MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **94** (2025) 97–134
ISSN: 0340–6253
doi: 10.46793/match.94-1.097A

A Complete and Bi-Continuous Invariant of Protein Backbones under Rigid Motion

Olga Anosova[a], Alexey Gorelov[b], William Jeffcott[a], Ziqiu Jiang[c], Vitaliy Kurlin[a,*]

[a] *Computer Science, University of Liverpool, Liverpool, L69 3BX, UK*

[b] *Université Grenoble Alpes, Institut Fourier, 38000 Grenoble, France*

[c] *Department of Surgery & Cancer, Faculty of Medicine, Imperial College London, London, W12 0NN, UK*

vitaliy.kurlin@liverpool.ac.uk

(Received November 12, 2024)

Abstract

Proteins are large biomolecules that regulate all living organisms and consist of one or several chains. The *primary* structure of a protein chain is a sequence of amino acid residues whose three main atoms (alpha-carbon, nitrogen, and carbonyl carbon) form a protein backbone. The *tertiary* structure is the rigid shape of a protein chain represented by atomic positions in 3-dimensional space. Because different geometric structures often have distinct functional properties, it is important to continuously quantify differences in rigid shapes of protein backbones. Unfortunately, many widely used similarities of proteins fail axioms of a distance metric and discontinuously change under tiny perturbations of atoms.

This paper develops a complete invariant that identifies any protein backbone in 3-dimensional space, uniquely under rigid motion. This invariant is Lipschitz bi-continuous in the sense that it changes up to a constant multiple of a maximum perturbation of atoms, and vice versa. The new invariant has been used to detect thousands of (near-)duplicates in the Protein Data Bank, whose presence inevitably skews machine learning predictions. The resulting invariant

[*] Corresponding author.

space allows low-dimensional maps with analytically defined coordinates that reveal substantial variability in the protein universe.

1 Motivations and the problem statement

A *protein* is a large biomolecule consisting of one or several chains of amino acid residues. The *primary structure* (*sequence*) of a protein chain is a string of residue labels (represented by one or three letters), each denoting one of (usually) 20 standard amino acids [37]. The *secondary* structure consists of frequent semi-rigid subchains such as α-helices and β-strands [31]. A sequence of a protein is easy to experimentally determine but important functional properties such as interactions with drug molecules depend on a 3-dimensional geometric shape (a *tertiary structure* or *fold*) represented by an embedding of all its atoms in $\mathbb{R}^3$ [45], see Fig. 1 (left).

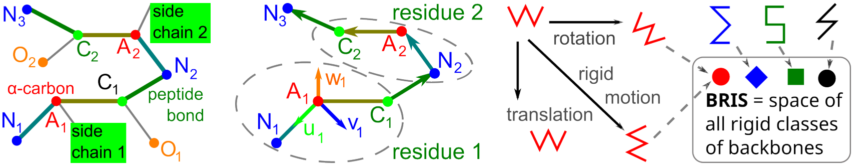


Figure 1. **Left**: all main atoms N_i, A_i, C_i of a protein chain form a *backbone* embedded in $\mathbb{R}^3$. **Middle**: each triangle $\triangle N_i A_i C_i$ defines an orthonormal basis $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i$. The coordinates of the bonds $\overrightarrow{C_i N_{i+1}}$, $\overrightarrow{N_{i+1} A_{i+1}}$, $\overrightarrow{N_{i+1} A_{i+1}}$ in this basis form the complete Backbone Rigid Invariant BRI. **Right**: All rigidly equivalent backbones form a single *rigid class*. All rigid classes of backbones form the *Backbone Rigid Space*.

In 1973, Nobel laureate Anfinsen conjectured that the sequence of any protein chain determines its 3D geometric shape [1]. Following this conjecture, neural networks such as AlphaFold2 and RosettaFold [3, 20, 34, 47] optimize millions of parameters to predict a protein fold from its sequence but need re-training [19] on the growing number of experimental structures in the Protein Data Bank (PDB) [6]. The reported accuracies of prediction are based on the LDDT (Local Distance Difference Test) [32, p. 2728] and TM-score [55], which fail the metric axioms. Then clustering algorithms can produce pre-determined clusters and may not be trustworthy [41].

Backbones of the same length (number of residues) can be optimally aligned to minimize the Root Mean Square Deviation (RMSD) between corresponding atoms [13]. This RMSD is slow to compute for all pairs of proteins and gives only distances without mapping the protein universe.

We develop a different approach by mapping the space of protein backbones in analytically defined coordinates similar to geographic-style maps of a new planet. The first question that we should ask about any real data such as protein tertiary structures is "same or different" [43].

Any embedded protein in $\mathbb{R}^3$ can be rigidly moved (translated or rotated), which changes all atomic coordinates but the underlying structure remains the same in the sense that different images of a protein under rigid motion have the same properties in a fixed environment. Though proteins are flexible molecules, it is important to distinguish their rigid shapes that can differently interact [14] with other molecules including medical drugs.

Definition 1.1 (Backbone Rigid Space BRIS_m)**.** A protein *backbone* is a sequence of m ordered triplets of main chain atoms (nitrogen N_i, α-carbon A_i, and carbonyl carbon C_i) given by their geometric positions in $\mathbb{R}^3$. A rigid *motion* is a composition of translations and rotations matching backbones in $\mathbb{R}^3$ (denoted by $S \cong Q$). The classes of all backbones of m triplets under rigid motion form the *Backbone Rigid Space* BRIS_m.

Rigid classes of backbones can be distinguished only by an *invariant* I defined as a descriptor preserved under any rigid motion. Any non-invariant descriptor J always has a *false negative* pair of backbones $S \cong Q$ with $J(S) \neq J(Q)$. The number of residues is invariant, while the center of mass moves together with a backbone and is not invariant.

Backbones were studied by incomplete invariants such as torsion angles, which allow *false positive* pairs of non-equivalent backbones $S \not\cong Q$ with $I(S) = I(Q)$. Because all atoms in a backbone S are ordered, their distance matrix determines $S \subset \mathbb{R}^3$ up to *isometry* (any distance-preserving transformation), but is large in size (quadratic, $O(m^2)$) and fails to distinguish mirror images. Adding a sign of orientation creates discontinuity for backbones that are almost (not exactly) mirror-symmetric.

Problem 1.2 formalizes the practically important conditions that were not all previously proved for earlier descriptors of proteins, see section 2.

Problem 1.2 (mapping the Backbone Rigid Space)**.** For any $m \geq 1$, design a map $I : \mathrm{BRIS}_m \to \mathbb{R}^k$ for some k satisfying the conditions below.

(a) **Completeness**: any backbones $S, Q \subset \mathbb{R}^3$ are rigidly equivalent if and only if $I(S) = I(Q)$, i.e. I has *no false negatives* and *no false positives.*

(b) **Reconstruction**: any protein backbone $S \subset \mathbb{R}^3$ can be reconstructed from its invariant value $I(S)$ uniquely under rigid motion.

(c) **Lipschitz continuity**: there is a distance d satisfying the metric axioms (1) $d(a, b) = 0$ if and only if $a = b$; (2) $d(a, b) = d(b, a)$; (3) triangle inequality $d(a, b) + d(b, c) \geq d(a, c)$ for all invariant values a, b, c; and a constant λ such that, for any $\varepsilon > 0$, if Q is obtained from S by perturbing every atom up to Euclidean distance ε, then $d(I(S), I(Q)) \leq \lambda\varepsilon$.

(d) **Atom matching**: there is a constant μ such that, for any backbones S, Q with $\delta = d(I(S), I(Q))$, all their atoms can be matched up to a distance $\mu\delta$ by a rigid motion.

(e) **Respecting subchains**: for any subchain of residues $R_i \cup \cdots \cup R_{i+j}$ in a backbone S, the invariant $I(R_i \cup \cdots \cup R_{i+j})$ can be obtained from $I(S)$ in linear time $O(j)$ with respect to the length of the subchain.

(f) **Linear time**: the invariant I, the metric d, a reconstruction in (b), and a rigid motion in (d) can be computed in time $O(m)$ for m residues.

The completeness in 1.2(a) means that I is the strongest possible invariant and hence *distinguishes all* protein backbones that cannot be exactly matched by rigid motion. The reconstruction in 1.2(b) is more practical because I may not allow an efficiently computable inverse map I^{-1} from an invariant value $I(S)$ to a backbone $S \subset \mathbb{R}^3$. The metric axioms for a distance d in 1.2(c) are essential because if the triangle axiom fails with any positive error, results of clustering can be made arbitrary [41].

The continuity in 1.2(c) fails for invariants based on principal directions that can discontinuously change (or become ill-defined) in degenerate cases when eigenvalues become equal. The atom matching in 1.2(d) says that, after finding a rigid motion f in $\mathbb{R}^3$, any atom $p \in S$ (say, α-carbon $A_i(S)$ in the i-th residue) has Euclidean distance at most $\mu\delta$ to the corresponding atom $q \in f(Q)$, also the α-carbon atom $A_i(Q)$ in the i-th residue of Q.

Conditions 1.2(c,d) guarantee the Lipschitz continuity of I and its inverse on the image $I(\mathrm{BRIS}_m) \subset \mathbb{R}^k$. New condition 1.2(e) is important for identifying secondary structures, which are subchains in full backbones.

The linear time in 1.2(f) makes all previous conditions practically useful because even the distance matrix needs $O(m^2)$ time and space, substantially slower than linear time $O(m)$ for thousands of residues.

The key contribution is the *Backbone Rigid Invariant* BRI, a map $\mathrm{BRIS}_m \to \mathbb{R}^{9m-6}$ that solves Problem 1.2. Conditions 1.2(d,e) are stated for the first time to the best of our knowledge. Section 6 will describe how BRI detected thousands of unexpected geometric duplicates in the PDB, some of which require corrections, already confirmed by their authors.

The numerical components of BRI play the role of geographic-style coordinates on the space BRIS_m, where any protein chain has a uniquely defined location. Sections 3 and 5 will discuss 2D projections of the full Backbone Rigid Space $\mathrm{BRIS} = \bigcup\limits_{m \geq 2} \mathrm{BRIS}_m$ and reveal substantial variability of traditional invariants in the PDB such as bond angles and lengths, which were previously expected to have fixed values for all proteins.

2 Past work on similarities of proteins

In the more general context of crystal structures, a canonical description in a reduced unit cell [38] can be achieved by the program TYPIX [39] for inorganic compounds and ACHESYM [24] for macromolecular crystals. Such conventional settings can be considered a complete invariant in the sense of condition (1.2a). However, a reduced cell discontinuously changes under almost any perturbation of atoms, which has been known at least since 1965 [29, p. 80] and was resolved only for generic crystals [49].

The majority of past approaches to quantify protein similarity use a geometric alignment by finding an optimal rigid motion that makes a given structure as close as possible to a template structure.

The widely used TM-score [55] TM $= \max\left\{\frac{1}{L_N}\sum\limits_{i=1}^{L_T}\frac{1}{1+(d_i/d_0)^2}\right\} \in [0,1]$ is maximized over all spatial alignments of two backbones, where $\frac{d_i}{d_0}$ is a normalized distance between aligned C_α atoms, L_T is the length of the template structure, L_N is the length of a given structure. Since any identical proteins (with all equal x, y, z coordinates) have TM-score 1, the simplest way to convert this similarity into a distance is to set TMD $= 1 -$ TM so that TMD$(S,S) = 0$ for any structure S. Unfortunately, this and many other conversions such as $-\log(\text{TM})$ fail the triangle inequality of a metric already for 3 atoms. Indeed, if $L_T = L_N = 1$ and d_i/d_0 are pairwise distances $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ between 3 atoms, which satisfy the triangle axiom, then $d = 1 - \text{TM}$ takes the values $\frac{1}{5}, \frac{1}{10}, \frac{1}{17}$, which fail this axiom, see 1.2(c), also for the (approximate) values $0.22, 0.11, 0.06$ of $d = -\log(\text{TM})$.

If the triangle axiom fails with any additive error, results of the clustering algorithms k-means and DBSCAN can be arbitrarily pre-determined [41]. The authors of another similarity LDDT (Local Distance Difference Test) concluded in [32, p. 2728] that “One disadvantage of the LDDT score is that it does not fulfill the mathematical criteria to be a metric. However, the same is true for most scores”. One metric satisfying all axioms is the Root Mean Square Deviation (RMSD) between optimally aligned ordered atoms [7]. This RMSD is slow to compute for all-vs-all comparisons in the PDB. As a result, many pairs with RMSD $= 0$ remained unnoticed.

If the order of atoms is ignored, the optimal alignment is NP-complete (provably too slow) [28]. Applying random rotations [10] creates many more structures that look different but should be considered rigidly equivalent. This ‘data augmentation’ makes the classification even harder.

The PDB recently implemented a structural superposition [11] of protein backbones by computing the score equal to the sum of absolute values in the upper triangle of the distance-difference matrix (DDM) for the distance matrices between all α-carbon atoms C_α. The description in [11] adds that "to account for possible gaps in the DDMs, caused by a lack of residue coordinates, these scores are multiplied by a scalar between 0-1, where 1 represents the absence of any gaps ... low scores represent chains with high structural similarity." This scaling by values less than 1 likely affects the triangle axiom, which needs checking in the light of the recent reviews [19, 35, 46] of protein folding prediction [20, 30, 34].

