Estimating the probability of coincidental similarity between atomic displacement parameters with machine learning

IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 035033 https://doi.org/10.1088/2632-2153/ac022d

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
30 June 2020

REVISED
27 April 2021

ACCEPTED FOR PUBLICATION
17 May 2021

PUBLISHED
14 July 2021

PAPER

Estimating the probability of coincidental similarity between atomic displacement parameters with machine learning

Viktor Ahlberg Gagner, Maja Jensen and Gergely Katona*

Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden
* Author to whom any correspondence should be addressed.

E-mail: gergely.katona@gu.se

Keywords: Bayesian, atomic displacement parameter, B-factor, x-ray crystallography, protein, Markov Chain Monte Carlo, Wishart distribution

Supplementary material for this article is available online

Abstract

High-resolution diffraction studies of macromolecules incorporate the tensor form of the anisotropic displacement parameter (ADP) of atoms from their mean position. The comparison of these parameters requires a statistical framework that can handle the experimental and modeling errors linked to structure determination. Here, a Bayesian machine learning model is introduced that approximates ADPs with the random Wishart distribution. This model allows for the comparison of random samples from a distribution that is trained on experimental structures. The comparison revealed that the experimental similarity between atoms is larger than predicted by the random model for a substantial fraction of the comparisons. Different metrics between ADPs were evaluated and categorized based on how useful they are at detecting non-accidental similarity and whether they can be replaced by other metrics. The most complementary comparisons were provided by Euclidean, Riemann and Wasserstein metrics. The analysis of ADP similarity and the positional distance of atoms in bovine trypsin revealed a set of atoms with striking ADP similarity over a long physical distance, and generally the physical distance between atoms and their ADP similarity do not correlate strongly. A substantial fraction of long- and short-range ADP similarities does not form by coincidence and are reproducibly observed in different crystal structures of the same protein.

1. Introduction

Experimental research often tests the question of whether or not properties of systems are the same. If the properties are discrete, one can decide this without much of a problem, but nature does not reveal continuous properties to arbitrary precision. The true nature of these properties is clouded from the experimenter's vision, and a statistical framework is indispensable for reaching a conclusion. Machine learning has many applications for macromolecular crystallography, and it has been used for improving the estimates of structure factor amplitudes [1] and their differences [2] with applications in experimental phasing [3], serial crystallography [1] and time-resolved structural studies [4]. These examples employed supervised machine learning tools. Clustering, as an unsupervised method, is also widespread in crystallographic analysis [5, 6]. One example of the usefulness of an unsupervised method is the hierarchical clustering of anisotropic displacement parameters (ADPs) in atomic resolution crystal structures [7].

The model of a crystal structure is inherently probabilistic. Rather than observing individual atomic positions that are later aggregated by descriptive multivariate statistics, atoms are described by the parameters of a distribution from the beginning. In this paper, we focus on the commonly used multivariate normal distribution description of atoms with spatial dimensions. Within this standard framework, there is no deeper layer of physical description. The analysis of the statistical relation between atoms has to start with their distribution parameters, the mean atomic coordinate vectors and scale matrices in the form of ADPs (equivalent to covariance matrices). A statistical model requires a distribution that can describe ADP

matrices and a likelihood function that estimates the probability of ADPs given the various *a priori* distribution parameters. Wishart's distribution is a multivariate gamma distribution [8], and it is frequently used as a conjugate prior to covariance matrices. Wishart's distribution can be achieved stepwise from simpler random variables with Bartlett's method [9]. There are other routes to generate covariance matrices. Cholesky decomposition [10] of a covariance matrix results in the vector of variances and a correlation matrix. For these components, suitable prior distributions can be assigned, for example, the Lewandowski–Kurowicka–Joe distribution [11], to the correlation matrix. This route can be computationally efficient, but its current implementation is incomplete in supporting libraries. Using a model of ADPs, it is possible to infer the joint posterior parameter distribution of a Wishart distribution using Bayesian methods. In turn, we can generate random predictions of ADPs through random sampling from the joint posterior distribution and propagating the samples through the deterministic steps of the model. The randomly generated ADPs can be used to predict the baseline probability of coincidental pairwise similarity between random matrices.

We also explore which type of comparison provides a detailed map between ADPs and analyzes what level of similarity does not simply arise in the random model of ADPs. To define similarity, different definitions of distance between the ADPs were used [12, 13]. To this end, there is a need to use proper metrics, which incorporate multiple aspects of ADPs. It should be noted that there is no universally better or worse definition of distance. Hence, the choice depends on what property one wishes to highlight, namely the direction of the displacements or their (relative) magnitude. Which definition one uses matters less when the ADPs are highly similar. Regardless of the metrics, we still have to determine the magnitude of the distance that can arise through coincidence.

In this short report, we used a machine learning model to predict the shortest distance where coincidental agreements can occur with a pre-defined probability. The Bayesian model is trained on the ADPs of a protein structure and tested with a number of different distance definitions. Finally, we take a closer look at the recurring properties among atoms, which show significant similarity to a selected atom.

2. Material and methods

2.1. Location of atoms in high-resolution crystal structures

The location of an atom is usually defined by its coordinate vector $\boldsymbol{\mu}$ and its symmetric (ideally) semi-definite ADP tensor **U** once the resolution of the crystallographic data is better than 1.2 Å and the atoms appear as separate electron density peaks. Crystallographic refinement algorithms may or may not enforce meaningful **U.** For the purpose of this analysis, non-positive definite atom records are ignored. B_{eq} and B_{est} are then defined as [14]:

$$B_{\mathrm{eq}} = \frac{8}{3}\pi^2 \left(E_1 + E_2 + E_3\right), \tag{1}$$

$$B_{\mathrm{est}} = \frac{8}{3}\pi^2 \sqrt{\frac{E_1 + E_2 + E_3}{E_1^{-1} + E_2^{-1} + E_3^{-1}}}, \tag{2}$$

where E_1, E_2 and E_3 are eigenvalues of **U. U** is equivalent to the covariance matrix of a multivariate normal distribution. Using the multivariate normal description of the atoms, the probability of their positions can be estimated in the crystal structure. The graphical representation of **U** is an ellipsoid. The eigenvectors and eigenvalues of the **U** tensor define the axis directions and their relative lengths, respectively. The surface of the ellipsoid is drawn at an arbitrary probability level of the multivariate normal distribution. The machine learning treatment of correlated diffraction intensity observations [2] and atom positions is very similar.

