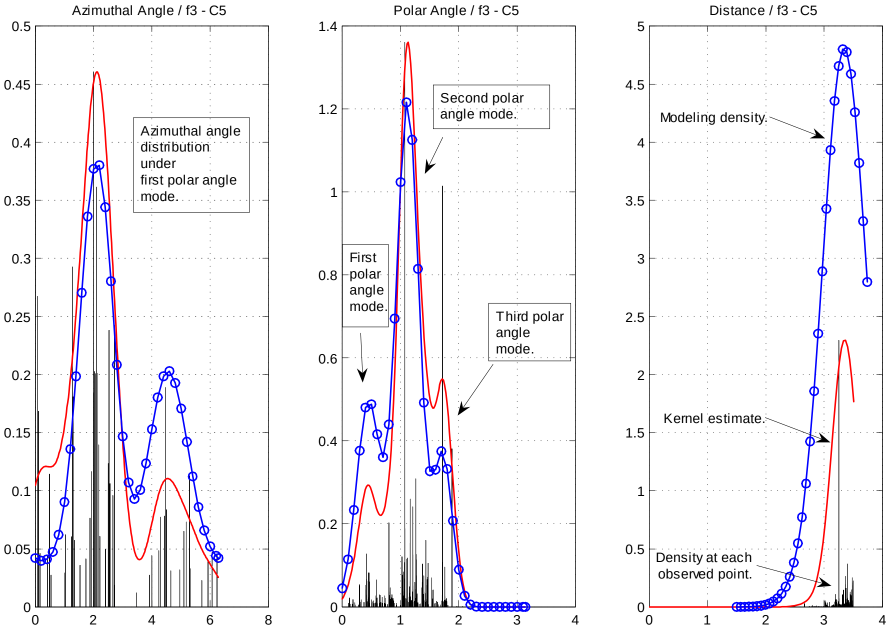
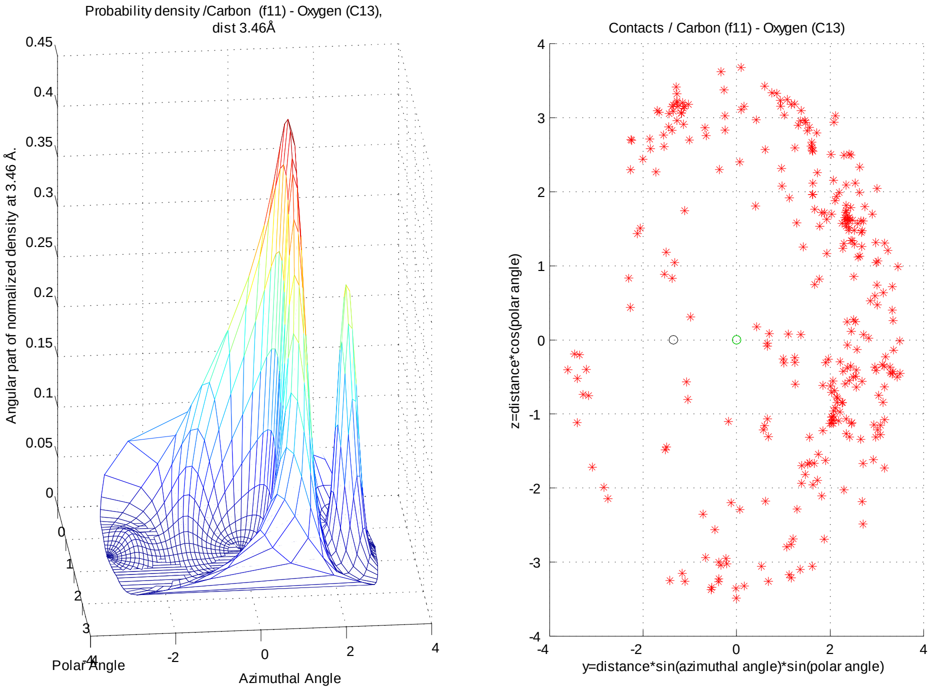


Figure 2.5: Estimation of the number of modes. The red line is the kernel estimate built on normal kernels with adapted weights and numerically defined band width. The blue circles joined with line represent the normalized modeling mixture density, which has either Normal (azimuthal angle and distance) or vonMises (polar angle) distributed components. The bars are density values defined for each observed point separately, i.e. a histogram based on intervals of varying length. The fragment class **f3** denotes a hydroxyl oxygen bonded to an aromatic structure and target class **C5** denotes a carbon in an aromatic structure.



Probability density /Carbon (f11) - Oxygen (C13), dist 3.46Å

Contacts / Carbon (f11) - Oxygen (C13)

Figure 2.6: Left: The form of the angular part of the modeling density at 3.46 Å. Right: Training data as a scatter plot.The circle in (0,0) is the main atom of the fragment. The left and right figures are oriented so that they could be matched visually, by thinking them superimposed. The fragment class **f11** denotes a carbon in an aromatic structure and target class **C13** denotes a hydroxyl oxygen.

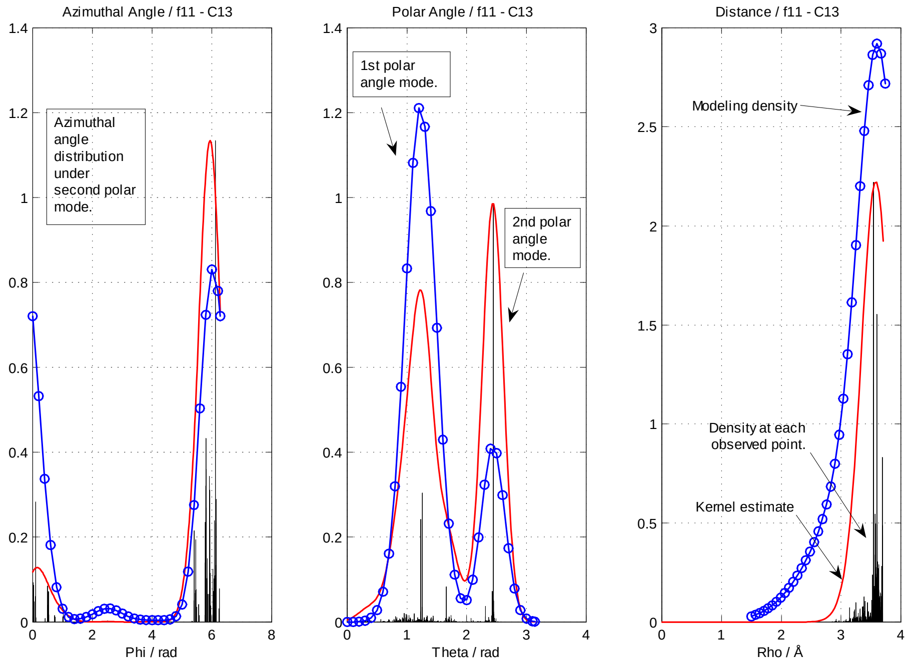


Figure 2.7: Estimating the number of modes. Red line is the kernel estimate built on normal kernels with adapted weights and numerically defined band width. The circles joined with line represent the normalized modeling mixture density, which has either Normal (azimuthal angle and distance) or vonMises (polar angle) distributed components. The bars are density values defined for each observed point separately, i.e. a histogram based on intervals of varying length. The fragment class **f11** denotes a carbon in an aromatic structure and target class **C13** denotes a hydroxyl oxygen.

density in the subfigure on the left is related to the RHS peak of the subfigure in the middle. The histograms in Figures 2.7 and 2.5 are constructed by calculating the size (length) of the surroundings for each observed value of the variable: for each point x_i is given density $N_i^{(x)} = (\frac{1}{2}(x_{i+1} - x_{i-1}))\ ^{-1}$, excluding the end points for which it is $N_i^{(x)} = (x_{i+1} - x_i)\ ^{-1}$. These densities $\{N_i^{(x)}\}_{i=1,\dots,n}$ are then used as weights in calculating the kernel estimate KE:

$$KE(x) = \frac{1}{h}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}N_i^{(x)}} \exp\{-\frac{1}{2}(\frac{x-\mu_i}{h})^2\}, \tag{2.19}$$

where h is the bandwidth and $\{\mu_i\}_{i=1,\dots,n}$ are the experimentally observed values of the variable x.

Computational cost

In this section so far, a procedure for forming the predictive density has been outlined. In order to ensure that the three-dimensional shape of the contact atom position density is captured as reliably as possible, a more advanced routine is needed, which inevitably increases the computational cost.

2.4.2 Comparison with results obtained using parametric densities

Given here next are the results obtained from calculating the same tasks using parametric and predictive forms of the model.

Calculating the contact preferences

The contact preferences are defined as probability masses confined to a certain spatial area in the environment of a specific type of molecular fragment, see Figure 2.8 and Table 4.2. The probability masses are calculated for each target class and then the results are normalized to give the fragment type specific target atom hierarchy, or the contact preferences. Details are given in the next two sections.

Reference points

The spatial area in which the probability mass is calculated, can be a small, or a large part of the surroudings of the fragment - from an effectively point like volume element to the entire spatial area where the target atoms are to be found. The former gives probability masses that approach zero as the volume approaches a single point, and the latter equals 'fragment class-target class' - specific prior probability. In the examples here, we have taken the volume to be a spatial area around a reference point (see Figure 2.8), and the spatial area is defined as intervals of the spherical polar coordinates: $\{[\rho_1, \rho_2], [\theta_1, \theta_2], [\phi_1, \phi_2]\}$.

The reference points, with corresponding intervals, used both in the article where this model is published [17], and here, are

$$\begin{aligned}\bar{r}_{ref}^{(1)} =& \; 2.8\mathring{A} * [\cos(\frac{\pi}{3}), \cos(\frac{19\pi}{10}) * \sin(\frac{\pi}{3}), \sin(\frac{19\pi}{10}) * \sin(\frac{\pi}{3})], \\ & \text{with intervals } [2.800 \pm 0.075\mathring{A}, \frac{\pi}{3} \pm \frac{\pi}{12}, \frac{19\pi}{10} \pm \frac{\pi}{10}]. \end{aligned} \quad (2.20)$$

$$\begin{aligned}\bar{r}_{ref}^{(2)} &= 3.40\mathring{A} \cdot [\cos(0), \cos(\phi_{arbitrary}) * \sin(0), \sin(\phi_{arbitrary}) * \sin(0)] = \\ &= 3.40\mathring{A} \cdot [1, 0, 0], \text{ with intervals } [3.400 \pm 0.15\mathring{A}, 0 + \frac{\pi}{5}, \pi \pm \pi]. \end{aligned} \quad (2.21)$$

$$\begin{aligned}\bar{r}_{ref}^{(3)} &= 2.8\mathring{A} * [\cos(\frac{\pi}{2}), \cos(\frac{3\pi}{2}) * \sin(\frac{\pi}{2}), \sin(\frac{3\pi}{2}) * \sin(\frac{\pi}{2})] = \\ &= 2.8\mathring{A} * [0, 0, -1], \text{ with intervals } [2.800 \pm 0.1\mathring{A}, \frac{\pi}{2} \pm \frac{\pi}{12}, \frac{3\pi}{2} \pm \frac{\pi}{12}]. \end{aligned} \quad (2.22)$$

$$\begin{aligned}\bar{r}_{ref}^{(4)} &= 3.3\mathring{A} * [\cos(\frac{\pi}{2}), \cos(\frac{3\pi}{2}) * \sin(\frac{\pi}{2}), \sin(\frac{3\pi}{2}) * \sin(\frac{\pi}{2})] = \\ &= 3.3\mathring{A} * [0, 0, -1], \text{ with intervals } [3.300 \pm 0.1\mathring{A}, \frac{\pi}{2} \pm \frac{\pi}{12}, \frac{3\pi}{2} \pm \frac{\pi}{12}]. \end{aligned} \quad (2.23)$$

The volumes are shown graphically in Figure 2.8. A method for evaluating the probability masses inside the volumes around the reference points is presented next.

Riemann sums

The predictive density has three-dimensional spatial form of the function (2.15), which is used for calculating the probability mass p corresponding to a certain contact atom type in a spatial area, or a volume, from the environment of a molecular fragment. The calculation is here done using a Riemann sum [28]:

$$\begin{aligned} p = &\left\{ \sum_{i=1}^{N_\rho} f(\frac{\rho_{i+1} + \rho_i}{2}) \cdot (\rho_{i+1} - \rho_i) \right\} \times \quad (2.24) \\ &\times \left\{ \sum_{j=1}^{N_\theta} g(\frac{\theta_{j+1} + \theta_j}{2}) \cdot (\theta_{j+1} - \theta_j) \right\} \times \left\{ \sum_{k=1}^{N_\phi} h(\frac{\phi_{k+1} + \phi_k}{2}) \cdot (\phi_{k+1} - \phi_k) \right\} = \\ &= \Delta\rho \cdot \Delta\theta \cdot \Delta\phi \cdot \sum_{i=1}^{N_\rho} f(\frac{\rho_{i+1} + \rho_i}{2}) \times \sum_{j=1}^{N_\theta} g(\frac{\theta_{j+1} + \theta_j}{2}) \times \sum_{k=1}^{N_\phi} h(\frac{\phi_{k+1} + \phi_k}{2}), \end{aligned}$$

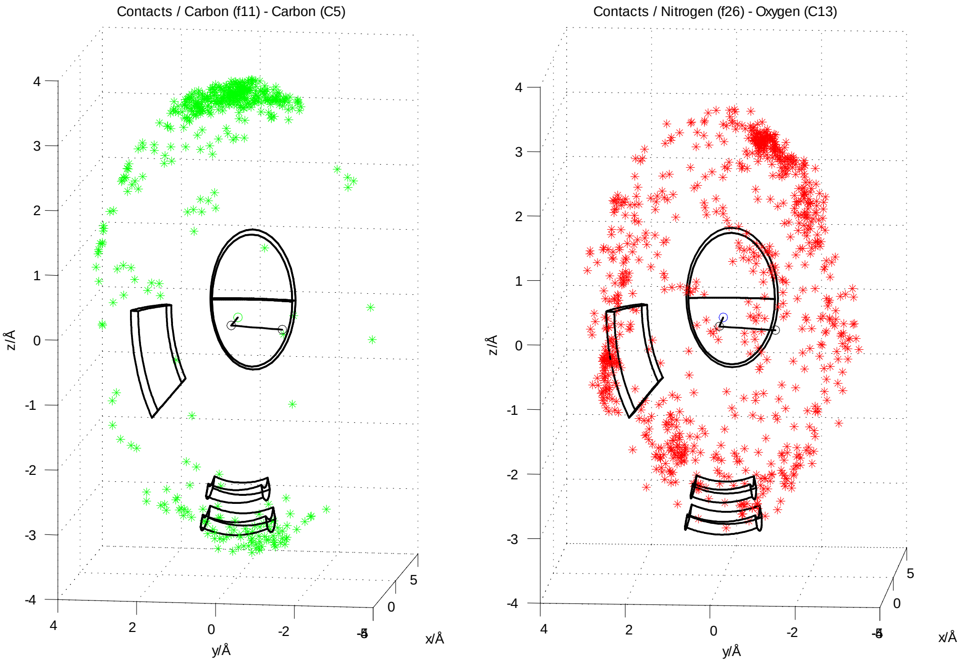


Figure 2.8: The volumes used in examples 1-4 of section The probability masses. The volumes are shown with respect to two different scatterplots. Left: Aromatic carbon contacts (**C5**) of the Aromatic carbon fragment (**f11**). Right: Hydroxyl oxygen contacts (**C13**) of the Primary amino nitrogen fragment (**f26**). Reference points are located in the centers of these volumes, both radially and with respect to the space angle. The target atoms and the fragments Main atom are color-coded: green for carbon, red for oxygen and blue for nitrogen. The two other atoms in the fragment (black) can in general be of several atom type, typically carbon (**C**), oxygen (**O**), nitrogen (**N**) or sulfur (**S**).

where $f(\rho)\cdot\, g(\theta)\cdot h(\phi) = p(\rho, \theta, \phi)$ and a uniform segmentation of the supports of $\{\rho, \theta, \phi\}$ is chosen. The numbers of terms in the Riemann sums (i.e. N_ρ, N_θ and N_ϕ) used in this predictive model study, were tens or hundreds, depending on the interval of integration. A Riemann sum is a definition of the definite integral [28], and is an easily implementable way of numerical integration for functions that are well behaved.

2.4.3 The probability masses

Results for fragment classes (see [17] or chapter Article for fragment class definitions) **f2, f5, f8, f11, f18, f22, f23** and **f27** around the first reference point (2.20) are given in Table 2.2. The probabilities are also shown graphically, together with results from the parametric method [17], which is called a direct method here. In Figure 2.9 are given probabilities for fragment classes **f2**, **f5**, **f23** and **f27** around the first reference point, and in Figure 2.10 are given the results for classes **f18, f20, f34** and **f36** around the second reference point (2.21). Figure 2.11 shows the results for classes **f2, f5, f11** and **f22** around the third and fourth reference points, i.e., points (2.22) and (2.23).

The results shown here for each reference point, eqs. (2.20) - (2.23), are normalized in order to be able to compare them with frequencies from reference data, i.e. not with probabilities for other reference points. The probabilities for a certain contact atom type determine the hierarchy among contact atom classes, or target classes, and essentially give the contact preferences for a fragment type.

2.4.4 Conclusions on the comparison of results

The difference between resulting probability masses from the predictive and the direct method is noticeable, but the results are similar, which is encouraging, because both give reasonable contact preferences. The presumption is that the predictive should, in a theoretical perspective, be more reliable, because it takes into account all the parameter values with corresponding weights (see eq. (2.14)) and not only the value corresponding to the peak of the parameter density, or the MAP estimate [18] [24]. Indicators of this can be seen for example in the contact preferences of an aromatic carbon (**f11)** and a secondary amino nitrogen (**f22**), see Figure 2.11. These methods will be further tested and also error analyses performed with more reference data collected. So far, due to too limited amount of reference data used, for example, for subsetting the data, the error limits have been defined only when testing the principle of using the model as a scoring function, like in Example 2 in chapter Article.

Other possibilities for modeling

One-dimensional densities used in this work, model the contact atom positions in required detail and are easily fit to data for each corresponding variable. In the

C\f	**f2**	**f5**	**f8**	**f11**	**f18**	**f22**	**f23**	**f27**
C3	0.0007	0.0009	0.0003	0.0069	0.0018	0.0245	0.0000	0.0006
C4	0.0068	0.0183	0.0132	0.0002	0.0540	0.0004	0.0005	0.0002
C5	0.0093	0.0239	0.0192	0.0002	0.0454	0.0096	0.0000	0.0001
C6	0.0007	0.0009	0.0003	0.2086	0.0687	0.0007	0.0000	0.0014
C7	0.0005	0.0011	0.0000	0.0357	0.0005	0.0029	0.0063	0.0000
C8	0.0951	0.0817	0.0834	0.0069	0.2516	0.0453	0.0246	0.0027
C9	0.0879	0.1696	0.2097	0.0209	0.1202	0.1207	0.0115	0.0083
C10	0.0595	0.1180	0.0947	0.2212	0.2659	0.0362	0.0000	0.0011
C11	0.1194	0.0030	0.0024	0.0252	0.0011	0.1744	0.1717	0.1956
C12	01774	0.1966	0.1373	0.0817	0.0020	0.0707	0.6488	0.4229
C13	0.1820	0.2443	0.2379	0.3719	0.0542	0.3421	0.0567	0.0800
C14	0.1325	0.0018	0.0020	0.0153	0.0025	0.0880	0.0726	0.2859
C15	0.1282	0.1400	0.1997	0.0053	0.1322	0.0846	0.0073	0.0013

Table 2.2: Predictive model based probabilities, calculated at the reference point 1 centered volume. The fragment classes in this table are a hydroxyl O bonded to an aliphatic structure (f2), a carbonyl oxygen (f5), a phosphate oxygen (f8), an aromatic carbon (f11), a fluorine bonded to an aliphatic structure (f18), a nitrogen in an aromatic ring (f22), a nitrogen in a non-aromatic planar ring (f23) and a primary nitrogen bonded to an aromatic structure (f27). The probabilities were calculated over all target classes (C3 - C15). All fragment and target classes are given in Tables 1.1 and 1.2.

