A complete and bi-continuous invariant of protein backbones under rigid motion

Olga Anosova[1], Alexey Gorelov[2], William Jeffcott[1], Ziqiu Jiang[3], and Vitaliy Kurlin (corresponding author)[1,4]

[1]Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
[2]Université Grenoble Alpes, Institut Fourier, 38000 Grenoble, France
[3]Department of Surgery & Cancer, Imperial College London, UK
[4]Materials Innovation Factory, University of Liverpool, Liverpool, L69 3NY, UK, vitaliy.kurlin@liverpool.ac.uk

(October 10, 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 (geometric) structure is the rigid shape of a protein chain represented by atomic positions in a 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 space, uniquely under rigid motion. This invariant is Lipschitz bi-continuous in the sense that it changes up to a constant multiple of any 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 space allows low-dimensional maps with analytically defined coordinates that reveal substantial variability in the protein universe.

1 Introduction: motivations and 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. A sequence is easy to experimentally detect but the important functional properties such as interactions with drug molecules depend on a 3D geometric shape (called a *tertiary structure* or *fold*) represented by an embedding of all its atoms in $\mathbb{R}^3$ [1], see Fig. 1 (left).

In 1973, Nobel laureate Anfinsen conjectured that the sequence of any protein chain determines its 3D geometric shape [2]. Neural networks for protein folding prediction such as AlphaFold2 [3–6] optimize millions of parameters and need re-training [7] on the growing number of experimental structures in the Protein Data Bank (PDB) [8].

Most importantly, the widely used similarities such as TM-score and LDDT [9, p. 2728] fail the axioms of a distance metric. Then clustering algorithms such as k-means and DBSCAN can produce pre-determined clusters [10]. Protein backbones of the same length (number of residues) can be optimally aligned to minimize the resulting Root Mean Square Deviation (RMSD) between corresponding atoms. This RMSD is slow to compute for all pairs of proteins and provides only distances without a mapping of the protein universe.

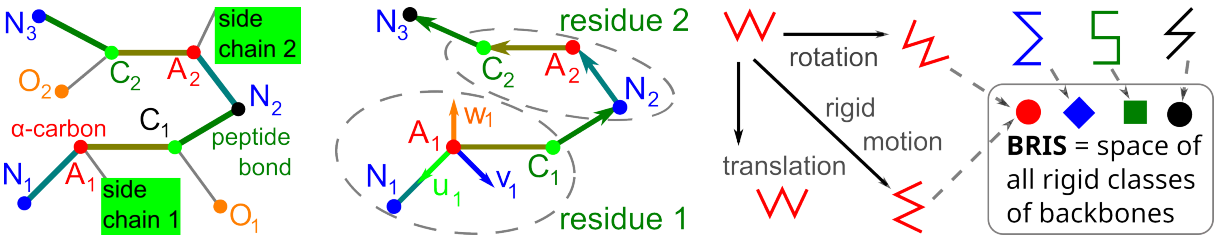

Figure 1. Left: a protein chain is a sequence of amino acid residues whose atoms N_i, A_i, C_i form a *backbone* embedded in $\mathbb{R}^3$. **Middle**: each triangle $\triangle N_iA_iC_i$ defines an orthonormal basis $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i$. The coordinates of the bonds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

We have developed a different approach to explicitly map 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 proteins is "same or different" [11]. Geometrically, the whole protein 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 under rigid motion have the same functional properties. Though proteins are flexible molecules, it is important to distinguish their rigid shapes that can differently interact [12] 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 $S \cong Q$). For any $m \geq 1$, 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 moments of inertia or torsion angles, which allow *false positive* pairs of non-equivalent backbones $S \ncong Q$ 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* (defined as 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 past descriptors of backbones, see the review of past work in section 2.

Problem 1.2 (mapping the Backbone Rigid Space BRIS_m by geographic-style invariant coordinates)**.** *Design a map* $I : \text{BRIS}_m \to \mathbb{R}^k$ *satisfying the following conditions.*

(a) ***Completeness****: any backbones* $S \cong Q$ *are rigidly equivalent if and only if* $I(S) = I(Q)$*, i.e. the invariant descriptor* 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 a linear time* $O(j)$*.*

(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 any backbone of* m *residues.*

The completeness in 1.2(a) means that I is the strongest possible invariant and hence distinguishes *all* backbones that can not be exactly matched by rigid motion.

The reconstruction in 1.2(b) is more practical because a complete invariant 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(d) are essential because if the triangle axiom fails with any positive error, results of clustering by k-means and DBSCAN based on d may not be trustworthy [10].

The continuity in 1.2(c) fails for invariants based on principal directions that can discontinuously change (or become ill-defined) in degenerate cases of high symmetry. 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 α-carbon $A_i(Q)$ in the i-th residue of Q. Conditions 1.2(c,d) guarantee the Lipschitz continuity of I and its inverse on the image $I(\text{BRIS}_m) \subset \mathbb{R}^k$.

New condition 1.2(e) is important for identifying *secondary* structures that are frequent semi-rigid subchains such as α-helices and β-strands [13]. 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 $O(m)$ for thousands of residues.

The key contribution is the *Backbone Rigid Invariant* $\text{BRI} : \text{BRIS}_m \to \mathbb{R}^{9m-6}$ that solves Problem 1.2. Conditions 1.2(d,e) are stated for the first time to the best of our knowledge. The numerical components of BRI play the role of geographic-style coordinates on the space BRIS_m of rigid classes of backbones consisting of m triplets of atoms from m residues. Geographic-style maps of the full Backbone Rigid Space $\text{BRIS} = \bigcup_{m \geq 2} \text{BRIS}_m$ will be visualized by 2D projections on pairs of averaged invariants.

