Bayesian Alignment of Continuous Molecular Shapes Using Random Fields

Irina Czogiel, Ian L. Dryden and Christopher J. Brignell
School of Mathematical Sciences, University of Nottingham

Abstract

Statistical methodology is proposed for comparing molecular shapes. In order to account for the continuous nature of molecules, classical shape analysis methods are combined with techniques used for predicting random fields in spatial statistics. Applying a modification of Procrustes analysis, Bayesian inference is carried out using Markov chain Monte Carlo methods for the pairwise alignment of the resulting molecular fields. Superimposing entire fields rather than the configuration matrices of nuclear positions thereby solves the problem that there is usually no clear one–to–one correspondence between the atoms of the two molecules under consideration. Using a similar concept, we also propose an adaptation of the generalised Procrustes analysis algorithm for the simultaneous alignment of multiple molecular fields. The methodology is applied to a dataset of 31 steroid molecules.

Keywords: Bioinformatics, Chemoinformatics, Geostatistics, Kriging, Markov chain Monte Carlo, Procrustes, Rotation, Shape, Size, Spatial, Steroids.

1 Introduction

A major goal in pharmaceutical research is the design of selective ligands for protein and DNA binding – an extremely difficult task because the space of ligands with a potential ben-

eficial effect on the human body is vast. Since in most practical cases the three–dimensional structure of a receptor is unknown, direct rational drug design techniques such as docking are not generally applicable. A way to tackle this problem is to make use of the fact that any chemical binding process requires some complementarity between the ligand and its receptor. Ligands which bind to the same target can therefore be expected to possess a certain degree of shape (and size) similarity. When designing new drug molecules, the converse of this concept is exploited. Here, the underlying conjecture is that molecules of a similar shape exhibit a similar biochemical activity and hence drug potency. In order to use this idea, methods for calculating molecular shapes and their similarities have to be available.

Molecular data are usually given in form of atomic coordinates and in most cases there is no clear correspondence between atoms of different molecules. From a statistical point of view, the task of comparing molecular shapes is therefore that of comparing unlabelled point sets which has been of recent interest in statistical shape and image analysis. For example Green & Mardia (2006) and Dryden *et al.* (2007) have proposed Bayesian approaches to the problem of comparing protein binding sites and small steroid molecules, respectively. Our alignment procedure differs substantially due to the use of continuous random fields which interpolate additional information measured at the point coordinates. In the context of molecules, the additional data usually comprise the values of molecular properties such as partial atomic charges or hydrophobicity associated with the individual atoms. As most of these properties are diffused in space rather than located at the discrete atoms positions, our random field approach captures the diffuse nature of molecular shapes better than the use of discrete point sets.

Our main application is a dataset comprising 31 steroid molecules which bind to the corticosteroid binding globulin (CBG) receptor. For each molecule, the xyz–coordinates of the atom positions as well as the atom types (e.g. carbon, oxygen, ...), the associated van der Waals radii and the partial atomic charge values at the atom positions are provided. The data has originally been compiled by Cramer *et al.* (1988), and Good *et al.* (1993) classified each steroid according to its binding activity towards the CBG receptor as 1 (high), 2 (intermediate), or 3 (low). A major feature of the dataset is that all molecules share a common

core structure consisting of four carbon rings. Figure 1 displays the two steroid molecules aldosterone and androstanediol. In this two–dimensional representation, the common ring structure is clearly visible. The main objective is to obtain the common features in each of the three groups which are associated with the type of binding activity.

INSERT FIGURE 1 ABOUT HERE

In Section 2, we motivate and describe our geostatistical model for molecular shapes and point out the relationship to existing models used in the chemoinformatics community. The Bayesian framework for the pairwise molecular alignment and similarity calculation is introduced in Section 3. An extension of this methodology to the simultaneous alignment of multiple molecular fields is described in Section 4. In Section 5, we apply our methods to the steroids data and assess the results with respect to their chemical relevance. Finally, Section 6 concludes the paper with a discussion.

2 Molecular Similarity Using Geostatistics

2.1 Molecular Similarity

In datasets for molecular alignment, each molecule M is usually represented by two matrices, namely its conformation matrix $\boldsymbol{X}^{\text{M}} = (\boldsymbol{x}_1^{\text{M}} \dots \boldsymbol{x}_{k_{\text{M}}}^{\text{M}})^T \in \mathbb{R}^{k_{\text{M}} \times 3}$ and a matrix of marks $\boldsymbol{Z}^{\text{M}} \in \mathbb{R}^{k_{\text{M}} \times p}$, where k_{M} denotes the number of atoms in M, $\boldsymbol{x}_i^{\text{M}} \in \mathbb{R}^3$ is the xyz–coordinate vector of the position of the ith atom, and $\boldsymbol{Z}^{\text{M}}$ row–wise contains p–dimensional vectors of molecular properties (e.g. partial charge, van der Waals radius, hydrophobicity, ...) observed at the atom positions.

We wish to develop a measure of similarity between two molecules which does not depend on their relative position. In particular, we are not interested in rotations $\boldsymbol{\Gamma} \in SO(3)$ and translations $\boldsymbol{\gamma} \in \mathbb{R}^3$ of a molecule B when matching it to a molecule A, say. As a member of the special orthogonal group $SO(3)$, the matrix $\boldsymbol{\Gamma}$ satisfies the $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I}_3$ and $|\boldsymbol{\Gamma}| = 1$, and can be described by three parameters. We will parameterise $\boldsymbol{\Gamma}$ using the Euler

angles in the so–called x–convention, where $\mathbf{\Gamma}$ is decomposed into the following elementary rotation matrices

$$\mathbf{\Gamma} = \mathbf{\Gamma}(\boldsymbol{\theta}) = \begin{pmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & \sin\theta_2 \\ 0 & -\sin\theta_2 & \cos\theta_2 \end{pmatrix} \begin{pmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

With the domains $-\pi \leq \theta_1, \theta_3 < \pi$ and $-\pi/2 \leq \theta_2 < \pi/2$, every $\mathbf{\Gamma} \in SO(3)$ is uniquely determined apart from a singularity at $\theta_2 = -\pi/2$. The probability measure for $SO(3)$ which is invariant under the group action is given by the Haar measure $d\mathbf{\Gamma} = 1/(8\pi^2)\cos(\theta_2)d\theta_1 d\theta_2 d\theta_3$ (e.g. Miles, 1965). The singularity therefore has a measure of zero although one must take care numerically in its vicinity.

Let us denote molecule A as $\left(\boldsymbol{X}^{\text{A}}, \boldsymbol{Z}^{\text{A}}\right)$ and a translated, rotated version of molecule B as $\left((\boldsymbol{X}^{\text{B}} - \mathbf{1}_{k_{\text{B}}}^T \boldsymbol{\gamma})\mathbf{\Gamma}, \boldsymbol{Z}^{\text{B}}\right)$, where $\mathbf{1}_{k_{\text{B}}}$ denotes the k_{B}–dimensional vector of ones.

The aim in molecule matching is to estimate $\mathbf{\Gamma}, \boldsymbol{\gamma}$ by maximizing a measure of similarity between the molecules. This procedure bears a clear resemblance to the ordinary partial Procrustes analysis well–known in statistical shape analysis (e.g. Dryden & Mardia, 1998, p.94) where analytical methods are applied to superimpose two configuration matrices of the same dimension by minimising the sum of the squared distances between corresponding landmarks. However, the optimisation problem at hand will in general involve numerical methods due to the lack of clear one–to–one correspondences between atoms in A and B, respectively. Moreover, not only the conformation matrices but also the matrices of observed molecular properties $\boldsymbol{Z}^{\text{A}}$ and $\boldsymbol{Z}^{\text{B}}$ should be taken into account when superimposing A and B. Another important difference from classical shape analysis is that viewing a molecule as a set of discrete landmarks implies a considerable simplification of the true nature of the molecules which are in fact fuzzy bodies of electronic clouds. To account for this, a continuous representation of molecular shapes is desirable.

2.2 Geostatistical Modelling of Molecular Shapes

In order to obtain a descriptor of the shape of a molecule M which captures its rather continuous nature, the values in $\boldsymbol{Z}^{\text{M}}$ are interpolated into $\mathbb{R}^3$ using spatial prediction (e.g. Cressie, 1993, Chapter 3). As the prediction is performed for each molecular property separately, it suffices to illustrate the procedure using the ith column of $\boldsymbol{Z}^{\text{M}}$, say, i.e. the k_{M}–dimensional vector $\boldsymbol{z}_i^{\text{M}} = \big(z_i(\boldsymbol{x}_1^{\text{M}}), \ldots, z_i(\boldsymbol{x}_{k_{\text{M}}}^{\text{M}})\big)^T$ containing the values of the molecular property Z_i ($i \in \{1, \ldots, p\}$) observed at the atom positions. For the sake of clarity, the indices M and i are thereby omitted in this section.