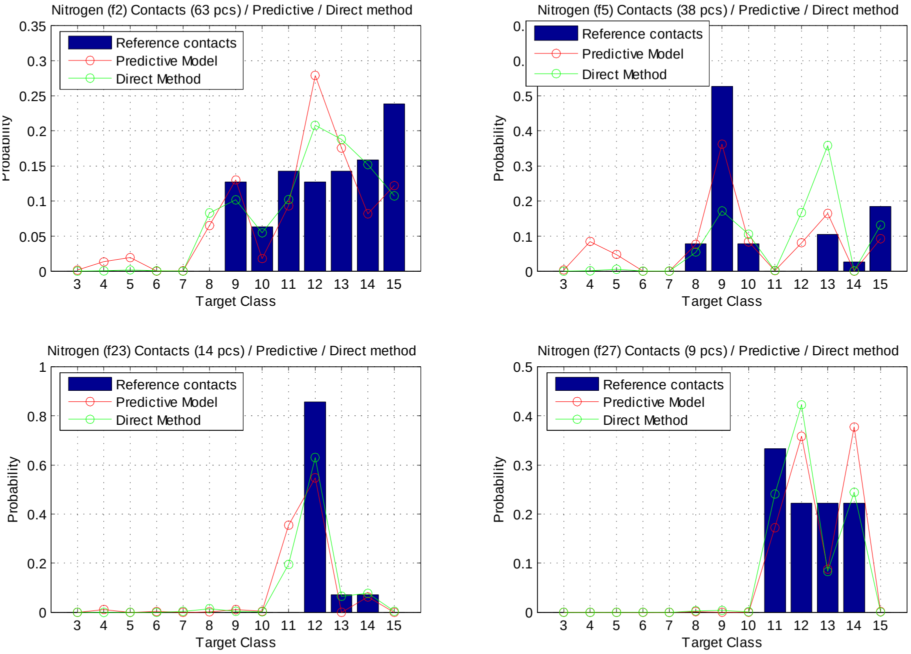


Figure 2.9: Reference point $\bar{r}^{(1)}_{ref}$: comparison of contact atom frequencies, and results from the direct method, with predictive model based probabilities for fragment classes **f2**, **f5**, **f23** and **f27.** Fragment classes denote a hydroxyl oxygen bonded to an aliphatic structure (f2), a carbonyl oxygen (f5), a nitrogen in a non-aromatic planar structure (f23) and a nitrogen bonded to an aromatic structure (f27). Hierarchy is given by the calculated probabilities, which are represented as circles. The circles are joined with a line to illustrate tendensies among target classes. The contact atom counts in reference data for **f2**, **f5**, **f23** and **f27** were 63, 38, 14 and 9, respectively.

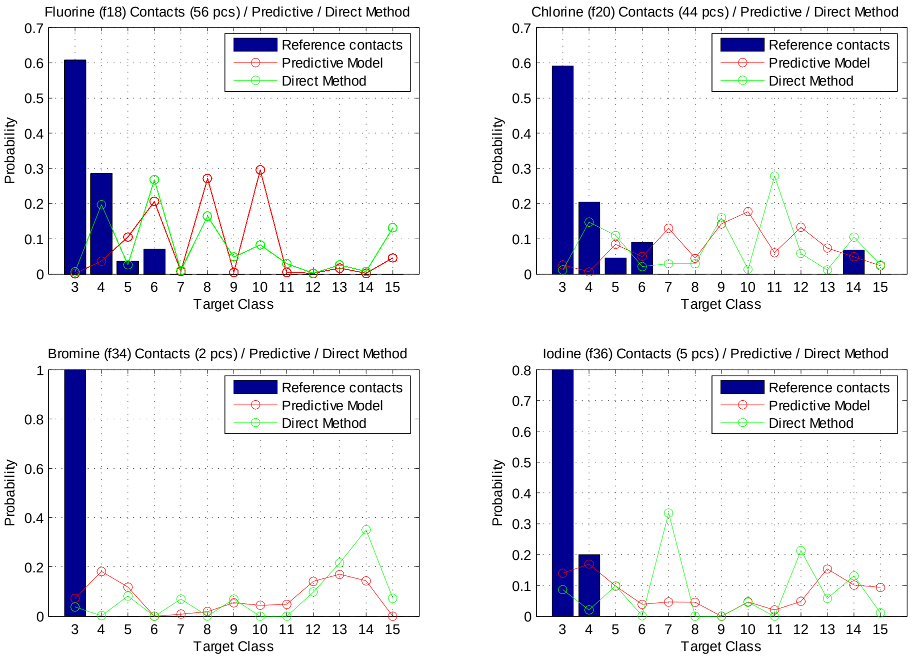


Figure 2.10: Reference point $\bar{r}^{(2)}_{ref}$: comparison of contact atom frequencies, and results from the direct method, with predictive model based probabilities for fragment classes **f18**, **f20**, **f34** and **f36.** Fragment classes denote a fluorine bonded to an aliphatic structure (f18), a chlorine bonded to an aromatic structure (f20), a bromine bonded to an aromatic structure (f34) and an iodine bonded to an aromatic structure (f36). Hierarchy is given by the calculated probabilities, which are represented as circles. The circles are joined with a line to illustrate tendensies among target classes. The contact atom counts in reference data for **f18**, **f20**, **f34** and **f36** are 56, 44, 2 and 5, respectively.

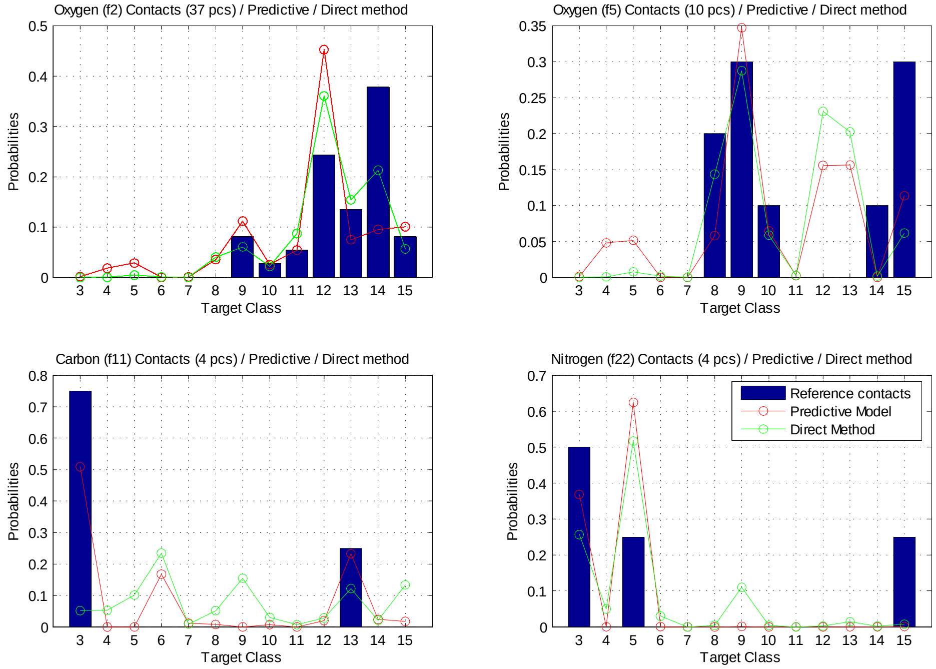


Figure 2.11: Reference points $\bar{r}_{ref}^{(3)}$ and $\bar{r}_{ref}^{(4)}$: comparison of contact atom frequencies, and results from the direct method, with predictive model based probabilities for fragment classes **f2**, **f5**, **f11** and **f22.** Fragment classes denote a hydroxyl oxygen bonded to an aliphatic structure (f2), a carbonyl oxygen (f5), a carbon in an aromatic structure (f11) and a nitrogen in an aromatic structure (f22). Hierarchy is given by the calculated probabilities, which are represented as circles. The circles are joined with a line to illustrate tendensies among target classes. The contact atom counts in reference data for **f2**, **f5**, **f11** and **f22** are 37, 10, 4 and 4, respectively.

3D probability density, they are combined in an interconnected way. Another possibility for modeling would be using a three-dimensional probability density from the start. A plausible method for this could be kernel density estimation [27], the 3D version of equation (2.19), used in estimation of the number of peaks in the 1D probability densities.

The three-dimensional kernel method can be called an unsupervised learning procedure, and would require a larger effort on fitting the model efficiently. This could be worthwhile, because the model fitting procedure does not include assumptions about the modeled distribution of contact atom positions, like the parametric densities as they have been used so far.

About calculations

The computational cost is significant when the conformations of both, ligand and residue side chain, are allowed to vary. An attempt to solve the problem with respect to the latter, could be modeling target atom positions in conformations of a side chain. The formed density would then be used in an overlap integral with a preference density for the target type. This reduces the calculation of contact preferences separately for each conformation to estimating numerically an integral. The approach takes automatically into account flexibility of the residues, therefore also incorporating spatial entropy in the model, taken that a representative set of energetically plausible conformations is included. This subject is further discussed and developed in section Amino acid side chain conformations, in the next chapter.

Chapter 3

Practical applications

The probabilistic model as a knowledge-based scoring function has at the present phase of the work been applied in two molecular environments, a small molecule binding site in a protein and the interface between two proteins forming a complex.

3.1 Knowledge-based scoring function for ligand binding

Characterizing and design of small molecules can include a stage where the binding strength for a set of molecules with respect to a binding site is computationally estimated. This is done to separate better from worse binding molecules, with the ultimate goal of defining the group of tight binders. Strength of the contact is partly defined by kinetic energy of the molecular structures and partly by the potential energies of the direct intermolecular contacts. As already discussed in the first chapter, the probability densities correspond to spatial forms of relative potential energies of the contact. A description of applying the model created in this work, and published in [17] is next reviewed.

3.1.1 Catechol-O-methyltransferase ligand binding site

A standard method for testing a model for its ability to predict preferred binders, is using a mixed set of molecular binders (ligands) and decoys. A recommended set is the Directory of Useful Decoys (DUD) [29]. One target site in DUD is the site for catechol group methylation in a catechol-O-methyltransferase protein (COMT). As an example of the models function in the task of a scoring function, a subset of physically, and partly chemically, similar ligands and decoys were chosen for testing the method. The rationale was to have such a set that the direct contact preferences in binding were emphasized. This was achieved so that all chosen molecules had a ring structure with two hydroxyl groups bonded to neighboring carbon atoms, similar to a catechol group, and therefore are in principle able to anchor to the magnesium ion (Mg^{2+}), that is considered as part of the binding site [30]. In addition, ligands and decoys had similar masses, with the exception of a known tight binder that was considerably heavier, and was added to the ligand group together with the natural ligand dopamine [30] as reference molecules.

The results showed that the contact preferences efficiently (receiver operating characteristic, ROC, related accuracy was 0.93) separated ligands from decoys. This is explored in detail in Example 6 of article [17]. Contents of the publication are presented also here, in chapter Article, where Example 2 corresponds to the above mentioned Example 6.

3.2 Amino acid residue mutations

A change in contact preferences for an amino acid residue position can alter the probability for the protein to perform its function. This is conceivable when the mutated residue is in a ligand binding site or, for example, in a position that affects the conformation or stability of the protein or protein complex. Before demonstrations of the scoring function use in this molecular environment, a preliminary subject is covered, the approach taken in this work to include amino acid residue side chain conformational variability, already referred to in the section About calculations of the second chapter.

3.2.1 Amino acid side chain conformations

Rotation around a bond between two sp^3 hybridized carbon atoms in a side chain is typically represented as distributions defining a rotamer, a rotational isomer [31] [32]. The hybridization state sp^3 refers to a carbon atom singly bonded to four other atoms. The rotamers are clearly separated and are centered around three potential energy minima, one of which is a global minimum. Another type of rotation, around bonds that connect an sp^3 and an sp^2 hybridized atom, the latter being in a side chain as part of a planar end group, has a distribution that is less articulate than the one for rotamers, but contains peaks too.

These distributions are obtained by collecting coordinate data for conformations from structure files. They are stored in rotamer libraries and presented as binned data forming discrete probabilities for rotation angle intervals [31], or possibly continous probability densities that are fitted to the observed data on conformations [32] [34]. The conformational properties of side chains are influenced, in addition to the electronic structure of the side chain, affected by the main chain conformation of the residue, by contact preferences imposed from the molecular environment. The latter means direct and water mediated contacts of the side chain to other residues and small molecules. These factors determine the functional form for the conformational potential energy of a particular side chain in a specific position and in given conditions like pH.

Conformation distributions of rotamer libraries, discrete or continous [32], give a lower probability to conformations that correspond to higher values of potential energy. This is somewhat misleading, because at typical temperatures around 300 K where biochemical reactions take place, there is thermal energy in the structure and a side chain dihedral, or torsion, angle has values over an interval with approximately classical harmonic oscillator, or even equal, probability for each value of the diheral angle. This approximately equal probability means steep potential energy walls at the ends of the intervals. The torsion angle intervals in an ensemble of side chains change with time, due to thermal energy transfer, and at a given point in time follow a distribution corresponding to thermal equilibrium, which is an argument directly based on thermodynamics [35].

With the above in mind, it is realized that intervals can be used to replace the static conformations, which then allows capturing the motion of a side chain in the statistical model of the type presented in this work. This way, the distribution of conformations can be built starting from thermal motion in a side chain, together with a functional form for the potential energy, and get the entropic contribution from side chain thermal motion incorporated. Technically this can be done by calculating overlap integrals, as described in the next section on residue mutations in Dengue virus envelope and pre-membrane proteins.

The overlaps could also be calculated using directly a rotamer library, but then the results would depend on the rotamer library and although the library would be main chain conformation dependent, what could not be taken into account is the position of an individual residue in the 3D structure. Position influences the likeliness of a level of thermal energy to be captured in the side-chain degrees of freedom. As an example, following from solvent accessibility, a residue located in a binding pocket or at a protein interface in a complex, is likely to occupy smaller portion of thermal energy spectrum than a residue on the outer surface of the biologically active unit, see for example [36] [37] [38]. Relating to this, it should be mentioned that data from all residues adhering to certain criteria are included in a rotamer library [31]. This bears significance when results of docking are scored, or an estimate for the effect of a point mutation is calculated, and also in determining the amount of entropy reduced in protein folding. When this torsion angle interval based method is compared to the traditional rotamer library approach, where side chain conformational space is seen as a collection of static structures with probability for occurrence, it is essential to note that a rotamer library corresponds to an ensemble of thermal energies and effects of potential energies from contacts with the environment. The interval approach then, can be used for modeling single side chains, taking into account their molecular environments.

Continuing this relationship, torsion angle values found in a database survey can be modeled with a density like von Mises [32], but they are not von Mises distributed during a cycle of side chain rotation. The connection between overlap integrals in the two approaches is demonstrated with equation

$$\int g(\bar{r})f(\bar{r})dV = \sum_i p_i \int h_i(\bar{r})f(\bar{r})dV \qquad (3.1)$$

where probability density g models side chain target atom positions, produced by some dihedral angle distributions. Function f is a contact preference density defined for the targets, density h is fitted to target positions obtained through a harmonic oscillator sample of angle values, and p_i a weight for the combination of kinetic energy levels that produced h_i. Probability p_i depends on probabilities of the energy levels, selected or sampled, with respect to each rotational angle. Equation (3.1) means $g(\bar{r}) = \sum_i p_i h_i(\bar{r})$, which can be stated so that the weighed angle value intervals generate the distribution of the side chain conformations, or starting from g, that a spatial distribution of an atom in a set of side chain conformations can be expressed as a linear combination of distributions for rotations confined to angle intervals. In this setting, the straightforward use of a rotamer library would correspond to the LHS of eq. (3.1), the equation being only approximately valid. The RHS shows that given the form of potential energy, what can still be adjusted in order to match the external conditions, is the distribution of angles over an interval and the weights for the intervals.

A set of overlap values for contact preference densities, and a torsion angle interval defined distribution of conformations, with weights like Boltzmann factors, covers both the consequences of internal motion with given energy, and the probability of occurrence for that energy. Energy levels with Boltzmann distributed net differences with respect to a mean value in a potential energy well is depicted in Figure 3.1. The functional form of the potential is taken from [33], where it is used in calculations of rotational energy contributions to side chain ensemble properties.

It has to be kept in mind that thermal motion for side chains with more than two dihedral angles is highly complex and the motion around one bond depends on the motion around another, when two or more rotatable bonds are involved. The approach described here captures the conformation space accessible for the side chain through internal rotations. In order to perform overlap calculations, the contact, or target atom cloud is represented as a probability density. It should still be mentioned that the X-ray structures are here taken as a sample, meaning that the measured structures are considered truly representative of the physical reality, especially with respect to variation of conformations of side chains. This can be taken as a reasonable assumption [31], [38].

Conformation dependent calculations

In the discussion of the previous section, it was envisioned that contact preference calculations were done using torsion angle intervals, determined with a known or approximated potential energy function, and a Boltzmann-distributed ensemple of kinetic energies with possible dependence on the residue position on the proteins 3D structure. Implementing that as calculations, is in this work left as a future challenge in order to avoid unfounded mean kinetic energy estimates and functional forms of potential energy.

Instead, the concepts developed are demonstrated in a computationally less demanding way. Namely, sp^3 to sp^3 hybridized atom bond torsion angles are given intervals according to their position in the side chain. The position is reflected by the moment of inertia (I) of the side chain fragment in rotational

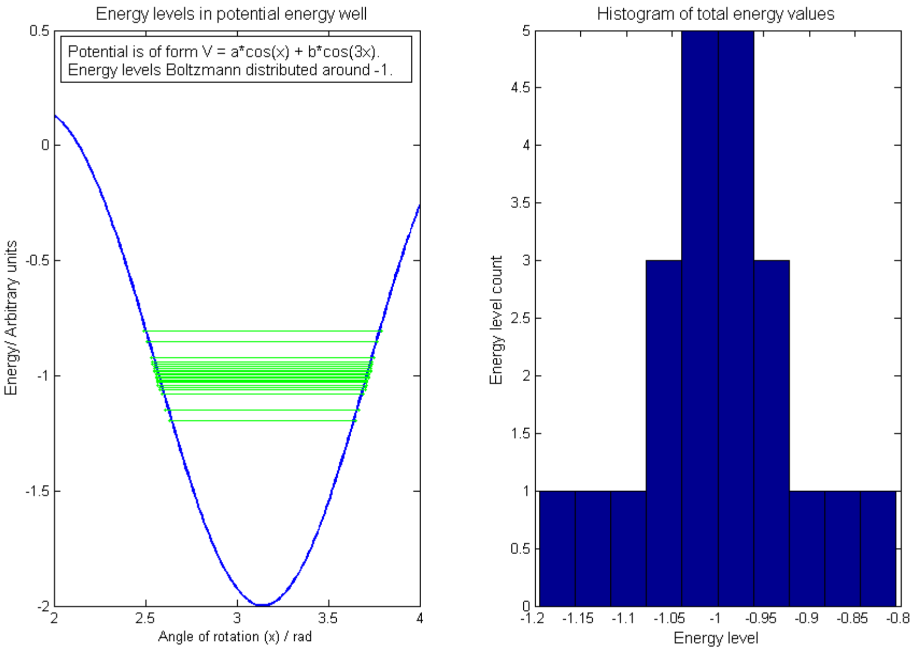


Figure 3.1: A theoretical distribution for levels of total energy in an ensemble of side chains, with respect to a dihedral angle. On the left: part of a potential energy function (blue line). Potential has the same functional form that in the literature has been shown to work in approximative quantum mechanical calculations of side chain rotational energies (see text for details). The total energy levels (green lines) are distributed according to Boltzmann distribution. The two intercepts of a green line with the curve define an interval. On the right: binned counts of energy levels generated with rejection sampling, presenting the distribution of the green lines in the left figure.

motion. Motivated by the equation for kinetic energy of a particle rotating around an axis, the intervals were made to shorten towards the main chain, inversely proportional to square root of moment of inertia. As an example, for a residue side chain with three consecutive sp^3 to sp^3 bonds, the sequence

$$(\Delta\chi_1, \Delta\chi_2, \Delta\chi_3) = \left(\Delta\chi_3\sqrt{\frac{I_3}{I_1}}, \Delta\chi_3\sqrt{\frac{I_3}{I_2}}, \Delta\chi_3\right), \tag{3.2}$$

is used. In equation (3.2), the interval of an internal rotation angle around the bond that is farthest from the main chain, $\Delta\chi_3$, has to be fixed first. The highest moment of inertia I_1 is around the bond closest to main chain, that is, bond between alpha and beta carbon ($C_\alpha - C_\beta$). Side chain dihedral angle intervals were centered around the canonical rotamers, known as gauche$^-$, gauche$^+$ and trans (see e.g. [31]).

The distribution of dihedral angles around bonds from sp^2 to sp^3 hybridized atoms, is taken as having relatively high probability for that the rotating group contains enough thermal energy, or potential energy from environment, to overcome barriers between rotational potential energy minima. That is, high probability relative to the dihedral angles around sp^3 to sp^3 bonds, where the side chain conformations have a rotameric structure. Therefore, the intervals for sp^2 - sp^3 bonds, in all cases here the sp^2 atom being in terminal end group of the side chain, were fixed on wide ranges of dihedral angle values. This was also based on visual inspection of single bond intervals in a rotamer library [32]. The extreme case was a bond to terminal carboxamide (-C(O)NH$_2$), which was given an interval of full 2π rotation, because this functional group can adopt all orientations to form hydrogen bonds with its environment [39].

Though the rotamer libraries do not give a direct reason for having different intervals for different side chain dihedral angles, doing so fits the scheme that both, main chain conformation, and preferences of the molecular environment, modify the form of the potential that gives boundaries to rotational motion [31]. In addition, the angular momentum transformed as impulse from thermal motion of the environment, depends on the relative direction of the impulse and the rotatable bond, and is therefore taken here to be on average evenly distributed on fragments having differing moments of inertia (I). Using this assumption, an interval dependent potential, like harmonic oscillator or stepwise changing rectangular, and the above mentioned equation for rotational kinetic energy $\frac{L^2}{2I}$, where L is the angular moment, one gets the interval sequences exemplified in equation (3.2).

The interval calculations are in the following demonstrated using a sum of contact preference density point values as a first approximation to the method. In practise, first a target atom position cloud corresponding to specific set of intervals for the torsion angles is generated. The distribution of the torsion angle values within an interval is taken, as another first approximation, to be uniform, which corresponds to a rectangular potential energy well with the walls at the interval end points. Second, given a backbone structure, the strength of a contact between two residues is estimated through calculating contact preference density point values for a fragment of one residue, in the generated target atom positions of the other. During the calculation, numeric values to characterize the contact are collected. These include sum, mean and maximum of the point values, together with the count of the points where probability density values where calculated. The information provided by these numbers can then be used to rank the suggested amino acid residues for the studied position in the protein backbone. Still more precisely, the ranking is based on contact preferences with respect to one of the target atoms in a residue.