2.2. Bayesian model of ADPs

The Bayesian model (equations (3)–(6)) approximated **U** matrices as fixed Wishart distributions [8]. The Wishart distribution is frequently used as a conjugate prior to covariance matrices. The prior distributions are described in equations (3) and (5). Equation (4) shows a deterministic transformation and equation (6) defines the probability of the observation **U**. Equations (5) and (6) represent the Wishart probability distributions:

$$p\left(\nu_0\right) = \sqrt{\frac{2}{\sigma^2\pi}} e^{-\frac{1}{2\sigma^2}\nu_0{}^2} = \sqrt{\frac{2}{9\pi}} e^{-\frac{1}{18}\nu_0{}^2}; \nu_0 \in [0, \infty], \tag{3}$$

wher σ is the scale parameter of the half-normal distribution ($\sigma = 3$ *a priori*),

$$\nu = \upsilon_0 + 5. \tag{4}$$

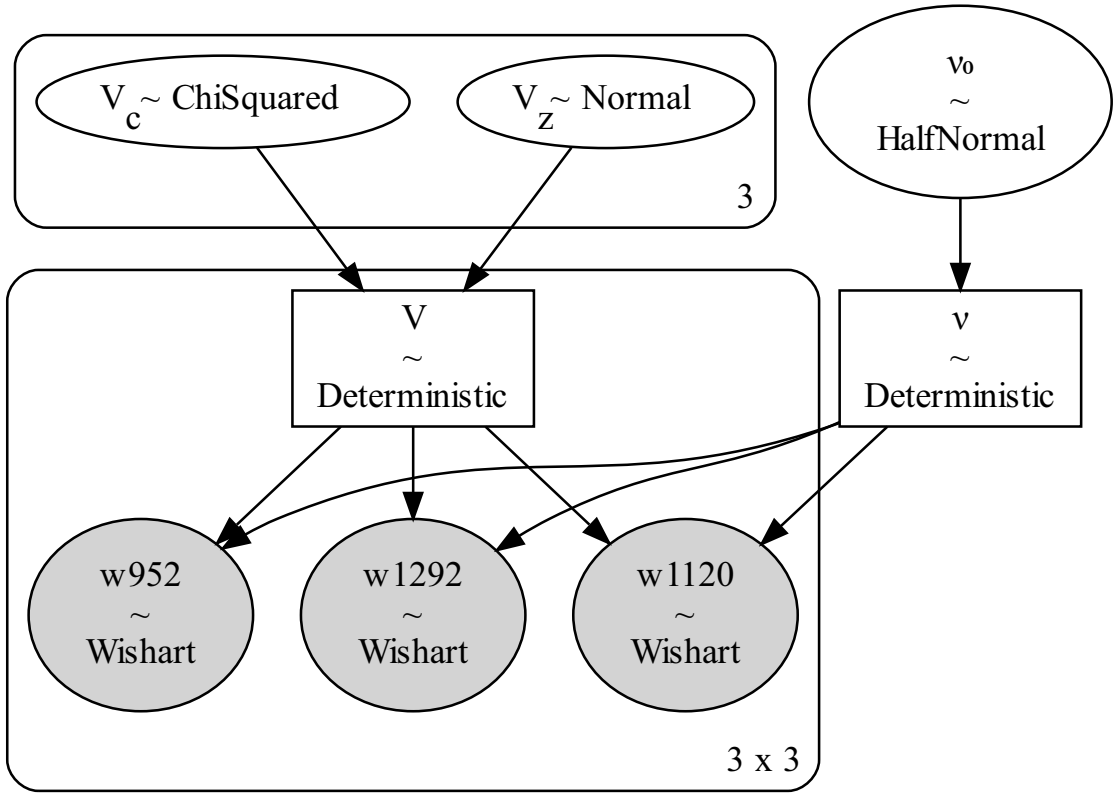

Figure 1. (a) Schematic representation of the Bayesian model used for machine learning. Variable names V_z and V_c are the off-diagonal and diagonal elements of the scale matrix **V** of the Wishart distribution, respectively. These are linked by Bartlett decomposition. (b) Degrees of freedom of the Wishart distribution are represented by ν, and ν_0 is its prior distribution before it is incremented by 5. For simplicity, only three of the 150 Wishart distributions are illustrated. Each corresponds to a randomly selected atom of the protein.

$$p(\mathbf{V}) = \frac{|\mathbf{V}|^{\frac{1}{2}}}{2^{\frac{15}{2}}\Gamma_p\left(\frac{5}{2}\right)} e^{-\frac{1}{2}tr(\mathbf{V})}; \mathbf{V} \succ 0 \tag{5}$$

$$p(\mathbf{U}|\nu, \mathbf{V}) = \frac{|\mathbf{U}|^{\frac{\nu-k-1}{2}}}{2^{\frac{\nu k}{2}}\left|\frac{\mathbf{V}}{\nu}\right|^{\frac{\nu}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} e^{-\frac{1}{2}tr\left(\left|\frac{\mathbf{V}}{\nu}\right|^{-1}\mathbf{U}\right)} = \frac{|\mathbf{U}|^{\frac{\nu_0+1}{2}}}{2^{\frac{3(\nu_0+5)}{2}}\left|\frac{\mathbf{V}}{\nu}\right|^{\frac{\nu_0+5}{2}}\Gamma_p\left(\frac{\nu_0+5}{2}\right)} e^{-\frac{1}{2}tr\left(\left|\frac{\mathbf{V}}{\nu}\right|^{-1}\mathbf{U}\right)}, \tag{6}$$

wher **U** are the ADP observations, **V** and ν are the positive definite-scale matrix and the degrees of freedom of the Wishart distribution, respectively. The rank of the matrix is $k = 3$. Γ_p is the multivariate gamma function ($p = 3$).

The relationship between the model parameters is illustrated in figure 1. In order to generate the ν_0 parameter a half-normal prior distribution was assumed with a σ parameter of 3. The prior scale matrix **V** is another Wishart prior distribution with an identity scaling matrix and ν parameter set to 5. The samples of the matrix **V** were generated by Bartlett decomposition [9] in order to maintain the symmetry and semi-definite positivity of its posterior samples. The ν_0 parameter of the observations was incremented by a fixed value of 5 to ensure that the Wishart likelihood function is always defined. The Wishart likelihood function in equation (6) has a degree of freedom ν and a scale parameter $\mathbf{V}/\nu$. The training set consisted of 150 protein atoms randomly selected from the crystal structure of a 0.93 Å resolution complex of bovine trypsin in a complex with a canonical inhibitor (PDB entry 2xtt) [15]. The remaining protein atoms formed the test set.

The model was implemented with the Python library PyMC3 [16], and the posterior distribution was estimated with the No-U-Turn Sampling (NUTS) algorithm [17] using the default parameters. The tuning was performed for 2000 Markov Chain Monte Carlo (MCMC) iterations, and the joint posterior distribution was represented by the last 1000 iterations. Four MCMC chains were generated to evaluate the reproducibility of convergence. The ADPs of the randomly selected atoms were assigned to separate 3×3 matrices, as vectorization of covariance matrices was not implemented in version 3.6 of PyMC3. This likely limited the

performance of the algorithm and increased the memory usage. A total of 150 observations were already sufficient for reproducible convergence to very similar posterior distributions even if a different set of atoms was selected. Another limitation of this implementation was that the MCMC algorithm could only use one processor core, and calculating four MCMC chains took 190 s on a Linux workstation (i7 3970X CPU at 3.50 GHz clock frequency). This runtime corresponds to 12 000 iterations and excludes initialization/compilation time.