In the geostatistical setting, $\boldsymbol{z} = \big(z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_k)\big)^T$ is viewed as a sample of one realisation $z(\boldsymbol{x})$ of the random field $\big\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^3\big\}$ which in the following is assumed to be second–order stationary with a positive definite, isotropic covariance function $\sigma(||\boldsymbol{h}||) = \text{Cov}\big(Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\big)$. As any molecular property gradually fades away with the distance from the molecular skeleton and therefore takes the value of zero in most parts of $\mathbb{R}^3$, we assume the constant mean to be zero. With these assumptions, simple kriging is appropriate to predict the value of the random field at a location of interest $\boldsymbol{x}_0$. Here, a weighted average of the form $\hat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^n u_i Z(\boldsymbol{x}_i)$ is sought so as to minimise the prediction mean squared error $\text{PMSE}(\boldsymbol{u}) = \text{E}\big[\big(\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\big)^2\big]$ with respect to the weight vector $\boldsymbol{u} = (u_1, \ldots, u_n)^T$. The resulting system of equations has the solution $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$ with predicted value for $Z(\boldsymbol{x}_0)$ given by $\hat{Z}(\boldsymbol{x}_0) = \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z} = \boldsymbol{u}^T\boldsymbol{z}$, where $\boldsymbol{\sigma} = \big(\sigma(\boldsymbol{x}_1 - \boldsymbol{x}_0), \ldots, \sigma(\boldsymbol{x}_n - \boldsymbol{x}_0)\big)^T$ and $(\boldsymbol{\Sigma})_{ij} = \sigma(\boldsymbol{x}_i - \boldsymbol{x}_j)$. By defining $\boldsymbol{\sigma}(\boldsymbol{x}) = \big(\sigma(\boldsymbol{x}_1 - \boldsymbol{x}), \ldots, \sigma(\boldsymbol{x}_n - \boldsymbol{x})\big)^T$, the above prediction equation can now be generalised to yield a field–based representation of molecular shape:

$$\hat{Z}(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z} = \boldsymbol{u}(\boldsymbol{x})^T\boldsymbol{z}. \tag{1}$$

Similar to other continuous definitions of molecular shape used in the structural alignment community, (e.g. Good *et al.*, 1992), $\hat{Z}(\boldsymbol{x})$ is a weighted average of the observed values of the considered molecular property with the weights depending on the position of $\boldsymbol{x}$ relative to the atom positions. However, the weights $\boldsymbol{u}(\boldsymbol{x})$ in (1) offer the advantage that they have a well–defined optimality property in that they are chosen to minimise the mean squared

prediction error.

A very useful descriptor of molecular shape can be obtained if equation (1) is seen as a weighted average of covariance functions centred at the atom positions, i.e.

$$\hat{Z}(\boldsymbol{x}) = \boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}(\boldsymbol{x}) = \sum_{i=1}^{k} w_i \sigma(\boldsymbol{x}_i - \boldsymbol{x}) \ , \tag{2}$$

where the vector of weights $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}\boldsymbol{z}$ does not depend on $\boldsymbol{x}$ and combines the information about the geometry of the molecule and the observed values of the quantity Z. As will be seen in the next section, (2) can directly be utilised for the structural alignment of two molecules.

2.3 The Kriged Carbo Index

A similarity index which is well–established in the literature on field–based molecular alignment is the Carbo index (Carbo *et al.*, 1980). In terms of the Carbo index, the similarity of two molecules A and B in a certain relative position with respect to the molecular property P is defined as

$$C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \frac{\int P_{\text{A}}(\boldsymbol{x}) P_{\text{B}}(\boldsymbol{x}) d\boldsymbol{x}}{\big(\int P_{\text{A}}^2(\boldsymbol{x}) d\boldsymbol{x} \big)^{1/2} \big(\int P_{\text{B}}^2(\boldsymbol{x}) d\boldsymbol{x} \big)^{1/2}}, \tag{3}$$

where $P_{\text{M}}(\boldsymbol{x})$ denotes the field of P for molecule M ($M \in \{A, B\}$) evaluated at point $\boldsymbol{x}$ in $\mathbb{R}^3$. The above index is a variant of Pearson's correlation coefficient. The numerator term measures the "overlap" of the molecular fields whereas the denominator is a normalising constant which ensures that $C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) \in [-1, 1]$. In situations where a discrepancy rather than a similarity measure is required, (3) can be uniquely mapped into the appropriate codomain using

$$D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \frac{1 - C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma})}{1 + C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma})} \in [0, \infty). \tag{4}$$

Due to the fact that $\int P_{\text{M}}^2(\boldsymbol{x}) d\boldsymbol{x}$ is invariant under translation and rotation, (4) is intimately linked to an alternative discrepancy measure, namely the integrated square error

$$\text{ISE}_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \int \big(P_{\text{A}}(\boldsymbol{x}) - P_{\text{B}}(\boldsymbol{x}) \big)^2 d\boldsymbol{x}. \tag{5}$$

The main difference between these measures is that (4) is invariant to the relative scales of the two fields whereas (5) not only depends on the scales but also on the extent of the molecules under study. In particular the latter is undesirable, so that we shall apply the Carbo–based discrepancy and similarity measures throughout this paper.

Written as (2), the kriged molecular fields of two molecules can directly be substituted into the Carbo index which then becomes

$$C_{\text{AB}}(\mathbf{\Gamma}, \boldsymbol{\gamma}) = \frac{\int \hat{Z}_{\text{A}}(\boldsymbol{x})\hat{Z}_{\text{B}}(\boldsymbol{x})d\boldsymbol{x}}{\big(\int \hat{Z}^2_{\text{A}}(\boldsymbol{x})d\boldsymbol{x}\big)^{1/2}\big(\int \hat{Z}^2_{\text{B}}(\boldsymbol{x})d\boldsymbol{x}\big)^{1/2}} = \frac{\sum_{i=1}^{k_{\text{A}}}\sum_{j=1}^{k_{\text{B}}} w_i^{\text{A}} w_j^{\text{B}} \int \sigma(\boldsymbol{x}-\boldsymbol{x}_i^{\text{A}})\sigma(\boldsymbol{x}-\boldsymbol{x}_j^{\text{B}})d\boldsymbol{x}}{N^{\text{A}}N^{\text{B}}}, \tag{6}$$

where

$$N^{\text{M}} = \Big(\sum_{i=1}^{k_{\text{M}}}\sum_{j=1}^{k_{\text{M}}} w_i^{\text{M}} w_j^{\text{M}} \int \sigma(\boldsymbol{x}-\boldsymbol{x}_i^{\text{M}})\sigma(\boldsymbol{x}-\boldsymbol{x}_j^{\text{M}})d\boldsymbol{x}\Big)^{1/2}, \quad M \in \{A, B\}$$

denotes the normalising constant associated with molecule M. Optimising the above expression with respect to rotation and translation then gives the required similarity measure, the Kriged Carbo Index

$$\hat{C}(A, B) = \sup_{\substack{\mathbf{\Gamma}\in SO(3)\\ \boldsymbol{\gamma}\in\mathbb{R}^3}} C_{\text{AB}}(\mathbf{\Gamma}, \boldsymbol{\gamma}), \tag{7}$$

which is invariant under the rigid body transformations.

In the case where more than one molecular property has been measured at the atom positions for each molecule, a multivariate version of the Carbo index is desirable. This can easily be obtained by first assessing the similarity of the two molecules in the given relative position for each property separately using (6), and then calculating a weighted average of the univariate Carbo indices. If the weights are positive and normalised to sum up to one, the resulting multivariate Carbo index takes values between minus one and one like its univariate equivalents and can therefore be optimised in the same way.

2.4 Relationship to Established Methods

Evaluating molecular similarity using (6) can be viewed as a generalisation of the SEAL (Steric and Electrostatic ALignment) method proposed by Kearsley & Smith (1990). Here,

two molecules A and B are aligned by maximising the similarity index

$$S_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \sum_{i=1}^{k_{\text{A}}} \sum_{j=1}^{k_{\text{B}}} w_{ij} \exp\left(-\alpha ||\boldsymbol{x}_i^{\text{A}} - \boldsymbol{x}_j^{\text{B}}||^2\right) \tag{8}$$

with respect to rotation and translation. The weights w_{ij} are thereby chosen to be weighted averages of the electrostatic and the steric properties of atom i in A and atom j in B, i.e. $\omega_{ij} = w_Q q_i^{\text{A}} q_j^{\text{B}} + w_S v_i^{\text{A}} v_j^{\text{B}}$, where q_i^{M} denotes the partial charge value associated with the ith nuclear position in molecule M and v_i^{M} denotes some power of the corresponding van der Waals radius r_i^{M}.