Interval based side chain conformations and the idea of the overlap method is next used in contact preference calculations. Some background information on the case studies is given first.

3.2.2 Significant residue positions in Dengue virus sequence

Dengue virus is a global health threat transmitted by certain species of mosquitos. There exists four genetically distinct serotypes [40], and infection with one serotype virus gives immunity to that serotype, but might cause an enhancement to the disease severity (antibody-dependent enhancement, ADE) following infection by another serotype [40]. Ongoing basic research and vaccine development require information on virus genome (RNA) that relate to severity of the disease caused by the infection. This information can then be translated to structural inspections, intended, for example, for finding antibody binding epitopes or functional differences of the viral proteins following point mutations.

There is a strong evolutionary selective pressure on Dengue virus populations [40] and therefore clustering the amino acid residue sequences of the virus gives clear signals for possible significance of a point mutation. Results of a clustering analysis on amino acid level sequences for Dengue virus were used as a starting point to find functional differences that would explain severity of the disease that follows infection. The analysis had been performed with a modified version of the Bayesian statistics based clustering tool K-Pax [41] and had resulted in identifying several significant positions along the expressed genome, each separating one or several sequence clusters from the other. Studies conducted in this work were limited to two viral surface proteins, the first especially significant with respect to immune response, namely envelope protein (E) [42], and the second, precursor membrane protein (prM), an important part of maturation and activation process of the virus [43]. Envelope protein being the immunologically central protein, i.e. containing the most effective epitopes for antibody binding, it is the main target in vaccine developement [40]. Figures 3.2 and 3.3 show two Dengue virus tree structures that represent the evolutionary relationships between clusters inside serotypes 1 and 2. The clustering had been performed with another Bayesian clustering tool BAPS [44].

Figures 3.2 and 3.3 contain information revealed by nucleic acid base level clustering analyses on the evolutionary history of Dengue virus serotypes 1 and 2. In the second diagram, Figure 3.3, several clusters show evolutionary features not seen when the clustering is done in the level of amino acid residues. One of these features is splitting of cluster 3, i.e. one cluster in amino acid

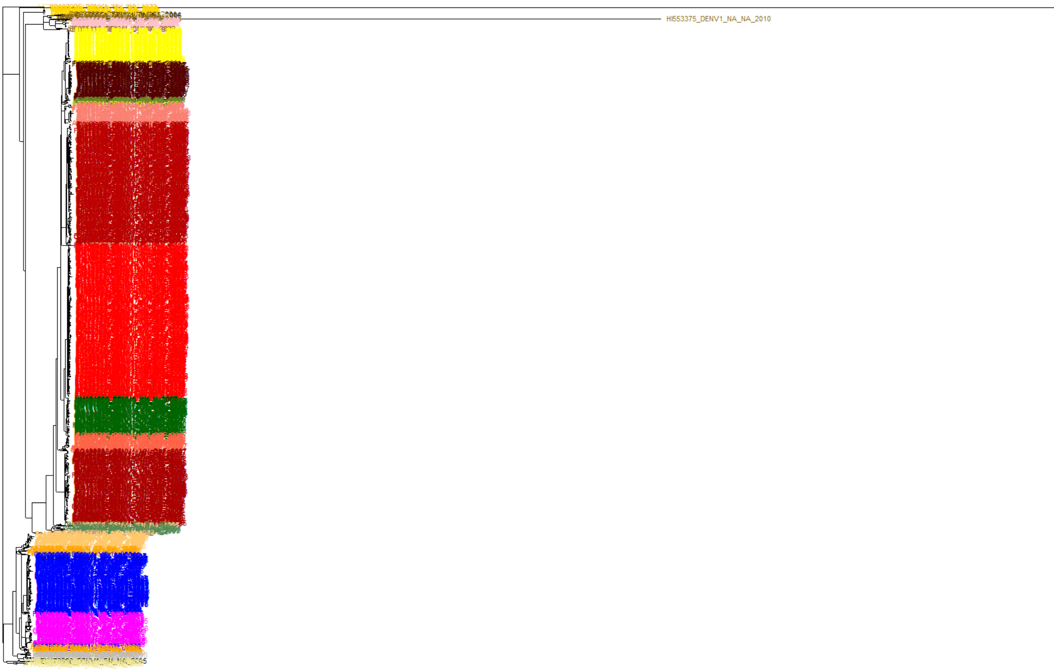


Figure 3.2: Tree diagram for Dengue virus serotype 1, estimated at the level of the nucleic acid bases of the RNA genome. Cluster 4 shown in red is discussed in the text. Due to the large number of sequences in the dataset, only cluster-level separation is presented in the figure to give an overview of the evolutionary relationships.

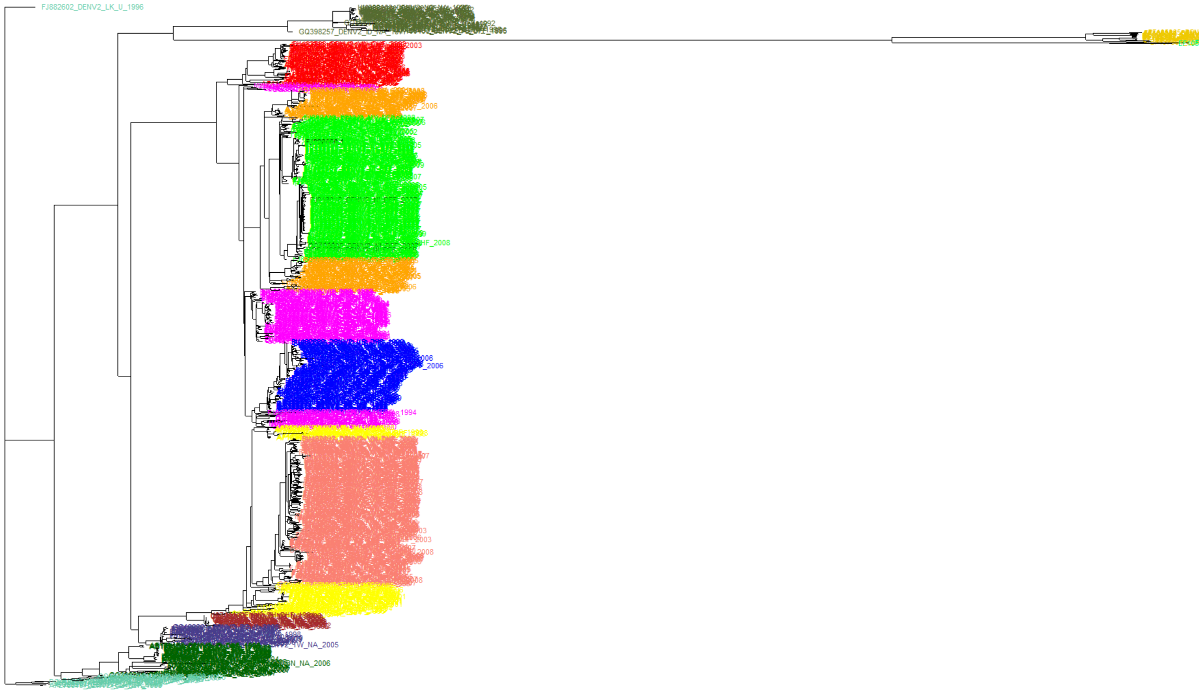


Figure 3.3: Tree diagram for Dengue virus serotype 2, estimated at the level of nucleic acid bases of the RNA genome. Clusters 1 and 6, shown in yellow and light rose, respectively, are discussed in the text. Due to the large number of sequences in the dataset, only cluster-level separation is presented in the figure to give an overview of the evolutionary relationships.

sequence analysis, to three groups (color coded magenta), each of which mainly correspond to a single country of occurrence. Almost all of the sequences in amino acid sequence cluster 3 are from time period 1985-1995, which suggests that this cluster represents an ancestral form for the virus populations of this serotype in these three countries. When mutations have taken place in a way, that a new genomic variant producing protein machinery more fit for the prevailing conditions, a new variant of the virus is created and starts to spread in the area. This may show in later virus samples as a new cluster and one of the clusters replacing number 3 is likely to be cluster 4 (blue in Figure 3.3), that occurred mainly in two areas where cluster 3 samples were registered, namely United States and Puerto Rico, and its appearance overlaps in time with the end of cluster 3 period, in the late 1990's.

Second detail to be pointed out here is that of clusters 1 (light rose) and 6 (yellow), where 1 replaced 6 in the area of Thailand and spread out in the early 2000's. The spread of a virus implicates an increase of its fitness-characteristic [40] and this particular case is discussed in more detail in the next section. Envelope/pre-membrane protein complex exists in a stage of the maturation process of the virus.

Selected positions in the amino acid sequence, that were interpreted as significant based on clustering, and seemed to match severity of disease in reported incidents collected by World Health Organization (WHO), were studied by calculating contact preferences of a potentially contacting amino acid residue on the other side of the interface. In other words, it was tested, which of the observed residues, based on the probabilistic model, was most preferred for a particular position in the structure of the protein complex. It was assumed that the contacting residue from the other side of the interface was unaltered, i.e. conserved, as can be verified from the analysed sequences.

Interface between envelope and pre-membrane proteins

Two positions in the amino acid sequence over the entire genome of the virus were studied, position 169 and position 363, both in the E/prM interface. The first was interpreted to correspond to number 55 in the pre-membrane protein sequence and the second to number 83 in the envelope protein sequence.

In order to quantify possible contacts across the interface, overlap integral values of contact preference probability densities were calculated for alleged water mediated contacts, and point value sums of the densities determined for direct contacts.

Mutation R55Q The first position studied was 55 in prM, where one cluster of Dengue virus serotype 1 had glutamine (GLN, Q) and other 20 clusters had arginine (ARG, R). This single cluster with Q contained less severe cases of the disease, i.e. febrile illnesses with relatively few hemorrhagic fevers or shock syndromes. This was considered a possible indicator for altered function of the proteins, and the changes in contact preferences were investigated. Figures 3.4 and 3.5 represent direct contact preferences over the interface as probability

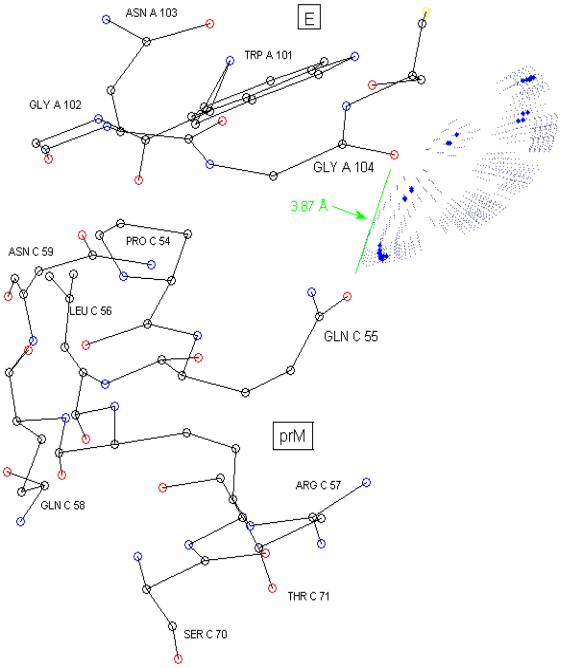


Figure 3.4: Contact preference density values plotted for glycine 104 from envelope side with respect to glutamine 55 (Q55) from pre-membrane side. Larger marker size means higher density values. Cutoff used for density values is 5 % of the maximum. Q55 is in a conformation illustrative of the dimensions of the contact site.

density value scatterplots, with some neighboring amino acid residues at the contact site.

Contact preferences were calculated using two structures from Protein Data Bank (PDB), containing the envelope/pre-membrane protein complex, both of which represent serotype 2, though the sequence data of this examplifying case study is for serotype 1. This inconsistency is because no corresponding structure was available for serotype 1, and it is assumed that the interface is structurally similar in different serotypes.

Results of the calculations are given in Table 3.1. Side chain conformations were generated for evaluation of the contacts, after which needed values were calculated for direct and water mediated bonds. A sum over the density point values can be large when the density has high values in fewer target atom positions and as large when the side chain conformations produce more target atom positions with lower density values. Cutoffs for the length of a water mediated contact were 3.3 Å and 6 Å. The higher end value was chosen because

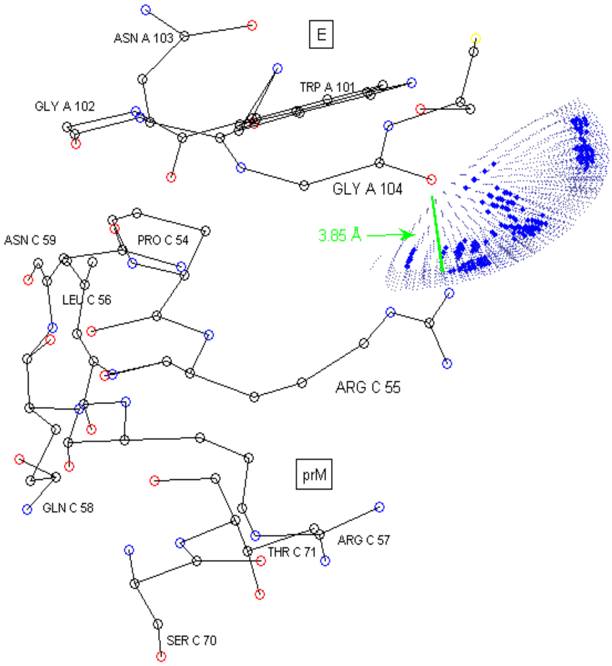


Figure 3.5: Contact preference density values plotted for glycine 104 from envelope side with respect to arginine 55 (R55) from pre-membrane side. Larger marker size mean higher density values. Cutoff used for density values is 5 % of the maximum. R55 is in a conformation illustrative of the dimensions of the contact site.

length 3 Å was taken as the upper limit for a hydrogen bond with water. In the other end, 3.3 Å is the upper limit used for direct hydrophilic contacts, so length intervals for direct and water mediated contacts do not overlap.

In calculations for direct contacts, a distribution of target atom positions represented a residue in the studied main chain position, which in this example is 55 of prM. Following the results from the clustering analysis, the residues tested for that position were glutamine and arginine. On the other side of the interface, the chosen contact residue was glycine (GLY, G) in position 104 in E, and it was represented by contact preference probability density of a fragment, with a main chain oxygen as the contact atom. The site is pictured in Figure 3.4. Water mediated contacts were calculated as overlaps of two contact preference densities, for a fragment in the residue tested for position 55 and the glycine in 104 on the other side of the interface. The preference densities were defined for a target classified as a generic hydroxyl oxygen. Results of the calculations are given in Table 3.1. Values corresponding to bonds with and without water are not directly comparable, because they describe different densities. The results without water are values of the spatial density and results with water are evaluated spatial overlap integrals, both having the units Å^{-3}. Also to be noted that, the side chain conformations for χ_2 of glutamine and (χ_2,χ_3) of arginine were generated only around the rotamers (t, t and g$^-$ respectively) of the ideal structures from PDB [45].

The numeric values in Table 3.1 suggest that, based on the model and given structures (PDB coordinate files), in the low pH structure glutamine and arginine in 55 of prM can be considered equally good direct targets for glycine in 104 of E. In the neutral pH structure, glutamine seems to be a better contact, but in the more likely case of a water mediated contact, arginine is the more preferred contact in both, low and neutral pH structures. The water bridge was considered more likely due to dimensions of the site, depicted in Figures 3.4 and 3.5. Overlaps were calculated with three different numbers of finite volume elements. Table 3.1 gives the mean values. Standard errors in units (10Å)$^{-3}$ are, for GLN 0.4*10^{-5} (3C5X) and 1.7*10^{-5} (3C6E), and for ARG 2.0*10^{-5} (3C5X) and 2.2*10^{-5} (3C6E). The difference in mean values was confirmed statistically significant with 95% confidence level using t-test. The test p-values were 0.01584 (3C5X) and 0.01980 (3C6E).

Tabulated in 3.1 are four values for each direct contact: the sum of the density point values and for completion, mean (Mean) and maximum (Max) point values with the number of points (Nr) as well. Arginine gets higher maximum, but lower mean values in both, low and neutral pH structures, suggesting that the point values of glutamine are more evenly distributed. Glutamine has a smaller amount of points, Nr, but the mean of the values evaluated in them is higher, corresponding to that the main result Sum is higher, especially for neutral structure 3C6E. The value for Sum is calculated so, that the varying finite volume element size produced by spherical polar coordinates, is taken into account, and therefore represents the probability density in the spatial area containing the target atom positions. Nevertheless, because the actual volume elements were not used, as in an integral, Sum depends on Nr, and the other

$1/(10Å)^{-3}$ **(χ_1 rmer)**	**GLN 55** **prM/E**	**ARG 55** **prM/E**	**pH**	**PDB** **(res.)**
GLY 104	0.344 (g^-) sum	0.333 (t) sum	low	3C5X (2.2Å)
GLY 104	0.551 (g^-) sum	0.324 (t) sum	neutr.	3C6E (2.6Å)
G104&H2O	$0.3*10^{-3}$ (g^-) overlap	$1.2*10^{-3}$ (t) overlap	low	3C5X (2.2Å)
G104&H2O	$0.5*10^{-3}$ (g^-) overlap	$1.4*10^{-3}$ (t) overlap	neutr.	3C6E (2.6Å)
Mean,Max **(Nr)/G104**	$7*10^{-5}$,$3*10^{-4}$ (4913)	$6*10^{-5}$,$8*10^{-4}$ (5265)	low	3C5X (2.2Å)
Mean,Max **(Nr)/G104**	10^{-4}, $4*10^{-4}$ (4913)	$6*10^{-5}$,$6*10^{-4}$ (5265)	neutr.	3C6E (2.6Å)

Table 3.1: Calculated highest values of contact preferences relating to premembrane protein position 55. Results given as sums of preference density point values (sum) or as evaluated preference density mean overlap values (overlap). Overlaps are calculated for water mediated contacts, while sums quantify direct contacts. Sum and overlap values can be compared separately, not with each other. The best scoring reference rotational isomer of rotation around alpha and beta carbon connecting bond given in parentheses after calculated value. PDB IDs refer to structures measured in neutral and acidic conditions given in column pH. Last two rows contain supplementary results for Sums. See text for more details.

two calculated values (Mean,Max) give clarifying information. The side chain conformations were generated in a process where the amount of torsion angle values included from the defined interval, depended on the rotatable bond count of the side chain. Arginine could therefore have a larger advantage in the number of target atom positions, and though the values for Mean, with similar Nr (5265 vs. 4913), is higher for glutamine, the result is not conclusive. Following from that a water bridge was considered the more likely form of interaction, a decisive result for this direct contact was not pursued. It is concluded, that in order to make interpretation of the results more straightforward, distribution of the target atom positions is to be modeled with a pobability density, which then allows calculating the results as overlap integrals, in a similar way as for a water bridge in this study.

The water mediated contacts were quantified using numerical integration with Riemann sums, where the amount of points is related to precision. The number of finite volume elements, $\Delta\rho\Delta\theta\Delta\phi$ in eq. (2.24), used in evaluating the integrals varied between 200 and 2000. An assumption used in this study is that the side chain can be in any rotamer, without considering the energy differences of the reference rotamers, for a quantum mechanical treatment of side chain conformational energies see, e.g., [46]. The lower energy rotamers could possibly be favoured in an ensemble, but other rotameric states are systematically observed and the molecular environment emphasized in this study is shown to have, together with main chain conformation, a significant influence on which side chain conformations are realized [31].

Numerically obtained results correspond with the severity loss in the one cluster with R55Q mutation in that, the presumed water bridge from position 55 in prM to position 104 in fusion loop of E would be weaker, therefore making the complex less stable.

Mutation N83K The second position here is 83 in envelope of dengue virus serotype 2, where two clusters from the k-pax clustering analysis had lysine (LYS, K) and 11 clusters, almost all the rest, had asparagine (ASN, N). There were two more clusters, with valine (VAL, V) in 83 of E, but for these, no metadata like severity of disease was available, and VAL was not included in the calculations. The two clusters with lysine had larger portions of severe cases. Figures 3.6 and 3.7 represent the contact site for ASN C 7 with a scatter plot of preference density values for the side chain end group fragment in ASN C 7, and both clustering based significant residue types (N and K) in representative conformations.