2.3. For ADP predictions, one posterior parameter sample was used for one prediction

2.3.1. Distance calculations

For the distance calculations, the ADPs were converted to dimensionless matrices by dividing the tensor element magnitudes by 1 Å^2. The atomic positions were converted into dimensionless vectors by dividing the elements by 1 Å. The metrics between the dimensionless $\mathbf{U_A}$ and $\mathbf{U_B}$ were defined as:

(a) Euclidean distance is Frobenius's norm of the difference matrix between $\mathbf{U_A}$ and $\mathbf{U_B}$:

$$d_{\mathrm{E}} = \|\mathbf{U_A} - \mathbf{U_B}\|_F, \tag{7}$$

(b) Riemann distance:

$$d_{\mathrm{R}} = \sqrt{\sum_i \log(\boldsymbol{\lambda_i})^2}, \tag{8}$$

wher $\boldsymbol{\lambda_i}$ are the joint eigenvalues of $\mathbf{U_A}$ and $\mathbf{U_B}$ (assuming common eigenvectors).

(c) Log Euclidean distance is Frobenius's norm of the difference matrix between $\log(\mathbf{U_A})$ and $\log(\mathbf{U_B})$:

$$d_{lE} = \|\log(\mathbf{U_A}) - \log(\mathbf{U_B})\|_F, \tag{9}$$

(d) Log det distance:

$$d_{lD} = \sqrt{\log\left(\det\left(\frac{\mathbf{U_A} + \mathbf{U_B}}{2}\right)\right) - \frac{1}{2} \times \log\left(\det(\mathbf{U_A})\det(\mathbf{U_B})\right)}, \tag{10}$$

(e) Wasserstein distance [18]:

$$d_{\mathrm{W}} = \left(tr\left(\mathbf{U_A} + \mathbf{U_B} - 2\left(\mathbf{U_A}^{1/2}\mathbf{U_B}\mathbf{U_A}^{1/2}\right)^{1/2}\right)\right)^{1/2}, \tag{11}$$

(f) Symmetric Kullback–Leibler distance [19]:

$$d_{\mathrm{KL}} = \frac{1}{2}\left(tr\left(\mathbf{U_B}^{-1}\mathbf{U_A}\right) + tr\left(\mathbf{U_A}^{-1}\mathbf{U_B}\right) - 6\right), \tag{12}$$

(g) B_{eq} distance:

$$d_{\mathrm{B}} = \frac{8}{3}\pi^2 \left|tr(\mathbf{U_A}) - tr(\mathbf{U_B})\right|, \tag{13}$$

(h) Eigenvalue distance:

$$d_\lambda = \|\boldsymbol{\lambda}_\mathbf{A} - \boldsymbol{\lambda}_\mathbf{B}\|, \tag{14}$$

where $\boldsymbol{\lambda}_\mathbf{A}$ and $\boldsymbol{\lambda}_\mathbf{B}$ are the eigenvalues of $\mathbf{U_A}$ and $\mathbf{U_B}$, respectively. The distance is defined as the vector norm of the differences between $\boldsymbol{\lambda}_\mathbf{A}$ and $\boldsymbol{\lambda}_\mathbf{B}$.

In addition, the distance between atom positions was defined as:

$$d_\mathrm{r} = \|\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\|, \tag{15}$$

where $\boldsymbol{\mu}_\mathbf{A}$ and $\boldsymbol{\mu}_\mathbf{B}$ are the coordinates of atoms A and B. The distance is defined as the vector norm of the differences between $\boldsymbol{\mu}_\mathbf{A}$ and $\boldsymbol{\mu}_\mathbf{B}$.

The matrix logarithm of $\mathbf{U}$ is defined as:

$$e^{\log(\mathbf{U})} = \mathbf{U}. \tag{16}$$

The metrics were implemented using the Python libraries NumPy [20, 21], SciPy [22] and pyRiemann [23]. The Python libraries Matplotlib [24], Graphviz [25] and Seaborn [24] were used for data visualization. Validation of the method using synthetic data is detailed in the supporting online information (figures S1–S4 and tables S1 and S2 (available online at stacks.iop.org/MLST/2/035033/mmedia)).

3. Results and discussion

Based on the ADP tensors of the protein structure the posterior probabilities of the Bayesian model parameters were estimated using the NUTS algorithm. Convergence was reached after tuning, and in the subsequent 1000 iterations the parameter trace shows low autocorrelation (figure 2). Using the posterior probabilities of ν and $\mathbf{V}$ parameters, it was possible to draw samples from the Wishart random distribution, which here we assumed as a simulated reference, equivalent to the ADP tensors. We compare the random pool of samples to the protein ADPs in figures 3–5. In figure 3, the ADP matrix elements are compared between the protein atoms and posterior predictions, and in figure 4 their B_{est} values are compared. Although the distribution of protein ADPs was more complex than what a single Wishart distribution could describe, it captured the central tendency of the distribution well. For example, the mean of B_{eq} of the real samples was 8.3 Å^2, while the mean of the predicted B_{eq} was 8.0 Å^2. The scale matrix is also adapted to the specific distribution of the off-diagonal elements in the ADPs. For example, U_{13} element was shifted towards more negative values in the observations and the predicted Wishart distribution displayed a similar tendency (means -0.010 and -0.011 Å^2, respectively). These off-diagonal elements have stronger tails than in the experimental ADPs. In addition to the successful scaling of selected tensor elements, the predicted tensors always maintained symmetry and positive definite properties. The tensors are best visualized as ellipsoids (figure 5). Both the real atoms and the machine learning model had a similar variety of sizes and shapes of ellipsoids and the principal axis did not show a strong directional preference.

By using a mixture model of multiple Wishart distributions, it may be possible to describe the ADPs of protein atoms more accurately and improve the similarities further. One should be cautious, however, as the increase in the number of parameters may lead to overfitting, which is why this type of treatment is not presented here.

3.1. Comparison of ADP metrics

We determine the distance between two samples from the same Wishart random distribution repeatedly and evaluate what distance corresponds to a certain level of cumulative probability. In order to compare covariance matrices, there are many metrics are available. All definitions assume that for identical covariance matrices the distance is 0 and most definitions are commutative, i.e. the distance is the same in both directions. Perhaps it is most computationally efficient to calculate the Euclidean distance (d_E). A general case of the Euclidean distance is the Riemann distance (d_R) defined between two points in a Riemann manifold. Locally, the geometry of a smooth Riemann manifold approaches Euclidean geometry, which makes the two interchangeable when very similar ADPs are compared. Unfortunately, d_E and d_R started to diverge rapidly. Atom pairs with similar d_E can have different d_R and vice versa. These metrics do not have to be particularly large to observe this tendency. Figure 6 shows a comparison of different covariance distances between pairs of randomly selected atoms in the protein structure and between two samples from the simulated Wishart distribution, respectively. For d_E, d_R, d_lE, d_W and d_lD, the two distance probability distributions differ to a large extent. Our null hypothesis is that there is no relationship between ADPs of