The relationship of the SEAL objective function with the similarity index based on the kriged molecular fields becomes clear when the Gaussian covariance function $\sigma^{\text{G}}(\boldsymbol{h}) = \sigma^2 \exp\{-||\boldsymbol{h}||^2/\rho^2\}$, is considered. The quantity σ^2 thereby denotes the variance of the random field and the value of the range parameter ρ governs the spatial dependence of neighbouring observations. If $\sigma^{\text{G}}(.)$ is substituted into (6) the integral can be calculated analytically, and the Carbo index becomes

$$C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^{k_{\text{A}}} \sum_{j=1}^{k_{\text{B}}} w_i^{\text{A}} w_j^{\text{B}} \exp\left(-\frac{1}{2\rho^2} ||\boldsymbol{x}_i^{\text{A}} - \boldsymbol{x}_j^{\text{B}}||^2\right)}{N^{\text{A}} N^{\text{B}}}, \tag{9}$$

where the normalising constant associated with molecule M can now be written as

$$N^{\text{M}} = \left(\sum_{i=1}^{k_{\text{M}}} \sum_{j=1}^{k_{\text{M}}} w_i^{\text{M}} w_j^{\text{M}} \exp\left(-\tfrac{1}{2\rho^2} ||\boldsymbol{x}_i^{\text{M}} - \boldsymbol{x}_j^{\text{M}}||^2\right)\right)^{1/2}, \quad M \in \{A, B\}.$$

If a bivariate version of (9) using the steric and electrostatic properties of the molecules under consideration is applied, the numerator of the Carbo index is very similar to the SEAL objective function described above. In fact, if the information about the geometry of the two molecules is neglected and the covariance matrices in (2) are replaced by the identity matrices of the appropriate dimension, the two objective functions are identical. However, the use of (9) instead of the SEAL objective function comprises several advantages: apart from allowing for spatial dependence of the molecular properties, the weights in (9) exhibit a well–defined optimality property in that they minimise the prediction mean squared error. Moreover, the results in SEAL highly depend on the choices of the adjustable parameters (Klebe *et al.*, 1994) which can be circumvented by the data–driven choice of the parameter values in the kriging–based approach.

3 MCMC for the Pairwise Alignment of Molecules

3.1 The Likelihood

We shall develop a Bayesian model for the alignment of two molecular fields. Using a Markov chain Monte Carlo (MCMC) scheme and posterior inference, a rotation/translation invariant molecular comparison can be carried out. Within this framework, it also is possible to introduce a mask parameter vector for each molecule to allow for the possibility that only parts of the molecules match. The underlying rationale for using masks is that most chemical binding processes only require a sufficient degree of complementarity between parts of the binding partners, whereas the rest of the molecules play at most a minor role. Let $\boldsymbol{\lambda}_{\text{A}} \in \mathbb{R}^{k_{\text{A}}}$ and $\boldsymbol{\lambda}_{\text{B}} \in \mathbb{R}^{k_{\text{B}}}$ denote the mask vectors whose entries are indicator functions, where $\lambda_i^{\text{M}} \in \{0, 1\}$ determines if the ith atom of molecule M ($M \in \{A, B\}$) is considered to contribute to the matching parts of the molecules $(\lambda_i^{\text{M}} = 1)$ or not $(\lambda_i^{\text{M}} = 0)$.

Following Dryden *et al.* (2007), we define a Bayesian model in which one molecule is viewed as random while the other one serves as a fixed reference molecule. Let A be the random molecule with an estimated field $\hat{Z}_{\text{A}}(\boldsymbol{x})$ and B the fixed molecule with field $\hat{Z}_{\text{B}}(\boldsymbol{x})$ We define the likelihood for the random molecule as

$$L(\hat{Z}_{\text{A}}(\boldsymbol{x})|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}, \tau, \xi, \hat{Z}_{\text{B}}(\boldsymbol{x})) \propto \tau^{\xi-1} \exp\big(-\tau\, D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}})\big), \tag{10}$$

where $\boldsymbol{\theta}$ denotes the vector of the Euler angles which specifies a rotation matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ and $\boldsymbol{\gamma}$ denotes a displacement vector between A and B. The mask vectors play a similar role as the labelling matrices in the MCMC schemes defined by Green & Mardia (2006) and Dryden *et al.* (2007). Due to the continuous representation of molecular shapes we use in our paper, however, there is no need to establish one–to–one (or many–to–one) correspondences between atoms in molecule A and molecule B, and it suffices to define two separate mask vectors. Further, $D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}})$ denotes a variant of the discrepancy measure (4) which depends on the mask vectors through a "partial" Carbo index of the form

$$C_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}) \;=\; \frac{\sum_{i:\lambda_i^{\text{A}}=1} \sum_{j:\lambda_j^{\text{B}}=1} w_i^{\text{A}}(\boldsymbol{\lambda}_{\text{A}}) w_j^{\text{B}}(\boldsymbol{\lambda}_{\text{B}}) \int \sigma(\boldsymbol{x} - \boldsymbol{x}_i^{\text{A}}) \sigma(\boldsymbol{x} - \boldsymbol{x}_j^{\text{B}}) d\boldsymbol{x}}{N^{\text{A}}(\boldsymbol{\lambda}_{\text{A}}) N^{\text{B}}(\boldsymbol{\lambda}_{\text{B}})},$$

where $N^{\text{M}}(\boldsymbol{\lambda}_{\text{M}})$ denotes the normalising constant associated with molecule M ($M \in \{A, B\}$). The term "partial" thereby reflects that the mask vectors determine which atoms (and associated quantities) are included in the molecular comparison. Throughout we shall use the Gaussian covariance function for the kriging, and hence the integral in the Carbo index is available analytically as in (9). The remaining parameter in (10) is a precision parameter $\tau \in \mathbb{R}^+$ which determines the mean and variance of the model.

3.2 Prior Distributions and Posterior Sampling

We do not have any prior information about the rigid body parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ so that they are treated as uniformly distributed on $SO(3)$ and on a large bounded region in $\mathbb{R}^3$, respectively. Let $\Lambda_{k_{\text{M}}}$ denote the space of all k_{M}–vectors with entries of either zero or one. To prevent the MCMC algorithm from converging to a solution where very few atoms are used in the field comparison, we introduce a penalty parameter $\zeta > 1$ and define the joint prior density of the mask vectors as

$$\pi(\boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}|\zeta) \propto \zeta^{\sum_i \lambda_i^{\text{A}} + \sum_i \lambda_i^{\text{B}}}, \quad (\boldsymbol{\lambda}_{\text{A}}^T, \boldsymbol{\lambda}_{\text{B}}^T) \in \Lambda_{k_{\text{A}}} \times \Lambda_{k_{\text{B}}}.$$

The penalty parameter therefore inherently comprises prior assumptions about the extent of the matching parts of A and B. With the further assumptions that the precision parameter is Gamma distributed *a priori*, i.e. $\tau \sim \Gamma(\alpha, \beta)$, and that all unknown parameters are independent *a priori*, their joint posterior density conditioned on the given data has the property

$$\begin{aligned}\pi\big(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}, \tau | \hat{Z}_{\text{A}}(\boldsymbol{x}), \hat{Z}_{\text{B}}(\boldsymbol{x}), \alpha, \beta, \xi, \zeta\big) \\ \propto \tau^{\xi+\alpha-2} \exp\big\{-\tau\left(D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}) + \beta\right)\big\} \cdot \zeta^{\sum_i \lambda_i^{\text{A}} + \sum_i \lambda_i^{\text{B}}} \cos(\theta_2).\end{aligned}$$

Note that this can be regarded as a mixture model over $\Lambda_{k_{\text{A}}} \times \Lambda_{k_{\text{B}}}$.

Bayesian inference can now be carried out in order to obtain a rotation/translation invariant notion of (dis)similarity between the molecular fields $\hat{Z}_{\text{A}}(\boldsymbol{x})$ and $\hat{Z}_{\text{B}}(\boldsymbol{x})$. In particular, we use MCMC to sample from the posterior distribution and obtain point estimates for the rigid

body parameters and the mask vectors which can then be substituted into $D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}})$. Within the MCMC scheme, τ is updated with a Gibbs step using its full conditional distribution. Updated versions of the other parameters are obtained in four blocks, each using a Metropolis–Hastings step. For the rigid body parameters, we use random walk proposals with normally distributed noise, and a proposal distribution for the masks vectors $\boldsymbol{\lambda}_{\text{A}}$ and $\boldsymbol{\lambda}_{\text{B}}$ can be obtained by choosing an entry at random and then switching its value from zero to one or *vice versa*.

The algorithm that is used ensures that the defined Markov chain is irreducible and aperiodic. Hence, the chain will converge and eventually the simulated value will be an approximate realisation from the posterior distribution. We will estimate the parameters using the posterior mode or posterior mean obtained over a large number of iterations.

Convergence to, and sampling from, the limiting distribution in practice results in an approximate stochastic minimisation of the discrepancy term, with the concentration τ being large for close molecule matches. In fact, if one is mainly interested in obtaining point estimates of the model parameters which provide a good superposition, a thorough exploration of the parameter space is redundant. Instead simulated annealing (Kirkpatrick *et al.*, 1983) can be included so that the MCMC algorithm simulatates from

$$\pi\big(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}}, \tau | \hat{Z}_{\text{A}}(\boldsymbol{x}), \hat{Z}_{\text{B}}(\boldsymbol{x}), \alpha, \beta, \xi, \zeta\big)^{1/T}, \tag{11}$$

where $T > 0$ is slowly reduced deterministically.