1/(10Å)$^{-3}$ (χ_1 rmer)	**ASN 83** **prM \ E**	**LYS 83** **prM \ E**	**pH**	**PDB** **(res.)**
ASN 7	0.2 (g$^+$) sum	23.4 (g$^+$) sum	low	3C5X (2.2Å)
ASN 7	0.1 (g$^+$) sum	2.1 (g$^+$) sum	neutr.	3C6E (2.6Å)
ASN7&H2O	0.026 (g$^+$) overlap	0.053 (g$^+$) overlap	low	3C5X (2.2Å)
ASN7&H2O	0.011 (g$^+$) overlap	0.035 (g$^+$) overlap	neutr.	3C6E (2.6Å)
Mean,Max **(Nr) / N 7**	6*10^{-4},3*10^{-3} (1089)	4*10^{-3},0.02 (6561)	low	3C5X (2.2Å)
Mean,Max **(Nr) / N 7**	10^{-4}, 5*10^{-4} (1089)	3*10^{-4},10^{-3} (6561)	neutr.	3C6E (2.6Å)

Table 3.2: Calculated highest values of contact preferences relating to envelope protein position 83. Results given as sums of preference density point values (sum) or as evaluated preference density mean overlap values (overlap). Overlaps are calculated for water mediated contacts, while sums quantify direct contacts. Sum and overlap values can be compared separately, not with each other. The best scoring reference rotational isomer of rotation around alpha and beta carbon connecting bond given in parentheses after calculated value. PDB IDs refer to structures measured in neutral and acidic conditions given in column pH. Last two rows contain supplementary results for Sums. See text for more details.

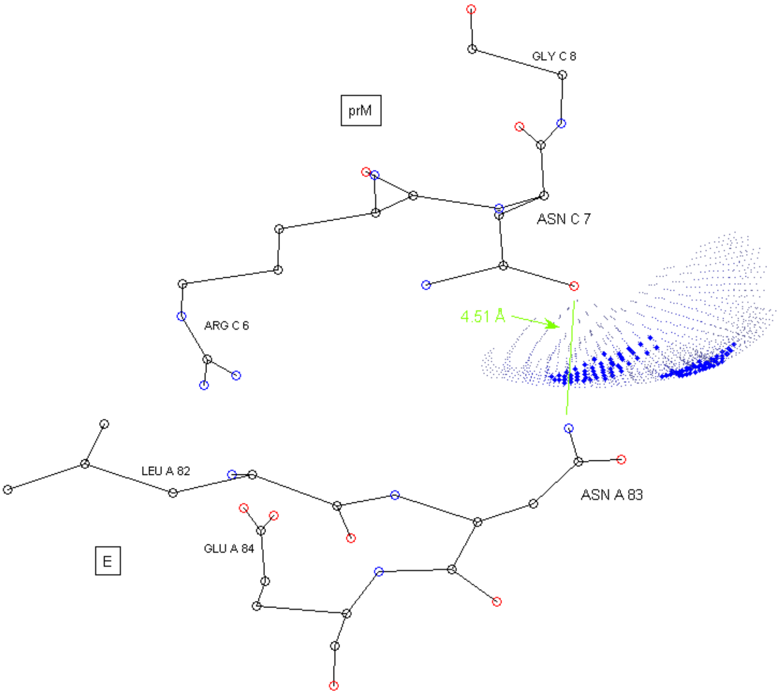


Figure 3.6: Contact preference density values plotted for asparagine 7 from pre-membrane side with respect to asparagine 83 from envelope side. Larger marker size means higher density values. Cutoff used for density values is 5 % of the maximum. Asparagine 83 is in a conformation illustrative of binding site dimensions.

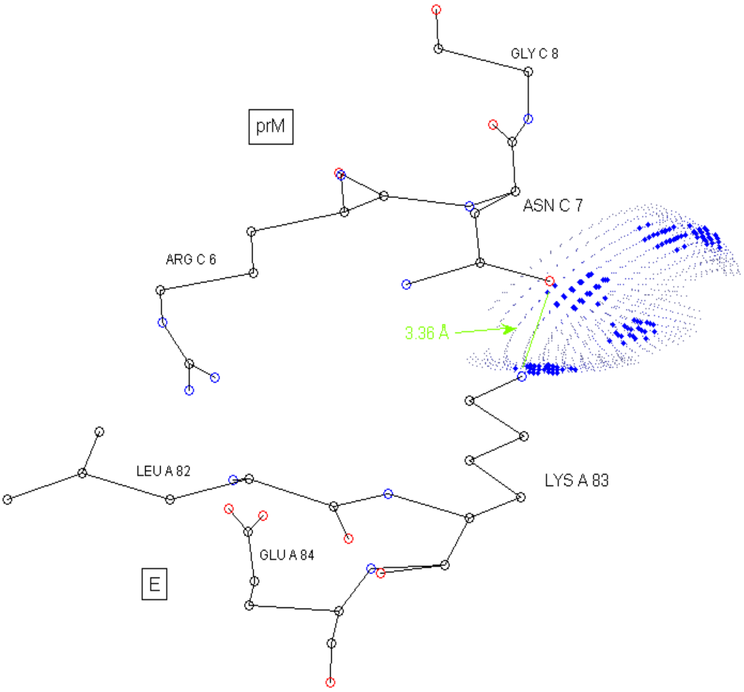


Figure 3.7: Contact preference density values plotted for asparagine 7 from premembrane side with respect to lysine 83 from envelope side. Larger marker size means higher density values. Cutoff used for density values is 5 % of the maximum. Lysine 83 is in a conformation illustrative of the binding site dimensions.

Overlaps were calculated with six (3C5X) and five (3C6E) different numbers of finite volume elements. Table 3.2 gives the mean values. Standard errors in units (10Å)$^{-3}$ are, for ASN 0.3*10^{-3} (3C5X) and 1.1*10^{-3} (3C6E), and for LYS 20.4*10^{-3} (3C5X) and 13.8*10^{-3} (3C6E). The difference in mean values was confirmed statistically significant with 95% confidence level using t-test. The test p-values were 0.02140 (3C5X) and 0.01984 (3C6E). The fragment class specific prior was defined for the residues involved in water mediated contacts so, that the relating results can be directly compared, also between studied positions: C55, A83 and 202.

Lysine in position 83 of E is the more preferred as direct contact for asparagine in position 7 of prM, and also more preferred contact in a water bridge. This can be explained by that lysine side chain is longer than asparagine side chain and has more rotatable bonds (4 vs. 2). This allows lysine to participate in more geometries that generate a preferred contact with asparagine on the other side of the interface, through a water molecule or directly. The values for the direct contact (Sum, Mean and Max), all indicate that lysine can make a stronger contact, because it does not only get highest Sum, but also Mean, despite larger number of points, Nr. In addition, lysine gets at least twice higher maximum probability values, indicative of deeper position in potential energy, with the *a priori* given estimate for fragment to target bond strength included, for details of fragment class specific prior probabilities, see chapter Article. It follows that lysine is more favoured, both in terms of bond strength and spatial entropy, the latter following from that there are more geometries that correspond to potential energies close enough to minimum, so that the contact is maintained. It should be noted that for lysine, alternative rotamers of three side chain dihedral angles were not considered, but all rotamers for asparagine were included in the calculations, and therefore lysine could get still stronger preference when all its degrees of freedom were fully mapped.

Numerically obtained results correspond with the severity rise in the two clusters with N83K mutation in that, the presumed water bridge from position 83 in E to position 7 in prM would be stronger, therefore making the complex more stable.

A hydrophobic pocket in domain II of E

This second case study is a residue pair that closes a hydrophobic pocket surrounding a tryptophan (TRP 206) residue of the envelope protein in domain II (DII). The residues are in positions 202 and 257 in the amino acid residue sequence of the envelope protein (E). Position 202 had been found significant in the clustering analysis performed with the program k-pax [41].

In the dimeric form of E, pockets of the two chains are close to one another (about 9 Å between residues of position 257 in chains A and B), that is, they are at the E/E interface. Dimer is the prevalent form before the virus particle is in an endosome of a host cell [43], where, in the lower than neutral pH, E protein complexes transform to trimeric, cone-like form. In the trimer, the pockets are on the outer surface of the cone. Tryptophan in position 206 (W206) and its environment could be of interest, because it is located in DII, which gives the strongest immunological response in E [40]. Table 3.3 presents numerical results for preference calculations quantifying the contact between residues in positions 202 (LYS or GLU) and 257 (GLU).

Overlap / (10Å)$^{-3}$	**4GSX** (ChainA)	**4GSX** (ChainB)	**1OK8**
E202-E257 (**χ_1 rmer**)	0.20 (g^{-})	0.05 (t)	0.39 (g^{-})
K202-E257 (**χ_1 rmer**)	0.02 (g^{-})	0.01 (t)	0.08 (g^{-})
Stage	late-stage fusion	late-stage fusion	postfusion
Serotype	1	1	2
Resolution	1.9 Å	1.9 Å	2.0 Å
Overlap / (100Å)$^{-3}$	**3C5X** (pH low)	**3C6E** (pH ntr.)	**3UAJ**
E202-E257 (**χ_1 rmer**)	0.11 (g^{+})	0.26 (g^{+})	0.74 (g^{+})
K202-E257 (**χ_1 rmer**)	0.01 (g^{+})	0.17 (g^{+})	0.17 (g^{+})
Stage	prM/E	prM/E	prefusion
Serotype	2	2	4
Resolution	2.2 Å	2.6Å	3.23 Å

Table 3.3: Probabilistic model based contact preferences calculated for a water mediated contact between a lysine, or a glutamate, and a glutamate using five PDB coordinate files, IDs shown. These structures represent different stages of the viral life cycle, for example, the first column in the lower row has data for a maturing virion in a lower than neutral pH. Note also serotypes and resolutions. Results are given as highest overlap mean values. See text for more details.

The results in Table 3.3 point to the direction that structural changes of the hydrophobic pocket occur in different stages of viral entry (3UAJ,1OK8,4GSX) and varying pH (3C5X,3C6E). In all cases glutamate in position 202 gets values indicating the stronger water mediated contact with glutamate in position 257 of the same protein chain. Conclusions based on these results are restricted to that there are structural changes, and that conditionally glutamate is the more preferred residue, because some factors, like contact with the neighboring residue in position 203, the main chain conformations and part of the side chain conformation space are neglected. The last point is related to that alternative rotamers are accounted only for rotations around the bond between alpha and beta carbons. Also, the spatial entropy of the side chains could be considered by quantifying the amount of conformation space allowing a water mediated contact, when suitable conditions on including minimum overlap value were determined.

Overlap values were calculated using three different numbers of finite volume elements in the Riemann sums. Statistical significance was tested for difference between glutamate-glutamate and lysine-glutamate water mediated contact overlap mean values for all six cases given in Table 3.3. Significance was confirmed with 95% confidence level using t-test with respect to all but 4GSX (chain B), where the test p-value 0.1044 did not allow rejecting the null hypothesis that overlap means were equal. The structure 4GSX contains two identical chains, but calculations, at the explained level, produce different rotamer preferences and overlap values, the latter especially for glutamate, see Table 3.3. Inspection of structure 4GSX reveals that distances between any two atoms, e.g. beta carbons, of the glutamate residue pair (E202,E257) in both chains, A and B, are very similar. This suggests that the difference in values calculated for a water mediated contact between GLU A/B 202 and GLU A/B 257, in the two chains, originates in the relative positions of the secondary structures the residues are part of, an alpha helix (257) and a beta structure (202). Results like this, that are not readily explained, can offer guidelines for further study with perhaps a more comprehensive description of the molecular environment, requiring then a heavier computational effort.

Residue contact results The highest values from contact preference calculations quantifying a water mediated bond were obtained for glutamate in position 202 of envelope, Table 3.3. The second highest ranked residue in this respect was lysine in position 83, located in pre-membrane/envelope protein interface, see Table 3.2 and Figure 3.7. This lysine was also calculated to form the strongest direct bond among those studied. The weakest water mediated bonds were estimated for glutamine in position 55 of pre-membrane protein, contact site depicted in Figure 3.4.

Though the overlap values of this Dengue virus related study can be directly compared, the relative energies of the contacts are at this stage not yet resolved. Preferences of the molecular environment with respect to residue atoms other than the target, influence the energetic cost for a residue to be in a higher than minimum energy rotamer, for a discussion see [47]. A treatment of all intermolecular contact preferences in a uniform setting, i.e. including standardized quantifiers and fragment class specific priors, with knowledge of the intramolecular energetic states, and thermal motion, can be used to estimate the relative energy of the molecular system when its components are varied. The results given by the probabilistic model include entropy in a conceptual scale so, that in the other end, there is large spatial areas of overlap with possibly modest density values, and therefore the entropy change with respect to free molecules is less negative, and in the other end, smaller areas with relatively high density values that correspond to bond strengths that are able to confine the motion. Both these situations correspond to a bound state scenario, where binding strength outweighs entropy loss in contact formation. The overlap can be of two contact preference densities, or of a contact preference density and a target atom distribution in a set of conformations.

Chapter 4

Article

The probabilistic method treated in the preceding chapters was published in article 'Probabilistic prediction of contacts in protein-ligand complexes' [17]. In the present chapter, the main parts of the publication are given in order to complete the contents of this book.

Tables and figures that contain the same information, both in the article and in other chapters of this monograph, are not presented separately, but instead are cross referenced in this chapter to the rest of the monograph. Additionally, first four of the six examples in the original article were excluded from this abridgement, because they were considered redundant with respect to the exposition here, and can be held as supplementary information accessible through the online publication [17]. The kept Examples five and six were numbered one and two, respectively. This chapter can be read as a separate entity preceding a discussion on overall conclusions about the model.

4.1 Abstract

We introduce a statistical method for evaluating atomic level 3D interaction patterns of protein-ligand contacts. Such patterns can be used for fast separation of likely ligand and ligand binding site combinations out of all those that are geometrically possible. The practical purpose of this probabilistic method is for molecular docking and scoring, as an essential part of a scoring function. Probabilities of interaction patterns are calculated conditional on structural X-ray data and predefined chemical classification of molecular fragment types. Spatial coordinates of atoms are modeled using a Bayesian statistical framework with parametric 3D probability densities. The parameters are given distributions *a priori*, which provides the possibility to update the densities of model parameters with new structural data and use the parameter estimates to create a contact hierarchy. The contact preferences can be defined for any spatial area around a specified type of fragment. We compared calculated contact point hierarchies with the number of contact atoms found near the contact

point in a reference set of X-ray data, and found that these were, in general, in close agreement. Additionally, using substrate binding site in cathechol-O-methyltransferase and 27 small potential binder molecules, it was demonstrated that these probabilities together with auxiliary parameters separate well ligands from decoys (true positive rate 0.75, false positive rate 0). A particularly useful feature of the proposed Bayesian framework is that it also characterizes predictive uncertainty in terms of probabilities, which have an intuitive interpretation from the applied perspective.

4.2 Introduction

Atomic level structures are an important source of information for inferring functional aspects about macromolecules and ligands binding to them. For instance, this is illustrated by the substantial amount of existing algorithms and structural data modeling software created for molecular docking and scoring purposes [48], [49], [50], [52], [53]. The Protein Data Bank (PDB) [20] offers the central public access to macromolecular structure files.

Although there is already a large amount of structural data available, it is by no means straightforward to model it reliably. There are several reasons for this, such as the inevitable errors present in experimental results and the "averaging" nature of the measurement process used in the construction of X-ray diffraction data. Moreover, along the conversion from a measurement to a structural coordinate file, several computational approximations and the subjective choices of experimentalists will influence the final outcome. Among the latter sources of variability, two major issues are flexibility of the molecules and computational constraints implemented in the refinement process. The first one is related to thermal motion and static disorder, and the second to biochemical *a priori* information that is always used in the refinement of a structure to create a coordinate file [53], [54]. These are accompanied by crystal packing effects, which also originate from the flexibility of the molecules, uncertainty in orientation and location of small molecules, including water.

It can be argued that for addressing the above-mentioned issues, statistical modeling provides the most promising approach, given its ability to capture uncertainties and errors in data. To meet these goals we introduce a Bayesian statistical method for evaluating atomic level 3D interaction patterns of protein-ligand contacts. Our work is motivated by the previous findings in Rantanen et al., [48], [49], [50] which showcased the usefulness of this kind of a multidisciplinary approach. However, given computational speed related constraints, it has not been possible to pursue these previous Bayesian methods further in contact preference exploration. Therefore, the method discussed here focuses on providing rapid means of computing, together with adjustability and robustness of the statistical model. The latter aspect refers in this context firstly to the constraint that two points in close proximity to each other (with respect to the system size) should not get very different contact preference hierarchies without an easily tractable reason. Secondly, in terms of robustness, the preference prediction model has to be balanced between adhering too closely to the possibly biased overall number of different types of contact atoms in the training data set and using a sole comparison of the probability densities defined for each contact atom type of a molecular fragment. Finally, the adjustability is concerned with both the model structure and the chemistry-based classification of molecular fragments. In our illustrations we consider 24 molecular fragment and 13 contact atom types, exhibiting interactions like hydrogen bonding, dispersion (e.g. aromatic-aromatic) and interactions between charged groups.

The main purpose of this paper is to show how a Bayesian statistical modeling approach can be utilized to make naturally ranked predictions about contact preferences, such that the model itself can be flexibly updated in the presence of novel data and other auxiliary information. Basically, this method is developed to retrieve information to be used in a knowledge-based scoring function. There exist several well performing scoring functions [2], [55] that utilize the experimental knowledge through inverse Boltzmann relation from statistical thermodynamics [35]. These functions depend only on distance between atoms, e.g., a ligand atom and a binding site atom. Our method differs from them in that also directional information is incorporated in the model, which has been shown in case of hydrogen bonding to still significantly improve evaluation of binding energetics from experimental data [56].

Three basic scoring function tasks have been defined [2], of which enrichment of ligands was tested with our method. The test was done through separating catechol-O-methyltransferase (COMT) ligands from decoys using logistic regression on a set of 27 small molecules having similar properties. The receiver operating characteristics (ROC) [57] for the results show that the probabilistic contact preferences give reliable information about the relative affinities in intermolecular contacts. These probabilities can be applied to intramolecular contacts as well. In practice, they are used as part of a molecular docking and scoring routine. The method described in this paper will be integrated as a functionality in the molecular modeling environment BODIL [51].

The structure of the article is as follows. First, data collection and the modeling approach are described. Thereafter, results from two case-studies are presented. Last, implications of the results and some future prospects are discussed.

4.3 Materials and methods

4.3.1 Data collection and processing

We used PDB as the main source of data in this work. Training data for the model was collected from a set of approximately 28000 structure files published before January 1^{st} 2009. The files were selected using the criteria presented in Table 4.1. A reference dataset for model validation was selected under the same criteria as the training set and contained 10361 structure files published between February 2^{nd} of 2009 and 31^{st} August 2011. X-ray structures form the

Has ligands:	Yes
Contains: Protein,DNA,RNA	Yes, No, No
Experimental method:	X-ray diffraction
Min. resolution	2.5 Å

Table 4.1: Criteria for selecting structure files from PDB

largest group of data present in PDB. The bulk of a structure file is the coordinate section, but there is also chemical and biological information, interpreted as metadata, which is necessary for constructing a sensible predictive model. The type of metadata that is most directly deducible from the experimental observations is the atom type. The atom type classification can be considered sufficiently reliable for the higher resolution (< 2.5 Å) structures, however, with at least one exception, which corresponds to the nitrogen (N) and oxygen (O) atoms in a carbamoyl group (-CO-NH_2). In this group, O and N cannot be distinguished solely on the basis of X-ray diffraction data, because of the symmetric structure of the group and very similar electronic densities around both O and N. This is a prime example of a regularly encountered error in the metadata, which, however, can be corrected by reversing the coordinates of O and N. The ligand metadata would impose this error when for instance hydrogen bonding with the ligand would require an acceptor (O) contact, but a donor (-NH_2) contact is given a closer coordinate location in the structure file.

A considerably more difficult problem to handle is the influence of the constraints used in the refinement of the protein structure from experimental data. These constraints generate some unreliability in the coordinates, because only conformations with restricted geometries are allowed for the amino acid chain, which together with limited resolution can lead to artificially distorted conformations of ligand structures. In practice this means that the refinement involves fitting an alleged structure to the experimentally determined electron density map, which does not define all structural features uniquely, especially when the resolution is low [54].

Molecular fragments of pre-defined types (see Tables 4.2 and 4.3) were searched from coordinate ligand structure files in PDB. The search was based on atom types, chemical connectivity and geometry, and the identified fragments were then labelled for use in the extraction of coordinate data from protein structure files. To obtain unique fragment orientations, atoms from within a functional group were, when possible, chosen for the fragment definitions. In order to build a predictive model, the set of 24 fragment classes in Table 4.2 was used while collecting a dataset of approximately 70,000 contacts, representing the 13 contact atom types, i.e. target classes in Table 4.3.

Regarding the contact classes in Table 4.3, for example, the class **C3** represents a pure van der Waals contact [58] and class **C4,** a hydrogen donor in a possible weak hydrogen bond in addition to a van der Waals contact [59], [60]. Aromatic carbons (**C5, f11**) can participate in both of the typical **C3**

Class	Description
f2	Hydroxyl oxygen bonded to a non-planar aliphatic structure
f3	Hydroxyl oxygen bonded to an aromatic structure
f5	Carbonyl oxygen (excluding those belonging to f9 and f10)
f6	Oxygen of a carboxyl group
f7	Carbamoyl oxygen
f8	Oxygen bonded to a phosphate group
f9	Amide group oxygen bonded to a non-aromatic structure
f10	Amide group oxygen bonded to an aromatic structure
f11	Secondary carbon in an aromatic structure
f12	Secondary carbon in a non-aromatic structure
f13	Primary carbon (with one hydrogen)
f17	Fluorine bonded to an aromatic structure
f18	Fluorine bonded to a non-aromatic structure
f20	Chlorine bonded to an aromatic structure
f21	Chlorine bonded to a non-aromatic structure
f22	Nitrogen in an aromatic structure (without a substituent)
f23	Nitrogen in a non-aromatic planar ring structure (without a substituent)
f26	Amino (primary) nitrogen singly bonded to a non-aromatic structure
f27	Amino (primary) nitrogen bonded to an aromatic structure
f29	Amino (primary) nitrogen singly bonded to a planar structure
f34	Bromine bonded to an aromatic structure
f35	Bromine bonded to an aliphatic structure
f36	Iodine bonded to an aromatic structure
f37	Iodine bonded to an aliphatic structure

Table 4.2: Fragment classes used in this study. Main forms of intermolecular interaction for these fragment types are hydrogen bonding, dispersion, charged group based electrostatic and halogen bonding. The fragment classification was partly adopted from the previous work of Rantanen et al. (see Introduction).