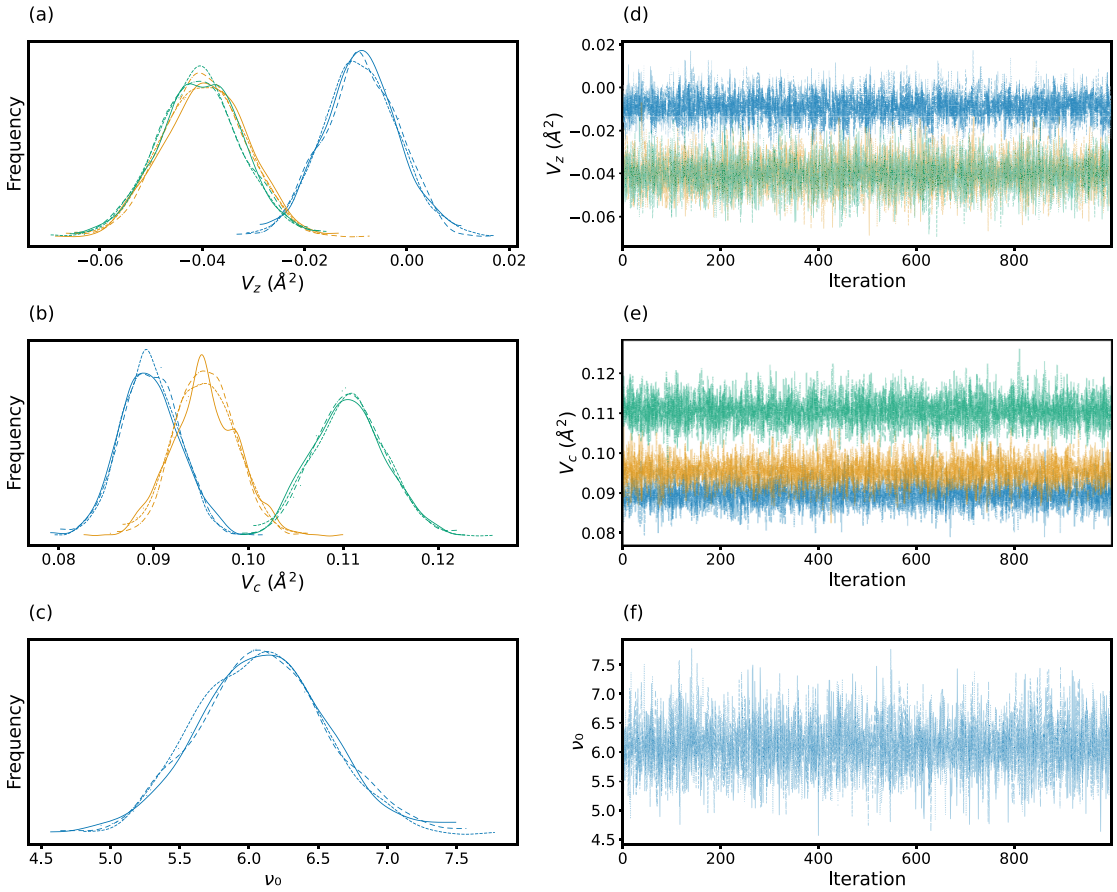

Figure 2. Posterior distribution of the model parameters after MCMC sampling. Kernel density estimates of parameters (a) V_z and (b) V_c and (c) ν, respectively. Parameter values of (d) V_z, (e) V_c and (f) ν_0 as a function of MCMC iterations. Four independent MCMC chains are plotted with different line styles.

atoms and their similarity can be predicted by our statistical random model. Table 1 compares distances at different significance levels, i.e. at which distance the pairwise distances derived from our random model reach 1% and 5% cumulative density. When protein atoms are compared, the probability of finding pairs with very short distances is much higher. For example, 1% cumulative density is reached at distance d_E of 0.052 in the random model, while 5% cumulative density of protein comparisons has already reached d_E of 0.034. For any of these 5% atom pairs we can reject the null hypothesis. By putting cumulative density into context, we can focus on pairwise comparisons to a single, generic atom from all other atoms in the structure. The crystal structure contains 1876 protein atoms. If we accept 1% false positive pairwise relationships, 27.7% (520 atoms) of the protein atoms are significantly more similar to a single atom than expected by chance. This number of atoms is much larger than the first and second coordination shell around a typical atom and a substantial fraction of similarities develops over a relatively long distance. Clearly, a protein is more inhomogeneous than this simple model suggests. While some atoms have almost no significantly similar partners, other atoms belong to an even larger network of significantly similar atoms.

The cumulative density of d_B at short distances is very similar when the experimental data and the random model are compared. It may be counterintuitive to realize that samples pulled from the same distribution do not have a maximum around zero for their pairwise distances and this is a characteristic of multivariate distributions. For the univariate B_{eq} distance (d_B) the most frequent pairwise distances are close to zero both for comparison of the random andexperimental data. This means that the high similarity of B_{eq} is not surprising and expected by chance, i.e. the similarity does not carry substantial information. The metric d_B is not good for detecting similarities. The distance distribution differences between the random and experimental model do diverge when distances are long, but in this study, we focus on the coincidental similarities not on coincidental differences.

3.2. Evaluation of ADP metrics and their correlation with the physical distance of atom pairs

Figure 7 shows the correlation between ADP metrics, and the distance between atom coordinates is also included (d_r). The diagonal plots represent the kernel density estimates of each distance distribution. The

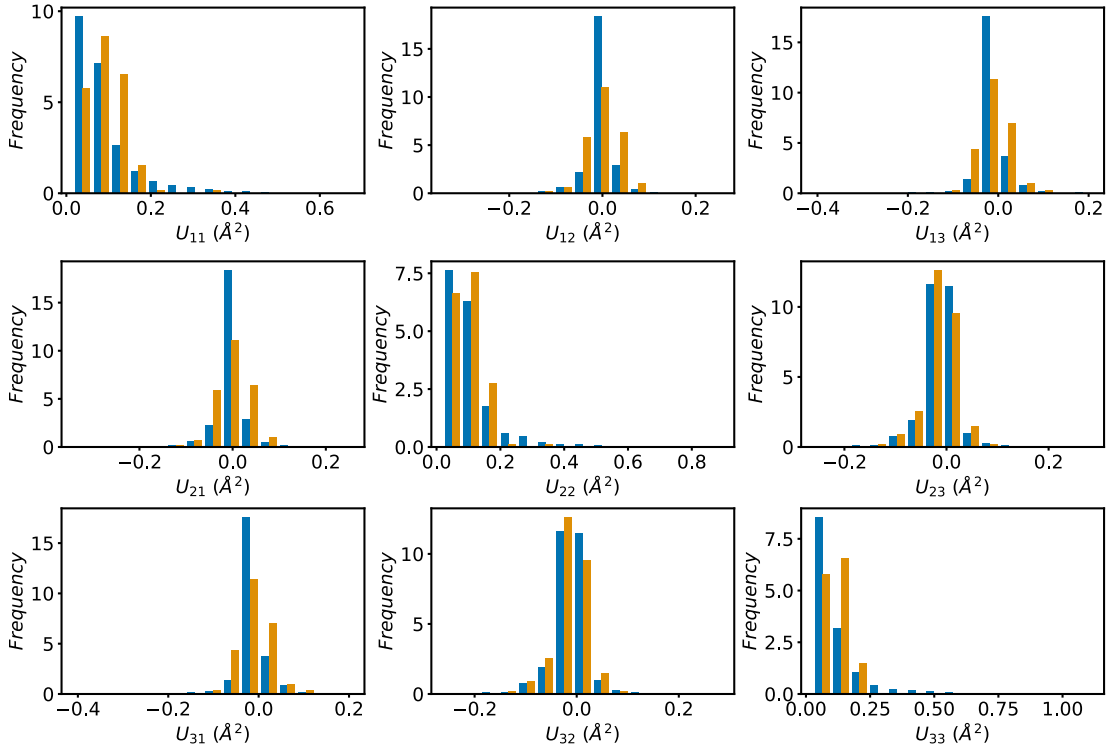

Figure 3. Comparison of observed (test set, blue) and predicted ADP elements (random model, orange).