2 Past work on similarities and invariants of proteins

In the more general context of crystal structures, a canonical description in a reduced unit cell [14] can be achieved by the program TYPIX [15] for inorganic compounds and ACHESYM [16] 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 experimentally known at least since 1965 [17, p. 80] and was resolved only for generic crystals [18].

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 [19] $\mathrm{TM} = \max \left\{ \frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1+(d_i/d_0)^2} \right\} \in [0,1]$ 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 $\mathrm{TMD} = 1 - \mathrm{TM}$ so that $\mathrm{TMD}(S, S) = 0$ for any structure S. Unfortunately, this and many other conversions such as $-\log(\mathrm{TM})$ 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 $1 - \mathrm{TM}$ 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 $-\log(\mathrm{TM})$.

If the triangle axiom fails with any additive error, results of the clustering algorithms k-means and DBSCAN can be arbitrarily pre-determined [10]. The authors of another similarity LDDT (Local Distance Difference Test) concluded in [9, 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 [20]. This RMSD is slow to compute for all-vs-all comparisons of proteins in the PDB so many pairs with RMSD= 0 were unnoticed. If the order of atoms is not respected, the optimal global alignment is NP-complete [21]. Any attempt to apply "random rotations to each protein domain structure as part of the model training routine" creates many more structures [22] that look different but should be considered rigidly equivalent.

More recently, the PDB implemented a structural superposition [23] 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. The description in [23] 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 protein folding [3, 5, 24] reviews [7, 25, 26].

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 latitude and longitude on Earth.

Protein backbones are traditionally represented by *torsion* (dihedral) angles φ_i, ψ_i visualized in Ramachandran plots [27]. 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 because a distance between any successive points can be arbitrarily changed while preserving all angles.

For proteins, even if all bond lengths and angles are fixed at ideal values, all torsion angles still should be ordered according to 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. AlphaFold2 [3] and more recent advances [28] 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 the 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 [29] 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 [30, 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 proteins consisting of thousands of atoms.

If a protein backbone is considered a cloud of unordered points, such clouds of different sizes can be visualized by eigenvalue invariants (or moments of inertia) characterizing the elongation of the cloud along principal directions In 1996, probably the first map of all 4K entries in the PDB appeared in [31, Fig. 5] by using the two largest eigenvalues, see the recent updates in [32, Fig. 2] and PDB-Explorer [33]. In 2020, Holm called for faster visualization of the protein space [34]: "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 volumes."

In 1977, Kendall [35] started to study configuration spaces of ordered points modulo rigid motion in $\mathbb{R}^n$ under the name of *size-and-shape spaces* [36]. 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 obstacle prevented a simple solution to Problem 1.2.

3 Completeness of 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 protein backbone $S \subset \mathbb{R}^3$ have m ordered triplets of atoms N_i, A_i, and C_i for $i = 1, \dots, m$. In the basis obtained by Gaussian orthogonalization of $\overrightarrow{A_iN_i}, \overrightarrow{A_iC_i}$, the vector $\overrightarrow{A_iN_i}$ has the coordinates $x(AN_i) = |\overrightarrow{A_iN_i}|$ and 0, while $\overrightarrow{A_iC_i}$ has $x(AC_i) = \dfrac{\overrightarrow{A_iC_i} \cdot \overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$ and $y(AC_i) = \Big|\overrightarrow{A_iC_i} - x(AC_i)\dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}\Big|$. The* triangular invariant TRIN(S) *is the $m \times 3$ matrix whose i-th row consists of $x(AN_i), x(AC_i), y(AC_i)$.*

The i-th row of TRIN(S) uniquely determines the shape of $\triangle N_i A_i C_i$. Many past approaches including AlphaFold2 [3] assumed that all these residue triangles are rigidly equivalent. To test this assumption on the PDB, we filter out unsuitable chains below.

Protocol 3.2 (selecting a subset of 707K+ chains in the PDB)**.** *A chain from any PDB entry is filtered out by these steps: (1) 4513 non-proteins (the _entity is not 'protein'); (2) 178153 disordered chains (atom occupancy less than 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.*

On May 4, 2024, the PDB had 213,191 entries with 1,091,420 chains, Protocol 3.2 produced $104,688 \approx 49\%$ entries including $707410 \approx 65\%$ chains within 4 hours 48 min 11 sec. All experiments were run on CPU Core i7-11700 @2.50GHz RAM 32Gb.

Example 3.3 (variability of residue triangles)**.** *Fig. 2 (row 1) shows the heat maps (in the logarithmic scale) of the invariants $x(AN_i), x(AC_i), y(AC_i)$ across all (about 110 million) residues in the 707K+ cleaned backbones obtained from the PDB by Protocol 3.2. Though standard deviations of these invariants are about* 0.01Å*, the maximum deviations of $x(AN_i), x(AC_i), y(AC_i)$ are about* $1.2, 1.7, 2.7$Å*, respectively.*

Table 1 below shows the TRIN and BRI (see Definition 3.4) invariants for the first 3 residues of two hemoglobin chains A in proteins 2hhb and 1hho (later referenced in Example 5.2). The overall TRIN and BRI average values (mean) are in the last rows.

Res	$x(AN)$	$x(CN)$	$y(CN)$	$x(N)$	$y(N)$	$z(N)$	$x(A)$	$y(A)$	$z(A)$	$x(C)$	$y(C)$	$z(C)$
V	1.45	-0.54	1.44	1.45	0	0	0	0	0	-0.54	1.44	0
L	1.47	-0.50	1.47	-0.91	0.25	-0.90	-0.64	1.32	0.02	-1.10	0.01	1.10
S	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	$x(AN)$	$x(CN)$	$y(CN)$	$x(N)$	$y(N)$	$z(N)$	$x(A)$	$y(A)$	$z(A)$	$x(C)$	$y(C)$	$z(C)$
V	1.48	-0.51	1.46	1.48	0.00	0.00	0.00	0.00	0.00	-0.51	1.46	0.00
L	1.49	-0.55	1.42	-0.14	0.66	1.16	-0.69	1.31	0.19	-1.51	-0.16	-0.03
S	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

Table 1. The TRIN and BRI invariants for the first 3 residues of 2hhb (top) and 1hho (bottom) along with their mean values for all 141 residues (as per Definition 5.1). Some values differ significantly between the two chains.