4 Multiple Alignment of Molecules

In the multiple alignment problem, the objective is to simultaneously superimpose a set of n molecules $M_1, \ldots M_n$. Previous approaches to this problem include Dryden *et al.* (2007) who extend their two–configuration matching approach to the multiple configuration situation and Ruffieux & Green (2008) whose approach is based on the model formulated by Green & Mardia (2006) (cf. Section 6 for a further discussion). Here, we adapt the generalised Procrustes analysis (GPA) algorithm for discrete landmark data (e.g. Dryden &

Mardia (e.g. 1998, p.90)) to our field–based approach. In the classical GPA context, it is of interest to find an alignment of the given objects which minimises the sum of their pairwise distances. A similar goodness of fit criterion for the multiple superposition of n molecular fields can be formulated in terms of their overall similarity as

$$C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \int \frac{\hat{Z}_i(\boldsymbol{x}) \hat{Z}_j(\boldsymbol{x})}{N_i N_j} d\boldsymbol{x}, \tag{12}$$

where $\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_n^T) \in \mathbb{R}^{\sum_i k_i}$, $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T) \in \mathbb{R}^{3n}$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_n^T) \in \mathbb{R}^{3n}$ denote the stacked vectors of the involved mask, rotation and translation parameters, respectively. As before, the field of the ith molecule depends on the position of the molecule and on the ith mask vector, i.e. $\hat{Z}_i(\boldsymbol{x}) = \hat{Z}_i(\boldsymbol{x}; \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$ whereas the associated normalising constant only depends on the mask vector, i.e. $N_i = N_i(\boldsymbol{\lambda}_i)$.

For the multiple alignment of $M_1, \ldots, M_n$ we want to maximise (12) with respect to the $6n + \sum_i k_i$ involved parameters. Note that the multiple Carbo index has the property

$$C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \propto \frac{1}{n} \sum_{i=1}^{n} \int \tilde{Z}_i(\boldsymbol{x}) \tilde{Z}_{(i)}(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{n} \sum_{i=1}^{n} C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i; \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\lambda}_{(i)}),$$

where $\tilde{Z}_i(\boldsymbol{x}) = \hat{Z}_i(\boldsymbol{x})/N_i$ denotes the normalised field of the ith molecule and $\tilde{Z}_{(i)}(\boldsymbol{x})$ denotes a "normalised mean field" of all but the ith molecule, i.e.

$$\tilde{Z}_{(i)}(\boldsymbol{x}) \;=\; \frac{1}{n-1} \sum_{j \neq i} \sum_{l:\lambda_l^j = 1} \frac{1}{N_j} w_l^j(\boldsymbol{\lambda}_j) \sigma(\boldsymbol{x}_l^j - \boldsymbol{x}),$$

where λ_l^j denotes the lth entry of the mask vector $\boldsymbol{\lambda}_j$, $\boldsymbol{x}_l^j$ is the xyz–coordinate vector of the lth atom in the jth molecule, and $w_l^j(\boldsymbol{\lambda}_j)$ denotes the corresponding kriging weight. Due to this decomposition, the optimisation can be carried out stepwise by maximising $C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i; \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\lambda}_{(i)})$ in turn. The vectors $\boldsymbol{\theta}_{(i)}^T = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_{i-1}^T, \boldsymbol{\theta}_{i+1}^T, \ldots \boldsymbol{\theta}_n^T)$, $\boldsymbol{\gamma}_{(i)}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{i-1}^T, \boldsymbol{\gamma}_{i+1}^T, \ldots \boldsymbol{\gamma}_n^T)$ and $\boldsymbol{\lambda}_{(i)}^T = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_{i-1}^T, \boldsymbol{\lambda}_{i+1}^T, \ldots \boldsymbol{\lambda}_n^T)$ are thereby kept fixed at each step.

Let $D_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$ denote the discrepancy measure which results from applying the distance transformation (4) to $C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$. A stepwise maximisation of $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is then equivalent to minimising $D_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$ in turn. To do so, we apply an optimisation version of the MCMC algorithm for the pairwise alignment at each step. The normalised mean

field $\tilde{Z}_{(i)}(\boldsymbol{x})$ thereby takes the role of the fixed reference molecule whereas $\tilde{Z}_i(\boldsymbol{x})$ acts as the random test molecule whose parameters $\boldsymbol{\theta}_i$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\lambda}_i$ are to be updated.

Our MCMC scheme can be used as an approximate optimisation algorithm due to the interplay of the precision parameter τ and the acceptance probability for "downhill moves". In particular, if we choose a prior distribution with a large mean for τ, the MCMC algorithm in practice pushes the estimates of the other model parameters towards the posterior mode, rather like using a low temperature parameter T in (11). An algorithm which updates the normalised fields $\tilde{Z}_i(\boldsymbol{x})$ in turn using a "large precision version" of the MCMC algorithm for the pairwise alignment and then uses the obtained MAP estimates to determine a new mean field will therefore in practice decrease $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ at every step. This procedure can then be repeated until a convergence criterion is met. The algorithm is displayed as Algorithm 1.

INSERT ALGORITHM 1 ABOUT HERE

As the objective of the multiple alignment is to find the molecular features common to all or most of the molecules under study, we initialise the algorithm by superimposing each molecule on the smallest (in terms of the number of atoms) steroid molecule in the dataset. Contrary to the pairwise alignment which started at a random place in the parameter space, this initialisation will be close to the global optimum which justifies the use of the large prior mean for the precision values. All the algorithms described in this paper have been written in R (R Development Core Team, 2006).

5 Application to Steroid Molecules

5.1 Pairwise Alignment

We first consider the pairwise alignment of the steroid molecules. As the alignment is asymmetric, in that one molecule is treated as random whereas the other one serves as a fixed reference molecule, we carry out each of the possible 930 pairwise superpositions.

For each superposition, 10,000 MCMC iterations are used, and each iteration contains five blocks updating rotation, translation, precision, and the two mask vectors, respectively. In an initial phase of the MCMC algorithm, we use the information about both the partial charge values and the (cubed) van der Waals radii by calculating a bivariate partial Carbo index. The univariate partial Carbo index for each property is thereby calculated assuming that the corresponding random field is very smooth and exhibits a Gaussian covariance structure. The range of the Gaussian covariance function associated with the electrostatic field is estimated from the data by visual inspection of a pooled empirical semivariogram function. The range for the steric field is taken to be the largest van der Waals radius in the dataset. The resulting covariance functions then have the form

$$\sigma_Q(\boldsymbol{h}) = \sigma_q^2 \mathrm{e}^{-\frac{||\boldsymbol{h}||^2}{\rho_q^2}} \quad \text{and} \quad \sigma_S(\boldsymbol{h}) = \sigma_s^2 \mathrm{e}^{-\frac{||\boldsymbol{h}||^2}{\rho_s^2}},$$

where $\rho_q^2 = 363$, $\rho_s^2 = 8.67$. As σ_q and σ_s cancel out when calculating the Carbo indices, they do not need to be estimated.

The initial phase comprises $N_I = 2,000$ MCMC iterations during which the relative weights for the partial charges and van der Waals radii are chosen dynamically: at the ith iteration they are defined as $w_Q = \frac{N_I - i}{N_I}$ and $w_S = \frac{i}{N_I}$, $i = 1, \ldots, N_I$. The electrostatic fields are therefore only used for an approximate alignment and their impact fades out as the algorithm proceeds. This mimics real–life molecular recognition where the long–range electrostatic attraction governs the initial approach of the molecules. As they get closer, however, the short–range repulsive steric forces take over and become the chief manipulator for the binding affinity (e.g. Richards, 1993). After the initial 2,000 iterations, the alignment is adjusted using the univariate partial Carbo for the cubed van der Waals radii only.

To choose the value for the likelihood parameter ξ, we exploit the fact that the likelihood for the data also has the form of a Gamma distribution for the precision parameter τ with shape parameter ξ and a variable scale parameter $D_{\text{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\text{A}}, \boldsymbol{\lambda}_{\text{B}})$ which changes from iteration to iteration. From pilot runs of the MCMC algorithm we therefore have to opportunity to estimate ξ empirically using standard probability plots. As a value of $\hat{\xi} = 18$ fits the

observed data well for all pilot runs, we use it throughout the analysis.

The hyperparameters associated with the prior distributions for the precision parameter chosen as $\alpha = 16$, $\beta = 0.04$. The choice of β is thereby based on the fact that the discrepancy measure D_{AB} for good matches typically takes values between 0.01 and 0.05. Due to the form of the posterior distribution for the precision parameter, larger values for β mask the impact of the discrepancy at each iteration on the proposed value for τ, which is undesirable. Even smaller values of β on the other hand increase the posterior mean for τ. Unless the initial alignment of the molecules is known to be close to the optimal one, this results in a spurious notion of precision and increases the probability of it getting trapped in a local mode. The same reasoning applies for the chosen value of α and overall, the combination of $\alpha = 16$ and $\beta = 0.04$ works well for our application.