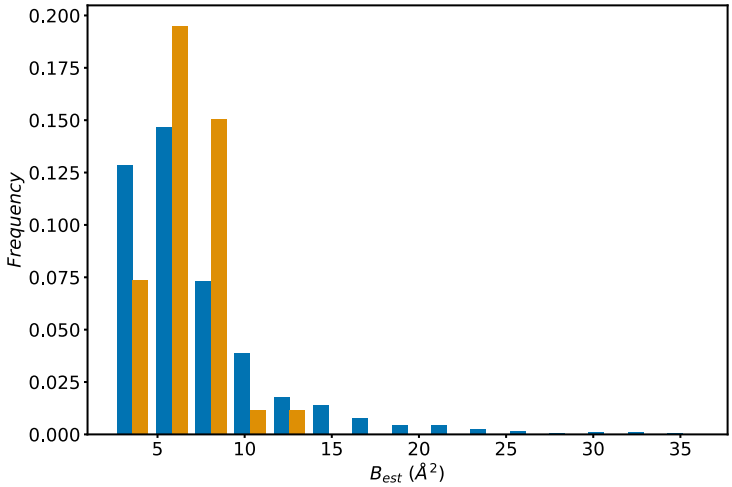

Figure 4. Histogram of B_{est} values of the observed (test set, blue) and predicted (random model, orange) ADPs.

d_{lD}, d_{lE}, d_{R} and d_{KL} metrics are highly correlated with one another. Therefore, only one of them, d_{R}, was retained in the final analysis. The kernel density estimate of d_{r} represents the pair distribution (p(r)) function, which provides information about the shape of the molecule. When d_{r} is plotted against different ADP metrics, it is apparent that atoms can be far apart and yet have nearly identical ADPs, and they are much more similar than expected by chance. The converse does not have to be true. When atoms are close together, their ADPs do have increased similarity. This can be observed due to the relative lack of points under the diagonal. There are almost no atom pairs within bonding distances (1–4 Å), which have highly dissimilar ADPs. This local similarity could originate from crystallographic restraints, rigid body motions and other types of local constraints. There is also a low frequency of short distances in the p(r) function. Notably, metrics that are less correlated with d_{B}, such as d_{R} show more diverse atom pairs at closer interatomic distances. B_{eq} is a pure measure of displacement amplitude without any directional information. This indicates that the local ADP similarity is mostly restricted to the displacement amplitudes rather than to

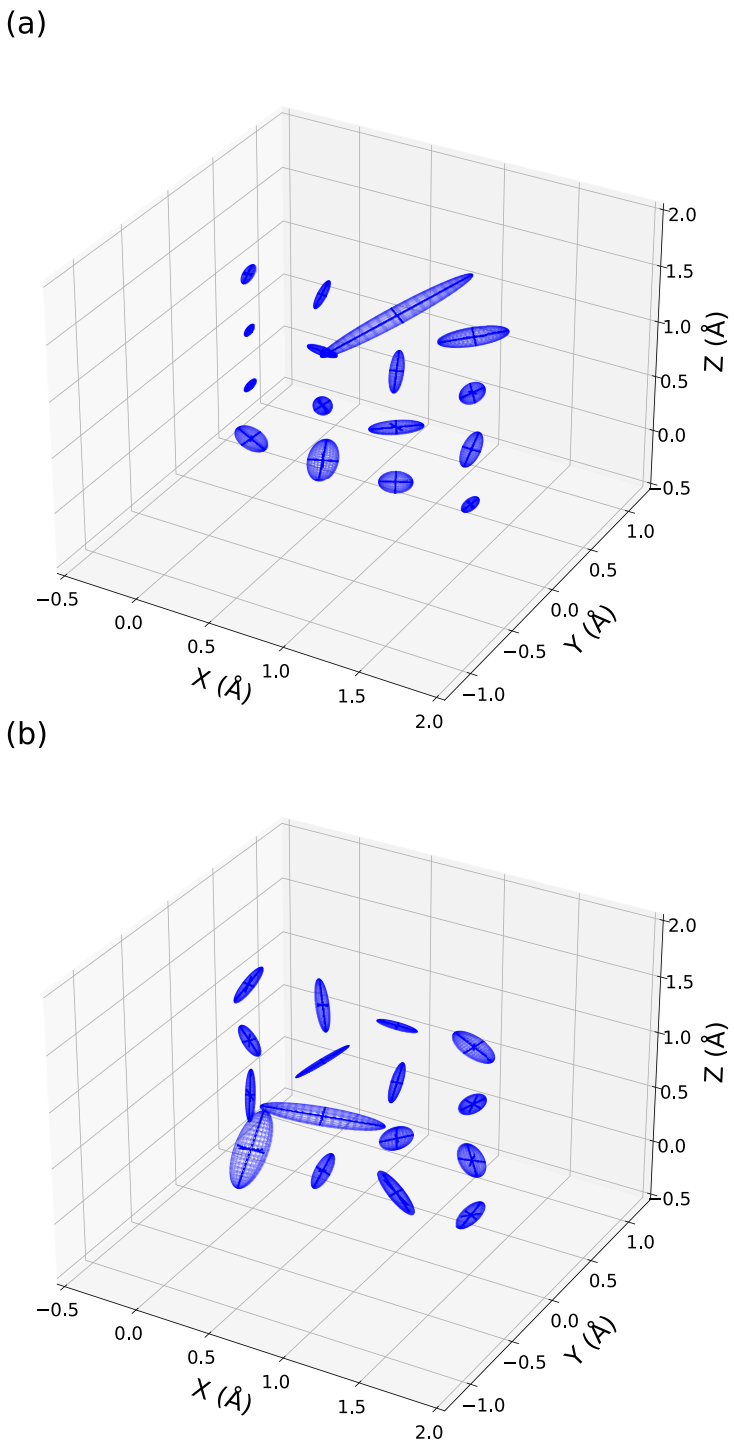

Figure 5. Comparison of 16 randomly selected (a) protein atoms from the test set and (b) 16 samples from the posterior predictions of the random model. ADPs are represented as ellipsoids where the axes are scaled by the corresponding eigenvalues of the matrices.

the displacement directions. This is in line with the general observation that the crystallographic B-factors tend to increase from the core towards the surface of proteins.

The isotropic displacement amplitude distance d_B and the metrics based on the shape parameters (eigenvalues) of $\mathbf{U}$ (d_λ) have poor correlation with other metrics, especially when the distances are short. This would normally be a favorable, complementary property, but figure 6 and table 1 show that short distances of d_λ and d_B occur easily by chance and do not carry substantial information.