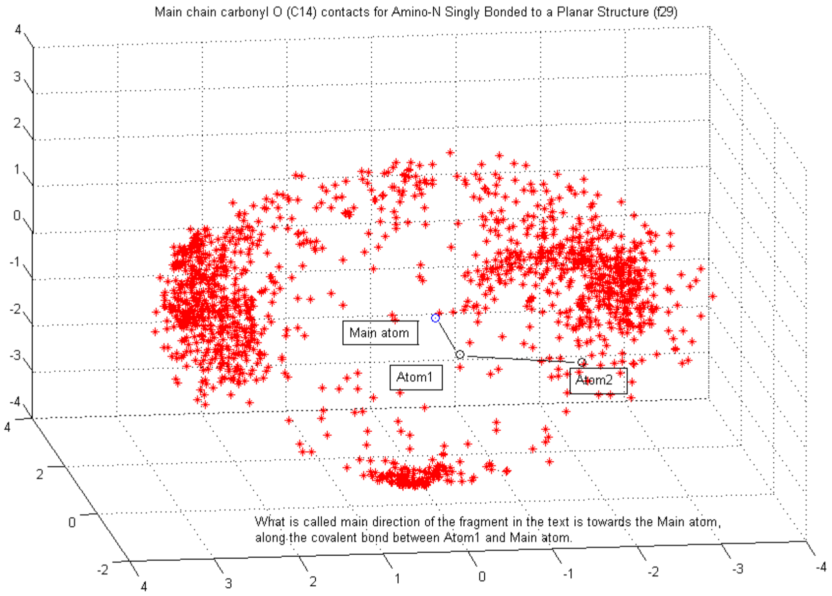


Figure 4.1: Contact atom (target) cloud formed by main chain carbonyl oxygens (**C14**) around fragment type (**f29**) (amino nitrogen singly bonded to a planar structure). In the reference frame where targets are modeled, polar angle value $\theta = 0$ corresponds to what is in this paper called the main direction of the fragment — the direction of the vector from Atom1 to Main-atom, and $\theta = \frac{\pi}{2}$ corresponds to the plane that includes Main-atom in the origo and to which the main direction is perpendicular. Azimuthal angle ϕ measures angular deviation from the plane of the fragment, so that the center of the smaller cluster below the fragment is (in the model frame) approximately in direction $[\theta = \pi, \phi = \frac{3\pi}{2} = -\frac{\pi}{2}]$. A fragment is defined by determining the characteristics of an atom triplet: Main-atom, Atom1 and Atom2. Main-atom is covalently bonded to Atom1, and Atom1 is covalently bonded to Atom2. Chemical properties of the Main-atom primarily determine the class of a fragment.

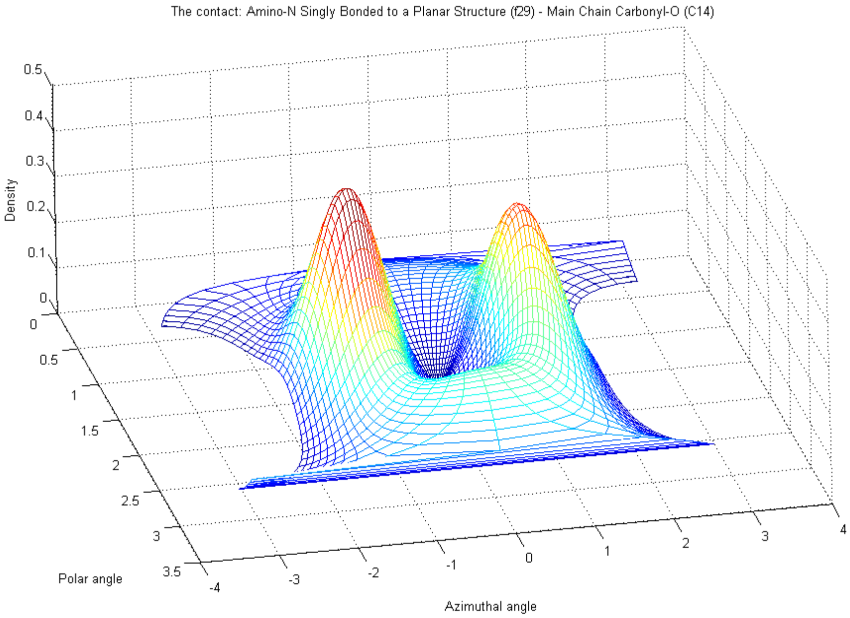


Figure 4.2: Probability density modeling the target cloud of Figure 4.1, which is depicted in the same reference frame with the density. They can be interpreted as overlayed such that elevations in the density correspond to dense areas in the cloud of data points. The main direction of the fragment, as described in the caption of Figure 4.1, is defined by $\theta = 0$ in the reference frame of the model, but corresponds in this figure to $[\theta = \frac{\pi}{2}, \phi = 0]$, where θ is the polar angle and ϕ is the azimuthal angle.

Class	Description
C3	Carbon of a methyl group
C4	Alpha carbon
C5	Carbon in an aromatic structure
C6	Sulfur of a thioether group
C7	Sulfur of a thiol group
C8	Nitrogen of an amide group
C9	Nitrogen of indole, imidazole and guanido groups
C10	Nitrogen of an amino group
C11	Oxygen of a carboxamide group
C12	Oxygen of a carboxyl group
C13	Oxygen of a hydroxyl group
C14	Main chain carbonyl oxygen
C15	Main chain amide nitrogen

Table 4.3: Classification of contact atoms, or targets. The target classification was partly adopted from the previous work of Rantanen et al. (see Introduction).

and **C4** interactions [61]. Halogen bonds have a role in biological processes [21] and therefore Fluorine [62], Chlorine, Bromine and Iodine are considered as so called fragment Main-atoms, as shown in Table 4.2. The target atoms in proteins, identified with three distance criteria ($\leq$ 3.3Å for alleged H-bonds and charged groups, $\leq$ 3.7Å for probable dispersion and $\leq$ 3.9Å for halogen bonds), were classified during the search using three criteria: 1) element, 2) amino acid residue and 3) side or main chain atom. The interaction was defined as between a fragment type and target type, or between nuclei, mediated by protons and/or electrons.

A fragment was defined by an atom triple: Main-atom, Atom1 and Atom2, and at least the Main-atom was given the following characteristics: element, covalent bond count, aromaticity and possibly functional group, see Tables 4.2 and 4.3. These characteristics were used in collecting data from PDB, resulting in coordinates with metadata. The aromaticity of an atom was decided using PDB Ligand Dictionary through PDBeChem [45].

In addition to classification, target atoms have to be put in one coordinate system, i.e. fragments are superimposed. This was done using an elementary translation-rotation: first the database coordinates of the Main-atoms were translated to origin, which creates a new dataset (F_{database} below in eq. 4.1), and then a rotation operation was defined to connect the fragments reference frame to a common coordinate system. This requires solving the following matrix equation:

$$F_{\text{target}} = R \cdot F_{\text{database}}, \tag{4.1}$$

where R refers to a 3×3 rotation; $F = [\bar{r}_1,\bar{r}_2,\bar{r}_3]$, $\bar{r}_i = [x_i, y_i, z_i]^T$ and $\bar{r}_3 = \bar{r}_1 \times \bar{r}_2$, i.e., $\bar{r}_3$ is the cross product of $\bar{r}_1$ with $\bar{r}_2$.

Thus, we have the equation,

$$R = F_{\text{target}} \cdot F_{\text{database}}^{-1}, \tag{4.2}$$

which was solved for each fragment. The resulting R was then used in the translation-rotations of the respective target atom position vectors to the common coordinate system. When the data is collected as mentioned above, as well as classified and coordinate systemized in this manner, the process results in a collection of three-dimensional distributions of points that present measured relative positions of specified atoms with respect to specified fragment types. These distributions were then modeled with 3D probability densities described below.

4.3.2 Statistical modeling

To obtain predictive distributions for contact preferences we utilize a Bayesian framework where the observed 3D coordinates in the training data are modeled with interconnected parametric 1D densities, such that the parameters are provided *a priori* uncertainty characterizations in terms of probability distributions. The prior distributions enable regulation of parameter estimates in order to prevent them from depending solely on the observed data, which is desirable especially under the circumstances where the data generation process is known to harbor intrinsic biases. Also, regularization of model parameter estimates with the prior information is most crucial when certain class pairs have only very sparse training data, in which case unsmoothed estimates can be strongly biased.

The core distribution we utilize for characterizing coordinate variability is the von Mises-Fisher distribution (vMF) which is widely applied for modeling directional data. Separate probability densities for all three coordinates were necessary in order to capture the properties of the target atom clouds in a uniform setting (details provided below). Spatially the most complex (multimodal) observed differences in target atom distributions are found around the main direction of the fragment, and to a somewhat lesser extent with respect to distributions of polar angle, i.e. angular deviation from the main direction, see Figures 4.1 and 4.2. The distance distributions are given *a priori* as many modes as the corresponding polar angle distributions have, though in most cases they practically form a unimodal density, but not always. This is explained more thoroughly later in this section. The variables and parameters of the densities used in our work are specified in Tables 4.4 and 4.5.

Symbol	**Variable**
ρ	distance
ϕ	azimuthal angle
θ	polar angle

Table 4.4: The spherical polar coordinates.

Symbol	Description	Treated as
μ_ρ	Mean of Normal density for the distance	Random variable
σ^2_ρ	Variance of Normal density for the distance	$Var(\rho) = \frac{1}{k}\sum_{l=1}^{k}\rho_l^2 - (\frac{1}{k}\sum_{l=1}^{k}\rho_l)^2$, for k observed distances.
λ_i	Expected direction, i:th von Mises mixture component for θ	Random variable
κ_i	Concentration parameter, i:th von Mises mixture component for θ	Random variable
μ_{ij}	Mean, j:th Normal mixture component for ϕ related to i:th von Mises mixture component	Random variable
σ^2_{ij}	Variance, j:th Normal mixture component for ϕ related to i:th von Mises mixture component	$Var(\phi)_{ij} = \frac{1}{k}\sum_{l=1}^{k}\phi_l^2 - (\frac{1}{k}\sum_{l=1}^{k}\phi_l)^2$, for k observed angles.

Table 4.5: Parameters for the model densities.

Let γ denote a generic set of parameters specifying a density in a 3D space. Then, the probability density in spherical polar coordinates $f_{fC}(\rho, \phi, \theta|\gamma)$ for fragment class f and target class C is assumed to be of the piece-wise defined form

$$\begin{aligned} & f_{fC}(\bar{r}|\{\mu_{i,\rho}\},\{\sigma^2_{i,\rho}\},\{\lambda_i\},\{\kappa_i\},\{\mu_j\},\{\sigma_j\}) \propto \\ & \sum_{i=1}^{N_{vM}} I(\rho \in c_i)I(\theta \in b_i) z_i f_{vM}(\theta|\lambda_i,\kappa_i)\cdot f_N(\rho|\mu_{i,\rho},\sigma^2_{i,\rho})\cdot \\ & \cdot[\sum_{j=1}^{N_{N,i}} I(\phi \in a_{ij}) w_{ij} f_N(\phi|\mu_{ij},\sigma_{ij})] \end{aligned} \tag{4.3}$$

where $\bar{r} = [\rho\sin(\theta)\cos(\phi), \rho\sin(\theta)\sin(\phi), \rho\cos(\theta)]^T$, the b_i divide $(0, \pi)$ and c_i divide $(0, \rho_{fC}^{(cutoff)})$ to non-overlapping intervals. The limit $\rho_{fC}^{(cutoff)}$ is the maximum distance used in collecting the target atom locations for a fragment class and target class pair (fC). The intervals b_i and c_i are associated with weight $z_i, \sum_{i=1}^{N_{vM}} z_i = 1$. The a_{ij} then, are non-overlapping intervals of $(0, 2\pi)$ associated with the weights $w_{ij}, \sum_{j=1}^{N_{N,i}} w_{ij} = 1$ (see below) and the different density components and parameters are defined using conventional nomenclature as:

$$f_N^{(i)}(\rho \mid \mu_{i,\rho}, \sigma_{i,\rho}^2) = N_{i,\rho} \cdot \exp(-\frac{1}{2\sigma_{i,\rho}^2}(\rho - \mu_{i,\rho})^2) \tag{4.4}$$

(i:th Normal component, $N_{i,\rho}$ normalizing term),

$$f_{vM}^{(i)}(\theta | \lambda_i, \kappa_i) = N_i \cdot \exp(\kappa \cos(\theta - \lambda_i))$$

(i:th von Mises component, N_i normalizing term) and

$$f_N^{(ij)}(\phi | \mu_{ij}, \sigma_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp(-\frac{1}{2\sigma_{ij}^2}(\phi - \mu_{ij})^2)$$

(j:th Normal component, relating to i:th von Mises component).

Equations 4.4 show the forms of the densities in the particular coordinate system, or reference frame, that is used for modeling and in which the main direction of the fragment coincides with the positive z-axis. A likelihood function is obtained as a product from the values of the components of the density (eq. 4.3) for n_i (or n_{ij}) points representing each of the included regions in the sample space,

$$\begin{gathered} L(\{\mu_{i,\rho}\}, \{\sigma_{i,\rho}\}, \{\lambda_i\}, \{\kappa_i\}, \{\mu_{ij}\}, \{\sigma_{ij}\}) = \\ \sum_{i=1}^{N_{vM}} L_i(\mu_{i,\rho}, \sigma_{i,\rho}, \lambda_i, \kappa_i) \cdot \left[\sum_{j=1}^{N_{N,i}} L_{ij}(\mu_{ij}, \sigma_{ij}) \right], \end{gathered} \tag{4.5}$$

where

$$L_i \propto z_i^{n_i} \cdot e^{\kappa_i \sum_{k=1}^{n_i} \cos(\theta_k - \lambda_i)} \cdot e^{-\frac{1}{2\sigma_{i,\rho}^2} \sum_{k=1}^{n_i} (\rho_k - \mu_{i,\rho})^2}, \tag{4.6}$$

$$L_{ij} \propto w_{ij}^{n_{ij}} \cdot e^{-\frac{1}{2\sigma_{ij}^2} \sum_{k=1}^{n_{ij}} (\phi_k - \mu_{ij})^2}. \tag{4.7}$$

The structure of the density (eq. 4.3) was chosen based on investigations of the forms of the target atom clouds. For example, use of normal densities for the distance data was supported by large p-values in Kolmogorov-Smirnov normality test and the numbers of components needed in the angle dependent part of the density (i.e. N_{vM} and $N_{N,i}$) were automatically chosen based on frequency distribution of binned data. This was done for both angles (θ and ϕ) by connecting the heights of the adjacent bars of the histogram, creating a sequence of values, in which every change of sign corresponds to a local minimum or a maximum. The number of maxima was restricted to the interval [1,5], was used to represent the number of modes in the density. The number of maxima was restricted by either reducing or increasing the number of bins in case the result would be outside the given interval.

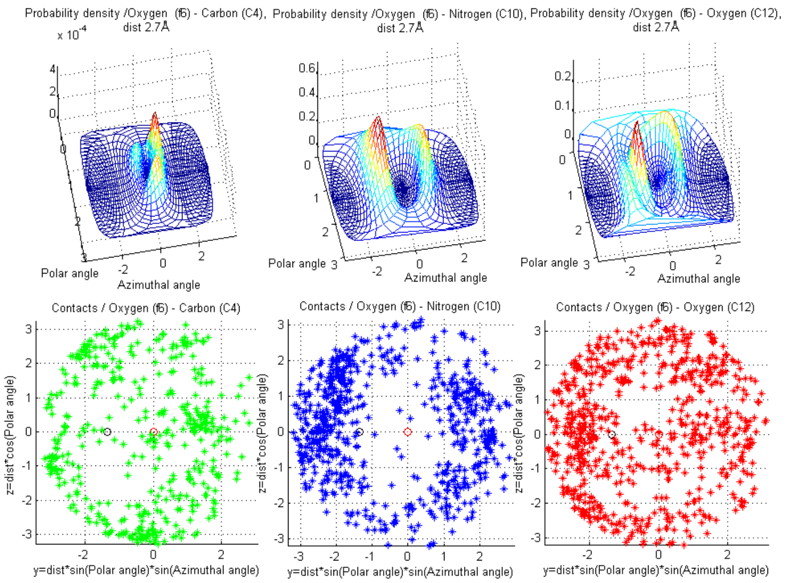


Figure 4.3: Contacts for carboxyl oxygen (**f6**), showing differences in spatial arrangement of contacts among three target classes - **C4** (alpha carbon), **C10** (amino nitrogen) and **C12** (carboxyl oxygen). The scatter plots contain all the target atoms found in the fragment class **f6** training data for these target classes.

In the separation of variables, the azimuthal angle and the distance are conditioned on a polar angle interval, see the equations in (4.4). The angular part reflects arc-like structures around the main direction of the fragment (for examples of this see Figures 4.3-4.5). The angular deviation from the main direction is the leading variable in the sense that a multimodal azimuthal angle distribution (i.e. around the main direction) and a unimodal distance distribution are defined separately inside each polar angle segment. The idea behind this is that the peak of a polar angle density is an indicator of the strength of the interaction between a fragment and a target. The smaller the angle, the stronger the interaction, and if there is any effective multimodality in the distance density, the modes should coincide with the modes of the polar angle density. The azimuthal angle distribution thus completes the directional structure within each polar angle mode.

The uncertainties of the distance and the azimuthal angle variances are difficult to model due to the data generation process, the limitations of which were discussed above, and therefore, we use the standard maximum likelihood estimates calculated marginally from observed coordinates. On the other hand,

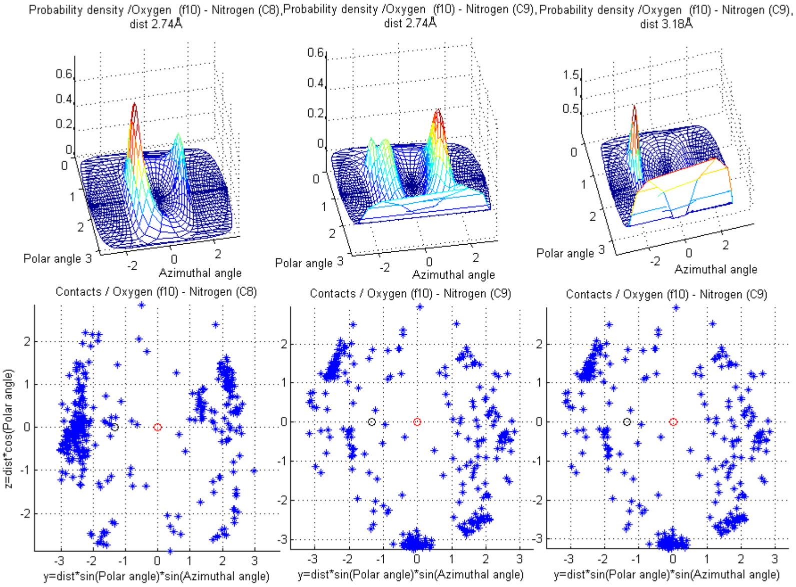


Figure 4.4: Contacts for amide oxygen (**f10**) showing differences in contact arrangements for two target classes - **C8** (carbamoyl nitrogen) and **C9** (imidazole, guanido or indole nitrogen), and also distance dependence for class **C9**. The scatter plots contain all contacts found in the training data for these fragment and target class pairs. Note that the target atom clouds in the middle and on the right are the same.

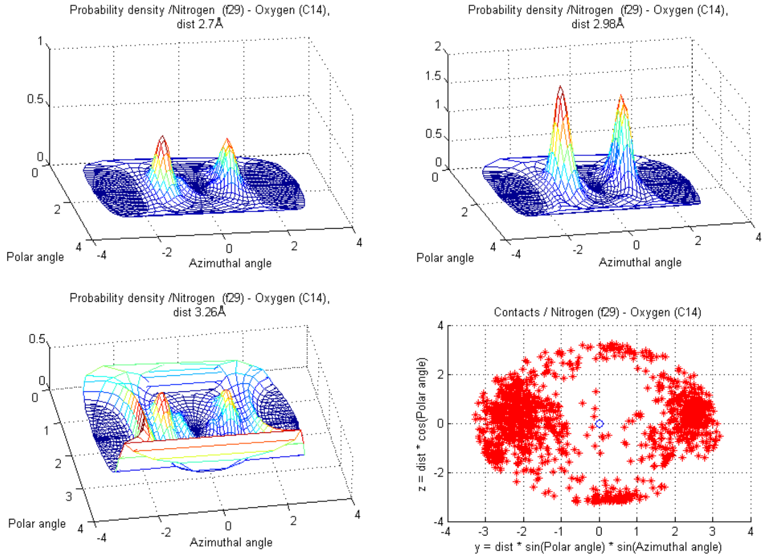


Figure 4.5: Contacts for nitrogen, singly bonded to a planar structure (**f29**) — e.g. in carbamoyl group, showing distance dependence for target class **C14** (carbonyl oxygen**).** The scatter plot contains all **C14** target atoms in the training data and the densities show which directions are emphasized at distances 2.7, 2.98 and 3.26 Å.

the means $\mu_{i,\rho}$, λ_i and μ_{ij} are central parameters representing a measure of the strength of the interaction between the fragment and the target.