More importantly, to efficiently navigate in the protein universe, in addition to distances, we need a map showing all known structures and also under-explored regions, where new proteins can be discovered. Such a geographic-style map needs a complete invertible and bi-continuous invariant I like the pair of latitude and longitude coordinates on Earth.

Protein backbones are traditionally represented by *torsion* (dihedral) angles φ_i, ψ_i visualized in Ramachandran plots [40]. For a general polygonal line on points $S \subset \mathbb{R}^3$, the sequence $\{\phi_i, \psi_i\}$ is invariant under rigid motion but incomplete. Indeed, for any successive points $p_i, p_{i+1} \in S$, we can shift all points $p_{i+1}, \dots, p_m$ by a vector $t(\vec{p}_{i+1} - \vec{p}_i)$ for any $t \neq 0$, which changes the overall rigid shape of S but keeps all relative angles between any straight segments and planes through successive points.

For protein backbones, even if all bond lengths and angles are fixed at ideal values, all torsion angles still should be ordered according to given residues to completely determine the rigid class of a backbone. Even if we keep all torsion angles in order, three invariants per residue cannot uniquely determine a rigid backbone having 3 atoms with 9 coordinates per residue in $\mathbb{R}^3$. AlphaFold2 [20] used 6 parameters per residue to define a rigid transformation on every i-th triplet (*residue triangle*) on the main atoms N_i, A_i, C_i to the next $(i+1)$-st residue triangle. However, the analysis in section 3 will show that rigid shapes of residue triangles substantially vary across the PDB. Our paper strengthens the past approach by defining 9 invariants per each of m residues, which gives $9m-6$ invariants in total after subtracting 6 parameters of a global rigid motion in $\mathbb{R}^3$.

If we consider a backbone S of $3m$ ordered atoms modulo isometry including reflections, the easier complete invariant known since 1935 [44] is the $3m \times 3m$ matrix $D(S)$ of all pairwise distances whose entry $D_{ij}(S)$ is the

Euclidean distance between the i-th and j-th points of S. Any backbone S can be reconstructed from $D(S)$ or, equivalently, from the Gram matrix of scalar products as in [9, Theorem 1], uniquely up to isometry in $\mathbb{R}^3$. The matrix $D(S)$ satisfies almost all conditions of Problem 1.2 apart from the linear time/size requirement, which is essential for large proteins.

If a protein is considered a cloud of unordered atoms (ignoring the order along a backbone), such clouds of different sizes can be visualized by eigenvalue invariants (or moments of inertia) characterizing the elongation of the cloud along its principal directions. In 1996, probably the first map of all 4K entries in the PDB appeared in [16, Fig. 5] based on the two largest eigenvalues, see the recent updates in [53, Fig. 2] and PDB-Explorer [18].

In 2020, Holm called for faster visualization of the protein space [17]: "It would be nice to restore the ability to move a lens across fold space in real-time ... this ability was based on pre-computed all-against-all structural similarities, which is not manageable with current data".

In 1977, Kendall [22] started to study configuration spaces of ordered points modulo rigid motion in $\mathbb{R}^n$ under the name of *size-and-shape spaces* [23]. If we consider sequences equivalent also under uniform scaling, the smaller *shape space* Σ_2^m of m ordered points in $\mathbb{R}^2$ can be described as a complex projective space $\mathbb{C}P^{m-1}$ due to the group SO(2) being identified with the unit circle in the complex space $\mathbb{C}^1 = \mathbb{R}^2$. However, there is no easy description of the space Σ_3^m of m-point sequences in $\mathbb{R}^3$, which has no multiplicative group structure similar to $\mathbb{R}^2 = \mathbb{C}^1$. This algebraic obstacle prevented a simple solution to Problem 1.2 in dimension $n = 3$.

3 The backbone rigid invariant (BRI)

We start with the simpler *triangular invariant* that describes the rigid shape of each residue triangle $\triangle N_i A_i C_i$ on three main atoms per each of m residues: nitrogen N_i, α-carbon A_i, and carbonyl carbon C_i, for $i = 1, \dots, m$, see Fig. 1 (middle). For any points $A, B \in \mathbb{R}^3$, let $|\overrightarrow{AB}|$ be the Euclidean length of the vector $\overrightarrow{AB}$ from A to B. We denote vectors by $\boldsymbol{u} \in \mathbb{R}^3$, their *scalar* and *vector* products by $\boldsymbol{u} \cdot \boldsymbol{v}$ and $\boldsymbol{u} \times \boldsymbol{v}$, respectively.

Definition 3.1 (triangular invariant TRIN)**.** Let a backbone $S \subset \mathbb{R}^3$ have $3m$ ordered atoms N_i, A_i, C_i, $i = 1, \dots, m$. In the plane of $\triangle N_i A_i C_i$, for the 2D basis obtained by Gaussian orthogonalization of $\overrightarrow{A_iN_i}, \overrightarrow{A_iC_i}$, the vector $\overrightarrow{A_iN_i}$ has the coordinates $x(A_iN_i) = |\overrightarrow{A_iN_i}|$, $y(A_iN_i) = 0$, while $\overrightarrow{A_iC_i}$ has $x(A_iC_i) = \dfrac{\overrightarrow{A_iC_i} \cdot \overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$ and $y(A_iC_i) = \left|\overrightarrow{A_iC_i} - x(A_iC_i)\dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}\right|$. The *triangular invariant* TRIN(S) is the $m \times 3$ matrix whose i-th row consists of the coordinates $x(A_iN_i), x(A_iC_i), y(A_iC_i)$ for $i = 1, \dots, m$.

The i-th row of TRIN(S) uniquely determines the shape of $\triangle N_i A_i C_i$. Many past approaches including AlphaFold2 [20] assumed that all these residue triangles are rigidly equivalent. To test this assumption on the PDB, we filter out unsuitable chains as follows. On May 4, 2024, the PDB had 213,191 entries with 1,091,420 chains. Protocol 3.2 below produced $104,688 \approx 49\%$ entries with $707410 \approx 65\%$ chains in 4 hours 48 min 11 sec. All experiments were run on CPU Core i7-11700 @2.50GHz RAM 32Gb.

Protocol 3.2 (selecting a subset of 707K+ chains in the PDB)**.** The PDB was filtered by removing the following entries and individual chains.
(1) 4513 non-proteins (the entity is labeled as 'not a protein').
(2) 178153 disordered chains whose some atoms have occupancies < 1.
(3) 201648 chains with residues having non-consecutive indices.
(4) 9941 incomplete chains missing one of the main atoms N_i, A_i, C_i.
(5) 4364 chains with non-standard amino acids.

Example 3.3 (variability of residue triangles)**.** Fig. 2 (row 1) shows the heatmaps of the invariants $x(A_iN_i), x(A_iC_i), y(A_iC_i)$ on a logarithmic scale from Definition 3.1 across all 110+ million residues from the 707K+ cleaned backbones obtained by Protocol 3.2. Though standard deviations of these invariants are about 0.01Å, the maximum deviations of $x(A_iN_i), x(A_iC_i), y(A_iC_i)$ have high values of $1.2, 1.7, 2.7$Å, respectively.

Table 1 below shows the coordinates of TRIN and BRI (see Definition 3.4) for the two hemoglobin chains A in proteins 2hhb and 1hho, which are shown in Fig. 3 (top middle) and discussed in Example 5.2.

To guarantee new condition 1.2(e) respecting subchains, Definition 3.4 will represent atoms $N_{i+1}, A_{i+1}, C_{i+1}$ in a basis of the previous i-th residue.

Table 1. Coordinates of TRIN and BRI for the first 3 residues of the chains A in 2hhb (top) and 1hho (bottom) with their means.

Res	$x(AN)$	$x(AC)$	$y(AC)$	$x(N)$	$y(N)$	$z(N)$	$x(A)$	$y(A)$	$z(A)$	$x(C)$	$y(C)$	$z(C)$
VAL	1.45	-0.54	1.44	1.45	0	0	0	0	0	-0.54	1.44	0
LEU	1.47	-0.50	1.47	-0.91	0.25	-0.90	-0.64	1.32	0.02	-1.10	0.01	1.10
SER	1.47	-0.48	1.45	-0.77	0.36	-0.98	-0.66	1.31	-0.05	-1.11	0.02	1.06
mean	1.47	-0.55	1.43	0.52	0.84	0.46	-0.48	1.38	0.05	0.01	0.65	-1.01

Res	$x(AN)$	$x(AC)$	$y(AC)$	$x(N)$	$y(N)$	$z(N)$	$x(A)$	$y(A)$	$z(A)$	$x(C)$	$y(C)$	$z(C)$
VAL	1.48	-0.51	1.46	1.48	0.00	0.00	0.00	0.00	0.00	-0.51	1.46	0.00
LEU	1.49	-0.55	1.42	-0.14	0.66	1.16	-0.69	1.31	0.19	-1.51	-0.16	-0.03
SER	1.44	-0.41	1.44	-0.63	0.27	-1.10	-0.36	1.36	-0.30	-1.43	0.14	0.40
mean	1.47	-0.53	1.43	0.56	0.81	0.44	-0.43	1.38	0.06	0.04	0.65	-1.02

The first residue needs only three invariants from Definition 3.1 to determine the rigid shape of $\triangle N_1 A_1 C_1$ in $\mathbb{R}^3$. Due to cleaning in Protocol 3.2, all consecutive atoms along any backbone have distances $d \geq 0.01$Å and all angles in any residue triangle $\triangle N_i A_i C_i$ are at least 3°, which makes the bases of all residue triangles well-defined in Definition 3.4 below.

Definition 3.4 (backbone rigid invariant BRI(S) of a protein backbone S)**.** In the notations of Definition 3.1, define the orthonormal basis vectors $\boldsymbol{u}_i = \dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$, $\boldsymbol{v}_i = \dfrac{\boldsymbol{h}_i}{|\boldsymbol{h}_i|}$ for $\boldsymbol{h}_i = \overrightarrow{A_iC_i} - b_i\overrightarrow{A_iN_i}$, $b_i = \dfrac{\overrightarrow{A_iC_i} \cdot \overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|^2}$, and $\boldsymbol{w}_i = \boldsymbol{u}_i \times \boldsymbol{v}_i$. The *backbone rigid invariant* BRI(S) is the $m \times 9$ matrix whose i-th row for $i = 2, \dots, m$ contains the coefficients x, y, z of the vectors $\overrightarrow{C_{i-1}N_i}$, $\overrightarrow{N_iA_i}$, $\overrightarrow{A_iC_i}$ in the basis $\boldsymbol{u}_{i-1}, \boldsymbol{v}_{i-1}, \boldsymbol{w}_{i-1}$. So the nine columns of BRI(S) contain the coordinates $x(N_i), y(N_i), z(N_i)$ of $\overrightarrow{C_{i-1}N_i}$, followed by the six coordinates $x(A_i), \dots, z(C_i)$. For $i = 1$, the first row of BRI(S) has only three non-zero coordinates $x(N_1) = x(A_1N_1)$, $x(C_1) = x(A_1C_1)$, $y(C_1) = y(A_1C_1)$ from the first row of TRIN(S) in Definition 3.1.

Fig. 2 shows heatmaps of the PDB cleaned by Protocol 3.2. We mapped each of 110+ million residues across all 707+ thousand chains to a pair of coordinates (x, y) from the invariants TRIN and BRI. When many points (x, y) were discretized to a single pixel, its color reflects the number of such points on a logarithmic scale in the color bars of all heatmaps.

For a backbone of m residues, the first row of the $m \times 9$ matrix BRI(S) contains only three non-zero coordinates. Hence the matrix BRI(S) can

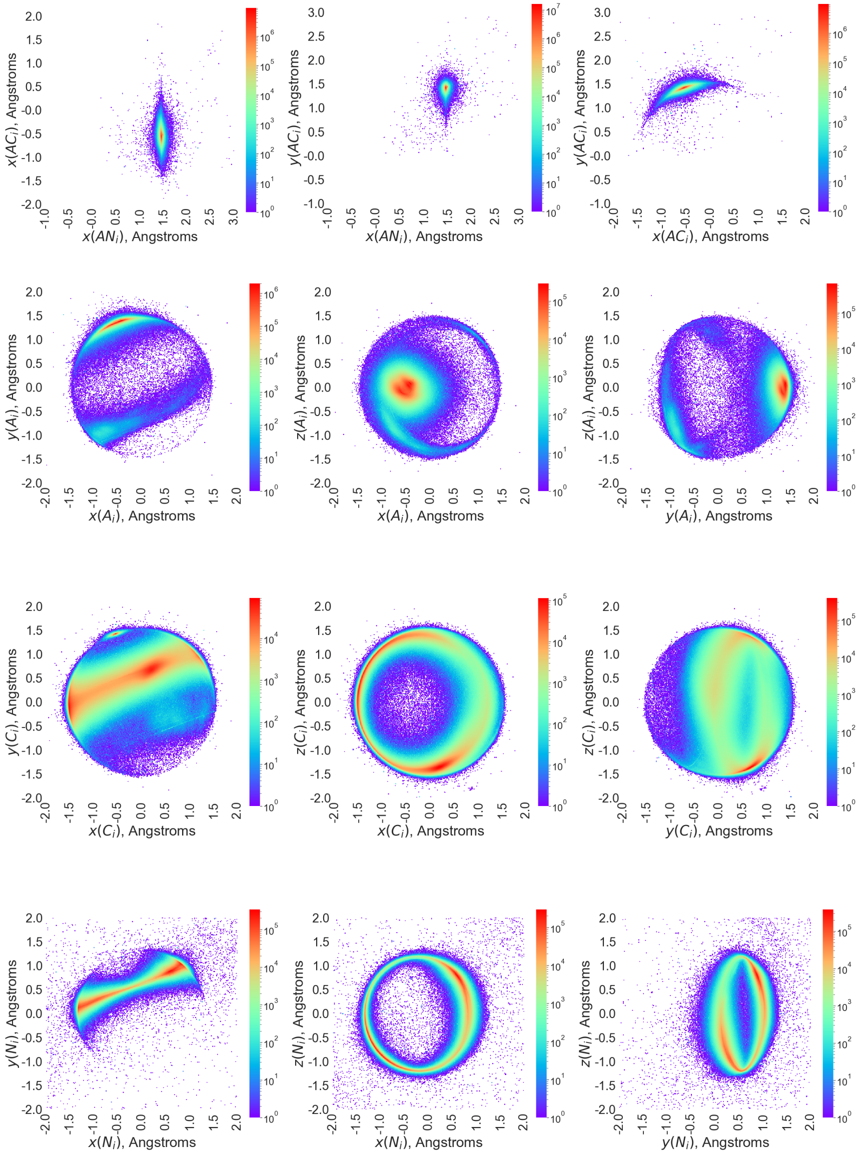


Figure 2. Heatmaps of the invariants TRIN and BRI from Definitions 3.1 and 3.4 for all 110M+ residues in the 707K+ chains obtained by Protocol 3.2. The color indicates the number of residues whose pair of invariants is discretized to each pixel.

be considered a vector of length $9(m-1)+3=9m-6$. The simplest metric on backbone rigid invariants as vectors in $\mathbb{R}^{9m-6}$ is L_∞ equal to the maximum absolute difference between all corresponding coordinates.