3.3. Qualitative patterns of ADP similarities

Our earlier analysis identified the reproducible similarity of Cβ atoms in the amino acid residues His-57 and Ser-195 using a hierarchical clustering approach in trypsin in a complex with a small molecular inhibitor

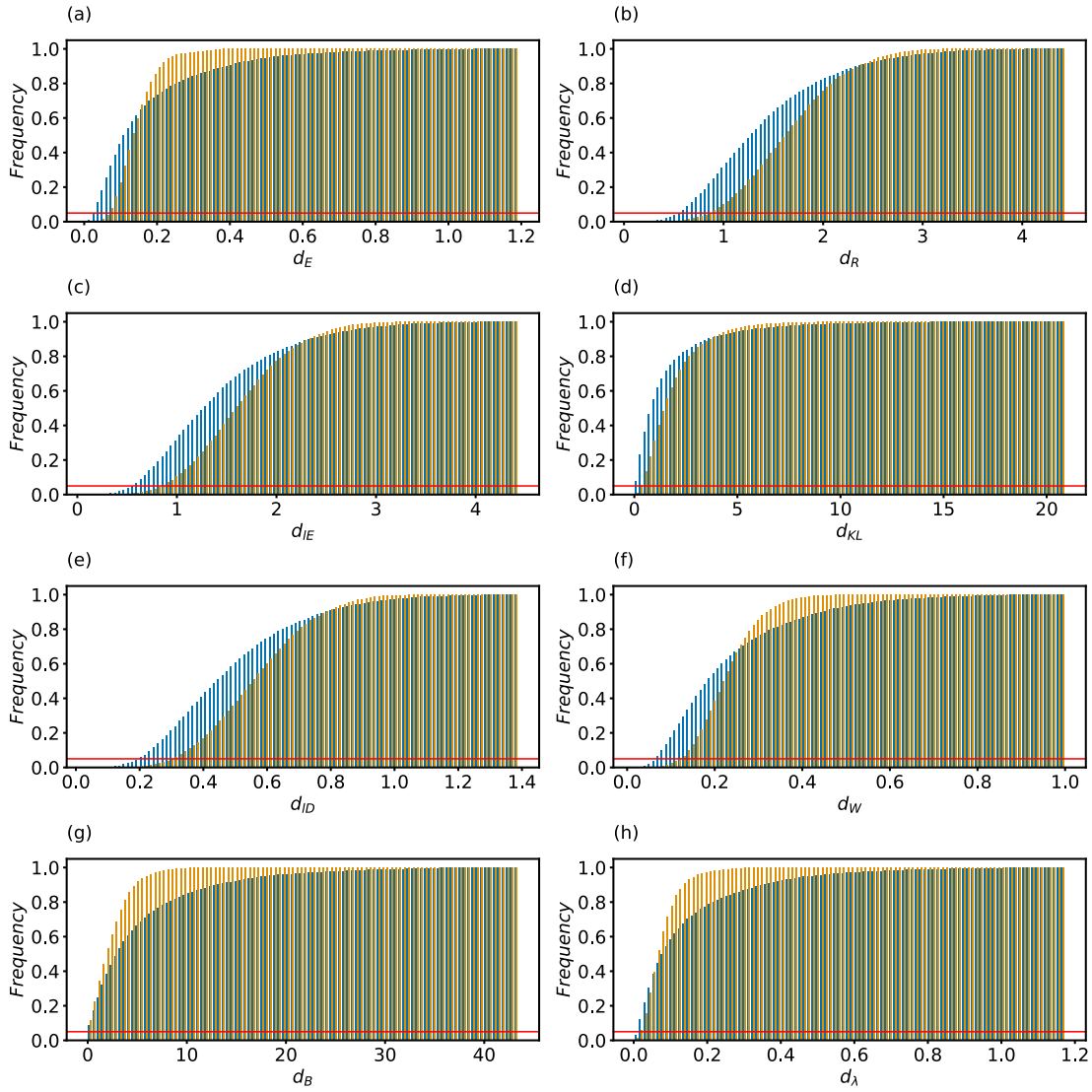

Figure 6. Cumulative normalized histogram of distances d_E, d_R, d_{IE}, d_{KL}, d_{ID}, d_W, d_B and d_λ (a)–(h) calculated by randomly selected pairs of atoms of the protein (blue) and two random samples of the posterior Wishart distribution (orange). Red line indicates 5% of the total observations.

Table 1. Statistical comparison of pairwise distances between ADPs of protein atoms and between posterior predicted ADPs. The abbreviation CDF indicates 'cumulative distribution function'.

Metric	Distance with CDF = 1% (random)	Distance with CDF = 5% (random)	Distance with CDF = 1% (protein)	Distance with CDF = 5% (protein)	Protein CDF at d with random CDF = 5% (%)	Protein CDF at d with random CDF = 1% (%)
d_E	0.052	0.073	0.022	0.034	27.7	15.1
d_R	0.639	0.882	0.403	0.583	21.7	7.1
d_{IE}	0.630	0.871	0.403	0.581	21.3	6.9
d_{KL}	0.210	0.405	0.082	0.173	21.6	7.2
d_W	0.091	0.125	0.046	0.071	25.8	11.8
d_{ID}	0.225	0.309	0.142	0.205	21.7	7.1
d_B	0.031	0.197	0.047	0.240	4.1	0.7
d_λ	0.013	0.023	0.009	0.017	10.1	2.4

benzamidine [7]. We can now verify the robustness of the claim in a different crystal structure, by looking at a different protein-inhibitor complex refined by different crystallographic software, and put the results into a statistical context. In our earlier analysis, we only used a variant of Euclidean distance and one type of clustering algorithm. Hierarchical clustering tends to commit an atom early to a cluster, making it unavailable for other plausible groupings. This is especially problematic when the ends of the branches