Parameter prior densities

A prior distribution in Bayesian statistics can either be used to model uncertainty about a parameter or to include *a priori* knowledge, or beliefs, in the model [63]. Both of these uses are necessary for our modeling purposes.

Prior densities can be chosen in various ways such that they are either conjugate distributions for a particular likelihood function, or some other probability densities possessing required statistical properties, such as an asymmetry. Priors utilized here for individual parameters in the model densities are summarized in Table 4.6. In Table 4.6, $I_0(.)$ is the zeroth order modified Bessel function of the first kind. For the von Mises and Normal distributions, conjugate priors are used (for the mathematical derivations related to these distributions see [64], [65], [66]).

Symbol	**Parameter Type**	**Functional Form of Prior**
$\mu_{i,\rho}$	Mean of ρ (i:th Normal component)	$e^{-\frac{1}{2\sigma_0^2}(\mu_{i,\rho}-\mu_0)^2}$; μ_0 and σ_0^2 are constants.
κ_i	Concentration of θ (i:th von Mises component)	$I_0^{-n_i}(\kappa_i - \kappa_0)$; κ_0 and n_i are constants.
λ_i	Mean of θ (i:th von Mises component)	$e^{\kappa_{\lambda_j} \cos(\lambda_j - \lambda_0)}$; λ_0 is a constant.
μ_{ij}	Mean of ϕ (j:th Normal component related to i:th von Mises component)	$e^{-\frac{1}{2\sigma_{\phi,0}^2}(\mu_{ij}-\mu_{\phi,0})^2}$; $\mu_{\phi,0}$ and $\sigma_{\phi,0}^2$ are constants.

Table 4.6: Prior densities for model parameters. More details can be found in section Parameter prior densities.

Our prior density for the parameters is a slightly modified version of the distributions considered in [22] and [67]. The density has the form:

$$p(\{\mu_{i,\rho}\}, \{\kappa_i\}, \{\lambda_i\}, \{\kappa_{\lambda_i}\}, \{\mu_{ij}\} | \mu_0, \sigma_0, \kappa_0, n_i, \lambda_0, c_i, \kappa_{\lambda,0}, \mu_{\phi,0}, \sigma_{\phi,0}) \propto$$

$$\prod_{i=1}^{N_{vM}} e^{-\frac{1}{2\sigma_0^2}(\mu_{i,\rho}-\mu_0)^2} \cdot I_0^{-n_i}(\kappa_i - \kappa_0) \cdot e^{\kappa_{\lambda_i} \cos(\lambda_i - \lambda_0)} \cdot I_0^{-c_i}(\kappa_{\lambda_i} - \kappa_{\lambda,0}) \cdot \tag{4.8}$$

$$\cdot \left[\prod_{j=1}^{N_{N,i}} e^{-\frac{1}{2\sigma_{\phi,0}^2}(\mu_{ij}-\mu_{\phi,0})^2} \right]$$

In equation (4.8) hyperparameters n_i and c_i represent measures of modeler's belief in the expected values of the concentrations. The larger the value of the hyperparameter, the more the prior is concentrated around κ_0 or $\kappa_{\lambda,0}$. A default choice for any of these hyperparameters is the number of observations available for calculating the estimates for κ_0 and $\kappa_{\lambda,0}$.

The posterior distribution

In Bayesian statistics, learning from observations takes place through the posterior distribution which is accessible from the joint probability density defined for the data and the parameters. For our piece-wise defined likelihood, the joint density is formally defined as

$$\begin{aligned} & p(\{\mu_{i,\rho}\},\{\lambda_i\},\{\kappa_i\},\{\mu_{ij}\}|\{\mu_{i,p}\},\{\lambda_{i,p}\},\{\kappa_{i,p}\},\{\mu_{ij,p}\}) = \\ & L(\{\mu_{i,\rho}\},\{\lambda_i\},\{\kappa_i\},\{\mu_{ij}\})\cdot \\ & \cdot\, p(\{\mu_{i,\rho}\},\{\lambda_i\},\{\kappa_i\},\{\mu_{ij}\}|\mu_0,\lambda_0,\kappa_0,\mu_{\phi,0}) \end{aligned} \tag{4.9}$$

The posterior density (eq. 4.9) has the same functional form as the prior density (eq. 4.8), but with updated parameters. As shown in the equation (4.9), the prior parameters μ_0, λ_0, κ_0 and $\mu_{\phi,0}$ are updated to $\mu_{i,p}$, $\lambda_{i,p}$, $\kappa_{i,p}$ and $\mu_{ij,p}$, respectively, in the usual manner in Bayesian inference. In contrary, the remaining parameters are determined either directly from the data or given a suitable value based on chemical knowledge.

In order to define contact preferences, every fragment class and target class pair has to be specified with some characteristics. Maximum *a posteriori* (MAP) estimates are in this respect suitable when the associated densities are (piecewise) unimodal. The MAP estimates of the parameters are defined and updated with new data according to the formulae in Table 4.7. The prior parameter σ_0 was given a constant value $0.01(\text{Å}^2)$ and R_0 was defined separately for each fragment type.

Updated parameters and the probability mass in a reference volume

In our method, to evaluate the plausibility of a contact atom type in a given spatial area, the probability mass within in this volume is evaluated. The mass is calculated using the model densities (4.4) with updated parameters, see Table 4.7. The spatial area, or volume, is defined by a distance interval and a solid angle (i.e. intervals polar and azimuthal angles), and can be arbitrarily located. The volume that contains all target locations is defined through the intervals $[0,\rho_{fC}^{(cutoff)}]$, $[0,\pi]$ and $[0,2\pi]$ for the distance, polar angle and azimuthal angle, respectively. The cutoff $\rho_{fC}^{(cutoff)}$is the maximum distance used when collecting data for a fragment class and target class pair (fC). The size

Posterior variable	**MAP estimate**	**Definitions and estimates**
Mean of distance	$\hat{\mu}_{i,p} = \frac{\bar{y} * \sigma_0^2 + \mu_0 * \sigma_{i,\rho}^2}{\sigma_0^2 + \sigma_{i,\rho}^2}$	$\bar{y} = \frac{1}{m_i} \sum_{l=1}^{m_i} y_l$, $\sigma_{i,\rho}^2 = \frac{1}{m_i} \sum_{l=1}^{m_i} \rho_l^2 - (\frac{1}{m} \sum_{l=1}^{m} \rho_l)^2$
Mean of polar angle	$\hat{\lambda}_{i,p} = \lambda_{i,p}$	$\lambda_{i,p} = \arctan(\frac{R_{o,i} * \sin(\lambda_0) + \sum_{l=1}^{k} \sin(\theta_l)}{R_{o,i} * \cos(\lambda_0) + \sum_{l=1}^{k} \cos(\theta_l)})$, $R_{o,i} = \kappa_{\lambda,i} / \kappa_i$ and $\lambda_0 = const.$
Concentration of polar angle	$\hat{\kappa}_{i,p} = \kappa_{i,p}$	$\kappa_{i,p}$ numerically by equalizing model and data variances.
Mean of azimuthal angle	$\hat{\mu}_{ij,p} = \frac{\bar{y} * \sigma_{\phi,0}^2 + \mu_{\phi,0} * \sigma_{ij}^2}{\sigma_{\phi,0}^2 + \sigma_{ij}^2}$	$\bar{y} = \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} y_l$, $\sigma_{ij}^2 = \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} \phi_l^2 - (\frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} \phi_l)^2$

Table 4.7: MAP estimates of parameters.

of the volume can be chosen to be large, when for example contact preferences on either side of the fragments plane are investigated. Alternatively, the size of the volume can also be small, depending on the situation under investigation.

The spatial information content of the model is coded in the particular functional form of the probability density, but if one would solely rely on probability densities, it could easily happen that a scarce contact atom type could obtain a hierarchically higher preference position than a relatively often encountered type. This could happen in a spatial area where all the few contacts of the former type are observed. These kinds of problems are avoided and results for different contact types made more directly comparable by supervising the model such that chemically more likely contacts are paralleled, as well as the less likely. The model supervision was here achieved by multiplying the probability masses with target type specific weights that are calculated from three parameters electronegativity, softness and mean distance. Softness of an element e, one from the group $\mathbf{G} = \{$C, N, O, S, F, CL, BR, I$\}$, was defined as twice the mean value of hardness among the elements in $\mathbf{G}$, minus hardness of the element e. Numerical values for absolute hardness were taken from Parr et al.[68]. The electronegativity and softness were used to represent the tendency of an element to obtain partial charge in a compound. The third parameter, mean distance, is a measure of the strength of the interaction between a fragment and a target, and the numerical value given to it was the arithmetic mean of the distances in the training data. These parameters are used to calculate the weights as proportional to Coulomb force between the partial charges at the mean distance, i.e.

$$p_{fC} = \frac{e_f \cdot e_C}{\mid r_{fC} \mid^2}, \tag{4.10}$$

where e_f and e_C are the obtained partial charges for a fragment Main-atom and a target atom, respectively, and $\mid r_{fC} \mid$ is the mean distance between the

f\C	C3	C4	C5	C6	C7	C8	C9
f2	0.0015	0.0077	0.1259	0.0012	0.0015	0.0970	0.1355
f3	0.0014	0.0069	0.1217	0.0016	0.0014	0.0990	0.1312
f5	0.0027	0.1316	0.0876	0.0020	0.0020	0.1233	0.1246
f8	0.0018	0.0884	0.0896	0.0020	0.0006	0.1297	0.1317
f11	0.0103	0.0542	0.1248	0.0119	0.0120	0.0777	0.0301
f18	0.0013	0.1002	0.1000	0.0014	0.0014	0.1391	0.1237
f20	0.0109	0.0691	0.0698	0.0740	0.0713	0.0864	0.0863
f22	0.0068	0.0365	0.1052	0.0079	0.0080	0.0594	0.0229
f23	0.0058	0.0313	0.1016	0.0062	0.0066	0.0103	0.0187
f26	0.0020	0.0176	0.0966	0.0025	0.0022	0.0037	0.1563
f27	0.0023	0.0130	0.1083	0.0027	0.0025	0.0040	0.0656
f34	0.0112	0.0698	0.0712	0.0266	0.0739	0.0884	0.0871
f36	0.0110	0.0661	0.0729	0.0696	0.0835	0.0814	0.0779

Table 4.8: Prior probabilities for fragment classes that were used in this study, target classes C3 - C9.

Main-atom and the target atom. The motivation for this prior is that the intermolecular interactions are mainly electrostatic, despite of the fact that they occur in many different forms, e.g., between a permanent dipole and an induced dipole, or between two induced dipoles, known as a London dispersion.

In the above formulation, it is assumed that a generic *a priori* information can be accurately utilized when modeling an interaction between f and C. It would also be possible to use calculated energies of some simplified fragment-target model as the *a priori* information, but the described approach is chosen because of its simplicity and independence of molecular details, which follows from utilizing element specific, measurable parameters, i.e. ionization energy and electron affinity [68]. Calculated prior probabilities relevant for this study are given in Tables 4.8 and 4.9.

4.3.3 Hierarchy calculations

In order to calculate the spatially dependent hierarchies around an arbitrary reference point

$$\bar{r}_{ref,z} = \rho_{ref} \cdot [\cos(\phi_{ref}) \cdot \sin(\theta_{ref}), \sin(\phi_{ref}) \cdot \sin(\theta_{ref}), \cos(\theta_{ref})], \qquad (4.11)$$

we defined intervals in spherical polar coordinates:

$$\Delta = [\rho_1, \rho_2, \theta_1, \theta_2, \phi_1, \phi_2], \qquad (4.12)$$

which define a volume that includes $\bar{r}_{ref,z}$, e.g. $\rho_{ref} = \frac{1}{2}(\rho_1 + \rho_2)$, $\theta_{ref} = \frac{1}{2}(\theta_1 + \theta_2)$ and $\phi_{ref} = \frac{1}{2}(\phi_1 + \phi_2)$. The reference point $\bar{r}_{ref,z}$ is defined in the reference frame that is used for modeling the data, and in which the fragment is in the

f\C	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**
f2	0.1003	0.1036	0.1125	0.1100	0.1068	0.0960
f3	0.1073	0.1026	0.1092	0.1152	0.1090	0.0934
f5	0.1279	0.0036	0.1306	0.1384	0.0017	0.1238
f8	0.1351	0.0037	0.1347	0.1519	0.0019	0.1288
f11	0.0787	0.0894	0.1718	0.1716	0.0911	0.0763
f18	0.1380	0.0022	0.1366	0.1350	0.0012	0.1198
f20	0.0882	0.0891	0.0896	0.0894	0.0901	0.0858
f22	0.0641	0.1193	0.1938	0.1922	0.1200	0.0638
f23	0.0050	0.2090	0.2059	0.1926	0.2020	0.0049
f26	0.0018	0.1822	0.1815	0.1749	0.1768	0.0020
f27	0.0018	0.2025	0.2013	0.1939	0.2002	0.0018
f34	0.0890	0.0955	0.1011	0.1025	0.0991	0.0846
f36	0.0841	0.0890	0.0871	0.1069	0.0903	0.0803

Table 4.9: Prior probabilities for fragment classes that were used in this study, target classes C10 - C15.

(-z)(-x)-plane. The Main-atom (see section Data collection and processing) is at the origin, Atom1 on the negative z-axis and Atom2 in a point $(- \mid x_2 \mid , - \mid y_2 \mid, 0)$, i.e. in the plane defined by the negative z- and x-axes. On the other hand, in the reference frame used for the graphical representations in this article, $\bar{r}_{ref,z}$ (eq. 4.11) is transformed to

$$\bar{r}_{ref} = \rho_{ref} \cdot [\cos(\theta_{ref}), \cos(\phi_{ref}) \cdot \sin(\theta_{ref}), \sin(\phi_{ref}) \cdot \sin(\theta_{ref})], \qquad (4.13)$$

which is in a reference frame where the fragment is in (-x)(-y)-plane, see Figures 2.8 and 4.1.

The probability masses in the volume defined by Δ (eq. 4.12) are evaluated using the model densities with the updated parameter values. Technically the calculations are done either with series expansions, see e.g. the equations 7.1.1., 7.1.7., 7.1.22 and 9.6.34 in [25] or directly as Riemann sums.

The fC-specific probability mass is the factor that gives a contact atom type C its rank, in the fragment class f related, and around a reference point $\bar{r}_{ref}$ defined hierarchy. Namely, the bigger the mass, the more probable the contact atom type. The volume can cover a larger portion of the neighborhood of the fragment, for example, the hemisphere on either side of the plane of the fragment.

A hierarchy can also be defined for example among the fragment types (f), with respect to a representative of a target class C around $\bar{r}_{ref}$. The motivation for choosing the probability density and the estimation procedures of the model parameters as described in Methods, is that they provide a rapid and flexible way to capture the relevant features of the target atom distributions, without relying on fine details of the target atom clouds, which are potentially misleading due to the intrinsic uncertainties in the data generation process.

4.4 Results

The calculated hierarchies are exemplified in the original article [17] (Examples 1-4), as well as in this thesis (chapter Statistical modeling), and are therefore not included in this section.

The functionality of the introduced Bayesian method of finding hierarchies is here illustrated by two case-studies. In order to assess the reliability of the results, standard errors for all results were determined with a bootstrap type procedure [69], which is described together with the corresponding results.

4.4.1 Example 1: Direct contacts of R-norepinephrine

Here we consider the hydrogen bonding and aromatic interaction preferences of norepinephrine (also known as noradrenaline; PDB ligand identifier: LT4). The molecular environment of this example is the norepinephrine binding site in chain B of human phenylethanolamine N-methyltransferase (PNMT) from PDB entry 3HCD. PNMT catalyses adrenaline synthesis with coenzyme S-adenosyl-L-methionine (AdoMet). In the structure of 3HCD AdoMet is replaced by its demethylated form S-adenosyl-L-homocysteine (AdoHcy) to study the binding mode of LT4 [71]. The X-ray resolution of the entry 3HCD is 2.39 Å, which is near the upper limit considered in our study (< 2.5 Å, see Table 4.1). As discussed previously, this means that some precaution is necessary while deducing interactions from the structure. Consequently, the statistical nature of our method is helpful, since the probability densities can indicate a certain relative location that is associated with what is experimentally observed in other structures. This enables the quantification of the relation in question as a probability.

LT4 has three hydroxyl groups, a terminal amino group and a six-carbon aromatic ring as its functional groups. The most preferred target class for any of these functional groups is defined as a class for which the product of the probability density peak value (eq. 4.3) and the class conditional prior probability (see Tables 4.8 and 4.9) has the highest numerical value.

Two of the hydroxyl oxygens are bonded to the aromatic ring (cathecol hydroxyl groups), and based on the model, prefer a nitrogen from histidine or arginine side chain as a contact. The aromatic ring carbons then prefer aromatic carbons, i.e. phenylalanine, histidine, tyrosine or tryptophan is a likely contact residue. The aromatic carbons also have strong contacts from the hydrophilic targets, for example carboxyl oxygens. The hydroxyl group of the aliphatic tail prefers lysine and the terminal amino group prefers glutamic acid/glutamate or aspartic acid/aspartate.

These *a priori* preferences are not related to LT4, instead only the 3D structure of the interactions is. Some directional aspects related to this example are illustrated in Figures 4.6 - 4.10. Figures 4.7 and 4.8 present hydrogen bonding contacts for the R-norepinephrine tail. There are two carboxyl oxygen (**C12**) contacts for the LT4 hydroxyl group, namely GLU B219 and ASP B267. The former is at a distance less than the maximum length used in this study for a oxygen donor and oxygen acceptor hydrogen bond, i.e. 3.14 Å < 3.30 Å.

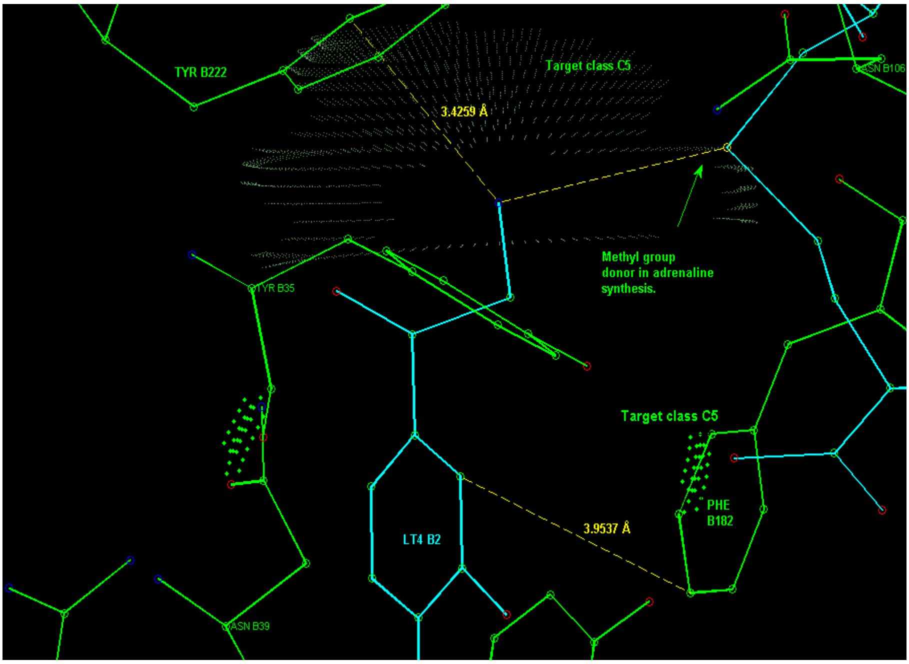


Figure 4.6: The aromatic phenylalanine contact PHE B182 in PNMT (see text of section Example 1) for the LT4 aromatic ring. Also depicted in the figure is the proximity (closest 3.4 Å) of the TYR B222 aromatic ring to the amino group of LT4. Though the distance and orientation of the ring fit well to the hydrogen bond donor - aromatic ring interaction scheme, the relative direction is such that TYR B222 corresponds to probability density values smaller than 20 % of the peak value. Therefore, this might not be a strong contact for the amino group, but possibly has a guiding task in the binding process. The depicted distance between amino group and the methyl group donating sulfur atom is 5.66 Å.