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. 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 all residue bases well-defined in Definition 3.4 below.

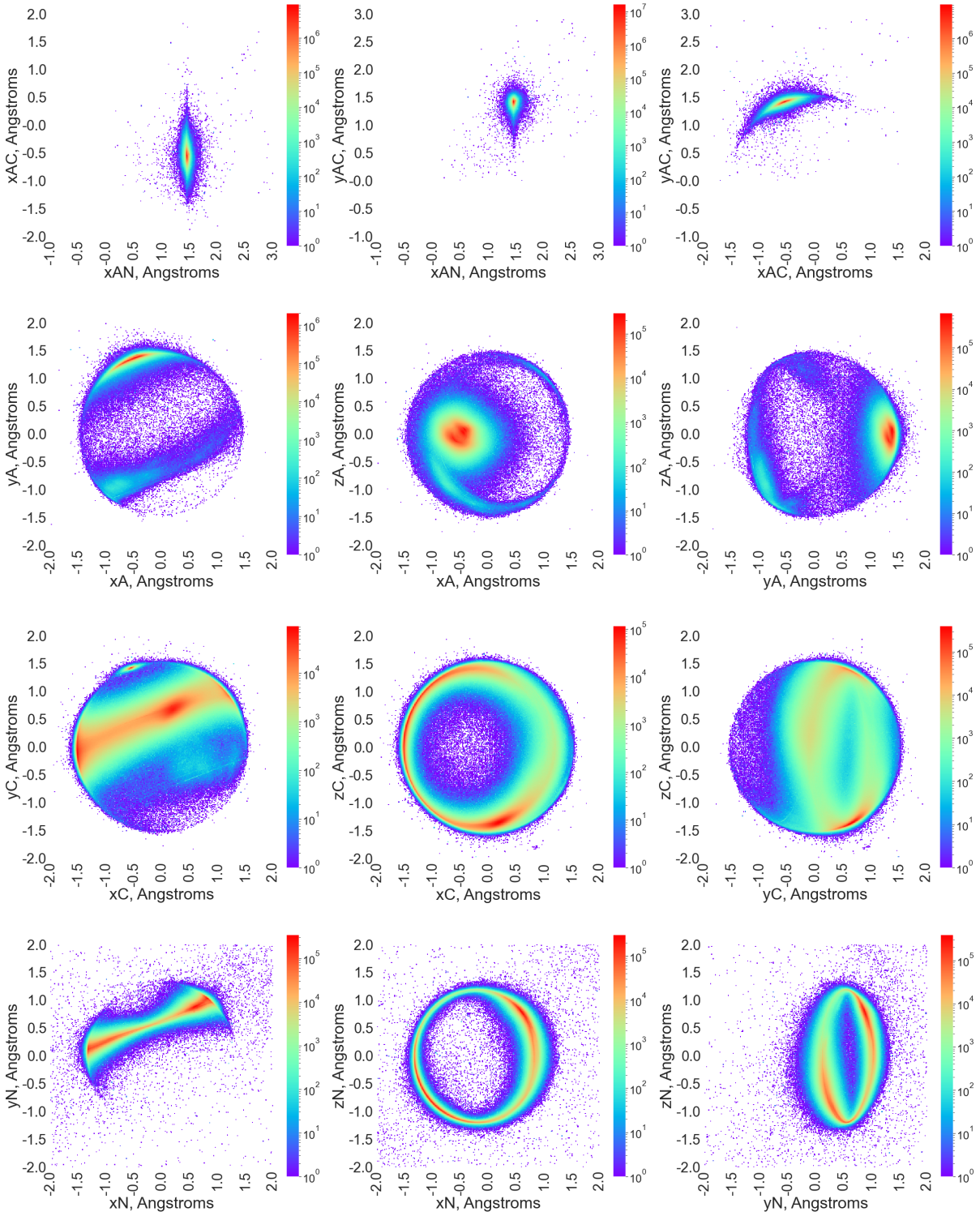

Figure 2. Heatmaps of the invariants TRIN and BRI across all residues in the 707K+ chains obtained by Protocol 3.2. The color indicates the number of residues whose pair of specified invariants is discretized to each pixel. **Row 2**: pairs of invariants $x(A_i), y(A_i), z(A_i)$ of the bonds N_iA_i from each nitrogen N_i to the α-carbon A_i in the same residue, see Definition 3.4. **Row 3**: pairs of invariants $x(C_i), y(C_i), z(C_i)$ of the bonds A_iC_i from each α-carbon to the carbonyl carbon C_i in the same residue. **Row 4**: pairs of invariants $x(N_i), y(N_i), z(N_i)$ of the bonds C_iN_{i+1}.

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* $i = 2, \ldots, m$ *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 the vector* $\overrightarrow{C_{i-1}N_i}$ *with the head* N_i*, followed by the six coordinates* $x(A_i), \ldots, z(C_i)$. *For* $i = 1$*, the first row of* BRI(S) *has only the three non-zero coordinates* $x(N_1) = x(AN_1)$, $x(C_1) = x(AC_1)$, $y(C_1) = y(AC_1)$ *from the first row of the triangular invariant* TRIN(S) *in Definition 3.1.*

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 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 $L_\infty(\text{BRI}(S), \text{BRI}(Q))$ 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 BRI(S).

Theorem 3.5 (completeness, reconstruction, and subchains)**.** ***(a)*** *The* $m \times 9$ *matrix* BRI(S) *in Definition 3.4 is a complete invariant under rigid motion, so any backbones* $S, Q \subset \mathbb{R}^3$ *of* m *residues are related by rigid motion if and only if* $\text{BRI}(S) = \text{BRI}(Q)$.