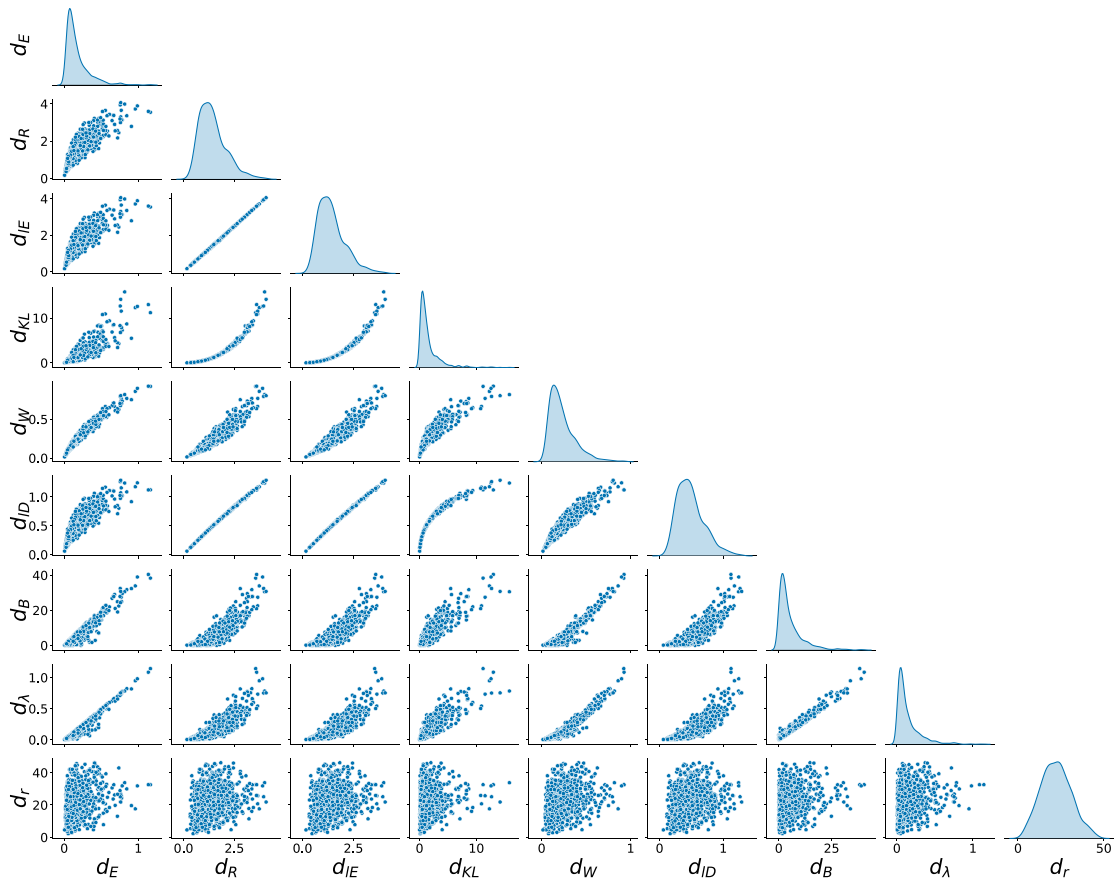

Figure 7. Comparison of distances and their kernel density estimates plotted by the Seaborn library. For clarity, only 1000 randomly selected atom pairs are visualized from the approximately two million combinations.

(small clusters) contain the most relevant information. Small changes in the metrics and algorithms used for clustering can result in a different tree. Thus, hierarchical clustering is excellent for capturing some patterns, but often misses links that are revealed in a more systematic analysis. In table 2, the similarities are ranked after increasing d_R. At this high level of similarity, d_E and d_W do not diverge strongly, but if the same atoms are sorted according to other metrics, the ranking changes slightly. Distances of these magnitudes are well below the 1% significance level (table 1). The metric d_λ and especially d_B correlate poorly with d_R, indicating that the displacement amplitudes alone do not contain sufficient information. The interatomic distance d_r varies greatly in this group of highly similar atoms to Cβ of His-57. At rank 1 and 2, we find atoms 16.2 and 24.9 Å apart, respectively. Only at rank 3, do we find a covalently bonded atom (Cα of His-57). All other similarities arise in the absence of covalent bonding and van der Waals interactions. We find Cβ of Ser-195 from our previous study [7] at rank 23 in the company of other carbon atoms, and in particular Cβ atoms. Recurring atoms are observed from the amino acid residues Ser-195, Ser-214, Lys-87 and Ile-89. Residue Ser-195 is the most important member of the catalytic triad as it performs the first nucleophilic attack on the substrate peptide or ester and forms covalent tetrahedral and acyl-enzyme intermediates during the catalytic cycle. Ser-214 is often attributed to the 'catalytic tetrad' [26] because its mutagenesis drastically hampers serine protease catalysis. Amide nitrogen of Ser-214 forms a hydrogen bond of variable strength to the P1 residue of the substrate [27]. Its side-chain hydroxyl group (Oγ) maintains an unconventional CH–O hydrogen bond to His-57; thus, Ser-214 maintains a link between the substrate and His-57 [28].

3.4. Origin of significant ADP similarities

We established a statistical framework for comparing ADPs and revealed that ADP similarities in bovine trypsin do not arise by coincidence. We have still not given a causal explanation. Unfortunately, we can only speculate about its origin if we consider the spatial distribution alone. We discuss here two kinematic processes: constrained displacements and coherence of motion. Both types of motion could result in identical positional distributions. We illustrate the difference with a theoretical experiment, where we assume that an atom is part of a planar chemical group. Suppose that this atom is constrained in its freedom. It cannot move in the plane direction and it displaces only in- and out-of-plane directions. Suppose that we

Table 2. The shortest ADP distances to Cβ of His-57. The color scale red-yellow-green illustrates the range between the lowest and highest values in the columns.

d_R rank	Atoms with ADP similar to Cβ of His-57	d_r	d_E	d_R	d_W	d_B	d_λ
1	106 ILE O	16.2	0.024	0.38	0.045	0.54	0.019
2	28 PRO Cβ	24.9	0.021	0.41	0.045	0.19	0.018
3	57 HIS Cα	1.5	0.024	0.43	0.047	0.60	0.021
4	195 SER Oγ	6.4	0.029	0.46	0.054	0.33	0.024
5	87 LYS Cα	16.8	0.025	0.46	0.052	0.36	0.014
6	214 SER Oγ	6.6	0.023	0.47	0.049	0.02	0.009
7	104 MET O	10.8	0.037	0.48	0.064	1.01	0.033
8	52 VAL O	13.9	0.038	0.48	0.066	1.07	0.033
9	51 TRP O	18.3	0.038	0.50	0.067	0.72	0.033
10	28 PRO Cα	23.5	0.036	0.50	0.064	0.55	0.028
11	87 LYS N	17.5	0.033	0.50	0.062	0.06	0.017
12	90 VAL N	10.5	0.034	0.50	0.063	0.23	0.013
13	87 LYS O	15.3	0.029	0.50	0.057	0.87	0.020
14	45 SER O	15.9	0.042	0.52	0.072	1.19	0.039
15	89 ILE Cβ	13.8	0.038	0.52	0.068	0.51	0.026
16	87 LYS C	15.3	0.032	0.52	0.061	0.19	0.013
17	100 ASN O	8.8	0.034	0.52	0.064	0.01	0.017
18	89 ILE Cα	12.5	0.039	0.52	0.068	0.95	0.034
19	84 SER C	19.0	0.038	0.53	0.069	0.38	0.026
20	121 ILE N	21.9	0.037	0.53	0.066	0.37	0.028
21	104 MET Cβ	8.4	0.033	0.53	0.063	0.53	0.029
22	214 SER Cβ	7.4	0.033	0.54	0.063	0.48	0.025
23	195 SER Cβ	6.7	0.038	0.54	0.069	1.13	0.033
24	89 ILE N	12.4	0.038	0.54	0.068	0.96	0.030
25	59 TYR Cδ1	9.5	0.033	0.55	0.062	0.60	0.019
26	50 GLN O	20.0	0.027	0.56	0.060	0.76	0.019
27	94 TYR Cγ	6.5	0.037	0.56	0.069	0.18	0.011
28	88 SER C	12.5	0.040	0.56	0.072	0.56	0.032
29	94 TYR Cε1	4.7	0.035	0.57	0.068	0.38	0.016
30	50 GLN C	20.4	0.033	0.57	0.066	0.02	0.027

have a number of identical chemical groups separated by long distances so that no direct contact between them is possible. If there is a yet unexplained process, which aligns the chemical groups parallel to one another, then the displacement of atoms will occur in similar directions. They do not have to move in phase; the constraints provided by the planar chemical group explain the uniform directional distribution. One problem with this model is that we do not have a universally accepted mechanism for protein folding, which explains why chemical groups should orient themselves in similar directions. This model also requires very strong constraints, for example, the out-of-plane displacement has to be perfectly perpendicular to the planar chemical group. Moreover, empirically we can observe that chemical groups are not perfectly aligned in the protein folded structure, and they appear to be affected by steric hindrances that prevent crystalline order. However, we cannot dismiss the presence of atoms with very similar displacements.