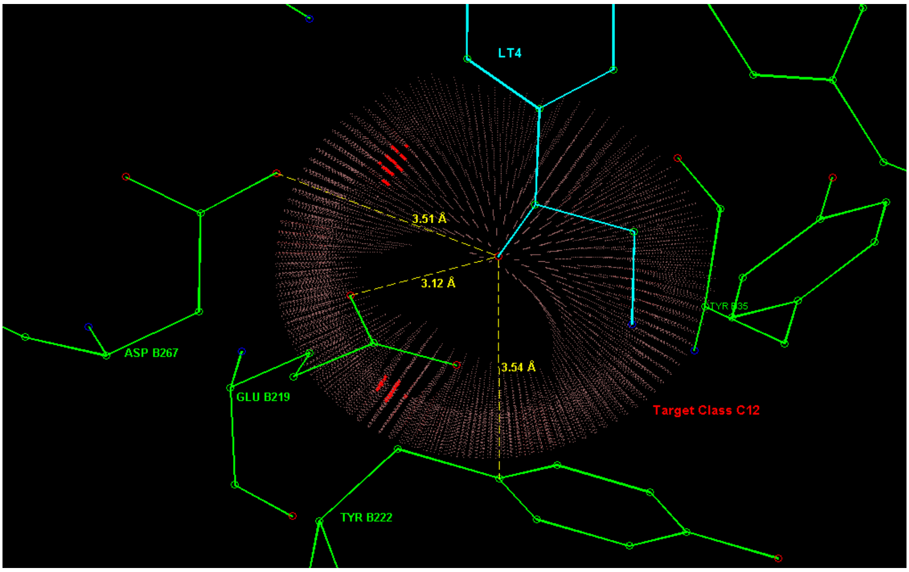


Figure 4.7: Contacts in the active site of PNMT (a methyltransferase) for R-noradrenaline tail hydroxyl group. Three amino acid residues (ASP B267, GLU B219 and TYR B222) were considered as contacts.

According to the model, its direction of approach is not typical for a hydrogen bond, but because the same GLU B219 residue is simultaneously a contact for the adjacent amino group, the actual preferred direction is such that it allows the carboxyl to bond with both of these functional groups in LT4. Therefore this is considered a direct contact. Regarding the amino group, the direction of approach of the GLU B219 carboxyl oxygen is typical for a hydrogen bond (almost optimal), only somewhat shifted to a direction that facilitates the double contact described above, see Figure 4.8. The latter carboxyl, ASP B267, is in a more typical direction, but even further apart, and it is confirmed from PDB entry 3HCD water locations that this contact is a bridged hydrogen bond, not a direct contact.

The TYR B222 contact for the hydroxyl of LT4 tail, see Figure 4.7, has an aromatic ring that can serve as a hydrogen bond acceptor. Therefore, in case the LT4 tail hydroxyl group would act as an acceptor in the above described hydrogen bonds (i.e. with water and GLU B219) it could in principle donate its hydrogen to a weak hydrogen bond with the aromatic ring of TYR B222, because the closest atom of the ring is at a distance of about 3.5 Å and the ring is facing toward the hydroxyl group. Consequently, as for the LT4 amino group this is not a strong contact, but perhaps has a guiding task in the binding process.

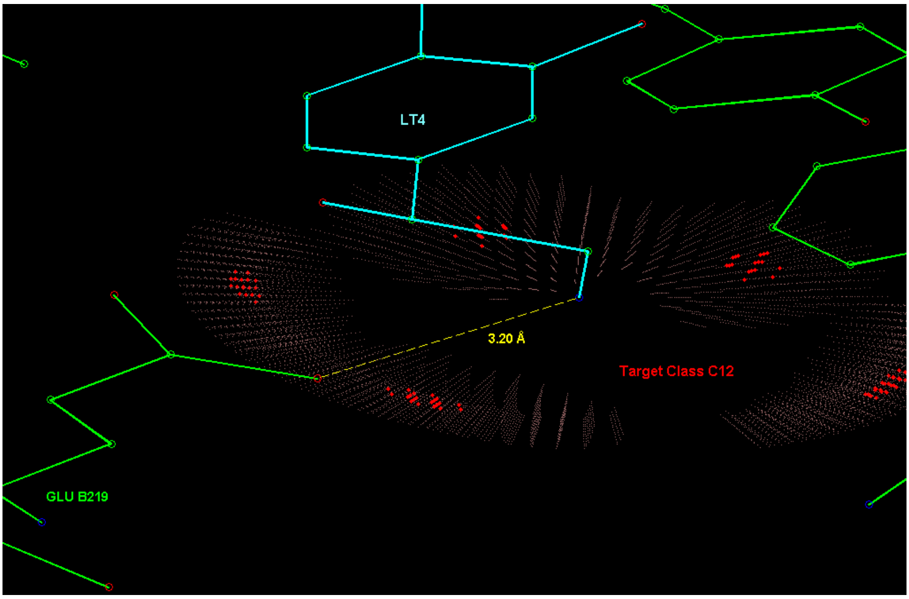


Figure 4.8: Carboxyl oxygen (C12) contact for the amino group of LT4. The residue GLU B219 carboxyl oxygen is located in a typical direction of a class C12 hydrogen bond acceptor for this functional group (fragment class **f26**).

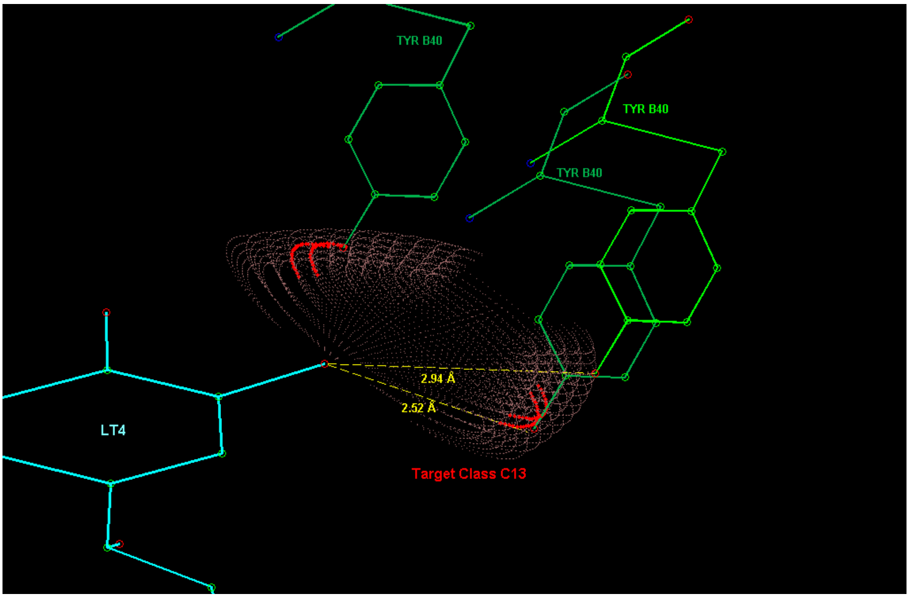


Figure 4.9: The contact between a catechol hydroxyl group and the tyrosine B40 residue of the phenylethanolamine N-methyltransferase (PNMT). The distance to both degenerate maxima is 2.52 Å.

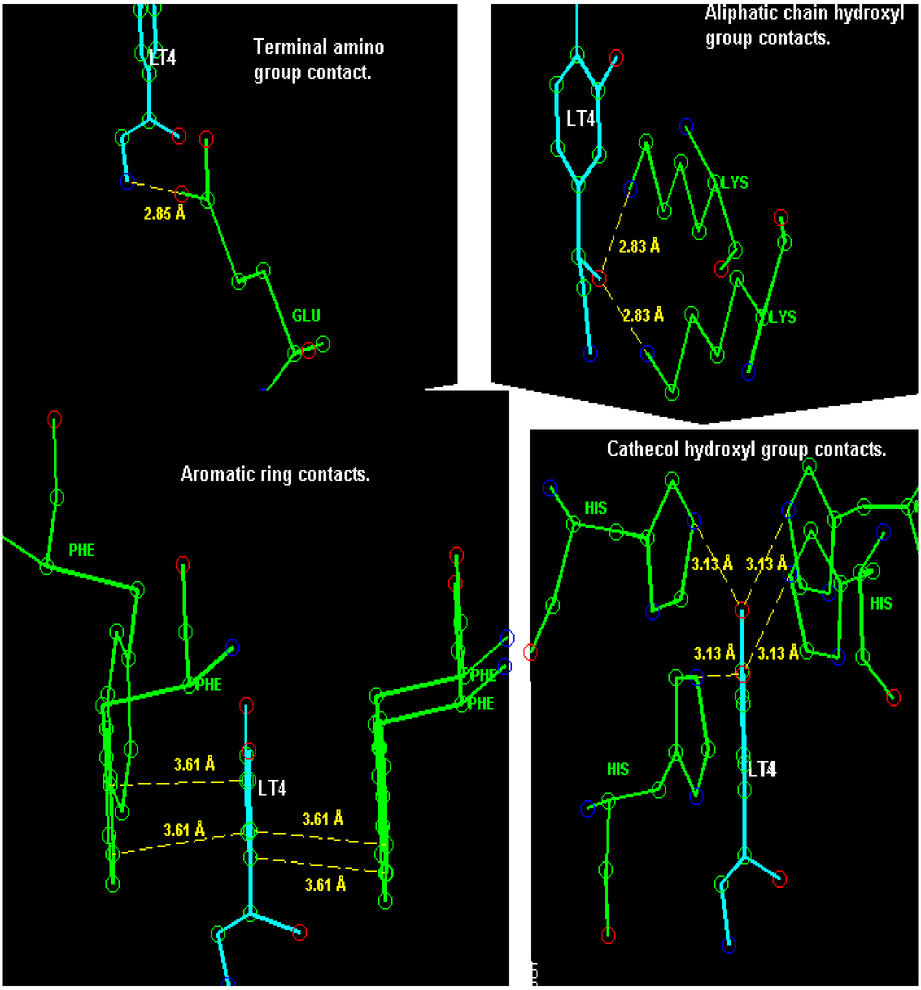


Figure 4.10: R-Norepinephrine (PDB ligand identifier: LT4) with amino acid residues that contain target atoms having highest probability density values in the model. In the figures on the right, the double contacts are created by a degeneracy that follows from the way the fragments are defined. Namely, the third atom (Atom2, see Section Data collection and processing) of a fragment in a molecule can frequently be chosen from more than one possible atom and each choice creates its own probability density, including a maximum. These densities are connected through a rotation around the covalent bond between the Main-atom and Atom1.

Index	ZINC	pmPerMass	RotBonds	Phob/Phil
1	21789	0.053	2	2.33
2	330141	0.071	0	6
3	3801154	0.038	3	6
4	3814483	0.011	2	2
5	3814484	0.049	1	6
6	3814485	0.051	1	6
7	33882	0.052	2	6
8	52627624	0.078	5	12

Table 4.10: Ligand properties in Example 2. Index values represent an ordering of the molecules used in this study.

4.4.2 Example 2: Separating COMT ligands from decoys in a subset of DUD

Here we demonstrate the usefulness of the estimated contact probability masses in discriminating between appropriate and poor binders using logistic regression. Details of logistic regression can be found, e.g., in [72]. In the current application context, logistic regression model connects a binary response variable (here ligand/decoy), with explanatory variables describing the modeled system. The outcome is a probability indicating how likely it is for a system to belong to either of the two response groups. Our testing was done by retrieving a set containing 6 out of the 11 Catechol-O-methyltransferase (COMT) ligands and 19 out of the total 468 decoys from the Directory of useful decoys (DUD) [29]. Two extra ligands were added, namely dopamine and BIA 3-335 (PDB Ligand identifiers LDP and BIA, respectively), by retrieving their structures from ZINC database [73]. These 27 small molecules (ZINC codes in Tables 4.10, 4.11 and 4.12) were chosen so that the DUD molecules have high mutual resemblance, especially so that the decoys have an aromatic ring with at least two primary oxygens bonded to neighboring carbons (in hydroxyl groups typically). This is because all except one COMT ligand in DUD have this type of a structure, and the exception is different only in that the ring is non-aromatic. The two extra ligands were included for reference, because they are known good binders, for BIA [74], and should have clearly higher preference to binding than an average decoy.

The search for the binding mode of the small molecule in the binding pocket (from PDB ID 1H1D) was started by orienting the molecule such that two of the primary oxygens would coordinate with the magnesium ion (Mg^{2+}), participating in the COMT function (see [74] for details), and here taken as part of the binding site. Then, predefined rotamers of the small molecule were rotated around the axis connecting the two coordinating Os, and to a lesser amount around a second axis. Direct contact probabilities between the small molecule and the binding site were calculated. Probabilities for the two coordinating Os were excluded to emphasize contacts for the rest of the molecule. Rotamers

Index	**ZINC**	**pmPerMass**	**RotBonds**	**Phob/Phil**
9	22831	0.049	0	3
10	366295	0.050	3	2
11	366296	0.044	3	2
12	370041	0.033	3	2.5
13	370042	0.036	3	2.4
14	370157	0.015	2	4.5
15	370162	0.029	2	4.5
16	402870	0.055	2	3
17	438536	0.032	2	3
18	1833085	0.010	1	3.67

Table 4.11: Properties of the first set of decoys in Example 2. Index values represent an ordering of the molecules used in this study.

Index	**ZINC**	**pmPerMass**	**RotBonds**	**Phob/Phil**
19	2519115	0.050	2	4.5
20	2990158	0.010	1	3.33
21	3836392	0.000	2	3
22	3871444	0.041	3	4.5
23	3836392	0.000	2	3
24	3995296	0.040	3	2
25	4000727	0.030	3	2
26	4404113	0.036	1	2.33
27	4443675	0.039	3	4.5

Table 4.12: Properties of the second set of decoys in Example 2. Index values represent an ordering of the molecules used in this study.

and orientations with close intra- or intermolecular contacts were removed using distance criteria, though the plausibility of a rotamer could be evaluated using probability masses for intramolecular contacts.

For each small molecule, the rotamer and orientation with highest probability were found and the probabilities were then used in a logistic regression model that represents a docking screening task. Two explanatory variables were used in the logistic regression model: the total probability mass of direct contacts, divided by the mass of the molecule (pmPerMass) and the ratio of the number of hydrophobic and hydrophilic fragments (Phob/Phil) in the molecule.

It is well known that in an actual binding affinity calculation for a ligand-protein pair in solution, one needs to consider energetics of direct contacts, water and metal mediated contacts, desolvation and entropy. The variable pmPerMass is here considered to be a measure of the binding energy of direct contacts, whereas the variable Phob/Phil reflects desolvation properties and perhaps tendency for water mediated contacts. Configurational entropy does not have in this study any obvious representative, because for the numbers of rotatable bonds (RotBonds) no predictive role was identified. Results of the predictions based on logistic regression are shown graphically in Figure 4.11. Values for the putative explanatory variables are given in Tables 4.10, 4.11 and 4.12.

The molecule that was most highly ranked, BIA 3-335, is a known tight binding inhibitor [74]. It is also heaviest of the 27 molecules included in the example; approximately 360 hydrogen masses compared to the more typical value that is between 150 and 250. The variable pmPerMass is intensive, i.e., the size of the molecule should not directly influence it's value, and it is assumed that success in predicting relative binding affinities for smaller candidate binders depends strongly on the accuracy of this variable.

Probabilities derived from the logistic regression are on average over 0.5 for the molecules in the alleged ligand group and below 0.5 for the alleged decoys, which represents a natural threshold between a ligand and a decoy in a screening process. Two exceptions in the ligand group are the low scoring molecules with index values 1 and 4. The ligand with index value 1 has the third highest pmPerMass, but quite low Phob/Phil value, while the ligand with index 4 has a low value for both (see Table 4.10). Hence, if both these molecules are considered as good binders, our approach does not contain enough information to reveal this. On the other hand, there are not necessarily experimental data available on the binding affinities for the decoys, which in this study were chosen to resemble the ligands as much as possible, each starting with the two, to Mg^{2+} anchoring primary oxygens bonded to an aromatic ring. This means that some decoys might be reasonably good binders. Nevertheless, based on the logistic regression model, molecules in the ligand group have on average clearly higher probabilities (0.62) to be ligands than the molecules in the decoy group (0.16). In summary, in the set of 27 chemically similar small molecules, containing 8 experimentally defined ligands, 6 highest ranked molecules were from the ligand group. Consequently, the receiver operating characteristics (ROC) [57] in the screening experiment are: true positive rate TPR=0.75, false positive

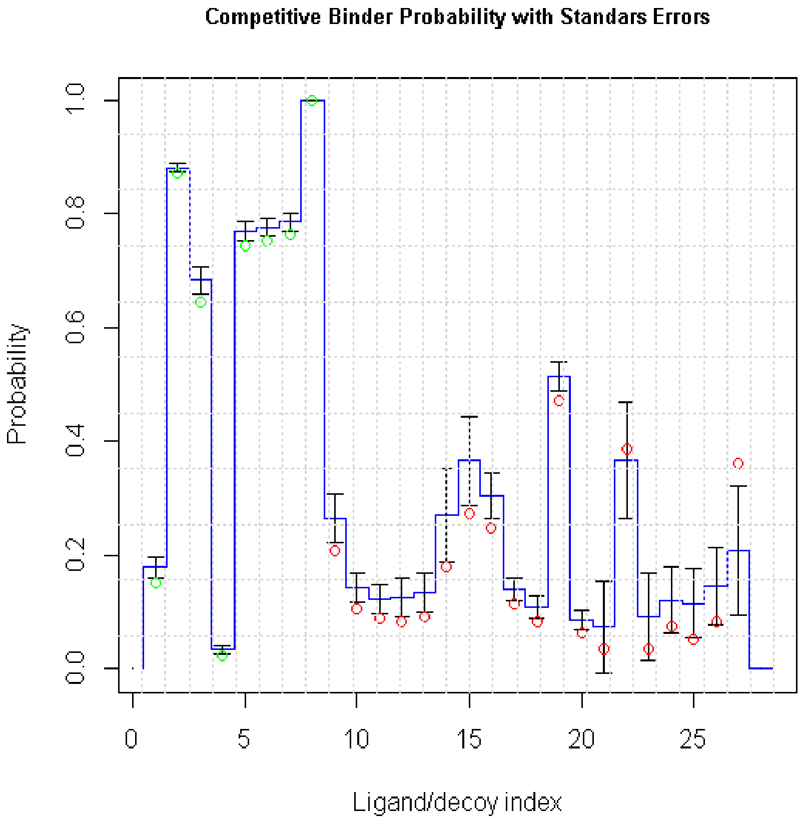


Figure 4.11: Probabilities calculated for separation of Catechol-O-methyltransferase (COMT) ligands from decoys in a subset of DUD. The green circles represent ligands and the red circles decoys, as they are classified in DUD, when all 27 considered molecules are included in the model. The two extra ligands (see text) have index values 7 and 8 (PDB identifiers LDP and BIA, respectively). Blue step curve gives the mean probability that was obtained from bootstrapping over the two small molecule subgroups to calculate standard errors for the logistic regression (error bars representing these are centered at the mean values).

rate FPR=0 and accuracy ACC=0.93. A ROC curve was produced by using discriminating thresholds having either different TPR or FPR. The ROC curve is given in Figure 4.12.

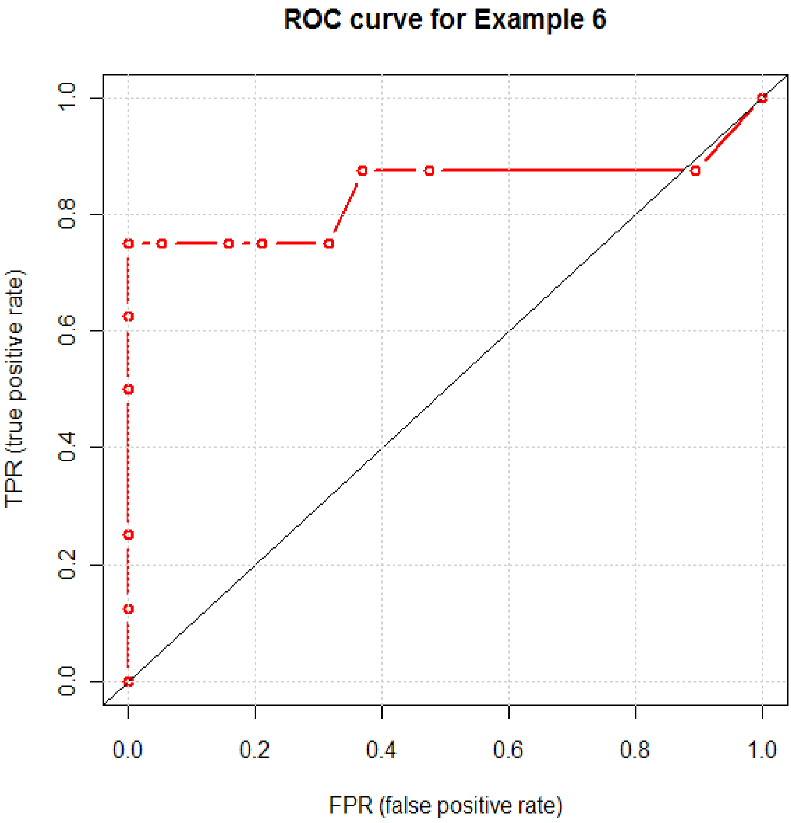


Figure 4.12: An ROC curve illustrating the functionality of the probabilistic model. The 12 threshold probabilities separating ligands from decoys that were used for calculating the characteristics are 0.05, 0.1, ..., 0.3, 0.4, 0.5, 0.65, 0.75, 0.8 and 0.9 . Example 6 in figure title refers to example 2 in this chapter.

One important aspect that has not been considered here, is whether the active form of COMT is a monomer or a multimer. If it is a multimer, it would be interesting to investigate how informative this characteristic is about the properties of suitable ligands. Additional potential molecular characteristics for further study are the flexibility of the binding site and the features of the binding modes having the highest probabilities (pmPerMass).