A small value δ of $L_\infty(\mathrm{BRI}(S),\mathrm{BRI}(Q))$ guarantees by Theorem 4.8 that backbones S,Q are closely matched by rigid motion. Another metric such as Euclidean distance or its normalization by the chain length has no such guarantees and can be small even for a few outliers that can affect the rigid shape and hence functional properties of a protein. Theorem 3.5 proves conditions 1.2(a,b,c,e,f) in Problem 1.2 for the invariant $\mathrm{BRI}(S)$.

Theorem 3.5 (completeness, reconstruction, and subchains)**. (a)** Under any rigid motion in $\mathbb{R}^3$, the matrix $\mathrm{TRIN}(S)$ in Definition 3.1 is invariant, $\mathrm{BRI}(S)$ in Definition 3.4 is a complete invariant, so any backbones $S,Q\subset\mathbb{R}^3$ are matched by rigid motion if and only if $\mathrm{BRI}(S)=\mathrm{BRI}(Q)$.

(b) The invariant $\mathrm{BRI}(S)$, metric L_∞ between invariants, and a reconstruction of $S\subset\mathbb{R}^3$ from $\mathrm{BRI}(S)$ can be computed in time $O(m)$.

(c) Let Q be a subchain of j consecutive residues in a backbone $S\subset\mathbb{R}^3$. If Q includes the first residue of S, then $\mathrm{BRI}(Q)$ consists of the first j rows of $\mathrm{BRI}(S)$. If Q starts from the i-th residue of S for $i>1$, the rows $2,\dots,j$ of $\mathrm{BRI}(Q)$ coincide with the rows $i+1,\dots,i+j-1$ of $\mathrm{BRI}(S)$, and the 1st row of $\mathrm{BRI}(Q)$ is computed from the i-th row of $\mathrm{BRI}(S)$ in a constant time. Hence $\mathrm{BRI}(Q)$ is computed from $\mathrm{BRI}(S)$ in time $O(j)$.

Proof of Theorem 3.5. **(a,b)** The formulae of the basis vectors in Definition 3.4 guarantee that all vectors have unit length $|\boldsymbol{u}_i|=|\boldsymbol{v}_i|=|\boldsymbol{w}_i|=1$ and are orthogonal to each other due to $\boldsymbol{u}_i\cdot\boldsymbol{v}_i=\boldsymbol{v}_i\cdot\boldsymbol{w}_i=\boldsymbol{w}_i\cdot\boldsymbol{u}_i=0$. Any rigid motion f acting on a backbone $S\subset\mathbb{R}^3$ has the form $f(p)=\vec{v}+R(\vec{p})$ for a fixed vector $\vec{v}\in\mathbb{R}^n$, an orthogonal map $R\in\mathrm{O}(\mathbb{R}^3)$, and any $p\in\mathbb{R}^3$. Then f maps every orthonormal basis $\boldsymbol{u}_i,\boldsymbol{v}_i,\boldsymbol{w}_i$ with the origin at a point $A_i\in\mathbb{R}^3$ to another orthonormal basis $R(\boldsymbol{u}_i)$, $R(\boldsymbol{v}_i)$, $R(\boldsymbol{w}_i)$ at the new origin $f(A_i)$. Hence the image of any vector $\overrightarrow{A_iP_i}=x\boldsymbol{u}_i+y\boldsymbol{v}_i+z\boldsymbol{w}_i$ under f has the same coordinates in the rigidly transformed basis: $f(\overrightarrow{A_iP_i})=R(\overrightarrow{A_iP_i})=xR(\boldsymbol{u}_i)+yR(\boldsymbol{v}_i)+zR(\boldsymbol{w}_i)$, so $\mathrm{BRI}(S)=\mathrm{BRI}(f(S))$.

For any residue having a fixed index i, Definition 3.4 needs only a constant time $O(1)$ to compute the basis vectors and coordinates of $\overrightarrow{C_{i-1}N_i}$ in the basis of the previous residue. The total time for computing the $m\times 9$ matrix $\mathrm{BRI}(S)$ is $O(m)$. The metric L_∞ has a linear time in the size $9m$.

The completeness will follow by showing that any backbone $S\subset\mathbb{R}^3$ can be efficiently reconstructed from $\mathrm{BRI}(S)$, uniquely after fixing the first residue whose shape is determined by the three non-zero values in the first row of $\mathrm{BRI}(S)$. In the first residue, the α-carbon A_1 can be moved to the origin $0\in\mathbb{R}^3$ by translation. Using $x(N_1)=|\overrightarrow{A_1N_1}|$, the N-terminal atom N_1 can be fixed in the positive x-axis by an orthogonal map from $\mathrm{SO}(3)$. A suitable rotation around the x-axis can move C_1 to the upper xy-plane. All these transformations preserve the lengths and scalar products. The final position of C_1 is uniquely determined by $x(C_1),y(C_1)$ in Definition 3.4.

After fixing $\triangle N_1,A_1,C_1$, it remains to prove that any other atom of S is uniquely determined by its x,y,z coordinates in $\mathrm{BRI}(S)$. Indeed, N_2 is obtained from C_1 by adding $\overrightarrow{C_1N_2}$, whose coordinates are the first three elements in the 2nd row of $\mathrm{BRI}(S)$. Then A_2 is obtained from N_1 by adding $\overrightarrow{N_2A_2}$, whose coordinates are the next three elements in the 2nd row of $\mathrm{BRI}(S)$. Then C_2 is obtained from A_2 by adding $\overrightarrow{A_2C_2}$ and so on.

(c) Since the complete invariant $\mathrm{BRI}(S)$ of a backbone S is locally defined by determining any i-th residue triangle $\triangle N_iA_iC_i$ in the basis of the previous $(i-1)$-st triangle, all rows of the matrix $\mathrm{BRI}(Q)$ for any subchain Q in the full backbone S coincide with the corresponding rows of $\mathrm{BRI}(S)$.

The only exception is the first row if Q starts from the i-th residue of S for $i>1$. In this case, the three non-zero invariants in the first row of Q can be obtained from the i-th row of $\mathrm{TRIN}(S)$ whose values are expressed in terms of the vectors $\overrightarrow{N_iA_i}$ and $\overrightarrow{A_iC_i}$ in Definition 3.1. This computation needs only a constant time independent of j because the coordinates of the vectors $\overrightarrow{A_iN_i}$ and $\overrightarrow{A_iC_i}$ are given in the i-th row of $\mathrm{BRI}(S)$. ■

Corollary 3.6 (completeness under isometry)**.** Any mirror image $\bar{S}$ of a backbone $S\subset\mathbb{R}^3$ has the invariant $\overline{\mathrm{BRI}}(S):=\mathrm{BRI}(\bar{S})$ obtained by reversing the signs in all z-columns of $\mathrm{BRI}(S)$. The unordered pair of $\mathrm{BRI}(S)$ and $\overline{\mathrm{BRI}}(S)$ is complete under isometry.

Proof of Corollary 3.6. To prove that $\overline{\mathrm{BRI}}(S) := \mathrm{BRI}(\bar{S})$ is obtained from $\mathrm{BRI}(S)$ by reversing the signs in all z-columns of $\mathrm{BRI}(S)$, consider the main atoms N_i, A_i, C_i in the i-th residue of S for any $i = 2, \dots, m$. The mirror image $\bar{S}$ has the corresponding atoms $\bar{N}_i, \bar{A}_i, \bar{C}_i$. There is a rigid motion f in $\mathbb{R}^3$ that matches these atoms so that $N_i = f(\bar{N}_i)$, $A_i = f(\bar{A}_i)$, $C_i = f(\bar{C}_i)$, and $f(\bar{S})$ is obtained from S by the reflection g in the plane of the residue triangle $\triangle N_i A_i C_i$. This reflection g preserves the basis vectors $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i$ from Definition 3.4 of the i-th residue of the backbone S.

In the orthonormal basis of u_i, v_i, $w_i = u_i \times v_i$, the coordinates of the vector $\overrightarrow{C_i N_{i+1}} = x(N_{i+1})\boldsymbol{u}_i + y(N_{i+1})\boldsymbol{v}_i + z(N_{i+1})\boldsymbol{w}_i$ determine the coordinates of the mirror image $f(\overrightarrow{\bar{C}_i \bar{N}_{i+1}}) = x(N_{i+1})\boldsymbol{u}_i + y(N_{i+1})\boldsymbol{v}_i - z(N_{i+1})\boldsymbol{w}_i$, where only the sign of the coefficient of $\boldsymbol{w}_i$ is reversed as required. Since the index $i = 2, \dots, m$ was arbitrarily chosen, it remains to notice that the first residue triangles $\triangle N_1 A_1 C_1$ and $\triangle \bar{N}_1 \bar{A}_1 \bar{C}_1$ can be matched by rigid motion, so all 3 non-zero invariants in the first rows of $\mathrm{BRI}(S)$ and $\mathrm{BRI}(\bar{S})$ coincide, while all z-coordinates are zeros. Finally, the unordered pair of $\mathrm{BRI}(S)$ and $\overline{\mathrm{BRI}}(S)$ is invariant under any rigid motion by Theorem 3.5(a) and under reflection, which swaps the invariants in this pair. By Theorem 3.5(b), any of $\mathrm{BRI}(S)$ and $\overline{\mathrm{BRI}}(S)$ suffices to reconstruct S or $\bar{S}$ up to rigid motion, hence S up to isometry in $\mathbb{R}^3$. ■

4 Lipschitz bi-continuity of the invariant BRI

Theorem 4.1 will prove the Lipschitz continuity of BRI in condition 1.2(c). For a given backbone S and its perturbation Q, let $l_{N,A}$ and $L_{N,A}$ denote the minimum and maximum bond length between any α-carbon A_i and nitrogen N_i in S, Q, respectively. The maximum bond lengths $L_{A,C}, L_{C,N}$ are similarly defined for other types of bonds.

Theorem 4.1 (Lipschitz continuity of BRI)**.** For any $\varepsilon > 0$, let Q be obtained from a backbone $S \subset \mathbb{R}^3$ by perturbing every atom of S up to Euclidean distance ε. Let $h = \min_i |y(A_i C_i)|$ be the minimum height in triangles $\triangle N_i A_i C_i$ at C_i for all residues in the backbones S, Q. Set $L = \max\{L_{C,N}, L_{N,A}, L_{A,C}\}$, $K = \dfrac{1}{l_{N,A}} + \dfrac{2}{h}\Big(1 + 2\dfrac{L_{A,C}}{l_{N,A}}\Big)$, and $\lambda = 2(1 + 2LK)$.

Then $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) \leq \lambda\varepsilon$.

Theorem 4.1 needs Lemmas 4.2, 4.3, 4.4, 4.5, and Proposition 4.6.

Lemma 4.2 (length difference)**.** Any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ satisfy $|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| \leq |\boldsymbol{u} - \boldsymbol{v}|$.

Proof. The triangle inequality for the Euclidean distance implies that $|\boldsymbol{u}| \leq |\boldsymbol{u} - \boldsymbol{v}| + |\boldsymbol{v}|$, so $|\boldsymbol{u}| - |\boldsymbol{v}| \leq |\boldsymbol{u} - \boldsymbol{v}|$. Swapping the vectors, we get $|\boldsymbol{v}| - |\boldsymbol{u}| \leq |\boldsymbol{u} - \boldsymbol{v}|$. Combining the inequalities $\pm(|\boldsymbol{u}| - |\boldsymbol{v}|) \leq |\boldsymbol{u} - \boldsymbol{v}|$, we conclude that $|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| \leq |\boldsymbol{u} - \boldsymbol{v}|$ as required. ■

Lemma 4.3 (perturbation of a vector)**.** Let A', B' be any ε-perturbations of points $A, B \in \mathbb{R}^n$, respectively, i.e. $|A - A'| \leq \varepsilon$, $|B - B'| \leq \varepsilon$. Then $|\overrightarrow{A'B'} - \overrightarrow{AB}| \leq 2\varepsilon$.