We prefer an alternative explanation because it does not require a speculative mechanism that orients the chemical groups over a long distance. Instead, we can assume that the discussed atom in the planar chemical group displaces coherently (in phase) with the same type of atoms in other groups, at least occasionally. Then, the planar constraints on the atoms would then work the other way around. The parallel orientation of the planar chemical groups becomes a consequence of the displacement alignment of a particular atom type.

4. Conclusion

Atomic displacement parameters of protein crystal structures contain useful information for machine learning analysis. We approximated ADPs with Wishart's distribution and used MCMC sampling of their joint posterior distribution. From the posterior distribution, we generated random ADP samples by forward prediction. We used the random samples to evaluate frequencies of coincidental similarities. To evaluate similarities, we tested eight metrics of which the Euclidean, Riemann and Wasserstein distances formed complementary, non-redundant sets of metrics sensitive to different aspects of ADPs. Comparisons of ADP

similarities with interatomic distances revealed poor correlation, indicating long-range constraints or coherence of protein dynamics. This study demonstrates that a substantial fraction of these long- and short-range ADP similarities does not form by coincidence. Finally, we showed further evidence that ADP similarities depend more on their chemical nature and bonding coordination.

Data availability statement

The data that support the findings of this study are openly available at DOIs. The implementation of the Bayesian model and data loading algorithm are available in the Zenodo depository, DOI: https://doi.org/10.5281/zenodo.4717293. The pairwise ADP distances of the test protein structure are available in the Zenodo depository, DOI: https://doi.org/10.5281/zenodo.4717613.

Acknowledgments

We would like to thank Christopher Hartl for helping to develop the Bayesian model. We are grateful to Ran Friedman and Stefano A Mezzasalma for their careful reading of the manuscript. This work was supported by the Röntgen-Ångström Cluster Framework (Grant No. 2015-06099). We acknowledge financial support from the LINXS—Lund Institute of Advanced Neutron and X-ray Science.

ORCID iD

Gergely Katona https://orcid.org/0000-0002-2031-8716

References

[1] Sharma A *et al* 2017 Asymmetry in serial femtosecond crystallography data *Acta Crystallogr.* A **73** 93–101
[2] Katona G, Garcia-Bonete M J and Lundholm I V 2016 Estimating the difference between structure-factor amplitudes using multivariate Bayesian inference *Acta Crystallogr.* A **72** 406–11
[3] Garcia-Bonete M J and Katona G 2019 Bayesian machine learning improves single-wavelength anomalous diffraction phasing *Acta Crystallogr.* A **75** 851–60
[4] Katona G *et al* 2016 Bayesian analysis of non-thermal structural changes induced by terahertz radiation in protein crystals *2016 41st Int. Conf. on Infrared, Millimeter, and Terahertz Waves (Irmmw-thz)* pp 1–2
[5] Foadi J *et al* 2013 Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography *Acta Crystallogr.* D **69** 1617–32
[6] Ortiz A R, Strauss C E M and Olmea O 2002 MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison *Protein Sci.* **11** 2606–21
[7] Gagnér V A *et al* 2019 Clustering of atomic displacement parameters in bovine trypsin reveals a distributed lattice of atoms with shared chemical properties *Sci. Rep.* **9** 19281
[8] Wishart J 1928 The generalised product moment distribution in samples from a normal multivariate population *Biometrika* **20A** 32–52
[9] Bartlett M S 1934 XX.—on the theory of statistical regression *Proc. R. Soc.* **53** 260–83
[10] Barnard J, McCulloch R and Meng X L 2000 Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage *Stat. Sin.* **10** 1281–311
[11] Lewandowski D, Kurowicka D and Joe H 2009 Generating random correlation matrices based on vines and extended onion method *J. Multivar. Anal.* **100** 1989–2001
[12] Merritt E A 1999 Comparing anisotropic displacement parameters in protein structures *Acta Crystallogr.* D **55** 1997–2004
[13] Zhou L and Liu Q 2014 Aligning experimental and theoretical anisotropic B-factors: water models, normal-mode analysis methods, and metrics. *J. Phys. Chem.* B **118** 4069–79
[14] Merritt E A 2011 Some B-eq are more equivalent than others *Acta Crystallogr.* A **67** 512–6
[15] Wahlgren W Y, Pál G, Kardos J, Porrogi P, Szenthe B, Patthy A, Gráf L and Katona G 2011 The catalytic aspartate is protonated in the Michaelis complex formed between trypsin and an *in vitro* evolved substrate-like inhibitor: a refined mechanism of serine protease action *J. Biol. Chem.* **286** 3587–96
[16] J Salvatier, Wiecki T V and Fonnesbeck C J 2016 Probabilistic programming in Python using PyMC3 *PeerJ Comput. Sci.* **2** e55
[17] Hoffman M D and Gelman A 2014 The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo *J. Mach. Learn. Res.* **15** 1593–623
[18] Olkin I and Pukelsheim F 1982 The distance between two random vectors with given dispersion matrices *Linear Algebra Appl.* **48** 257–63
[19] Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86
[20] McKinney W 2012 *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (Sebastopol, CA: O'Reilly Media, Inc.)
[21] Walt S V D, Colbert S C and Varoquaux G 2011 The NumPy array: a structure for efficient numerical computation *Comput. Sci. Eng.* **13** 22–30
[22] Virtanen P *et al* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72
[23] Congedo M, Barachant A and Bhatia R 2017 Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review *Brain-Comput. Interfaces* **4** 155–74
[24] Bisong E 2019 Matplotlib and Seaborn *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (Springer) pp 151–65

[25] Ellson J, Gansner E, Koutsofios L, North S C and Woodhull G 2001 Graphviz—open source graph drawing tools *Int. Symp. on Graph Drawing* (Springer) pp 483–4
[26] Krem M M and Di Cera E 2001 Molecular markers of serine protease evolution. *EMBO J.* **20** 3036–45
[27] Fodor K, Harmat V, Neutze R, Szilágyi L, Gráf L and Katona G 2006 Enzyme:substrate hydrogen bond shortening during the acylation phase of serine protease catalysis *Biochemistry* **45** 2114–21
[28] Derewenda Z S, Derewenda U, Kobos P M 1994 (His)C-epsilon-H…O=C hydrogen-bond in the active-sites of serine hydrolases *J. Mol. Biol.* **241** 83-93