The value for the penalty parameter is chosen applying the decision theoretical approach described in Green & Mardia (2006). From pilot runs of our MCMC scheme we found that a penalty parameter value of $\zeta = 3$ gives the best distinction between included and excluded atoms in terms of the marginal posterior inclusion probabilities p_i. This value $\zeta = 3$ gives desirable robustness against changes of cost ratio $K = l_{01}/(l_{01} + l_{10})$, where l_{01} is the cost for falsely treating an atom as part of the matching parts of the molecules and l_{10} is the cost for a false negative. The optimal mask vector $\boldsymbol{\lambda}^{\text{opt}}$ for a given cost ratio $K \in [0, 1]$ is given by $\hat{\lambda}_i^{\text{opt}} = I_{\{p_i > K\}}$, where $I_{\{\text{E}\}}$ denotes the indicator function of an event E.

As standard deviations of the proposal distributions we choose $\eta_1 = 3.25^\circ$ for the rotation parameters and $\eta_2 = 0.5$Å for the translation parameters, and these values ensure acceptance rates for the associated parameters between 20% and 40%. The standard deviation for the rotation parameters is thereby in line with previously described proposal distributions for rotation parameters, e.g. with $\eta_1 = 3.25^\circ$, roughly 92% of the proposed rotation values fall into the limits of the uniform proposal distribution on $[-0.1, 0.1]$ which Green & Mardia (2006) use for a Metropolis update of θ_2.

Finally, for each run we define the initial relative position of the two molecules under study by first aligning the reference molecule along its principal axes and transforming the test

molecule in the same way to preserve the relative position. We then translate the random test molecule using a translation vector $\boldsymbol{\gamma}_0$, where $\gamma_{0i}\,(i = 1, 2, 3)$ is uniformly distributed on $[-5\text{Å}, 5\text{Å}]$. A further rotation using a rotation matrix $\mathbf{\Gamma}(\boldsymbol{\theta}_0)$, where $\theta_{0i}\,(i = 1, 2, 3)$ is uniformly distributed on $[-90^\circ, 90^\circ]$, then transforms the test molecule to its random initial position.

INSERT FIGURES 2-4 ABOUT HERE

An example run of the MCMC algorithm is illustrated in Figures 2-4. Here, aldosterone is taken to be the random test molecule which is to be superimposed onto androstanediol (cf. Figure 1). The initial relative position and the relative position according to the maximum a–posteriori (MAP) estimates of the rigid body parameters after a burn–in period of 2,000 iterations are displayed in Figure 2. Figure 3 shows the trace plots for the number of atoms which are involved in the field calculations and are hence considered as belonging to the common part of the molecules and a (post burn–in) summary of the masks vectors for the two molecules. For each atom, the average value of the corresponding entry (big circles) and the MAP estimate (small circles) are displayed. Figure 4 shows the trace plots of the other variable quantities.

In the majority of the 930 superpositions, a similar behaviour of the trace plots can be observed. Simultaneous inference about the rigid body parameters, the precision parameter and the mask vectors, however, is a difficult task and due to the complexity of the problem it is not surprising that the MCMC algorithm sometimes gets trapped in a local mode. As described in Dryden *et al.* (2007), the local modes for the steroid application essentially correspond to alignments along the principal axes, and one of these alignments is correct. To overcome the difficulty of this multimodality, we restart the algorithm by generating another random initial position for the test molecule if the sum of the 10% smallest distances between atoms of the test and the reference molecule exceeds 400 Å after 1,500 iterations or if the mean of the Carbo distance values between iteration 3,000 and 4,000 exceeds 0.1. These criteria are based on the experience we gained from pilot runs of the MCMC algorithm. The latter can thereby interpreted as a convergence criterion whereas the first is merely used as an early detector for an alignment along the wrong principal axis.

INSERT TABLE 1 ABOUT HERE

To investigate the sensitivity of the analysis to the prior distributions, we again consider the alignment of the two molecules aldosterone and androstanediol. Table 1 shows how different values of the penalty parameter ζ affect the empirical (post burn–in) 95% credibility intervals for the number of included atoms for both molecules. As expected, the total number of included atoms increases with ζ. As the two molecules in the example run are structurally very similar, they can be aligned more closely if more atoms are included so that credibility interval for the precision parameter is shifted towards higher values as ζ increases. After a certain threshold, however, even larger values for the penalty parameter force the algorithm to include more atoms in the similarity calculations than desired and the precision decreases. Table 1 shows that, in terms of the number of included atoms, the algorithm is robust against changes of α. Also as the posterior mean and variance of the precision parameter directly depends α, the credibility intervals for τ become wider and get shifted towards higher values as α increases.

The pairwise distances which result from the superpositions can be regarded as chemically relevant if they reflect the membership of the steroid molecules to the three activity classes, i.e. if steroids within an activity class can be aligned more closely than those from different activity classes. To asses this, we perform two cluster analyses using Ward's (1963) method as implemented in the R function `hclust`. To account for the asymmetry in our alignment method, the applied pairwise dissimilarity measures for two molecules A and B are thereby based on both the MCMC run which superimposes A on B and the MCMC run which superimposes B on A. In particular, we use

$$D_{\text{mean}}(A,B) = \sqrt{D_{AB}\big(\mathbf{\Gamma}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\lambda}}_{\text{A}}, \bar{\boldsymbol{\lambda}}_{\text{B}}\big) D_{BA}\big(\mathbf{\Gamma}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\gamma}}, \bar{\boldsymbol{\lambda}}_{\text{B}}, \bar{\boldsymbol{\lambda}}_{\text{A}}\big)}$$

and

$$D_{\text{MAP}}(A,B) = \sqrt{D_{AB}\big(\mathbf{\Gamma}(\boldsymbol{\theta}^{\text{MAP}}), \boldsymbol{\gamma}^{\text{MAP}}, \boldsymbol{\lambda}_{\text{A}}^{\text{MAP}}, \boldsymbol{\lambda}_{\text{B}}^{\text{MAP}}\big) D_{BA}\big(\mathbf{\Gamma}(\boldsymbol{\theta}^{\text{MAP}}), \boldsymbol{\gamma}^{\text{MAP}}, \boldsymbol{\lambda}_{\text{B}}^{\text{MAP}}, \boldsymbol{\lambda}_{\text{A}}^{\text{MAP}}\big)},$$

where $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\gamma}}$ denote the (post burn–in) estimates of the marginal posterior mean vectors of the rigid body parameters, and $\bar{\boldsymbol{\lambda}}_{\text{A}}$ and $\bar{\boldsymbol{\lambda}}_{\text{B}}$ denote thresholded mean mask vectors. The cost ratio is thereby chosen as $K = 0.7$ which is based on the fact that values of $\bar{\lambda}_i^{\text{M}}$ below

0.7 appear as outliers in the majority of graphs of the type of Figure 3. From a decision theoretical point of view, $K = 0.7$ indicates that we consider a false inclusion of an atom as worse than a false exclusion which is readily justified by the fact that including atoms in the distance calculation which do not contribute to the binding affinity towards the common receptor can distort an alignment more severely than falsely omitting relevant atoms. The second cluster analysis is based on a similar distance measure but using the MAP estimates of the parameters.

INSERT FIGURE 5 ABOUT HERE

Figure 5 shows the dendrograms resulting from the cluster analyses. The graph on the left–hand side is based on $D_{\text{mean}}(.)$, and the right–hand side shows the dendrogram calculated using $D_{\text{MAP}}(.)$. The labels on both sides correspond to the activity classes of the steroid molecules. It is notable that both distance measures lead to a very good separation of high and low activity steroids. In particular, the cluster analysis based on $D_{\text{MAP}}(.)$ is at the highest level able to separate these two activity classes completely. Overall, both dendrograms are more homogeneous than the one in Dryden *et al.* (2007) which is comparable to the ones in Figure 5 in that it uses the geometrical information only, i.e. the dendrogram on the right–hand side of the top row of Figure 5 in their paper.

5.2 Multiple Alignment

The pairwise superpositions used to initialise the field GPA algorithm (step 1) are carried out in exactly the same way as the superpositions described in the previous section. Only the penalty parameter ζ is reduced to $\zeta = 2$ to incorporate the knowledge that the reference molecule in all superpositions has a small number of atoms. Whereas in step 1 the electrostatic fields of the molecules are used for an approximate alignment, the superpositions on the mean fields (step 7) are obtained using only the discrepancies of the steric fields. Like in the pairwise alignment, the steric fields are thereby assumed to exhibit a Gaussian covariance structure with a practical range of $\sqrt{3}\rho_s = 1.7$. As the initial molecular fields obtained in step 1 are good approximations of the fields which minimise the multiple Carbo index,

we use $\alpha = 600$ and $\beta = 0.0001$ to ensure that full conditional of the precision parameter has a large mean value at each iteration, and we reduce the standard deviations of the proposal distributions for the rigid body parameters to $\eta_1 = 0.75$ Å and $\eta_2 = 0.03°$. Moreover, we set the number of iterations for each MCMC run in step 7 to 500, and the tolerance value to $tol = 0.0001$. Therefore here, the algorithm is used as a stochastic optimizer.