(b) *For any backbone* S *of* m *residues,* BRI(S)*, the metric* L_∞ *between invariants, and a reconstruction of* $S \subset \mathbb{R}^3$ *from* 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* BRI(Q) *consists of the first* j *rows of* BRI(S). *If* Q *starts from the* i*-th residue of* S *for* $i > 1$*, the rows* $2, \ldots, j$ *of* BRI(Q) *coincide with the rows* $i+1, \ldots, i+j-1$ *of* BRI(S)*, and the 1st row of* BRI(Q) *are computed from the* i*-th row of* BRI(S) *in a constant time. Hence* BRI(Q) *is computed from* BRI(S) *in time* $O(j)$.

Proof of Theorem 3.5. **(a)** 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 given backbone $S \subset \mathbb{R}^3$ is a linear map acting on every orthonormal basis $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i$. Hence the image under f of any vector $\boldsymbol{p} = x\boldsymbol{u}_i + y\boldsymbol{v}_i + z\boldsymbol{w}_i$ has the same coordinates in the rigidly transformed basis: $f(\boldsymbol{p}) = xf(\boldsymbol{u}_i) + yf(\boldsymbol{v}_i) + zf(\boldsymbol{w}_i)$.

(b) For any residue of 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 BRI(S) is $O(m)$. The metric L_∞ has a linear time in the length of vectors.

The completeness will follow by showing that any backbone $S \subset \mathbb{R}^3$ can be efficiently reconstructed from BRI(S), uniquely after fixing the first residue whose shape is determined by the three non-zero values in the first row of 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 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 its coordinates $x(C_1)$ and $y(C_1)$ written in Definition 3.4.

After fixing the first three atoms 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 BRI(S). Indeed, the position of the next atom N_2 is obtained from C_1 by adding the vector $\overrightarrow{C_1N_2}$, whose coordinates are the first three elements in the 2nd row of BRI(S). Then A_2 is obtained from N_1 by adding $\overrightarrow{N_2A_2}$, whose coordinates are the second three elements in the 2nd row of BRI(S). Then C_2 is obtained from A_2 by adding $\overrightarrow{A_2C_2}$ and so on.

(c) Since the complete invariant BRI(S) of a backbone S is locally defined by determining any i-th residue triangle in the basis of the previous $(i-1)$-st triangle, all rows BRI(Q) of any subchain Q in S coincide with the corresponding rows of 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 3 non-zero invariants in the first row of Q can be obtained from the i-th row of t TRIN(S) whose values are expressed in terms of the vectors $\vec{N_iA_i}$ and $\vec{A_iC_i}$ in Definition 3.1. This computation needs only a constant time independent of j because the coordinates of the vectors $\vec{A_iN_i}$ and $\vec{A_iC_i}$ are given in the i-th row of BRI(S). □

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

Proof of Corollary 3.6. To prove that $\overline{\text{BRI}}(S) := \text{BRI}(\bar{S})$ is obtained from BRI(S) by reversing the signs in all z-columns of 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 $\triangle N_iA_iC_i$. This mirror 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 $\overrightarrow{C_iN_{i+1}} = x(N_{i+1})\boldsymbol{u}_i + y(N_{i+1})\boldsymbol{v}_i + z(N_{i+1})\boldsymbol{w}_i$ 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 $\triangle N_1A_1C_1$ 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 BRI(S) and BRI($\bar{S}$) coincide, while all z-coordinates are zeros. Finally, the unordered pair of BRI(S) and $\overline{\text{BRI}}(S)$ 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 the invariant BRI(S) and $\overline{\text{BRI}}(S)$ 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).

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 across all real protein backbones, respectively. The maximum bond lengths $L_{A,C}, L_{C,N}$ are similarly defined for other types of bonds in a backbone.

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 be the minimum height in the triangle $\triangle N_i A_i C_i$ at the atom 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 $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) \le \lambda\varepsilon$.*

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

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

Proof. The triangle inequality for the Euclidean distance implies that $|\boldsymbol{u}| \le |\boldsymbol{u} - \boldsymbol{v}| + |\boldsymbol{v}|$, so $|\boldsymbol{u}| - |\boldsymbol{v}| \le |\boldsymbol{u} - \boldsymbol{v}|$. Swapping the vectors, we get $|\boldsymbol{v}| - |\boldsymbol{u}| \le |\boldsymbol{u} - \boldsymbol{v}|$. Combining the inequalities $\pm(|\boldsymbol{u}| - |\boldsymbol{v}|) \le |\boldsymbol{u} - \boldsymbol{v}|$, we conclude that $|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| \le |\boldsymbol{u} - \boldsymbol{v}|$ 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. $|A - A'| \le \varepsilon$, $|B - B'| \le \varepsilon$. Then $|\overrightarrow{A'B'} - \overrightarrow{AB}| \le 2\varepsilon$.*

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

Lemma 4.4 (perturbation of a normalized vector)**.** *Let $\boldsymbol{u}$ be a δ-perturbation of a vector $\boldsymbol{v} \in \mathbb{R}^n$, i.e. $|\boldsymbol{u} - \boldsymbol{v}| \le \delta$. Then $\left|\dfrac{\boldsymbol{u}}{|\boldsymbol{u}|} - \dfrac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| \le \dfrac{2\delta}{l}$, 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 $\left|\dfrac{\boldsymbol{u}}{|\boldsymbol{u}|} - \dfrac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| \le \dfrac{4\varepsilon}{l_{N,A}}$, 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 $\left|\dfrac{\boldsymbol{u}}{|\boldsymbol{u}|} - \dfrac{\boldsymbol{v}}{|\boldsymbol{v}|}\right| =$

$$= \left|\frac{|\boldsymbol{v}|\boldsymbol{u} - |\boldsymbol{u}|\boldsymbol{v}}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|}\right| = \frac{|\,(|\boldsymbol{v}| - |\boldsymbol{u}|)\boldsymbol{u} + |\boldsymbol{u}|(\boldsymbol{u} - \boldsymbol{v})\,|}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|} \le \frac{|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| \cdot |\boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{u}| \cdot |\boldsymbol{v}|} =$$

$$= \frac{|\,|\boldsymbol{u}| - |\boldsymbol{v}|\,| + |\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{v}|} \le \frac{2|\boldsymbol{u} - \boldsymbol{v}|}{|\boldsymbol{v}|} \le \frac{2\delta}{|\boldsymbol{v}|} = \frac{2\delta}{l},$$