4.5 Discussion

Predictions about unresolved binding sites, or ligands, can be made by building the preferred contact patterns from the molecules included in a set of functionally classified fragments. In our method, these contact patterns are composed of probability masses calculated for the fragments to have a specific kind of contact in a spatial area. When, for example, the binding affinity of a molecule is studied and the probability masses are defined for an entire molecule, they can be used in a docking and scoring procedure. The absolute binding affinity would be given by the total energetics of the binding process in a thermodynamic setting, including direct and bridged contacts, desolvation and entropy. It is presumed, that using a fragmentation where the fragments have distinct and unique contact patterns, the probability densities described here contain information beyond the chemical complementarity, namely on energetics (for results in this direction, see [75]). This is reasoned out by an analogy with quantum mechanics, because it can be argued that the probability masses are proportional to the amount of binding energy, which are needed in evaluating the binding affinities. In our setting this means relative binding affinities, i.e. rankings over a set of ligands and decoys.

The results obtained in section Example 2, show a level of reliability that is typical for a successful scoring function, see e.g. reviews [2], [55]. Our experiment revealed that potentially very reliable information could be retrieved when our probabilistic method is combined with an effective search routine. An important aspect is that the ligand and decoy molecules were similar, i.e. the decoys used were 'drug-like' [69]. This is based on that they typically had masses between 150 to 250 hydrogen masses, contained both hydrophobic and hydrophilic fragments throughout the structure and were chosen so that each can be anchored to the magnesium ion in the COMT binding site. This should make separation of ligands from decoys challenging and be ultimately based on finer details of the binding affinity, because no decoy was readily rejectable. In a docking and scoring routine such a method can also be used to find the most favourable orientation for the most favourable rotamer, or conformer, of a small molecule in a binding site, i.e. the pose. When adjusting the method for calculations of the absolute binding affinities, the same difficulties will be faced as for any knowledge-based scoring function, described in [2].

In addition to the quality of the fragmentation, the reliability of the data is a central issue in the prediction of contact preferences and some issues related to this were discussed in Introduction and section Materials and methods. When choosing the structure for a prediction model, it is essential to understand the data generation process; in principle from the experimental measurement to the coordinate file. Regarding the special characteristics of the experimental method, X-ray diffraction is sensitive to thermal motion in the crystal. This weakens locally the electron density map, and since electron density maps are precisely the starting point in structure refinement, such an effect should preferably be assessed. In the further refinement, constraints are used in order to keep the protein structure within chemically acceptable boundaries. It thus follows that the ligand atom positions have uncertainty which is not straightforward to quantify. Possible approaches to quantification could be exploration of the effects of the constraints on a theoretical basis or using structures refined with different constraints. On the other hand, PDB files contain substantial amounts of metadata that could potentially also be used in modeling. An example of this are the b-factors, which can be used for incorporating thermal motion in the model. Another example is provided by the occupancies that are needed to take into account the more long lasting local displacements, i.e. alternative conformations in the crystallized protein-ligand complexes [54].

Though the results in the example sections are given with standard errors [69], performance of our model in predicting favourable intermolecular contacts could be more quantitatively verified in the future when more extensive reference sets of sufficiently high quality become available. The approximately 10,000 structure files from PDB used as reference data did only give a preliminary test for certain fragment types. This is mainly because of the 3D nature of the problem, since in order to obtain a good spatial resolution, the frequencies need to be defined in less extensive volumes.

4.6 Conclusions

The examples show that, tentatively our approach can be used to study structural aspects of biochemical reactions or as a tool in predicting the most favourable binding modes and separating ligands from decoys. A plausible future test would be to create a hierarchy among a group of ligands and compare their binding probabilities to experimentally measured binding affinities, e.g. those of KiBank database referred to in DUD. Test on each stage of the docking and scoring procedure has to be successfully conducted, before it is shown that the method is applicable for the purpose. Then it can be directly compared, e.g., with the knowledge-based potentials that are only distance dependent.

Reliable evaluation of binding affinities for potential ligands of a binding site, would be a desirable feature of a virtual drug design screen, see for example [2], [12]. As discussed, the distance and direction dependent probability masses obtained with the approach described here, are taken to provide direct information on relative binding affinity, which is supported by the results in section Example 2. Regarding further development of our modeling approach, both statistics- and chemistry-based generalizations and improvements are possible, including the obvious expansion to all imaginable molecular fragment types.

Bayesian predictive modeling in the normative sense as defined in [24] provides a potential approach to representing contact preference distributions. Such a predictive model could exploit directly the 3D structures of the probability densities (eq. 4.3) that model the contact atom positions, instead of considering density parameters as the main characterization of the spatial information. The obvious disadvantage of such an approach is the considerably increased computational effort needed to derive approximations to the sought after predictive distributions.

The reliability of an inferred order of preference depends, on one hand, on how successfully the error from experimental methods and structure refinement is quantified in terms of the used probability densities. On the other hand it depends on how realistically the chemical likelihood of a contact atom type and the bias in the data set are taken into account. The latter are here incorporated as prior information, see equation (4.10), which guides the model with chemistry-based knowledge. A third fundamental area for chemistry-based improvements are the classifications (Tables 4.2 and 4.3). For example, a classification can be envisioned where the covalent bond count of Atom1 (see section Data collection and processing) would be used as one of the characteristics defining the fragment, which would then remove the degeneracy described in section Example 1. This kind of more structural way of defining the fragments would expand the classifications, but should also give fragment definitions that are closer to being unique.

Chapter 5

Some generic conclusions

The probabilistic model developed in this work, has been tested for several case studies and has proven to give reasonable and robust estimates for the intended purpose. That is, a scoring function for ranking small molecules in contact with a protein binding site, and assessing the effect of an amino acid residue mutation on the preference of a protein-protein contact. The model based probabilities can be calculated in time scales that allow routine use in finding molecular complex candidates for further, for example, experimental analyses.

In the field of molecular modeling, this probabilistic method is placed between first principles theoretical models and statistical methods modeling collected experimental data. The former produce results having best correspondence with measured molecular properties, and require computations that still today are too time-consuming to be used, e.g., in modeling and visualization environments for rapid inspection of structural aspects following an induced change, like a residue mutation. Results from statistical methods then, depend on amount and quality of the data used as input, and can be fast enough to evaluate when estimates are needed in seconds, but their import does not necessarily transfer from one task to another, because they do not automatically have a strict chemical interpretation.

Compared to another knowledge-based scoring function, the statistical potential that uses a probability distribution function to convert observed relative distances to an estimate of the thermodynamic free energy, output of our probabilistic method have a more straightforward interpretation. This is due to the feature that, in addition to distance, also directional data are used to build the model, which makes correspondence between the densities and quantum mechanical effective potentials for the motion of nucleii possible. In order to add further physical realism to the scoring, theoretical components independent of training data are incorporated into the model. Examples of these components are the chemical element specific estimates for relative partial charges and deformabilities of electron cloud, both of which are used in defining the fragment class conditional prior probabilities. In addition we use theoretical dihedral angle intervals to describe thermal motion generated amino acid side chain conformations for overlap integrals representing the noncovalent bond strengths. A modeling approach like this can also take advantage of quantum mechanical electronic structure calculations, results of which could be used, for example, in defining the mentioned dihedral angle intervals and finding theoretical foundations for a more advanced molecular fragment classification.

In addition to finalising the molecular fragment classification, development of the method can be pursued in numerous other aspects as well. These include formulating a contact as a fragment-fragment interaction, testing other functional forms for the probability densities, attempting analytical derivations to reduce computational load and implementing the method into a modeling and visualization environment. Computational challenges associated with the knowledge-based scoring function are larger than for a traditional statistical potential, since also directional aspects are taken into account. The load is further increased by modeling flexibility of the molecules. This has so far been planned for amino acid residue side chains, but it is considered possible to treat protein main chains and small molecules with the same approach, i.e., building distributions of conformations for overlap calculations, starting from ensembles of intervals for internal rotation angles. Algorithm development is plausibly needed to achieve sufficient computational speed for fast structural inspections without compromising the discussed implementation of chemistry and physics.

5.1 Relation to drug design methods

The function of this probabilistic method has been outlined in this thesis as a scoring function for estimating the relative strength of molecular contacts in general, however, also a discussion about drug design as a possible area of application is to the purpose.

A pharmacophore model, as described in the review [76], is either ligand-based or structure-based. The latter analyse ligand binding sites through the experimentally measured structures of molecular complexes, or binding sites without a bound small molecule. Ligand-based methods aim at quantifying similarities of features that guide the binding process and are conducted in sets of small molecules that are considered as possible ligands for a binding site. The ligand-based approach has two major challenges that are not satisfactorily resolved currently. These challenges are molecular alignment and modeling structural flexibility of the alleged ligands. Solving the alignment problem is attempted with either a point-based or property-based model. The property-based approach, to some extent, models the molecular flexibility through the tolerance radius related to the so called properties, which are spatially defined chemical characteristics like hydrophobicity. These spatial characteristics are not necessarily ranked, nor are they internally weighed, and a probabilistic 3D model could provide both these features. Consequently, defining the properties in a pharmacophore model is a potential application of the probabilistic contact preferences.

A second widely used drug design method is the three-dimensional quantitative structure-activity relationship (3D-QSAR), for a review of this methodological field see for example [77]. A QSAR method seeks to find correlation between all structural aspects of a molecule and experimental data corresponding to its biological activity. Part of the 3D-QSAR model is a description of non-covalent interaction fields around the molecules, and the energies of these fields are calculated, e.g., with force fields [12], [77]. Relative interaction energies among a group of molecules can also be estimated with the probability densities of the contact preferences and the efficacy of this approach would be worthwhile testing.

Bibliography

[1] Snyder DA et al. (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination *J Am Chem Soc* **127**(47):16505-11

[2] Huang S-Y, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions *Phys Chem Chem Phys* **12:** 12899-12908

[3] Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set *J Chem Inf Model* **49**: 1079-1093

[4] Englebienne P, Moitissier N (2009) Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? *J Chem Inf Model **49**: 1568-1580*

[5] Stone AJ (2013) *The Theory of Intermolecular Forces (*2nd Ed.*)*, Oxford University Press

[6] Israelachvili JN (2011) *Intermolecular and Surface Forces (*3rd Ed.*)*, Academic Press

[7] Nelson DL, Cox MM (2008) *Lehninger Principles of Biochemistry (5th Ed.),* W. H. Freeman & Co.

[8] Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications *Nature reviews. Drug discovery* **3** (11): 935–49

[9] Davydov AS (1991) *Quantum Mechanics* (2nd Ed.), Great Britain: Pergamon Press

[10] Atkins P, Friedman R (2010) *Molecular Quantum Mechanics (*5th Ed.*),* Oxford University Press, USA

[11] Pauling L, Wilson EB (1935) *Introduction to Quantum Mechanics, With Applications to Chemistry*, McGraw-Hill

[12] Gilson MK, Zhou H-X (2007) Calculation of protein ligand binding affinities *Annu Rev Biophys Biomol Struct* **36**: 21–42

[13] Schmuttenmaer CA (2004) Exploring Dynamics in the Far-Infrared with Terahertz Spectroscopy *Chem Rev* **104**: 1759-79

[14] Nagai N, Kumazawa R, Fukasawa R (2005) Direct evidence of intermolecular vibrations by THz spectroscopy *Chem Phys Lett* **413:** 495-500

[15] Baum et al. (2010) Non-additivity of functional group contributions in protein–ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry *J Mol Biol* **397(**4**):** 1042-54

[16] Bransden BH, Joachain CJ (2000) *Quantum Mechanics,* Prentice Hall

[17] Hakulinen R, Puranen S, Lehtonen JV, Johnson MS, Corander J (2012) Probabilistic prediction of contacts in protein-ligand complexes *PLoS ONE* **7**(11): e49216

[18] Gelman A, Carlin JB, Stern HS, Rubin DB (2003) *Bayesian Data Analysis* New York: Chapman & Hall/CRC Texts in Statistical Science

[19] Bernardo JM, Smith AFM (1994) *Bayesian Theory* Chichester, England: Wiley

[20] Berman H et al. (2000) The Protein Data Bank *Nucl Acids Res* **28**: 235-242

[21] Auffinger P (2001) Halogen Bonds in Biological Molecules *Proc Natl Acad Sci U S A* **101**: 16789–6794

[22] Guttorp P, Lockhart RA (1988) Finding the location of a signal: A Bayesian analysis *JASA* **83**(402): 322-30

[23] Nuñez-Antonio G, Gutiérrez-Peña E (2005) Bayesian Analysis of Directional Data Using the von Mises–Fisher Distribution *Comm Stat - Simulation and Computation* **34:** 989-99

[24] Geisser S (1993) *Predictive Inference: An Introduction,* London: Chapman & Hall

[25] Abramowitz M, Stegun IA (eds.) (1972) *Handbook of mathematical functions* (10th pr.) National Bureau of Standards

Applied Mathematics Series - 55

[26] Man O (2005) Is the Fisher Distribution Addive? *Stud Geophys Geod* **49**: 561-572

[27] Wand MP, Jones MC (1995) *Kernel Smoothing*, London: Chapman & Hall/CRC

[28] Thomas GB Jr, Finney RL (1996) *Calculus and Analytic Geometry* (9th ed.), Addison Wesley

[29] Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking Sets for Molecular Docking *J Med Chem* **49**: 6789–6801

[30] Männistö PT, Kaakkola S (1999) Catechol-O-methyltransferase (COMT): Biochemistry, Molecular Biology, Pharmacology, and Clinical Efficacy of the New Selective COMT Inhibitors *Pharm Rev* **51** (4): 593—628

[31] Dunbrack RL (2002) Rotamer Libraries in the 21st Century *Curr Opin Struct Biol* **12(**4**):** 431-40

[32] Harder T et al. (2010) Beyond rotamers: a generative, probabilistic model of side chains in proteins *BMC Bioinformatics* **11:** 306

[33] Doig AJ (1996) Thermodynamics of amino acid side-chain internal rotations *Biophys Chem* **61***:* 131-141

[34] Shapovalov MV, Dunbrack RL Jr (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions *Structure* **19**(6): 844-58

[35] Mandl F (1988) *Statistical Physics* (2nd Ed.) Chichester, England: John Wiley & Sons

[36] Rhodes G (2006) *Crystallography made crystal clear (*6th *Ed.),* Elsevier

[37] Martz E, Rhodes G, Thompson L *Nature of 3D Structural Data* Available: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/nature_of_3d_structural_data.html. Accessed: 18 June 2013.

[38] Understanding PDB Data. Looking at Structures: Dealing with Coordinates Available: http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/coordinates.html. Accessed: 18 June 2013.

[39] De Maeyer J, Desmet IL (1997) All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination *Fold Des* **2:** 53-66

[40] OhAinle M et al. (2011) Dynamics of Dengue Disease Severity Determined by the Interplay Between Viral Genetics and Serotype-Specific Immunity *Sci Transl Med* **3(**114**):**114ra128

[41] Marttinen P, Corander J, Törönen P, Holm L (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues *Bioinformatics* **22**: 2466-2474

[42] Pierson TC, Fremont DH, Kuhn RJ, Diamond MS (2008) Structural insights into the mechanisms of antibody-mediated neutralization of flavivirus infection: implications for vaccine development *Cell Host Microbe* **4**(3): 229-38

[43] Cockburn JJ et al. (2012) Mechanism of dengue virus broad cross-neutralization by a monoclonal antibody *Structure **20**(2): 303-14*

[44] Cheng L, Siren J, Connor TR, Aanensen DM, Corander J (2013) Hierarchical and spatially explicit clustering of DNA sequences with BAPS software *Mol Biol Evol* **30**(5): 1224-8

[45] Dimitropoulos D, Ionides J, Henrick K (2006) Using MSDchem to Search the PDB Ligand Dictionary. Current Protocols in Bioinformatics UNIT 14.3. Available: http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1403s15/full. Accessed 12 June 2013.

[46] Zhu X, Lopes PE, Shim J, MacKerell AD Jr. (2012) Intrinsic energy landscapes of amino acid side-chains *J Chem Inf Model* **52**(6):1559-72

[47] Renfrew PD, Butterfoss G, Kuhlman B (2008) Using quantum mechanics to improve estimates of side chain rotamer energies *Proteins* **71**(4): 1637-46

[48] Rantanen V-V, Denessiouk KA, Gyllenberg M, Koski T, Johnson MS (2001) A Fragment Library Based on Gaussian Mixtures Predicting Favorable Molecular Interactions *J Mol Biol* **313**: 197–214

[49] Rantanen V-V, Gyllenberg M, Koski T, Johnson MS (2003) A Bayesian Molecular Interaction Library *J Comput Aided Mol Des* **17**: 435–461

[50] Rantanen V-V, Gyllenberg M, Koski T, Johnson MS (2005) A Priori Contact Preferences in Molecular Recognition *J Comput Biol Bioinform Res* **3**: 861–890

[51] Lehtonen JV, Still D-J, Rantanen V-V, Ekholm J, Björklund D, et al. (2004) BODIL: A Molecular Modeling Environment for Structure-Function Analysis and Drug Design *J Comput Aided Mol Des* **18**: 401–419

[52] Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications *Nat Rev Drug Discov* **3**: 935–949

[53] Scapin G (2006) Structural Biology and Drug Discovery *Curr Pharm Des* **12**: 2087–2097

[54] Ravi Acharya K, Lloyd MD (2005) The Advantages and Limitations of Protein Chrystal Structures *Trends Pharmacol Sci* **26**: 10–14

[55] Huang S-Y, Zou X (2010) Advances and Challenges in Protein-Ligand Docking *Int J Mol Sci* **11**: 3016-3034

[56] Kortemme T, Morozov AV, Baker D (2003) An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes *J Mol Biol* **326**: 1239-1259

[57] Egan JP (1975) *Signal Detection Theory and ROC Analysis* New York: Academic Press

[58] Finkelstein AV, Ptitsyn OB (2002) *Protein Physics: A Course of Lectures* London: Academic Press

[59] Desiraju G (1996) The $CH{\cdots}O$ Hydrogen Bond: Structural Implications and Supramolecular Design *Acc Chem Res* **29**: 441–449

[60] Scheinert S, Kar T, Gu Y (2001) Strength of the $C^{\alpha}H \cdot \cdot O$ Hydrogen Bond of Amino Acid Residues *J Biol Chem* **276**: 9832–9837

[61] Grimme S (2008) Do Special Non-covalent $\pi - \pi$ Stacking Interactions Really Exist? *Angew Chem Int Ed Engl* **47**: 3430–3434

[62] Howard JAK, Hoy VJ, O'Hagan D, Smith GT (1996) How Good is Fluorine as a Hydrogen Bond Acceptor? *Tetrahedron* **52**: 12613–12622

[63] Gelman A, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis* New York: Chapman and Hall

[64] Bernardo JM, Smith AFM (1994) *Bayesian Theory* Chichester, England: Wiley

[65] Mardia KV, Jupp PE (1999) *Directional Statistics* (2nd ed.) New York: Wiley

[66] Mardia KV, Kent JT, Bibby JM (1979) *Multivariate Analysis* London: Academic Press

[67] Nuñez-antonio A, Gutiérrez-peña E (2005) A Bayesian Analysis of Directional Data Using the von Mises-Fisher Distribution *Communications in Statistics — Simulation and Computation* **34**: 989–999

[68] Parr RG, Pearson RG (1983) Absolute Hardness: Companion Parameter to Absolute Electronegativity *J Am Chem Soc* **105**: 7512–7516

[69] Nicholls A (2008) What Do We Know and When Do We Know It? *J Comput Aided Mol Des* **22**: 239–255

[70] Strömbergsson H, Kleywegt GJ (2009) A Chemogenomics View on Protein Ligand Spaces *BMC Bioinf* **10**(Suppl 6)*:* S13

[71] Drinkwater N, Gee CL, Puri M, Criscione KR, McLeish MJ, et al. (2009) Molecular Recognition of Physiological Substrate Noradrenaline by the Adrenaline-Synthesizing Enzyme PNMT and Factors Influencing Its Methyltransferase Activity *Biochem J* **422**: 463–471

[72] Bewick V, Cheek L, Ball J (2005) Logistic regression *Crit. Care* **9**: Statistics review 14. Available: http://ccforum.com/content/9/1/112. Accessed 4 June 2012.

[73] Irwin JJ, Shoichet BK (2005) ZINC–a Free Database of Commercially Available Compounds for Virtual Screening *J Chem Inf Model* **45**: 177-182

[74] Bonifácio MJ, Archer M, Rodriques ML, Matias PM, Learmonth DA, et al. (2002) Kinetics and Crystal Structure of Catechol-O-Methyltransferase Complex with Co-Substrate and a Novel Inhibitor with Potential Therapeutic Application *Mol Pharmacol* **62**:795-805

[75] Morozov AV, Kortemme T, Baker D (2004) Close Agreement Between the Orientation Dependence of Hydrogen Bonds Observed in Protein Structures and Quantum Mechanical Calculations *Proc Natl Acad Sci U S A* **101**: 6946-6951

[76] Yang S-Y (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances *Drug Discov Today* **15**: 11/12

[77] Verma J, Khedkar VM, Coutinho EC (2010) 3D-QSAR in drug design - A review *Curr Top Med Chem* **10**: 95-115

Riku Hakulinen

Probabilistic contact preferences in protein-ligand and protein-protein complexes

This thesis presents a Bayesian statistical model for computational studies of molecular complexes. Case studies of testing the method in the most relevant molecular environments are described. Plausible steps to further develop the model are discussed.

Åbo Akademi University Press | ISBN 978-952-12-2954-1