INSERT FIGURE 6 ABOUT HERE

The algorithm converges quickly. After the 4th field GPA iteration, the improvement of the multiple alignment ceases to exceed the tolerance threshold and the algorithm stops. Figure 6 shows orthographic views of the resulting overlays. The superposition after step 1 of the field GPA algorithm is displayed in the top row, and the bottom row shows the final overlay after 4 iterations.

Although the field GPA is not a posterior simulation algorithm in the strict sense, it is still worth investigating the effect of the used values for α and β: in step 7 of the field GPA algorithm, the Carbo indices measuring the overlap of the field of an individual molecule M_i and the mean field of the remaining 30 molecules take very high values of around 0.97 so that the corresponding discrepancy values are very small (around 0.015). During the course of the algorithm, these distance values decrease down to values around 0.002. For the distance to have an impact, the value of β should therefore be below this value. With this restriction, the result of the field GPA is fairly robust against changes of α and β. In particular, with our choice of $\beta = 0.0001$, we ran the algorithm for nine values of α between 100 and 900 and observed only marginal differences between these runs in terms of the resulting entries of the masks vectors and the molecular coordinates. Merely the convergence rate is affected by the choice of α, and lower values yield a slower convergence.

The relative positions obtained in the field GPA provide the best overall alignment of the 31 steroid molecules and can therefore be used as basis for a global comparison of the steric properties of the molecules. It is, for example, of interest to explore whether there are significant differences between the mean steric fields of the three activity groups. However, the field GPA described above is designed to find the overall mean field and extracts only

features common to all molecules so that the resulting mask vectors are not suitable for this comparison. We therefore perform the generalised field matching within each group separately to obtain mask vectors which reflect the steric properties common to all molecules within a group but with the features of the individual molecules removed. Using these mask vectors and the relative positions obtained in the overall field GPA, we then calculate the mean fields for each groups.

INSERT FIGURE 7 ABOUT HERE

Figure 7 displays xy–cross–sections of the mean fields for different values of z. Light points thereby correspond to locations where the displayed steric field takes a large value whereas dark points show field values close to zero. Due to the fact that the common ring structure of the molecules is almost planar, the middle row ($z = 0$) essentially depicts the ring atoms of the mean fields and is similar for all three activity groups. At $z = 1.5$ and $z = -1.5$, however, differences occur and, as expected, these differences are most pronounced between the mean field of the high and low activity groups. The objective now is to assess whether the differences are statistically significant or not.

For each pair (C_a, C_b) of activity classes ($a, b = 1, 2, 3;\ a \neq b$), we want to test the null hypothesis that there are no differences between the observed mean fields. We consider a (two sample) t–field of the form

$$t_{ab}(\boldsymbol{x}) = \frac{\bar{Z}_a(\boldsymbol{x}) - \bar{Z}_b(\boldsymbol{x})}{s^*_{\text{pool}}(\boldsymbol{x})\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}, \quad \boldsymbol{x} \in \mathbb{R}^3, \tag{13}$$

where n_a and n_b denote the number of molecules in activity class C_a and C_b, respectively, and $\bar{Z}_a(\boldsymbol{x})$ and $\bar{Z}_b(\boldsymbol{x})$ denote the corresponding mean fields, and $s^{*2}_{\text{pool}}(\boldsymbol{x}) = s^2_{\text{pool}}(\boldsymbol{x}) + d$, is the pooled variance (with $d = 0.001$ a small offset to avoid spurious significance in regions far away from the centre where all predictions are essentally zero).

For each pairwise comparison of the given average fields (low, medium, high), we define a three–dimensional grid G and calculate a t–value of the form (13) at a large number of

points (142598 here). The residual process for jth molecule has the form

$$\begin{aligned} r_j(\boldsymbol{x}) &= \hat{Z}_j(\boldsymbol{x}) - \bar{Z}_{a_j}(\boldsymbol{x}) \\ &= \sigma_s^2 \left(\frac{n_{a_j}-1}{n_{a_j}} \sum_{l:\lambda_l^j=1} w_l^j(\boldsymbol{\lambda}_j)\, \mathrm{e}^{\frac{||\boldsymbol{x}_l^j-\boldsymbol{x}||^2}{\rho_s^2}} - \frac{1}{n_{a_j}} \sum_{\substack{k:M_k\in C_{a_j} \\ k\neq j}} \sum_{l:\lambda_l^k=1} w_l^k(\boldsymbol{\lambda}_k)\, \mathrm{e}^{\frac{||\boldsymbol{x}_l^k-\boldsymbol{x}||^2}{\rho_s^2}} \right), \end{aligned}$$

where C_{a_j} $(a_j \in \{1,2,3\})$ denotes the activity class of M_j. The mean of the variances of the standardised residual processes across the grid of interest G serves as an estimate for λ. Applying this procedure we obtain $\hat{\lambda} = 0.031$.

Using results from Cao & Worsley (2001), the above estimates can be used to approximate the probability that, under the null hypothesis, the maximum $T_{\max}$ of the random t–field under study exceed a threshold t. For the two–sided t–tests in our example, a threshold of $t = 5.26$ can be considered as significant at the 0.01 significance level. This critical value is conservative in the sense that it the largest of the critical values associated with the three pairwise comparisons. Figure 8 shows the regions in which significant differences between the mean steric fields of the three activity classes could be found which occur in the bottom right and/or top left of G. These findings are in line with Figure 7 and they are also supported by Figure 9 in Dryden *et al.* (2007) which is the equivalent figure for the atom–based alignment (although no significance tests were carried out). These findings support the conjecture that the steric properties of the steroid molecules have a discriminating effect with respect to the binding affinity towards the CBG receptor.

INSERT FIGURE 8 ABOUT HERE

6 Discussion

A major advantage of our procedure is that point correspondences do not need to be estimated when matching molecules. Another approach which does not require correspondences has been formulated by Durrleman *et al.* (2007) who view the given sets of point coordinates as segmented lines and formulate a distance between the point sets in terms of

a distance between the lines using "currents" and reproducing kernel Hilbert spaces. However, they do not incorporate the possibility that only subsets of the given point sets match but they do use non–rigid deformations.

In our examples we have used interpolation in the kriging step. An alternative would be to include a nugget effect in the covariance function, and the kriging would result in smooth predictions. This would be particularly appropriate in applications where there is more measurement noise present.

Our methodology has been developed in the context of aligning and comparing molecules in chemoinformatics. Although kriging has been mentioned before in the literature on molecular similarity (e.g. Fang *et al.*, 2004), its application to the estimation of a molecular field provides a novel tool for determining a field–based structural alignment. However, the fact that our alignment procedure can be seen as a probabilistic framework and generalisation of the SEAL algorithm which is well–established in the field of rational drug design, provides an indication of the suitability of our approach.

Our multiple alignment approach is related to the Bayesian model proposed by Dryden *et al.* (2007) which uses a similar concept but formulated in terms of the point locations. Contrary to that, a hidden point configuration in the fully model–based Bayesian approach by Ruffieux & Green (2008) is integrated out and the multiple alignment of n point sets involves all $2^n - n - 1$ possible types of matches. The fact that our field–based approach provides the opportunity to naturally incorporate additional information is of particular advantage in the multiple alignment setting as the resulting mean fields allow straightforward post–processing like significance testing.

When an alignment is to be carried out using more than one molecular property, a way to possibly improve the superposition results is to introduce separate mask vectors for each property. With separate masks, one could account for covariances between the field using cokriging (e.g. Subramanyam & Pandalai, 2004), which would be computationally demanding.

Our alignment methodology is based on continuous representation of shapes. As molecules

are fuzzy bodies of electronic clouds rather than discrete sets of atoms, it is particularly suitable in the problems described in this paper. However, it is not restricted to the molecular context and applicable for any situation where marked, unlabelled point sets are to be compared. In fact, as it does not require any predefined atom–by–atom correspondence, the field–based superposition of continuous shapes could be an approach to resolve the alignment problem for a fairly broad range of applications. Examples include matching organs in medical images, or matching objects in images of real–world scenes (e.g. faces).

Acknowledgements

This work is supported by an EPSRC/University of Nottingham studentship and a Leverhulme Research Fellowship. The authors would like to thank Jonathan Hirst and James Melville for motivating discussions about this work. The steroids dataset is available at `http://www2.ccc.uni-erlangen.de/services/steroids/`.

References

Cao, J. & Worsley, K. J. (2001): Applications of random fields in human brain mapping. In: M. Moore (ed.) *Spatial Statistics: Methodological Aspects and Applications*, volume 159, 169–182. New York: Springer Lecture Notes in Statistics.

Carbo, R., Leyda, L., & Arnau, M. (1980): An electron density measure of the similarity between two compounds. *International Journal of Quantum Chemistry*, 17, 1185–1189.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988): Comparative molecular field analysis CoMFA. 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110, 5959–5967.