Proof. Apply the triangle inequality:
$|\overrightarrow{A'B'} - \overrightarrow{AB}| = |\overrightarrow{A'A} + \overrightarrow{BB'}| \leq |\overrightarrow{A'A}| + |\overrightarrow{BB'}| \leq 2\varepsilon$. ■

Lemma 4.4 (a normalized vector)**.** Let $\boldsymbol{u}$ be a δ-perturbation of a vector $\boldsymbol{v} \in \mathbb{R}^n$, i.e. $|\boldsymbol{u} - \boldsymbol{v}| \leq \delta$. Then $\left|\dfrac{\boldsymbol{u}}{|\boldsymbol{u}|} - \dfrac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| \leq \dfrac{2\delta}{l}$, where $l = \max\{|\boldsymbol{u}|, |\boldsymbol{v}|\}$. Hence if $\boldsymbol{u} = \overrightarrow{AN}$ and $\boldsymbol{u}' = \overrightarrow{A'N'}$ are vectors between atoms A_i, N_i and their ε-perturbations, then $\left|\dfrac{\boldsymbol{u}}{|\boldsymbol{u}|} - \dfrac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| \leq \dfrac{4\varepsilon}{l_{N,A}}$, where $l_{N,A}$ is the minimum bond length between N_i, A_i.

Proof. Assume that $\max\{|\boldsymbol{u}|, |\boldsymbol{v}|\} = |\boldsymbol{v}|$, which we denote by l. Then

$$\left|\frac{\boldsymbol{u}}{|\boldsymbol{u}|} - \frac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| = \left|\frac{|\boldsymbol{v}|\boldsymbol{u} - |\boldsymbol{u}|\boldsymbol{v}}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|}\right| = \frac{|\,(|\boldsymbol{v}| - |\boldsymbol{u}|)\boldsymbol{u} + |\boldsymbol{u}|(\boldsymbol{u} - \boldsymbol{v})\,|}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|} \leq$$
$$\frac{|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| \cdot |\boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|} = \frac{|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| + |\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{v}|} \leq \frac{2|\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{v}|} \leq \frac{2\delta}{l},$$

where we used the triangle inequality, Lemma 4.2, and $|\boldsymbol{u} - \boldsymbol{v}| \leq \delta$. The second inequality follows for $\delta = 2\varepsilon$ from Lemma 4.3 and $l_{N,C} \leq \max\{|\boldsymbol{u}|, |\boldsymbol{v}|\}$. ■

Lemma 4.5 (products)**.** For any $\boldsymbol{u}, \boldsymbol{u}', \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$, if $|\boldsymbol{v}'| = |\boldsymbol{v}| = 1$, then

(a) $|(\boldsymbol{u}' \cdot \boldsymbol{v}') - (\boldsymbol{u} \cdot \boldsymbol{v})| \leq |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$,

(b) $|(\boldsymbol{u}' \times \boldsymbol{v}') - (\boldsymbol{u} \times \boldsymbol{v})| \leq |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$,

(c) $|(\boldsymbol{u}'\cdot\boldsymbol{v}')\boldsymbol{v}'-(\boldsymbol{u}\cdot\boldsymbol{v})\boldsymbol{v}|\leq|\boldsymbol{u}'-\boldsymbol{u}|+2|\boldsymbol{u}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|$.

Proof. **(a)** Any scalar and vector product has the upper bound $|\boldsymbol{u}|\cdot|\boldsymbol{v}|$.

$$\begin{aligned}&|(\boldsymbol{u}'\cdot\boldsymbol{v}')-(\boldsymbol{u}\cdot\boldsymbol{v})|=|(\boldsymbol{u}'-\boldsymbol{u})\cdot\boldsymbol{v}'+\boldsymbol{u}\cdot(\boldsymbol{v}'-\boldsymbol{v})|\leq\\&\leq|(\boldsymbol{u}'-\boldsymbol{u})\cdot\boldsymbol{v}'|+|\boldsymbol{u}\cdot(\boldsymbol{v}'-\boldsymbol{v})|\leq\leq|\boldsymbol{u}'-\boldsymbol{u}|\cdot|\boldsymbol{v}'|+|\boldsymbol{u}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|=\\&=|\boldsymbol{u}'-\boldsymbol{u}|+|\boldsymbol{u}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|\text{ due to }|\boldsymbol{v}'|=1.\end{aligned}$$

(b) Prove as (a) with the vector product instead of the scalar product.

(c) It follows by using $|\boldsymbol{v}|=1$ and part (a):

$$\begin{aligned}&|(\boldsymbol{u}'\cdot\boldsymbol{v}')\boldsymbol{v}'-(\boldsymbol{u}\cdot\boldsymbol{v})\boldsymbol{v}|=|(\boldsymbol{u}'\cdot\boldsymbol{v}'-\boldsymbol{u}\cdot\boldsymbol{v})\boldsymbol{v}'+(\boldsymbol{u}\cdot\boldsymbol{v})(\boldsymbol{v}'-\boldsymbol{v})|\leq\\&\leq|\boldsymbol{u}'\cdot\boldsymbol{v}'-\boldsymbol{u}\cdot\boldsymbol{v}|\cdot|\boldsymbol{v}'|+|\boldsymbol{u}\cdot\boldsymbol{v}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|\leq\\&\leq|\boldsymbol{u}'\cdot\boldsymbol{v}'-\boldsymbol{u}\cdot\boldsymbol{v}|+|\boldsymbol{u}|\cdot|\boldsymbol{v}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|\leq|\boldsymbol{u}'-\boldsymbol{u}|+2|\boldsymbol{u}|\cdot|\boldsymbol{v}'-\boldsymbol{v}|\end{aligned}$$

as required. ■

Recall that $l_{N,A}$ denotes the minimum bond length between N_i and A_i, and $L_{A,C}$ is the maximum distance between A_i and C_i, while h is the minimum height in $\triangle N_iA_iC_i$ at C_i for all residues in given backbones.

Proposition 4.6 (perturbations of a basis)**.** In the conditions of Theorem 4.1, if any atom is perturbed up to ε, the basis vectors from Definition 3.4 are perturbed as follows:

(a) $|\boldsymbol{u}_i'-\boldsymbol{u}_i|\leq\dfrac{4\varepsilon}{l_{N,A}}$;

(b) $|\boldsymbol{v}_i'-\boldsymbol{v}_i|\leq\dfrac{8\varepsilon}{h}(1+2\dfrac{L_{A,C}}{l_{N,A}})$;

(c) $|\boldsymbol{w}_i'-\boldsymbol{w}_i|\leq 4\varepsilon K$, where $K=\dfrac{1}{l_{N,A}}+\dfrac{2}{h}\Big(1+2\dfrac{L_{A,C}}{l_{N,A}}\Big)$ for all $i=1,\dots,m$.

Proof. **(a)** In Definition 3.4 the vector $\boldsymbol{u}_i=\dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$ satisfies the inequality $|\boldsymbol{u}_i'-\boldsymbol{u}_i|\leq\dfrac{4\varepsilon}{l_{N,A}}$ by Lemma 4.4.

(b) The second vector is $\boldsymbol{v}_i = \dfrac{\boldsymbol{h}_i}{|\boldsymbol{h}_i|}$ for $\boldsymbol{h}_i = \overrightarrow{A_iC_i} - b_i\overrightarrow{A_iN_i}$ and $b_i = \dfrac{\overrightarrow{A_iC_i} \cdot \overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|^2}$. Set $\boldsymbol{p}_i = \overrightarrow{A_iC_i}$, $\boldsymbol{q}_i = \dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$, so $|\boldsymbol{q}_i| = |\boldsymbol{q}'_i| = 1$, where any dash denotes a perturbation of a point or a vector. Also, $|\boldsymbol{p}_i| = |\overrightarrow{A_iC_i}|$ has the upper bound $L_{A,C}$. By Lemma 4.5(c):

$$|b'_i\overrightarrow{A'_iN'_i} - b_i\overrightarrow{A_iN_i}| = |(\boldsymbol{p}'_i \cdot \boldsymbol{q}'_i)\boldsymbol{q}'_i - (\boldsymbol{p}_i \cdot \boldsymbol{q}_i)\boldsymbol{q}_i| \leq$$
$$|\boldsymbol{p}'_i - \boldsymbol{p}_i| + 2|\boldsymbol{p}_i| \cdot |\boldsymbol{q}'_i - \boldsymbol{q}_i| \leq 2\varepsilon + 2L_{A,C}\frac{4\varepsilon}{l_{N,A}},$$

where we used $|\boldsymbol{p_i}| \leq L_{A,C}$ and $|\boldsymbol{q}'_i - \boldsymbol{q}_i| \leq \dfrac{4\varepsilon}{l_{N,A}}$ by Lemma 4.4. Then

$$|\boldsymbol{h}'_i - \boldsymbol{h}_i| = |\boldsymbol{p}'_i - b'_i\overrightarrow{A'_iN'_i} - (\boldsymbol{p}_i - b_i\overrightarrow{A_iN_i})| \leq$$
$$\leq |\boldsymbol{p}'_i - \boldsymbol{p}_i| + |b'_i\overrightarrow{A'_iN'_i} - b_i\overrightarrow{A_iN_i}| \leq 2\varepsilon + 2\varepsilon + \varepsilon\frac{L_{A,C}}{l_{N,A}} = 4\varepsilon(1 + 2\frac{L_{A,C}}{l_{N,A}}).$$

The vectors $\boldsymbol{h}_i$, $\boldsymbol{p}_i$, and $b_i\overrightarrow{A_iN_i} = (\boldsymbol{p}_i \cdot \boldsymbol{q}_i)\boldsymbol{q}_i$ form a right-angled triangle with the hypotenuse $|\boldsymbol{p}_i|$. The length $|\boldsymbol{h}_i| = |\overrightarrow{A_iC_i}|\sin\angle N_iA_iC_i$ is the height in $\triangle N_iA_iC_i$ at the atom C_i. Using the given minimum height $h \leq |\boldsymbol{h}_i|$, Lemma 4.4 for $\delta = 4\varepsilon(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$ implies that

$$|\boldsymbol{v}'_i - \boldsymbol{v}_i| = \left|\frac{\boldsymbol{h}'_i}{|\boldsymbol{h}'_i|} - \frac{\boldsymbol{h}_i}{|\boldsymbol{h}_i|}\right| \leq \frac{2\delta}{h} \leq \frac{8\varepsilon}{h}(1 + 2\frac{L_{A,C}}{l_{N,A}}).$$

(c) The perturbation of $\boldsymbol{w}_i = \boldsymbol{u}_i \times \boldsymbol{v}_i$ is estimated by Lemma 4.5(b):

$$|\boldsymbol{w}'_i - \boldsymbol{w}_i| = |(\boldsymbol{u}'_i \times \boldsymbol{v}'_i) - (\boldsymbol{u}_i \times \boldsymbol{v}_i)| \leq |\boldsymbol{u}'_i - \boldsymbol{u}_i| + |\boldsymbol{u}_i| \cdot |\boldsymbol{v}'_i - \boldsymbol{v}_i| \leq$$
$$\leq \frac{4\varepsilon}{l_{N,A}} + \frac{8\varepsilon}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big) = 4\varepsilon K, \text{ where } K = \frac{1}{l_{N,A}} + \frac{2}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big)$$

as required. ■

Proof of Theorem 4.1. In the perturbed backbone Q, let N'_i, A'_i, C'_i denote ε-perturbations of atoms $N_,A_i, C_i$ from the backbone S for $i = 1, \dots, m$.

We prove that any coordinate of $\text{BRI}(S)$ changes by at most $\lambda\varepsilon$ for the given Lipschitz constant λ. The first coordinate $x(N_1)$ changes by at most 2ε because $|x(N_1') - x(N_1)| = \big|\,|\overrightarrow{A_1'N_1'}| - |\overrightarrow{A_1N_1}|\,\big| \leq 2\varepsilon$ by Lemma 4.2. For the coordinate $x(C_1) = \dfrac{\overrightarrow{A_1C_1}\cdot\overrightarrow{A_1N_1}}{|\overrightarrow{A_1N_1}|}$, set $\boldsymbol{u} = \overrightarrow{A_1C_1}$ and $\boldsymbol{v} = \dfrac{\overrightarrow{A_1N_1}}{|\overrightarrow{A_1N_1}|}$, so $|\boldsymbol{v}| = 1$. We write the perturbed versions of all vectors with a dash.

Then $|x(C_1') - x(C_1)| = |\boldsymbol{u}'\cdot\boldsymbol{v}' - \boldsymbol{u}\cdot\boldsymbol{v}| \leq |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}|\cdot|\boldsymbol{v}' - \boldsymbol{v}|$ by Lemma 4.5(a). Lemma 4.3 implies that $|\boldsymbol{u}' - \boldsymbol{u}| \leq 2\varepsilon$. Lemma 4.4 for $u = \overrightarrow{A_1'N_1'}$ and $v = \overrightarrow{A_1N_1}$ implies that $|\boldsymbol{v}' - \boldsymbol{v}| \leq \dfrac{4\varepsilon}{l_{N,A}}$, where $l_{N,A}$ is the minimum length of the bond between an α-carbon A_i and N_i across all backbones. Also, the Euclidean length $|\boldsymbol{u}| = |\overrightarrow{A_1C_1}|$ has the upper bound $L_{A,C}$ equal to the maximum length of the bond between A_i and C_i across all backbones. Then $|x(C_1') - x(C_1)| \leq 2\varepsilon\Big(1 + 2\dfrac{L_{A,C}}{l_{N,A}}\Big)$.