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

Lemma 4.5 (product perturbations)**.** *For any $\boldsymbol{u}, \boldsymbol{u}', \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$, if $|\boldsymbol{v}'| = |\boldsymbol{v}| = 1$, then*
(a) $|(\boldsymbol{u}' \cdot \boldsymbol{v}') - (\boldsymbol{u} \cdot \boldsymbol{v})| \le |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$,
(b) $|(\boldsymbol{u}' \times \boldsymbol{v}') - (\boldsymbol{u} \times \boldsymbol{v})| \le |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$,
(c) $|(\boldsymbol{u}' \cdot \boldsymbol{v}')\boldsymbol{v}' - (\boldsymbol{u} \cdot \boldsymbol{v})\boldsymbol{v}| \le |\boldsymbol{u}' - \boldsymbol{u}| + 2|\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$.

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

$$|(\boldsymbol{u}' \cdot \boldsymbol{v}') - (\boldsymbol{u} \cdot \boldsymbol{v})| = |(\boldsymbol{u}' - \boldsymbol{u}) \cdot \boldsymbol{v}' + \boldsymbol{u} \cdot (\boldsymbol{v}' - \boldsymbol{v})| \le |(\boldsymbol{u}' - \boldsymbol{u}) \cdot \boldsymbol{v}'| + |\boldsymbol{u} \cdot (\boldsymbol{v}' - \boldsymbol{v})| \le$$

$\le |\boldsymbol{u}' - \boldsymbol{u}| \cdot |\boldsymbol{v}'| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}| = |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$ due to $|\boldsymbol{v}'| = 1$, which proves (a).

(b) is proved as (a) after replacing the scalar product with the vector product.

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

$$|(\boldsymbol{u}' \cdot \boldsymbol{v}')\boldsymbol{v}' - (\boldsymbol{u} \cdot \boldsymbol{v})\boldsymbol{v}| = |(\boldsymbol{u}' \cdot \boldsymbol{v}' - \boldsymbol{u} \cdot \boldsymbol{v})\boldsymbol{v}' + (\boldsymbol{u} \cdot \boldsymbol{v})(\boldsymbol{v}' - \boldsymbol{v})| \le |\boldsymbol{u}' \cdot \boldsymbol{v}' - \boldsymbol{u} \cdot \boldsymbol{v}| \cdot |\boldsymbol{v}'| + |\boldsymbol{u} \cdot \boldsymbol{v}| \cdot |\boldsymbol{v}' - \boldsymbol{v}| \le$$

$$\le |\boldsymbol{u}' \cdot \boldsymbol{v}' - \boldsymbol{u} \cdot \boldsymbol{v}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}| \cdot |\boldsymbol{v}' - \boldsymbol{v}| \le |\boldsymbol{u}' - \boldsymbol{u}| + 2|\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|.$$ □

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) $|\boldsymbol{u}'_i - \boldsymbol{u}_i| \le \dfrac{4\varepsilon}{l_{N,A}}$*, where $l_{N,A}$ is the minimum bond length between N_i, A_i for all residues;*

(b) $|\boldsymbol{v}'_i - \boldsymbol{v}_i| \le \dfrac{8\varepsilon}{h}(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$*, where $L_{A,C}$ is the maximum bond length between atoms A_i and C_i, while h is the minimum height in $\triangle N_iA_iC_i$ at C_i for all residues;*

(c) $|\boldsymbol{w}'_i - \boldsymbol{w}_i| \le 4\varepsilon K$*, 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 $\boldsymbol{u}_i = \dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$ by Lemma 4.4 satisfies $|\boldsymbol{u}'_i - \boldsymbol{u}_i| \le \dfrac{4\varepsilon}{l_{N,A}}$.

(b) The second vector is $\boldsymbol{v}_i = \dfrac{\boldsymbol{h}_i}{|\boldsymbol{h}_i|}$ for $\boldsymbol{h}_i = \overrightarrow{A_iC_i} - b_i\overrightarrow{A_iN_i}$, $b_i = \dfrac{\overrightarrow{A_iC_i} \cdot \overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|^2}$. Set $\boldsymbol{p}_i = \overrightarrow{A_iC_i}$ and $\boldsymbol{q}_i = \dfrac{\overrightarrow{A_iN_i}}{|\overrightarrow{A_iN_i}|}$, 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 $L_{A,C}$. Lemma 4.5(c) implies that

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

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

$$|\boldsymbol{h}'_i - \boldsymbol{h}_i| = |\boldsymbol{p}'_i - b'_i\overrightarrow{A'_iN'_i} - (\boldsymbol{p}_i - b_i\overrightarrow{A_iN_i})| \le |\boldsymbol{p}'_i - \boldsymbol{p}_i| + |b'_i\overrightarrow{A'_iN'_i} - b_i\overrightarrow{A_iN_i}| \le$$

$\le 2\varepsilon + 2\varepsilon + \varepsilon\dfrac{L_{A,C}}{l_{N,A}} = 4\varepsilon(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$. 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 $h \le |\boldsymbol{h}_i|$, Lemma 4.4 for $\delta = 4\varepsilon(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$ implies that $|\boldsymbol{v}'_i - \boldsymbol{v}_i| = \left|\dfrac{\boldsymbol{h}'_i}{|\boldsymbol{h}'_i|} - \dfrac{\boldsymbol{h}_i}{|\boldsymbol{h}_i|}\right| \le \dfrac{2\delta}{h} \le \dfrac{8\varepsilon}{h}(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$.