Cressie, N. A. C. (1993): *Statistics for Spatial Data*. Chichester: Wiley.

Dryden, I. L., Hirst, J. D., & Melville, J. M. (2007): Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics. *Biometrics*, 63, 237–251.

Dryden, I. L. & Mardia, K. V. (1998): *Statistical Shape Analysis*. Chichester: Wiley.

Durrleman, S., Pennec, X., Trouvé, A., & Ayache, N. (2007): Measuring brain variability via sulcal lines registration: a diffeomorphic approach. In: N. Ayache, S. Ourselin, & A. Maeder (eds.) *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 4791 of LNCS. Brisbane.

Fang, K.-T., Ying, H., & Liang, Y.-Z. (2004): New approach by kriging models to problems in QSAR. *Journal of Chemical Information and Computer Sciences*, 44, 2106–2113.

Good, A. C., Hodgkin, E. E., & Richards, W. G. (1992): The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. *Journal of Chemical Information and Computer Sciences*, 32, 188–191.

Good, A. C., So, S., & Richards, W. G. (1993): Structure–activity relationships from molecular similarity matrices. *Journal of Medicinal Chemistry*, 36, 433–438.

Green, P. J. & Mardia, K. V. (2006): Bayesian alignment using hierarchical models, with application in protein bioinformatics. *Biometrika*, 93, 235–254.

Kearsley, S. K. & Smith, G. M. (1990): An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlaps. *Tetrahedron Computer Methodology*, 3, 315–633.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983): Optimization by simulated annealing. *Science*, 220, 671–680.

Klebe, G., Meitzner, T., & Weber, F. (1994): Different approaches toward an automatic structural alignment of drug molecules: Applications to sterol mimics, thrombin and thermolysin inhibitors. *Journal of Computer–Aided Molecular Design*, 8, 751–778.

Miles, R. E. (1965): On random rotations in $\mathbb{R}3$. *Biometrika*, 52, 636–639.

R Development Core Team (2006): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards, W. G. (1993): Computers in drug design. *Pure and Applied Chemistry*, 65, 231–234.

Ruffieux, Y. & Green, P. J. (2008): Alignment of multiple configurations using hierarchical models. *submitted for publication*.

Subramanyam, A. & Pandalai, H. S. (2004): On the equivalence of the cokriging and kriging systems. *Mathematical Geology*, 36, 507–523.

Ward, J. H., Jr. (1963): Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

ζ	95% CI for τ	95% CI for $\sum_j \lambda_j^{\text{A}}$	95% CI for $\sum_j \lambda_j^{\text{B}}$
2	(226.62, 543.78)	(34, 46)	(34, 45)
3	(230.93, 543.30)	(37, 49)	(38, 48)
4	(250.69, 562.65)	(40, 51)	(40, 49)
5	(244.67, 548.41)	(41, 51)	(42, 51)
α	95% CI for τ	95% CI for $\sum_j \lambda_j^{\text{A}}$	95% CI for $\sum_j \lambda_j^{\text{B}}$
3	(102.53, 315.95)	(36, 48)	(37, 48)
13	(221.14, 515.13)	(38, 49)	(38, 49)
23	(344.68, 770.30)	(38, 48)	(39, 49)
33	(432.36, 1010.77)	(35, 48)	(37, 50)

Table 1: The impact of the penalty parameter (first four rows) and α (last four rows) on the marginal posterior distribution of the parameters of interest. The credibility intervals are based on every 20th value of the parameters recorded after the burn–in period.

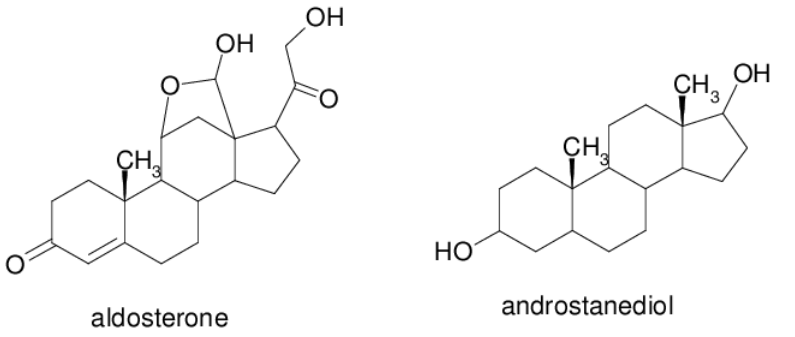

Figure 1: Two–dimensional representations of two steroid molecules from the dataset. The molecules are structurally similar in that their core structure consists of four carbon rings.

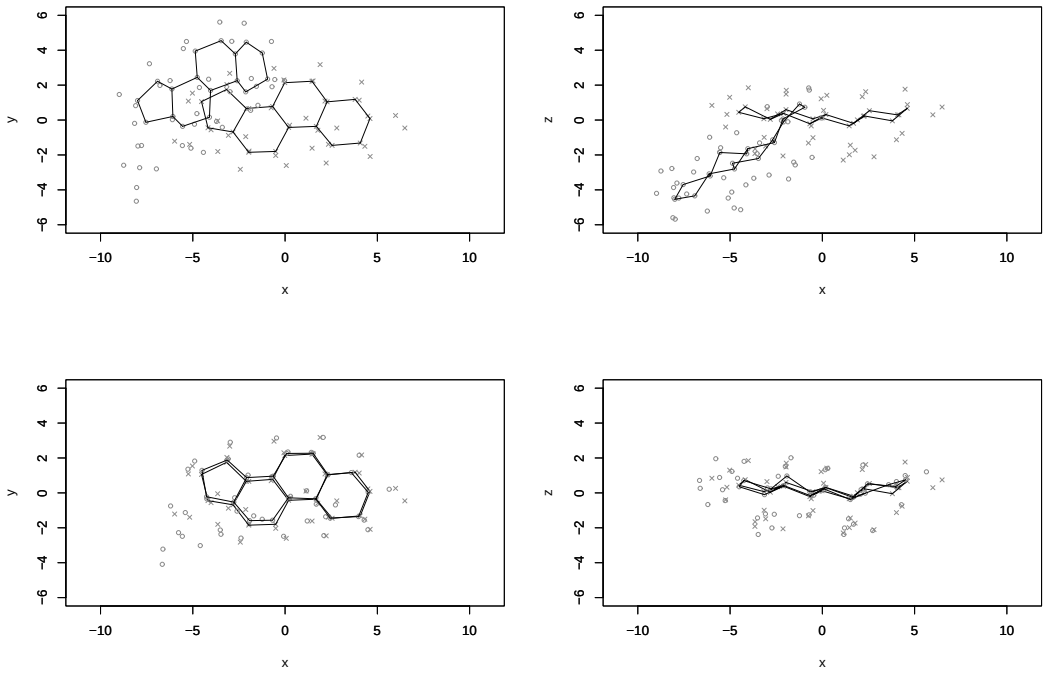

Figure 2: Orthographic views of the carbon rings in the starting position and the MAP position for the alignment of aldosterone and androstanediol. The unit of all axes is Ångström (Å).

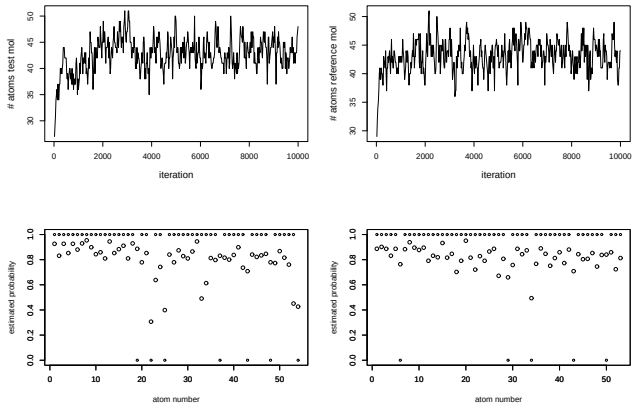

Figure 3: Top Row: Trace plots of the number of atoms which are involved in the kriging procedure. Bottom Row: Two possible point estimates for the mask vector of test molecule (left) and the reference molecule (right), respectively. The big circles show the mean values of the (0,1)–entries for the masks vectors, and the small circles display the observed mask vectors at the MAP iteration. The total number of atoms in test molecule is 54. The reference molecule has 53 atoms in total.