In the notations above, the last non-zero coordinate in the first row of $\text{BRI}(A)$ is $y(C_1) = |\overrightarrow{A_1C_1} - x(C_1)\dfrac{\overrightarrow{A_1N_1}}{|\overrightarrow{A_1N_1}|}| = |\boldsymbol{u} - x(C_1)\boldsymbol{v}|$. We estimate the perturbation first by Lemma 4.2:

$$\begin{aligned}
&|y(C_1') - y(C_1)| = \big|\,|\boldsymbol{u}' - x(C_1')\boldsymbol{v}'| - |\boldsymbol{u} - x(C_1)\boldsymbol{v}|\,\big| \leq \\
&\leq |\boldsymbol{u}' - x(C_1')\boldsymbol{v}' - (\boldsymbol{u} - x(C_1)\boldsymbol{v})| \leq |\boldsymbol{u}' - \boldsymbol{u}| + |x(C_1')\boldsymbol{v}' - x(C_1)\boldsymbol{v}| \leq \\
&\leq 2\varepsilon + |(x(C_1') - x(C_1))\boldsymbol{v}' + x(C_1)(\boldsymbol{v}' - \boldsymbol{v})| \leq \\
&\leq 2\varepsilon + |x(C_1') - x(C_1)| + |x(C_1)|\cdot|\boldsymbol{v}' - \boldsymbol{v}| \leq \\
&\leq 2\varepsilon + 2\varepsilon\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big) + |\overrightarrow{A_1C_1}|\frac{4\varepsilon}{l_{N,A}} \leq 4\varepsilon\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big),
\end{aligned}$$

where we substituted $|x(C_1') - x(C_1)| \leq 2\varepsilon(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$ and $|\boldsymbol{v}' - \boldsymbol{v}| \leq \dfrac{4\varepsilon}{l_{N,A}}$.

In any i-th row for $i = 2, \dots, m$, we estimate by Proposition 4.6(a):

$$\begin{aligned}
&|x(N_i') - x(N_i)| = |\overrightarrow{C_{i-1}'N_i'}\cdot\boldsymbol{u}_i' - \overrightarrow{C_{i-1}N_i}\cdot\boldsymbol{u}_i| \leq |\overrightarrow{C_{i-1}'N_i'} - \overrightarrow{C_{i-1}N_i}| + \\
&+ |\overrightarrow{C_{i-1}N_i}|\cdot|\boldsymbol{u}_i' - \boldsymbol{u}_i| \leq 2\varepsilon + L_{C,N}\frac{4\varepsilon}{l_{N,A}} = 2\varepsilon(1 + 2\frac{L_{C,N}}{l_{N,A}})
\end{aligned}$$

due to the upper bound $|\overrightarrow{C_{i-1}N_i}| \leq L_{C,N}$. For the other coordinates y, z, similarly use Proposition 4.6(b,c):

$$\begin{aligned}
&|y(N'_i) - y(N_i) = |\overrightarrow{C'_{i-1}N'_i} \cdot \boldsymbol{v}'_i - \overrightarrow{C_{i-1}N_i} \cdot \boldsymbol{v}_i| \leq \\
&\leq |\overrightarrow{C'_{i-1}N'_i} - \overrightarrow{C_{i-1}N_i}| + |\overrightarrow{C_{i-1}N_i}| \cdot |\boldsymbol{v}'_i - \boldsymbol{v}_i| \leq \\
&\leq 2\varepsilon + L_{C,N} \cdot \frac{8\varepsilon}{h}(1 + 2\frac{L_{A,C}}{l_{N,A}}) = 2\varepsilon\Big(1 + 4\frac{L_{C,N}}{h}(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big). \\
&|z(N'_i) - z(N_i)| = |\overrightarrow{C'_{i-1}N'_i} \cdot \boldsymbol{w}'_i - \overrightarrow{C_{i-1}N_i} \cdot \boldsymbol{w}_i| \leq \\
&\leq |\overrightarrow{C'_{i-1}N'_i} - \overrightarrow{C_{i-1}N_i}| + |\overrightarrow{C_{i-1}N_i}| \cdot |\boldsymbol{w}'_i - \boldsymbol{w}_i| \leq \\
&\leq 2\varepsilon + L_{C,N} \cdot 4\varepsilon K = 2\varepsilon(1 + 2L_{C,N}K), \text{ for } K = \frac{1}{l_{N,A}} + \frac{2}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big).
\end{aligned}$$

For the atoms A_i, C_i, we get similar upper bounds by replacing the factor $L_{C,N}$ with $L_{N,A}$, $L_{A,C}$, respectively. Taking into account all upper bounds above, the overall upper bound for the L_∞ metric on invariants is $L_\infty(\text{BRI}(S), \text{BRI}(Q)) \leq \lambda\varepsilon$, where $\lambda = 2(1 + 2LK)$ for $L = \max\{L_{C,N}, L_{N,A}, L_{A,C}\}$ and $K = \dfrac{1}{l_{N,A}} + \dfrac{2}{h}\Big(1 + 2\dfrac{L_{A,C}}{l_{N,A}}\Big)$ as required. ■

Example 4.7 (continuity in practice)**.** For all 707K+ cleaned chains, the median upper bound for λ is about 34.5 but the real values are smaller as in the example below. Consider the backbone S of the chain A (141 residues) from the standard hemoglobin 2hhb in the PDB.

We perturb S to Q by adding to each coordinate x, y, z of all atoms in S some uniform noise up to various thresholds $\varepsilon = 0.01, 0.02, \ldots, 0.1$Å. Fig. 3 (top left) shows how the distance $L_\infty(\text{BRI}(S), \text{BRI}(Q))$ averaged over 20 perturbations depends on ε As expected by Theorem 4.1, the metric L_∞ is perturbed linearly up to $\lambda\varepsilon$, where $\lambda \approx 4$.

Because the metric L_∞ between invariants BRI ($m \times 9$ matrices) can be computed in linear time $O(m)$, Theorem 4.1 also completes condition (1.2f) in Problem 1.2. Theorem 4.8 will prove condition in 1.2(d).

Theorem 4.8 (inverse continuity of BRI)**.** For any $\delta > 0$ and backbones $S, Q \subset \mathbb{R}^3$ with $L_\infty(\text{BRI}(S), \text{BRI}(Q)) < \delta$, there is a rigid motion f of $\mathbb{R}^3$ such that any atom of S is $\mu\delta$-close to the corresponding atom of $f(Q)$ for

$\mu = \sqrt{3}\dfrac{(8LK)^{m-1}-1}{8LK-1}$. Let $\widehat{\mathrm{BRI}}(S)$ be $\mathrm{BRI}(S)$ after multiplying the i-th row by $\dfrac{(8LK)^{i-1}-1}{8LK-1}$ for $i = 2, \dots, m$. Then $L_\infty(\widehat{\mathrm{BRI}}(S), \widehat{\mathrm{BRI}}(Q)) < \delta$ guarantees a rigid motion f of $\mathbb{R}^3$ such that any atom of the backbone S is $\sqrt{3}\delta$-close to the corresponding atom of $f(Q)$.

Proof of Theorem 4.8. Choose the origin of $\mathbb{R}^3$ at the first alpha-carbon atom A_1 of the backbone S, the positive x-axis through the vector $\overrightarrow{A_1N_1}$, and the y-axis so that the triangle $N_1A_1C_1$ belongs to the upper half of the xy-plane. Shift another backbone Q so that its first alpha-carbon atom A_1' coincides with the origin A_1. Rotate the image of Q so that its first nitrogen atom N_1 is in the x-axis through the atoms A_1, N_1 of S and the next carbon C_1' of Q is in the upper xy-plane.

For the resulting motion f, we will prove that the atoms of S are $\mu\delta$-close to the corresponding atoms of the image of Q, which we still denote by N_i', A_i', C_i' for simplicity. Because the atom N_1' is in the x-axis through $\overrightarrow{A_1N_1}$, the first basis vectors of length 1 coincide ($\boldsymbol{u}_1' = \boldsymbol{u}_1$) and hence also uniquely define the other basis vectors ($\boldsymbol{v}_1' = \boldsymbol{v}_1$, $\boldsymbol{w}_1' = \boldsymbol{w}_1$).

Then $|x(N_1') - x(N_1)| \le \delta$ implies that the atom N_1' is δ-close to N_1 in the x-axis. The atoms C_1, C_1' are $\delta\sqrt{2}$-close due to

$$|C_1' - C_1| = \sqrt{|x(C_1') - x(C_1)|^2 + |y(C_1') - y(C_1)|^2} \le \sqrt{\delta^2 + \delta^2} = \delta\sqrt{2}.$$

Because the first bases coincide, we consider the second residue:

$$\begin{aligned}
&|N_2' - N_2| = \\
&= |x(N_2')\boldsymbol{u}_1 + y(N_2')\boldsymbol{v}_1 + z(N_2')\boldsymbol{w}_1 - x(N_2)\boldsymbol{u}_1 - y(N_2)\boldsymbol{v}_1 - z(N_2)\boldsymbol{w}_1| = \\
&= \sqrt{|x(N_2') - x(N_2)|^2 + |y(N_2') - y(N_2)|^2 + |z(N_2') - z(N_2)|^2} \le \\
&\le \sqrt{\delta^2 + \delta^2 + \delta^2} = \delta\sqrt{3}.
\end{aligned}$$

Similarly, we get the upper bound $\varepsilon = \delta\sqrt{3}$ for the deviations $|A_2' - A_2|$ and $|C_2' - C_2|$. We will prove the following upper bound on deviations of atoms by induction on the number $m \geq 2$ of residues.

$$\max\{|N'_m - N_m|, |A'_m - A_m|, |C'_m - C_m|\} \leq \sqrt{3}(1+8LK+\dots+8(LK)^{m-2})\delta,$$

$$\text{where } L = \max\{L_{C,N}, L_{N,A}, L_{A,C}\}, \quad K = \frac{1}{l_{N,A}} + \frac{2}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big).$$

The base $m = 2$ was completed above. The inductive assumption says that the upper bound $\varepsilon = \sqrt{3}(1 + 8LK + \dots + (8LK)^{i-2})\delta$ holds for a single value of $i \geq 2$. The inductive step below is for the next value $i+1$.

Proposition 4.6 estimates deviations of vectors in the second basis:

$$|\boldsymbol{u}'_2 - \boldsymbol{u}_2| \leq \frac{4\varepsilon}{l_{N,A}}, \quad |\boldsymbol{v}'_2 - \boldsymbol{v}_2| \leq \frac{8\varepsilon}{h}\big(1 + 2\frac{L_{A,C}}{l_{N,A}}\big), \quad |\boldsymbol{w}'_2 - \boldsymbol{w}_2| \leq 4\varepsilon K.$$

For nitrogens, we split the deviations in the $(i+1)$-st residue into the deviations proportional to the differences in coordinates and the deviations proportional to the differences in basis vectors as follows:

$$\begin{aligned}
&|N'_{i+1} - N_{i+1}| = |x(N'_{i+1})\boldsymbol{u}'_i + y(N'_{i+1})\boldsymbol{v}'_i + z(N'_{i+1})\boldsymbol{w}'_i - \\
&- x(N_{i+1})\boldsymbol{u}_i - y(N_{i+1})\boldsymbol{v}_i - z(N_{i+1})\boldsymbol{w}_i| = |(x(N'_{i+1})\boldsymbol{u}'_i - x(N_{i+1})\boldsymbol{u}_i) + \\
&+ (y(N'_{i+1})\boldsymbol{v}'_i - y(N_{i+1})\boldsymbol{v}_i) + (z(N'_{i+1})\boldsymbol{w}'_i - z(N_{i+1})\boldsymbol{w}_i)| = \\
&= \Big|\big(x(N'_{i+1}) - x(N_{i+1})\big)\boldsymbol{u}'_i + x(N_{i+1})(\boldsymbol{u}'_i - \boldsymbol{u}_i) + \big(y(N'_{i+1}) - y(N_{i+1})\big)\boldsymbol{v}'_i + \\
&+ y(N_{i+1})(\boldsymbol{v}'_i - \boldsymbol{v}_i) + \big(z(N'_{i+1}) - z(N_{i+1})\big)\boldsymbol{w}'_i + z(N_{i+1})(\boldsymbol{w}'_i - \boldsymbol{w}_i)\Big| \leq \\
&\leq \Big|\big(x(N'_{i+1}) - x(N_{i+1})\big)\boldsymbol{u}'_i + \big(y(N'_{i+1}) - y(N_{i+1})\big)\boldsymbol{v}'_i + \\
&+ \big(z(N'_{i+1}) - z(N_{i+1})\big)\boldsymbol{w}'_i\Big| + \Big|x(N_{i+1})(\boldsymbol{u}'_i - \boldsymbol{u}_i)\Big| + \\
&+ \Big|y(N_{i+1})(\boldsymbol{v}'_i - \boldsymbol{v}_i)\Big| + \Big|z(N_{i+1})(\boldsymbol{w}'_i - \boldsymbol{w}_i)\Big|.
\end{aligned}$$