(c) The third basis vector $\boldsymbol{w}_i = \boldsymbol{u}_i \times v_i$ has a perturbation estimated by Lemma 4.5(b):

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

where $K = \dfrac{1}{l_{N,A}} + \dfrac{2}{h}\Big(1 + 2\dfrac{L_{A,C}}{l_{N,A}}\Big)$ as required. □

Proof of Theorem 4.1. In the backbone Q, let N'_i, A'_i, C'_i denote ε-perturbations of atoms $N_iA_iC_i$ from the backbone S for $i = 1, \dots, m$. We prove that any coordinate of BRI(S) changes by at most $\lambda\varepsilon$ for the given Lipschitz constant λ. The first coordinate $x(N_1)$ changes up at most 2ε because $|x(N'_1) - x(N_1)| = \big|\,|\overrightarrow{A'_1N'_1}| - |\overrightarrow{A_1N_1}|\,\big| \le 2\varepsilon$ 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 $|x(C_1') - x(C_1)| = |\boldsymbol{u}' \cdot \boldsymbol{v}' - \boldsymbol{u} \cdot \boldsymbol{v}| \le |\boldsymbol{u}' - \boldsymbol{u}| + |\boldsymbol{u}| \cdot |\boldsymbol{v}' - \boldsymbol{v}|$ by Lemma 4.5(a). Lemma 4.3 implies that $|\boldsymbol{u}' - \boldsymbol{u}| \le 2\varepsilon$. Lemma 4.4 for $u = \overrightarrow{A_1'N_1'}$ and $v = \overrightarrow{A_1N_1}$ implies that $|\boldsymbol{v}' - \boldsymbol{v}| \le \dfrac{4\varepsilon}{l_{N,A}}$, 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 $|x(C_1') - x(C_1)| \le 2\varepsilon(1 + 2\dfrac{L_{A,C}}{l_{N,A}})$. In the notations above, the last non-zero coordinate in the first row of 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:

$$|y(C_1') - y(C_1)| = |\,|\boldsymbol{u}' - x(C_1')\boldsymbol{v}'| - |\boldsymbol{u} - x(C_1)\boldsymbol{v}|\,| \le |\boldsymbol{u}' - x(C_1')\boldsymbol{v}' - (\boldsymbol{u} - x(C_1)\boldsymbol{v})| \le$$

$$\le |\boldsymbol{u}' - \boldsymbol{u}| + |x(C_1')\boldsymbol{v}' - x(C_1)\boldsymbol{v}| \le 2\varepsilon + |(x(C_1') - x(C_1))\boldsymbol{v}' + x(C_1)(\boldsymbol{v}' - \boldsymbol{v})| \le$$

$$2\varepsilon + |x(C_1') - x(C_1)| + |x(C_1)| \cdot |\boldsymbol{v}' - \boldsymbol{v}| \le 2\varepsilon + 2\varepsilon(1 + 2\frac{L_{A,C}}{l_{N,A}}) + |\overrightarrow{A_1C_1'}|\frac{4\varepsilon}{l_{N,A}} \le 4\varepsilon(1 + 2\frac{L_{A,C}}{l_{N,A}}),$$

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

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

$$|x(N_i') - x(N_i)| = |\overrightarrow{C_{i-1}'N_i'} \cdot \boldsymbol{u}_i' - \overrightarrow{C_{i-1}N_i} \cdot \boldsymbol{u}_i| \le |\overrightarrow{C_{i-1}'N_i'} - \overrightarrow{C_{i-1}N_i}| + |\overrightarrow{C_{i-1}N_i}| \cdot |\boldsymbol{u}_i' - \boldsymbol{u}_i| \le$$

$$\le 2\varepsilon + L_{C,N}\frac{4\varepsilon}{l_{N,A}} = 2\varepsilon(1 + 2\frac{L_{C,N}}{l_{N,A}}), \text{ due to the upper bound } |\overrightarrow{C_{i-1}N_i}| \le L_{C,N}.$$

For the other coordinates y, z, similarly use Proposition 4.6(b,c), respectively, as follows:

$$|y(N_i') - y(N_i)| = |\overrightarrow{C_{i-1}'N_i'} \cdot \boldsymbol{v}_i' - \overrightarrow{C_{i-1}N_i} \cdot \boldsymbol{v}_i| \le |\overrightarrow{C_{i-1}'N_i'} - \overrightarrow{C_{i-1}N_i}| + |\overrightarrow{C_{i-1}N_i}| \cdot |\boldsymbol{v}_i' - \boldsymbol{v}_i| \le$$

$$\le 2\varepsilon + L_{C,N} \cdot \frac{8\varepsilon}{h}(1 + 2\frac{L_{A,C}}{l_{N,A}}) = 2\varepsilon(1 + 4\frac{L_{C,N}}{h}(1 + 2\frac{L_{A,C}}{l_{N,A}}).$$

$$|z(N_i') - z(N_i)| = |\overrightarrow{C_{i-1}'N_i'} \cdot \boldsymbol{w}_i' - \overrightarrow{C_{i-1}N_i} \cdot \boldsymbol{w}_i| \le |\overrightarrow{C_{i-1}'N_i'} - \overrightarrow{C_{i-1}N_i}| + |\overrightarrow{C_{i-1}N_i}| \cdot |\boldsymbol{w}_i' - \boldsymbol{w}_i| \le$$

$$\le 2\varepsilon + L_{C,N} \cdot 4\varepsilon K = 2\varepsilon(1 + 2L_{C,N}K), \text{ where } K = \frac{1}{l_{N,A}} + \frac{2}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big).$$