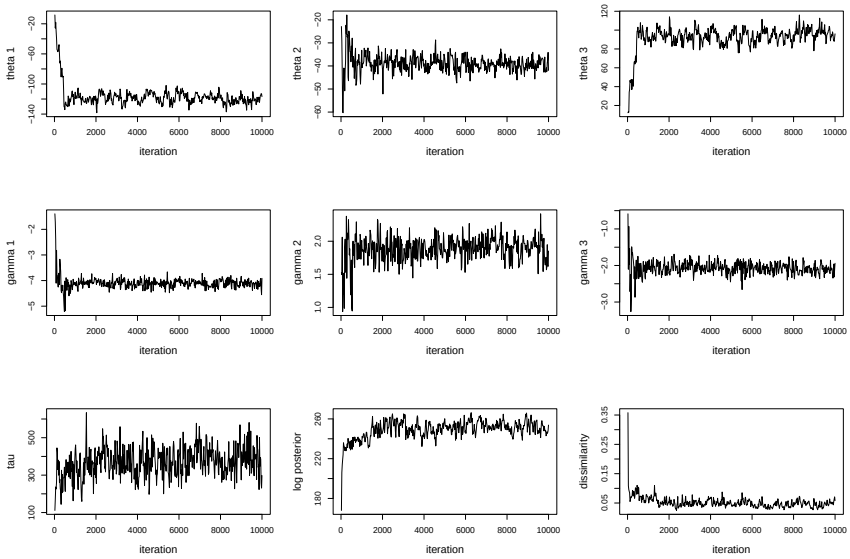

Figure 4: Top Row: trace plot of the rotation parameters $\theta_i\,(i = 1, 2, 3)$ in degrees. Middle Row: trace plots of the translation parameters $\gamma_i\,(i = 1, 2, 3)$. Here, all rigid body parameters are defined in terms of the initial relative position of the two molecules under consideration. Bottom row: Trace plots of the precision parameter, the log–posterior (up to a constant) and the Carbo distance. In all plots, every 20th simulated value is displayed.

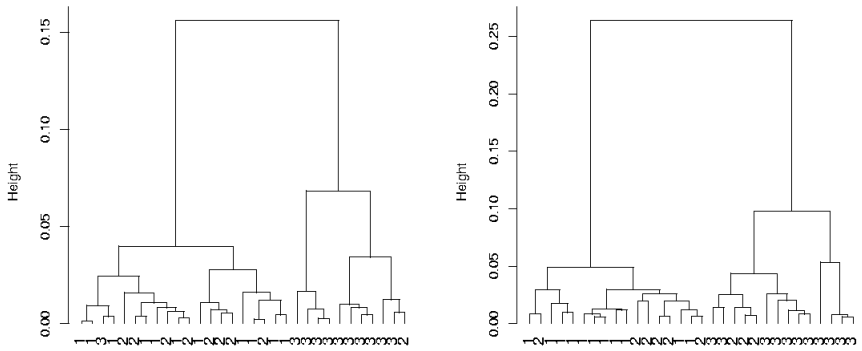

Figure 5: Cluster Analysis Using Ward's Method: The left–hand side dendrogram is based on $D_{\text{mean}}(.)$, and the dendrogram on the right–hand side is calculated using $D_{\text{MAP}}(.)$. The labels correspond to the activity classes of the steroids (1=high, 2=intermediate, 3=low).

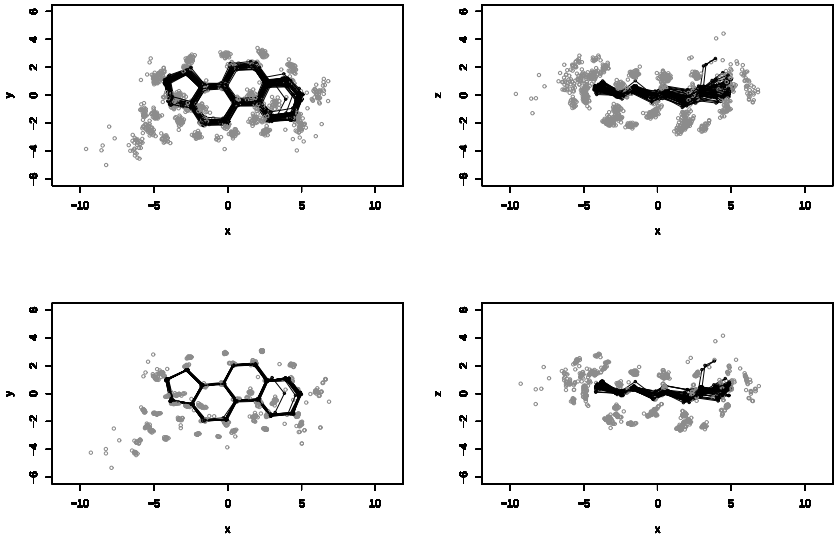

Figure 6: Top row: Orthographic projections of the the initial relative position of the 31 steroid molecules. Bottom Row: Orthographic projections of the final relative position of the 31 steroid molecules.

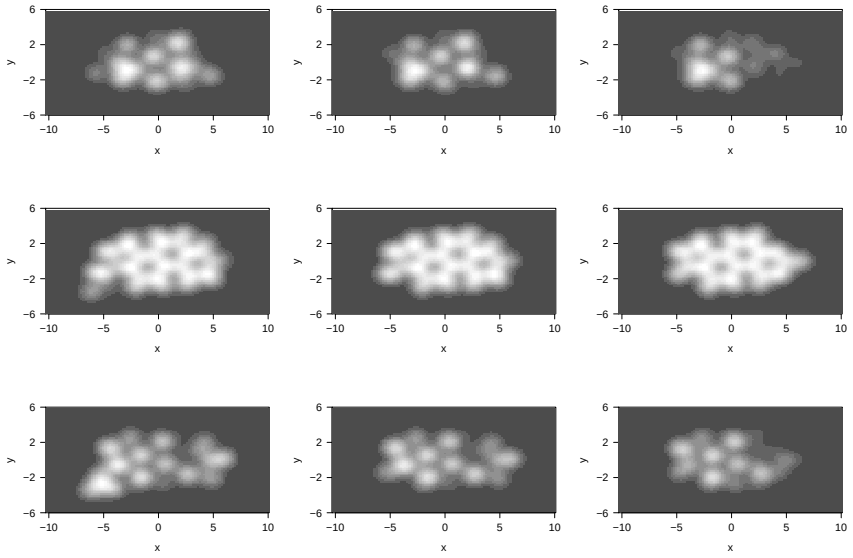

Figure 7: Cross–sections of the mean steric fields of the three activity groups (left column: high activity, middle column: medium activity, right column: low activity). The different rows display cross sections at $z = -1.5$ (top row), $z = 0$ (medium row), and $z = 1.5$ (bottom row). Light points correspond to locations with large value of the displayed field whereas dark values show points with values close to zero.

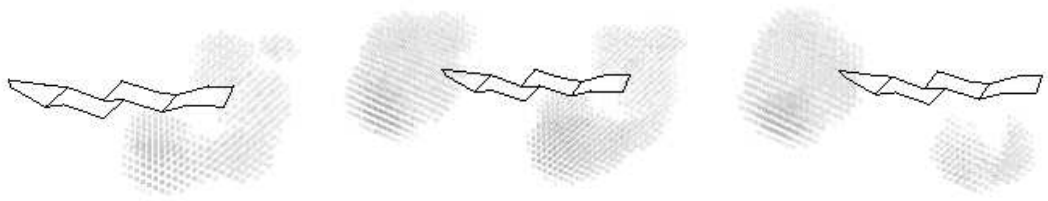

Figure 8: Thresholded t–Fields Resulting from Pairwise Comparison of the Steric Mean Fields of the Three Activity Classes. Left–Hand Side: Low vs. Medium Activity Class, Middle: Low vs. High Activity Class, Right–Hand Side: Medium vs. High Activity Class. The shaded areas display statistically significant regions. For orientation, the mean ring structure resulting from the overall GPA is displayed as well.

Algorithm 1 Stochastic GPA for Molecular Fields

choose the smallest molecule as reference molecule and superimpose the $n-1$ remaining molecules onto it using the MCMC algorithm for the pairwise alignment; the relative positions of the resulting $n-1$ MAP fields and the field calculated from the unchanged data for the smallest molecule then constitute the starting point for the generalised superposition

define $d \leftarrow d_0$, where $d_0 > tol$ and tol is a positive tolerance threshold

calculate the multiple Carbo index $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \int \frac{\hat{Z}_i(\boldsymbol{x}) \hat{Z}_j(\boldsymbol{x})}{N_i N_j} d\boldsymbol{x}$

while $d > tol$ **do**

  for i in $(1:n)$ **do**

    using the current parameter values for rotation, translation and mask vectors, calculate a normalised mean field $\tilde{Z}_{(i)}(\boldsymbol{x})$ omitting the ith molecule

    based on the discrepancy $D_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$, superimpose the ith molecular field onto $\tilde{Z}_{(i)}(\boldsymbol{x})$ using a large precision version of the MCMC algorithm for the pairwise alignment; $\tilde{Z}_{(i)}(\boldsymbol{x})$ thereby takes the role of the reference molecule and $\boldsymbol{\lambda}_{(i)}$, $\boldsymbol{\theta}_{(i)}$ and $\boldsymbol{\gamma}_{(i)}$ are treated as fixed

    record the MAP estimates for position and mask of the ith molecule

  end for

  calculate the updated $C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \int \frac{\hat{Z}_i^*(\boldsymbol{x}) \hat{Z}_j(\boldsymbol{x})}{N_i N_j} d\boldsymbol{x}$

  $d \leftarrow C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) - C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$

  $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \leftarrow C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$

end while