In the last upper bound, the first big modulus is the Euclidean length of a vector written in the orthonormal basis $\boldsymbol{u}'_i, \boldsymbol{v}'_i, \boldsymbol{w}'_i$. Since the coordinates of this vector have absolute values at most δ, this length has the upper bound $\delta\sqrt{3}$. In the second row of the matrix BRI, we estimate each term by replacing absolute values of coordinates with the maximum bond lengths and by using $|x(N_{i+1})| \leq L_{C,N}$ and Proposition 4.6 as follows:

$$|x(N_{i+1})| \cdot |\boldsymbol{u}_i' - \boldsymbol{u}_i| \leq L_{C,N}\frac{4\varepsilon}{l_{N,A}},$$
$$|y(N_{i+1})| \cdot |\boldsymbol{v}_i' - \boldsymbol{v}_i| \leq L_{C,N}\frac{8\varepsilon}{h}\Big(1+2\frac{L_{A,C}}{l_{N,A}}\Big),$$
$$|z(N_{i+1})| \cdot |\boldsymbol{w}_i' - \boldsymbol{w}_i| \leq L_{C,N}\cdot 4\varepsilon K, \text{ where } K = \frac{1}{l_{N,A}} + \frac{2}{h}\Big(1+2\frac{L_{A,C}}{l_{N,A}}\Big).$$

Taking the sum of these estimates, the final deviation is

$$|N_{i+1}' - N_{i+1}| \leq$$
$$\sqrt{3}\delta + 4\varepsilon L_{C,N}\Big(\frac{1}{l_{N,A}} + \frac{2}{h}(1+2\frac{L_{A,C}}{l_{N,A}}) + \frac{1}{l_{N,A}} + \frac{2}{h}(1+2\frac{L_{A,C}}{l_{N,A}})\Big) =$$
$$= \sqrt{3}\delta + 8L_{C,N}\varepsilon K \leq \sqrt{3}(1 + 8LK(1+\dots+(8LK)^{i-2})\delta =$$
$$= \sqrt{3}(1+\dots+(8LK)^{i-1})\delta.$$

For the atoms A_{i+1}, C_{i+1} in the $(i+1)$-st residue, we get the same bound by replacing $L_{C,N}$ with $L_{N,A}, L_{A,C} \leq L$. The bound for $i = m$ is $\sqrt{3}(1+\dots+(8LK)^{m-2}))\delta = \sqrt{3}\dfrac{(8LK)^{m-1}-1}{8LK-1}\delta$.

Now consider the modified invariant $\widehat{\mathrm{BRI}}(S)$ obtained by multiplying the i-th row of $\mathrm{BRI}(S)$ by $\dfrac{(8LK)^{i-1}-1}{8LK-1}$ for $i = 2, \dots, m$. Then the δ-closeness of the corresponding invariant components in the metric L_∞ means smaller deviations $|x(N_i') - x(N_i)| \leq \delta\dfrac{8LK-1}{(8LK)^{i-1}-1}$, similarly for other components. This extra multiplicative factor gives the upper bound $|N_{i+1}' - N_{i+1}| \leq \sqrt{3}\delta$, similarly for all other atoms. ■

A Lipschitz constant μ plays no significant role because any metric on invariant values can be divided by μ, which makes this constant 1. The second part of Theorem 4.8 offers a smarter adjustment of $\mathrm{BRI}(S)$ to the modified invariant $\widehat{\mathrm{BRI}}(S)$ depending on a row index of $\mathrm{BRI}(S)$ to guarantee the smaller Lipschitz constant $\sqrt{3}$.

5 Average invariant, diagrams, barcodes

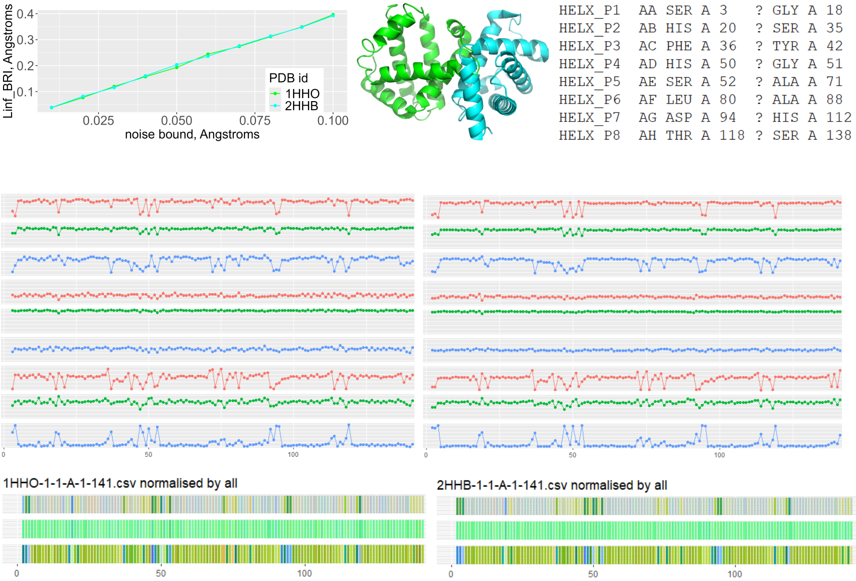
This section simplifies the complete invariant BRI to its average vector in $\mathbb{R}^9$ and introduces the diagram and barcode that visually represent BRI.

Definition 5.1 (average invariant Brain, standard deviation of invariants, diagram BID, and barcode BIB)**.** For any protein backbone S of m residues, the *backbone rigid average invariant* $\text{Brain}(S) \in \mathbb{R}^9$ is the vector of nine column averages in $\text{BRI}(S)$ excluding the first row. The standard deviation can be computed in a similar way. The *backbone invariant diagram* $\text{BID}(S)$ consists of nine polygonal curves going through the points $(i, c(i))$, $i = 2, \dots, m$, where c is one of the coordinates (columns) of $\text{BRI}(S)$, see Fig. 3 (middle). For each atom type such as N, the coordinates $(x(N_i), y(N_i), z(N_i))$ are linearly converted into the RGB color value for $i = 1, \dots, m$. The resulting three color bars for the ordered atoms N, A, C form the *backbone invariant barcode* $\text{BIB}(S)$, see Fig. 3 (bottom).

Example 5.2 (hemoglobins)**.** The PDB contains thousands of hemoglobin structures. We consider here the structure 2hhb as a standard, and compare it with oxygenated 1hho, which contains an extra oxygen whose transport is facilitated by hemoglobin. In both cases, we considered the main chains (entity 1, model 1, chain A) of 141 residues. Table 1 showed the TRIN and BRI invariants for the first 3 residues of 2hhb and 1hho.

The top left image in Fig. 3 (top) shows that the Lipschitz constant from Theorem 4.1 is $\lambda \approx 4$ for both hemoglobins. Fig. 3 (middle) illustrates the complexity of identifying similar proteins that can be given with very distant coordinates. The similarity under rigid motion becomes clearer by comparing their diagrams and barcodes in Fig. 3 (rows 2, 3).

More importantly, a rigidly repeated pattern such as α-helix or β-strand has constant invariants over several residue indices, which are easily detectable in BID and visible in BIB as intervals of uniform color. The PDB uses the baseline algorithm DSSP (Define Secondary Structure of Proteins) [21], which depends on several manual parameters and sometimes outputs α-helices of only two residues. For instance, the PDB files 1hho and 2hhb in Fig. 3 (right) include HELX_P4 consisting of only residues 50



```
HELX_P1   AA SER A 3   ? GLY A 18
HELX_P2   AB HIS A 20  ? SER A 35
HELX_P3   AC PHE A 36  ? TYR A 42
HELX_P4   AD HIS A 50  ? GLY A 51
HELX_P5   AE SER A 52  ? ALA A 71
HELX_P6   AF LEU A 80  ? ALA A 88
HELX_P7   AG ASP A 94  ? HIS A 112
HELX_P8   AH THR A 118 ? SER A 138
```

Figure 3. **Row 1**: the Lipschitz continuity of BRI from Theorem 4.1 is illustrated on the left by perturbing hemoglobins in Example 4.7, whose main chains A of 141 residues are shown in the middle (oxygenated 1hho in green, standard 2hhb in cyan) and eight α-helices found by [21] on the right. **Row 2**: the Backbone Invariant Diagram (BID) of the hemoglobins 1hho vs 2hhb in the PDB, see Definition 5.1. **Row 3**: the Backbone Invariant Barcode (BIB), see Example 5.2.

and 51, and HELX_P5 of length 20 over residue indices $i = 52, \ldots, 71$. Fig. 3 shows that a 'constant' interval of little noise appears only for $i = 54, \ldots, 70$. Hence new invariants allow a more objective detection of secondary structures, which will be explored in future work.

While the complete BRI(S) can be used to compare backbones of the same length, the average invariant Brain$(S) \in \mathbb{R}^9$ and the standard deviation invariant can help to visualize all backbones of different lengths on the same heatmap. In each image of Fig. 4, any protein backbone is represented by a single point (x, y) whose coordinates are the two simplest statistics (average and standard deviation) of a fixed invariant across all residues in a fixed chain. The top images in Fig. 4 show that the deviations of all three invariants from Definition 3.1 can be as large as 0.25Å. So the shapes of residue triangles can substantially vary even for a fixed backbone, while AlphaFold2 [20] assumes that they all have identical shapes.

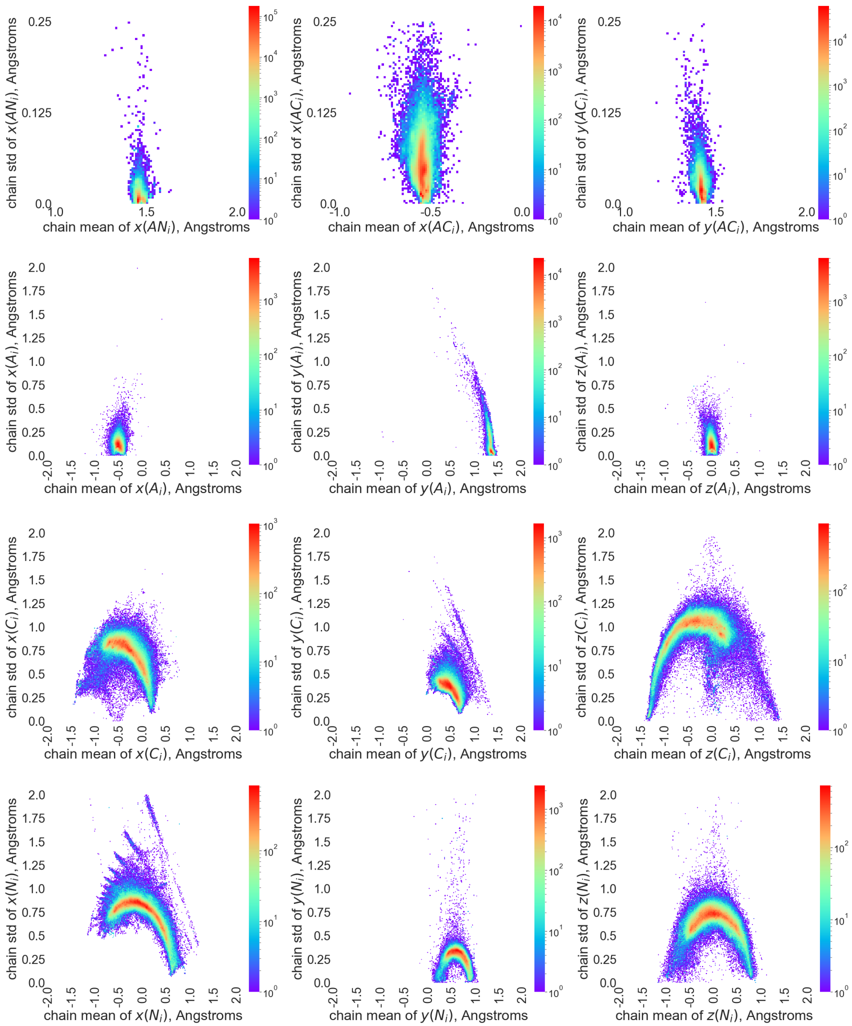


Figure 4. Heatmaps of average and standard deviations of the invariants TRIN and BRI across all 707K+ chains obtained by Protocol 3.2. The color indicates (on the logarithmic scale) the number of chains whose pairs (x, y) of the average x and the standard deviation y are discretized to each pixel.

6 Duplicates with identical coordinates

The linear time of the complete invariant BRI(S) has enabled all-vs-all comparisons for all tertiary structures in the PDB, which was additionally cleaned by Protocol 3.2. To speed up comparisons, Lemma 6.1 proves that the metric $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$ between complete invariants is not smaller than the much faster distance $L_\infty(\mathrm{Brain}(S), \mathrm{Brain}(Q))$ between the averaged invariants (vectors of 9 coordinates) from Definition 5.1.

Lemma 6.1 (relation between metrics on the invariants BRI and Brain)**.** Any protein backbones S, Q of the same number of residues satisfy the inequality $L_\infty(\mathrm{Brain}(S), \mathrm{Brain}(Q)) \leq L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$.