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 $L_\infty(\text{BRI}(S), \text{BRI}(Q)) \le \lambda\varepsilon$, 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)**.** *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 $\varepsilon = 0.01, 0.02, \dots, 0.1$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 $\lambda \approx 4$ in this experiment.*

Because the metric L_∞ between $m \times 9$ matrices can be computed in $O(m)$ time, Theorem 4.1 also completes condition (1.2f) in Problem 1.2. Theorem 4.8 will prove the atom matching 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 $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q)) < \delta$, 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 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 estimate the deviations of atoms in the second residue:

$$|N_2' - N_2| = |x(N_2')\boldsymbol{u}_1 + y(N_2')\boldsymbol{v}_1 + z(N_2')\boldsymbol{w}_1 - x(N_2)\boldsymbol{u}_1 - y(N_2)\boldsymbol{v}_1 - z(N_2)\boldsymbol{w}_1| =$$

$$= \sqrt{|x(N_2') - x(N_2)|^2 + |y(N_2') - y(N_2)|^2 + |z(N_2') - z(N_2)|^2} \le \sqrt{\delta^2 + \delta^2 + \delta^2} = \delta\sqrt{3}.$$

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 upper bound on deviations of atoms by induction on $m \ge 2$.

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

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

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 $i \ge 2$. The inductive step is for $i + 1$. Proposition 4.6 estimates deviations of vectors in the second basis:

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

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

$$|N_{i+1}' - N_{i+1}| = |x(N_{i+1}')\boldsymbol{u}_i' + y(N_{i+1}')\boldsymbol{v}_i' + z(N_{i+1}')\boldsymbol{w}_i' - x(N_{i+1})\boldsymbol{u}_i - y(N_{i+1})\boldsymbol{v}_i - z(N_{i+1})\boldsymbol{w}_i| =$$

$$= |(x(N'_{i+1})\boldsymbol{u}'_i - x(N_{i+1})\boldsymbol{u}_i) + (y(N'_{i+1})\boldsymbol{v}'_i - y(N_{i+1})\boldsymbol{v}_i) + (z(N'_{i+1})\boldsymbol{w}'_i - z(N_{i+1})\boldsymbol{w}_i)| =$$

$$= \Big|\big(x(N'_{i+1}) - x(N_{i+1})\big)\boldsymbol{u}'_i + x(N_{i+1})(\boldsymbol{u}'_i - \boldsymbol{u}_i) + \big(y(N'_{i+1}) - y(N_{i+1})\big)\boldsymbol{v}'_i +$$

$$+ y(N_{i+1})(\boldsymbol{v}'_i - \boldsymbol{v}_i) + \big(z(N'_{i+1}) - z(N_{i+1})\big)\boldsymbol{w}'_i + z(N_{i+1})(\boldsymbol{w}'_i - \boldsymbol{w}_i)\Big| \le$$

$$\le \Big|\big(x(N'_{i+1}) - x(N_{i+1})\big)\boldsymbol{u}'_i + \big(y(N'_{i+1}) - y(N_{i+1})\big)\boldsymbol{v}'_i + \big(z(N'_{i+1}) - z(N_{i+1})\big)\boldsymbol{w}'_i\Big| +$$

$$\Big|x(N_{i+1})(\boldsymbol{u}'_i - \boldsymbol{u}_i)\Big| + \Big|y(N_{i+1})(\boldsymbol{v}'_i - \boldsymbol{v}_i)\Big| + \Big|z(N_{i+1})(\boldsymbol{w}'_i - \boldsymbol{w}_i)\Big|.$$

In the last expression, the first row contains 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 $|x(N_{i+1})| \le L_{C,N}$ and Proposition 4.6 as follows:

$$|x(N_{i+1})| \cdot |\boldsymbol{u}'_i - \boldsymbol{u}_i| \le L_{C,N}\frac{4\varepsilon}{l_{N,A}}, \quad |y(N_{i+1})| \cdot |\boldsymbol{v}'_i - \boldsymbol{v}_i| \le L_{C,N}\frac{8\varepsilon}{h}\Big(1 + 2\frac{L_{A,C}}{l_{N,A}}\Big),$$

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

Taking the sum of the above estimates, the final deviation of nitrogens is

$$|N'_{i+1} - N_{i+1}| \le \sqrt{3}\delta + 4\varepsilon L_{C,N}\Big(\frac{1}{l_{N,A}} + \frac{2}{h}\big(1 + 2\frac{L_{A,C}}{l_{N,A}}\big) + \frac{1}{l_{N,A}} + \frac{2}{h}\big(1 + 2\frac{L_{A,C}}{l_{N,A}}\big)\Big) =$$

$$= \sqrt{3}\delta + 8L_{C,N}\varepsilon K \le \sqrt{3}(1 + 8LK(1 + \cdots + (8LK)^{i-2})\delta = \sqrt{3}(1 + \cdots + (8LK)^{i-1})\delta.$$

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 $L_{N,A}, L_{A,C} \le L$. The bound for $i = m$ is $\sqrt{3}(1 + \cdots + (8LK)^{m-2})\delta = \sqrt{3}\dfrac{(8LK)^{m-1} - 1}{8LK - 1}\delta$. Now consider the modified invariant $\widehat{\mathrm{BRI}}(S)$ obtained by multiplying the i-th row of BRI(S) by $\dfrac{(8LK)^{i-1} - 1}{8LK - 1}$ for $i = 2, \ldots, m$. Then the δ-closeness of the corresponding invariant components in the metric L_∞ means smaller deviations $|x(N'_i) - x(N_i)| \le \delta\dfrac{8LK - 1}{(8LK)^{i-1} - 1}$, similarly for other components. This extra multiplicative factor gives the bound $|N'_{i+1} - N_{i+1}| \le \sqrt{3}\delta$, 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 BRI(S) to the modified invariant $\widehat{\mathrm{BRI}}(S)$ depending on a row index of BRI(S) to guarantee the smaller Lipschitz constant $\sqrt{3}$.