Proof of Lemma 6.1. If protein backbones S, Q have m residues and $\delta = L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$, then any corresponding elements of the $m \times 9$ matrices $\mathrm{BRI}(S), \mathrm{BRI}(Q)$ differ by at most δ. For any $j = 1, \ldots, 9$, their averages of the j-th columns differ by at most δ because

$$\left| \frac{1}{m} \sum_{i=1}^{m} \mathrm{BRI}_{ij}(S) - \frac{1}{m} \sum_{i=1}^{m} \mathrm{BRI}_{ij}(Q) \right| \leq$$
$$\frac{1}{m} \sum_{i=1}^{m} |\mathrm{BRI}_{ij}(S) - \mathrm{BRI}_{ij}(Q)| \leq \frac{1}{m} \sum_{i=1}^{m} \delta = \delta.$$

Hence $L_\infty(\mathrm{Brain}(S), \mathrm{Brain}(Q)) \leq \delta$ as required. ■

The complete invariants and their statistical summaries (averages and deviations) were computed in 3 hours 18 min 21 sec. After comparing all (888+ million) pairs of same-length backbones within 1 hour, we found 13907 pairs S, Q with the *exact zero-distance* $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) = 0$ between complete invariants meaning that all these backbones S, Q are related by rigid motion, but they may not be geometrically identical.

However, 9366 of these pairs turned out to have x, y, z coordinates of all main atoms *identical to the last digit* despite many of them (763) coming from *different PDB entries.* Table 2 lists nine pairs whose geometrically identical chains unexpectedly differ in the sequences of amino acids.

Table 2. Chains with identical backbones but different sequences.

PDB id1 & chain	method and resolutions, Å	PDB id2 & chain	all atoms have identical x, y, z	different residues
1a0t-B	X-ray, 2.4, 2.4	1oh2-B	all 3×413	9
1ce7-A	X-ray, 2.7, 2.7	2mll-A	all 3×241	1, GLY≠HIS
1ruj-A	X-ray, 3, 3	4rhv-A	all 3×237	1, GLY≠SER
1gli-B/D	X-ray, 2.5, 1.7	3hhb-B/D	all 3×146	1, MET≠VAL
2hqe-A	X-ray, 2, 2	2o4x-A	all 3×217	1, GLN≠GLU
5adx-T	EM, 4, 8.2	5afu-Z	all 3×165	1, ILE≠VAL
5lj3-O	EM, 3.8, 10	5lj5-P	all 3×252	1, ALA≠VAL
8fdz-A	X-ray, 2.5, 2.2	8fe0-A	all 3×200	1, THR≠SER

In a similar case [49], when five pairs of unexpected duplicates were found in the Cambridge Structural Database (CSD). Their integrity office agreed that a single atomic replacement should perturb geometry at least slightly, so all coordinates cannot remain the same. Five journals started investigations into the data integrity of the relevant publications [8].

We e-mailed all authors of the experimental structures listed in Table 2 whose contacts we found. Two authors replied with details and confirmed that their PDB entries should be corrected, see appendix A.

The duplicates from Table 2 were shown to the PDB validation team, who did not know about the found coincidences (in coordinates) and differences (in amino acids) because the PDB validation is currently done for an individual protein (checking atom clashes, outliers etc).

The recently published method [12] didn't report any duplicates. Right now anyone can download the PDB files from Table 2 and see all coincidences of x, y, z coordinates with their own eyes without any computations. Here are the links to the identical files in the first row of Table 2, where the 4-letter PDB id can be replaced with any other id:
https://files.rcsb.org/download/1A0T.cif and
https://files.rcsb.org/download/1OH2.cif.

The histogram in Fig. 5 reveals the scale of near-duplicates among 707K+ cleaned chains up to small distances $L_\infty \leq 0.01$Å on the horizontal axis. Each of 10 vertical bins over an interval of length 0.001Å indicates the number of pairs (on the logarithmic scale) of backbones S, Q whose distance $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$ is within this interval. Since all atomic coordinates in the PDB have 3 decimal places, all distances were rounded to 0.001Å. The bound of 0.01Å is considered noise because the smallest inter-atomic distance is about 100 times larger at $1\text{Å} = 10^{-10}$ m.

The physical meaning of distances follows from the bi-continuity conditions (c,d) in Problem 1.2. If every atom of a backbone S is shifted up to Euclidean distance ε, then $\mathrm{BRI}(S)$ changes up to $\lambda\varepsilon$ in L_∞. The Lipschitz constant λ was expressed in Theorem 4.1 and estimated as $\lambda \approx 4$ for the hemoglobin chains in Example 5.2. So any small perturbation of atoms yields a small value of L_∞ in Angstroms. The inverse Lipschitz continuity in (1.2d) implies that a small distance $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) = \delta$ guarantees that all atoms of S, Q can be matched (after a suitable rigid motion) up to Euclidean distance $\mu\delta$, see Theorem 4.8. For all 775K+ pairs in Fig. 5, the median of the maximum atomwise deviation (of optimally aligned chains) divided by $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$ is about 0.4. So the closeness of BRIs practically guarantees the closeness by RMSD.

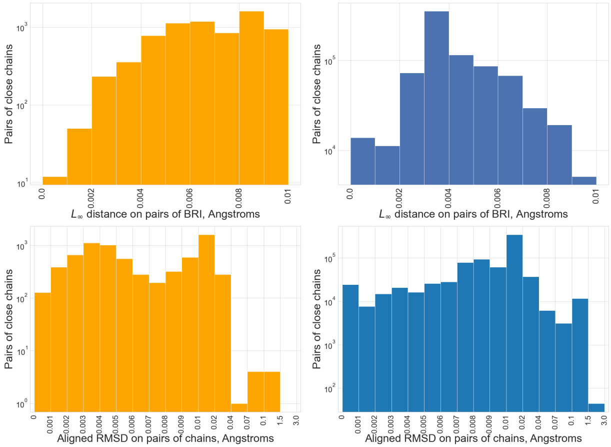


Figure 5. Histograms of near-duplicate chains (of the same length) on the log scale. **Top row**: 783075 pairs with $L_\infty \leq 0.01$Å on BRIs including 13907 pairs of exact duplicates with $L_\infty = 0$. **Bottom row**: the same pairs with traditional RMSD. **Left column**: 7151 pairs with different sequences of amino acids. **Right column**: 775852 pairs with identical sequences.

One potential explanation of identical coordinates is the molecular replacement method [42], which uses an existing protein structure, often a previous PDB deposit or part thereof, to solve a new structure. If the newly calculated electron density map does not allow for further refinement then the coordinates may (reasonably) remain unchanged.

The same coincidences can happen with lower-quality cryo-EM maps in which an existing PDB structure may be placed but where the resolution may not allow for further refinement of atomic coordinates [15,36].

We have checked that the found duplicate backbones also have identical distance matrices on $3m$ ordered atoms, which were slower to compute in time $O(m^2)$ over two days on a similar machine. The widely used DALI server [13] also confirmed the found duplicates by the traditional Root Mean Square Deviation (RMSD) through optimal alignment. The DALI took about 30 min on average to find a short list of nearest neighbors of one chain in the whole PDB. Extrapolating this time to all pairwise comparisons for 707K+ cleaned chains yields 40+ years, slower by orders of magnitude than 6 hours needed for all comparisons of the complete invariants BRI on the same desktop computer. Our implementation of RMSD for Fig. 5 has the median time of 534 microseconds per pair of chains (of the same length), about 10 times slower than L_∞ on BRIs.

The FoldSeek algorithm [47] is claimed to be 4000 times faster than RMSD by Dali due to optimal alignments of 3-residue subchains instead of full original chains, which takes 3.65 days by the estimates above, still an order of magnitude slower than L_∞ on BRIs. But any similarity distance needs a proof of all metric axioms for trustworthy clustering [41].

The ultra-fast speed of all-vs-all comparisons by BRI is explained by the hierarchical nature of this complete invariant. To find near-duplicates in the PDB, we first compared only average invariants $\mathrm{Brain}(S) \in \mathbb{R}^9$. By Lemma 6.1 the full comparisons by BRI are needed only for a tiny proportion of backbones with the closest vectors $\mathrm{Brain}(S)$. This hierarchical speed-up is unavailable for any distance without underlying invariants.

7 Discussion of the PDB and data integrity

Using protein structures as an important example, this paper advocates a justified approach to any real data objects. The first and often missed step is to define an *equivalence relation* for given data because real objects can be digitally represented in (usually infinitely) many different ways.

For instance, a human can be recognized in a huge number of digital photos but science progressed to discover the human genome and other biometric data, which are being included even in passports. All other objects (protein backbones for example) similarly need complete invariants for unambiguous identification because a distance metric alone is insufficient to understand deeper relations beyond pairwise similarities.

There is little sense in distinguishing most objects (including flexible molecules) under rigid motion because translations and rotations preserve their properties in the same environment. Hence the input of all prediction algorithms should be invariant, ideally a complete continuous invariant.

The Lipschitz bi-continuity is essential because adding a small noise to input should not lead to a drastically different output and vice versa. Earlier versions of Problem 1.2 with weaker conditions were solved for 2D lattices [5, 25], periodic crystals [48, 49], and finite clouds of unordered points [26, 50] within the new area of Geometric Data Science [2, 51].

The crucial novelty in the proposed approach is treating (the rigid class of) any experimental structure (protein backbone) from the PDB as an *objective ground truth* instead of labels assigned manually or by earlier algorithms with many parameters. Problem 1.2 asked for an analytically defined invariant I whose explicit formula should remain unchanged for any new data without required re-training in machine learning.

While traditional approaches explored finite datasets within infinite spaces in a 'horizontal' way, solutions to Problem 1.2 and its analogs for other data [52] provide 'vertical' breakthroughs by building 'geographic' maps of continuous data spaces as viewed from a satellite [4].

Fig. 2 and 4 can be zoomed at any spot and mapped by using further invariants. Such navigation maps with invertible coordinates enable inverse design while any dimensionality reduction to a latent space was proved [27] to be discontinuous (making close points distant) or collapsing an unbounded region to a point (losing an infinite amount of data).

The main contributions are Theorems 3.5, 4.1, and 4.8, which solved Problem 1.2 for protein backbones, detected thousands of (near-)duplicates in the PDB and enabled a justified exploration of the protein universe.

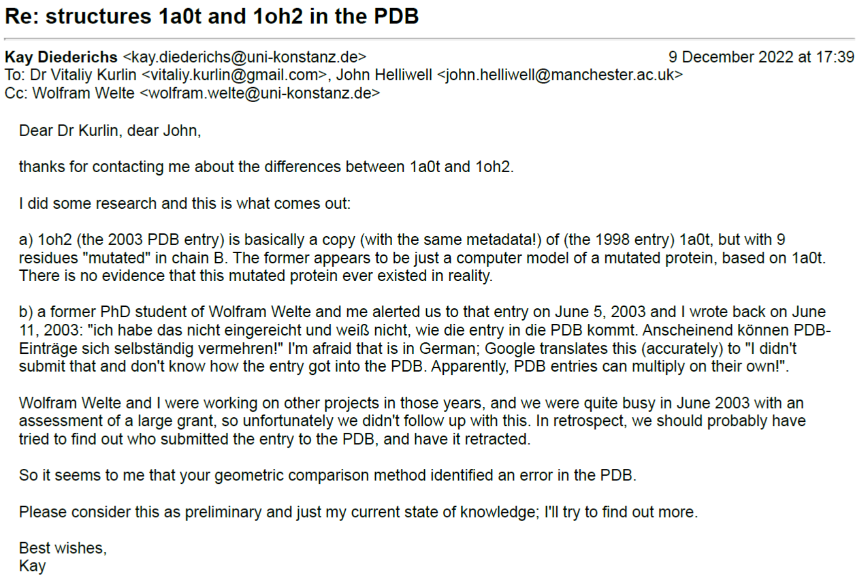
The supplementary data (available by request) include the Python code and a table of exact duplicates whose corresponding coordinates coincide in all decimal places and hence might need further refinement. Improving the PDB validation is needed to avoid unjustified predictions and claims of 'solutions' based on skewed data [33]. The recent analysis of the PDB revealed large numbers of waterless structures [54] and raised concerns.

***Acknowledgment*:** This work was supported by the Royal Academy of Engineering Fellowship IF2122/186, EPSRC New Horizons EP/X018474/1, and Royal Society APEX fellowship APX/R1/231152. The authors thank Mariusz Jaskolski, Alex Wlodawer, and Daniel Rigden for their helpful comments on early drafts and all other reviewers for their valuable time.

A Appendix: updates on PDB duplicates

This appendix discusses several duplicates that were found by the new invariants and later confirmed by their authors, and subsequent updates in the PDB. After finding the first duplicates in Table 2, we emailed the authors of the underlying publications whose contact details were still possible to find. The common author of the PDB entries 1a0t and 1oh2, Kay Diederichs, confirmed the error in December 2022 (see Fig. 6).

John Helliwell studied our duplicates including those with the same sequences of amino acids. After finding his pair of duplicates, he e-mailed us to confirm this error on 15th February 2023 (see Fig. 7). After meeting with the PDB validation team on February 27, 2023, where John was also present, the webpage https://www.rcsb.org/structure/removed/3UNR was updated without any reference to our work reporting the error: "Entry 3UNR was removed from the distribution of released PDB entries (status Obsolete) on 2023-03-01. It has been replaced (superseded) by 4YTA".



Re: structures 1a0t and 1oh2 in the PDB

Kay Diederichs <kay.diederichs@uni-konstanz.de> 9 December 2022 at 17:39
To: Dr Vitaliy Kurlin <vitaliy.kurlin@gmail.com>, John Helliwell <john.helliwell@manchester.ac.uk>
Cc: Wolfram Welte <wolfram.welte@uni-konstanz.de>

Dear Dr Kurlin, dear John,

thanks for contacting me about the differences between 1a0t and 1oh2.

I did some research and this is what comes out:

a) 1oh2 (the 2003 PDB entry) is basically a copy (with the same metadata!) of (the 1998 entry) 1a0t, but with 9 residues "mutated" in chain B. The former appears to be just a computer model of a mutated protein, based on 1a0t. There is no evidence that this mutated protein ever existed in reality.

b) a former PhD student of Wolfram Welte and me alerted us to that entry on June 5, 2003 and I wrote back on June 11, 2003: "ich habe das nicht eingereicht und weiß nicht, wie die entry in die PDB kommt. Anscheinend können PDB-Einträge sich selbständig vermehren!" I'm afraid that is in German; Google translates this (accurately) to "I didn't submit that and don't know how the entry got into the PDB. Apparently, PDB entries can multiply on their own!".

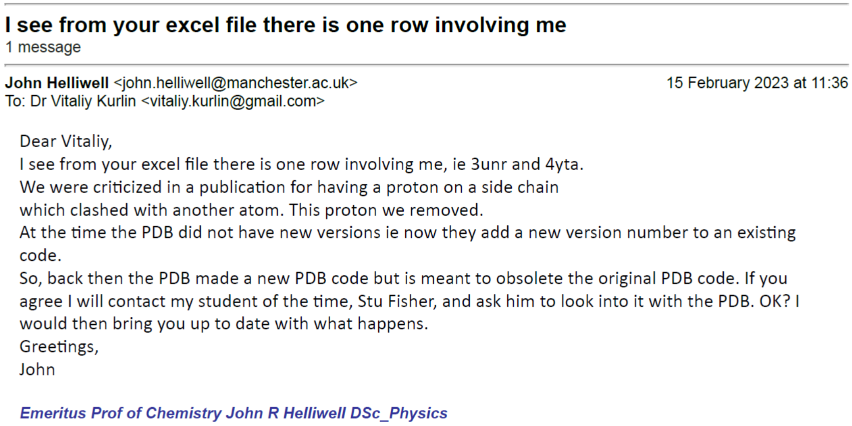
Wolfram Welte and I were working on other projects in those years, and we were quite busy in June 2003 with an assessment of a large grant, so unfortunately we didn't follow up with this. In retrospect, we should probably have tried to find out who submitted the entry to the PDB, and have it retracted.

So it seems to me that your geometric comparison method identified an error in the PDB.

Please consider this as preliminary and just my current state of knowledge; I'll try to find out more.

Best wishes,
Kay

Figure 6. Author's confirmation of the duplicates 1a0t and 1oh2.



I see from your excel file there is one row involving me
1 message

John Helliwell <john.helliwell@manchester.ac.uk> 15 February 2023 at 11:36
To: Dr Vitaliy Kurlin <vitaliy.kurlin@gmail.com>

Dear Vitaliy,
I see from your excel file there is one row involving me, ie 3unr and 4yta.
We were criticized in a publication for having a proton on a side chain
which clashed with another atom. This proton we removed.
At the time the PDB did not have new versions ie now they add a new version number to an existing code.
So, back then the PDB made a new PDB code but is meant to obsolete the original PDB code. If you agree I will contact my student of the time, Stu Fisher, and ask him to look into it with the PDB. OK? I would then bring you up to date with what happens.
Greetings,
John

Emeritus Prof of Chemistry John R Helliwell DSc_Physics

Figure 7. John Helliwell's confirmation of the duplicates 3unr and 4yta.

The PDB validation team confirmed that PDB entries are updated only by authors' request or by their permission. After we e-mailed all authors of the first found duplicates in December 2022, five entries from our list were updated in the PDB without acknowledging our work, see Table 3.

Table 3. These five PDB entries had duplicates similar to Table 2 but were modified after our initial contacts in December 2022. All original and updated files are still accessible online.

PDB entry	date of modification	reason of modification
4rhv	2023-01-18	Remediation
1ruj	2023-01-18	Remediation
1gli	2023-02-08	Remediation
3hhb	2023-02-08	Remediation
1cov	2023-04-19	Remediation

The older versions of the PDB files are still available via the web link ftp://snapshots.rcsb.org/20230102. Other duplicates in Table 2 were not previously reported, so their PDB files show the duplication of geometry with differences in sequences on December 19, 2024.

Because all x, y, z coordinates in the PDB are given with three decimal places relative to 1Å, a distance of less than 0.01Å can be considered negligible, especially due to floating point errors.

References

[1] C. B. Anfinsen, Principles that govern the folding of protein chains, *Science* **181** (1973) 223–230.

[2] O. Anosova, V. Kurlin, M. Senechal, The importance of definitions in crystallography, *IUCrJ* **11** (2024) 453–463.

[3] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network, *Science* **373** (2021) 871–876.

[4] M. J. Bright, A. I. Cooper, V. A. Kurlin, Geographic-style maps for 2-dimensional lattices, *Acta Cryst. A* **79** (2023).

[5] M. J. Bright, A. I. Cooper, V. A. Kurlin, Continuous chiral distances for 2-dimensional lattices, *Chirality* **35** (2023) 920–936.

[6] S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, S. Velankar, Protein Data Bank (PDB): the single global macromolecular structure archive, *Protein Cryst. Meth. Protoc.* **1607** (2017) 627–641.

[7] O. Carugo, How root-mean-square distance (rmsd) values depend on the resolution of protein structures that are compared, *J. Appl. Cryst.* **36** (2003) 125–128.

[8] D. S. Chawla, Crystallography databases hunt for fraudulent structures, *ACS Central Sci.* **9** (2023) 1853–1855.

[9] B. V. Dekster, J. B. Wilker, Edge lengths guaranteed to form a simplex, *Arch. Math.* **49** (1987) 351–366.

[10] E. J. Draizen, S. Veretnik, C. Mura, P. E. Bourne, Deep generative models of protein structure uncover distant relationships across a continuous fold space, *Nature Commun.* **15** (2024) #8094.

[11] J. Ellaway, Structural superposition by the Protein Data Bank in Europe, `https://github.com/PDBe-KB/pdbe-kb-manual/wiki/Structural-superposition`.

[12] D. Guzenko, S. K. Burley, J. M. Duarte, Real time structural search of the Protein Data Bank, *PLoS Comput. Biol.* **16** (2020) e1007970.

[13] L. Holm, Dali: Protein structure comparison server, `http://ekhidna2.biocenter.helsinki.fi/dali`.

[14] A. Heifetz, M. Eisenstein, Effect of local shape modifications of molecular surfaces on rigid-body protein–protein docking, *Protein Engin.* **16** (2003) 179–185.

[15] M. Hekkelman, A. Perrakis, R. Joosten, PDB-redo: updated and optimised crystallographic structures, `http://https://pdb-redo.eu`.

[16] L. Holm, C. Sander, Mapping the protein universe, *Science* **273** (1996) 595–602.

[17] L. Holm, Dali and the persistence of protein shape, *Protein Sci.* **29** (2020) 128–140.

[18] X. Jin, M. Awale, M. Zasso, D. Kostro, L. Patiny, J.-L. Reymond, PDB-explorer: a web-based interactive map of the Protein Data Bank in shape space, *BMC Bioinf.* **16** (2015) 1–15.

[19] D. T. Jones, J. M. Thornton, The impact of AlphaFold2 one year on, *Nature Meth.* **19** (2022) 15–20.

[20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. C. B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, A. Bridgland, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with alphafold, *Nature* **596** (2021) 583–589.

[21] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolym. Origin. Res. Biomol.* **22** (1983) 2577–2637.

[22] D. Kendall, The diffusion of shape, *Adv. Appl. Probab.* **9** (1977) 428–430.

[23] D. Kendall, D. Barden, T. Carne, H. Le, *Shape and Shape Theory*, Wiley, New York, 2009.

[24] M. Kowiel, M. Jaskolski, Z. Dauter, Achesym: an algorithm and server for standardized placement of macromolecular models in the unit cell, *Acta Cryst. D* **70** (2014) 3290–3298.

[25] V. Kurlin, Mathematics of 2-dimensional lattices, *Found. Comput. Math.* **24** (2024) 805–863.

[26] V. Kurlin, Polynomial-time algorithms for continuous metrics on atomic clouds of unordered points, *MATCH Commun. Math. Comput. Chem.* **91** (2024) 79–108.

[27] P. S. Landweber, E. A. Lazar, N. Patel, On fiber diameters of continuous maps, *Am. Math. Monthly* **123** (2016) 392–397.

[28] R. Lathrop, The protein threading problem with sequence amino acid interaction preferences is NP-complete, *Protein Engin. Des. Sel.* **7** (1994) 1059–1068.

[29] S. L. Lawton, R. A. Jacobson, The reduced cell and its crystallographic applications, Tech. rep., Ames Lab., Iowa State Univ. of Science and Tech., 1965.

[30] J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, et al., Macromolecular modeling and design in Rosetta: recent methods and frameworks, *Nature Meth.* **17** (2020) 665–680.

[31] K. U. Linderstrøm-Lang, *Lane Medical Lectures: Proteins and Enzymes*, vol. 6, Stanford Univ. Press, Stanford, 1952.

[32] V. Mariani, M. Biasini, A. Barbato, T. Schwede, lddt: a local superposition-free score for comparing protein structures and models using distance difference tests, *Bioinf.* **29** (2013) 2722–2728.

[33] R. McDonnell, A. Henderson, A. Elcock, Structure prediction of large rnas with AlphaFold3 highlights its capabilities and limitations, *J. Mol. Biol.* **436** (2024) #168816.

[34] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, Colabfold: making protein folding accessible to all, *Nature Meth.* **19** (2022) 1–4.

[35] P. B. Moore, W. A. Hendrickson, R. Henderson, A. T. Brunger, The protein-folding problem: Not yet solved, *Science* **375** (2022) 507–507.

[36] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, Refmac5 for the refinement of macromolecular crystal structures, *Acta Cryst. D* **67** (2011) 355–367.

[37] S. B. Needleman, *Protein Sequence Determination: a Sourcebook of Methods and Techniques*, vol. 8, Springer, 2012.

[38] P. Niggli, *Krystallographische und Strukturtheoretische Grundbegriffe*, vol. 1, Akademische verlagsgesellschaft mbh, 1928.

[39] E. Parthé, L. Gelato, B. Chabot, M. Penzo, K. Cenzual, R. Gladyshevskii, *TYPIX Standardized Data and Crystal Chemical Characterization of Inorganic Structure Types*, Springer, 2013.

[40] G. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins, *Adv. Protein Chem.* **23** (1968) 283–437.

[41] S. Rass, S. König, S. Ahmad, M. Goman, Metricizing the euclidean space towards desired distance relations in point clouds, *IEEE Trans. Inf. Forens. Sec.* **19** (2024) 7304–7319.

[42] M. G. Rossmann, The molecular replacement method, *Acta Cryst. A* **46** (1990) 73–82.

[43] P. Sacchi, M. Lusi, A. J. Cruz-Cabeza, E. Nauha, J. Bernstein, Same or different - that is the question: identification of crystal forms from crystal structure data, *CrystEngComm* **22** (2020) 7170–7185.

[44] I. Schoenberg, Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de Hilbert, *Ann. Math.* (1935) 724–732.

[45] L. R. Scott, A. Fernández, *A Mathematical Approach to Protein Biophysics*, Springer, 2017.

[46] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, P. D. Adams, AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination, *Nature Meth.* **21** (2024) 110–116.

[47] M. Van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with foldseek, *Nature Biotech.* **42** (2024) 243–246.

[48] D. Widdowson, M. M. Mosca, A. Pulido, A. I. Cooper, V. Kurlin, Average minimum distances of periodic point sets - foundational invariants for mapping all periodic crystals, *MATCH Commun. Math. Comput. Chem.* **87** (2022) 529–559.

[49] D. Widdowson, V. Kurlin, Resolving the data ambiguity for periodic crystals, *Adv. Neur. Inf. Process. Sys.* **35** (2022) 24625–24638.

[50] D. E. Widdowson, V. A. Kurlin, Recognizing rigid patterns of unlabeled point clouds by complete and continuous isometry invariants with no false negatives and no false positives, in: *Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 1275–1284.

[51] D. E. Widdowson, V. A. Kurlin, Continuous invariant-based maps of the cambridge structural database, *Crystal Growth Design* **24** (2024) 5627–5636.

[52] D. E. Widdowson, V. A. Kurlin, Navigation maps of the material space for automated self-driving labs of the future, *arXiv:2410.13796* (2024).

[53] M. Wirth, A. Volkamer, V. Zoete, F. Rippmann, O. Michielin, M. Rarey, W. H. Sauer, Protein pocket and ligand shape comparison and its application in virtual screening, *J. Comput. Aided Mol. Design* **27** (2013) 511–524.

[54] A. Wlodawer, Z. Dauter, P. Rubach, W. Minor, J. I. Loch, D. Brzezinski, M. Gilski, M. Jaskolski, Waterless structures in the Protein Data Bank, *IUCrJ* **11** (2024) 966–976.

[55] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, *Proteins Struct. Func. Bioinf.* **57** (2004) 702–710.