5 Averaged invariants, diagrams, and barcodes

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

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 Brain$(S) \in \mathbb{R}^9$ *is the vector of nine column averages in* BRI(S) *excluding the first row. The standard deviation can be computed in a similar way. The* backbone invariant diagram BID(S) *consists of nine polygonal curves going through the points $(i, c(i))$, $i = 2, \ldots, m$, where c is one of the coordinates (columns) of* 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 $i = 1, \ldots, m$. The resulting three color bars for the ordered atoms N, A, C form the* backbone invariant barcode BIB(S)*, see Fig. 3 (bottom).*

While the complete invariants 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 map, see Fig.4.

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. The TRIN and BRI invariants are shown for the first 3 residues of 2hhb and 1hho in Table 1.*

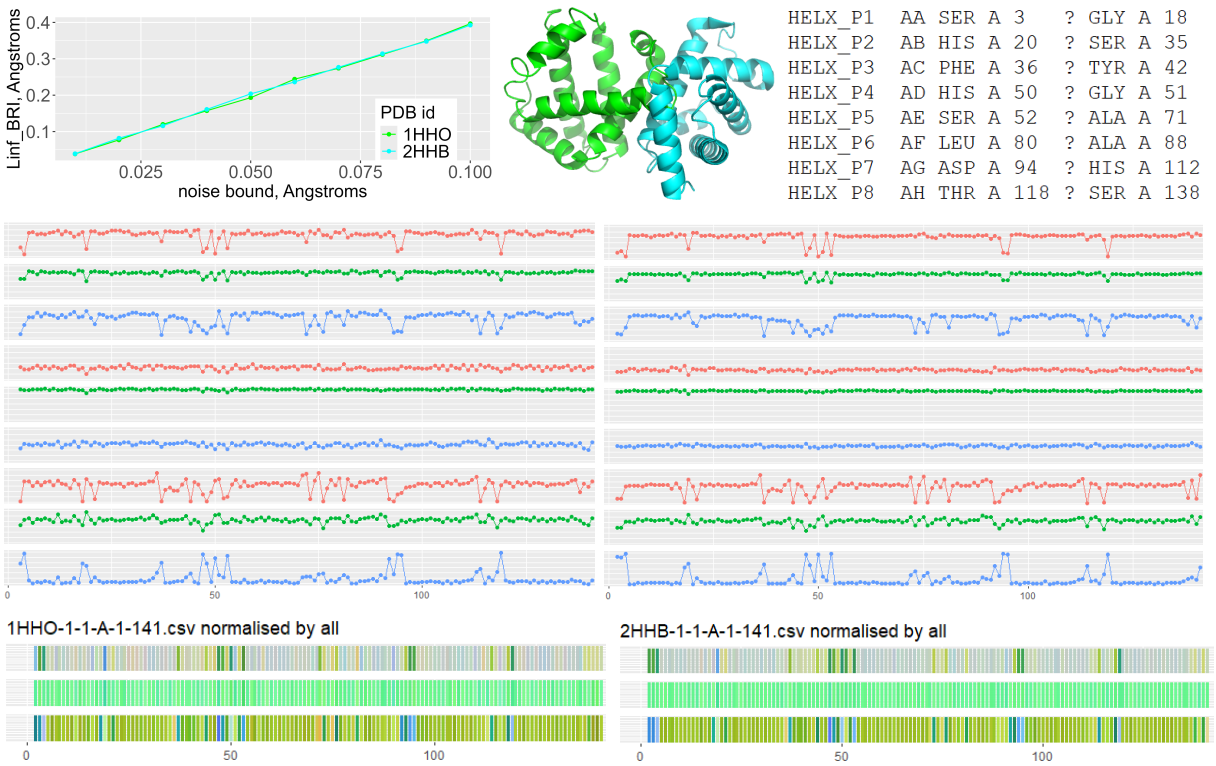

Figure 3. **Row 1**: the Lipschitz continuity of BRI from Theorem 4.1 is illustrated on the left by perturbing hemoglobins in Example 4.7, whose main chains A of 141 residues are shown in the middle (oxygenated 1hho in green, standard 2hhb in cyan) and eight α-helices found by [37] and extract from the PDB 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.

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 and 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) [37], 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 and 51, and HELX_P5 of length 20 over residue indices $i = 52, \dots, 71$. Fig. 3 shows that a 'constant' interval of little noise appears only for $i = 54, \dots, 70$. Hence new invariants allow a more objective detection of secondary structures, which will be explored in future work.*

6 Duplicates with identical coordinates in the PDB

The linear time of the complete invariant $\mathrm{BRI}(S)$ 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 provides a faster computable lower bound for the metric $L_\infty(\mathrm{BRI}(S), \mathrm{BRI}(Q))$. in terms of the average invariant Brain.

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

Proof of Lemma 6.1. If 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 $j = 1, \dots, 9$, their averages of the j-th columns differ by at most δ because

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

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

The complete invariants and their summaries (averages and deviations) were computed in 3 hours 18 min 21 sec. After comparing all (883+ million) pairs of same-length backbones within 2.5 hours, we found 13403 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.*

In nine pairs, geometrically identical chains unexplainably differ in the sequences of amino acids, see Fig. 5 (left). In a similar case [18], when five pairs of unexpected duplicates were found in the Cambridge Structural Database (CSD), all involved crystallographers concluding 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 [38]. We e-mailed all authors of structures marked in Fig. 5 (left) whose contacts we found. Two authors replied with details and confirmed that their PDB entries should be corrected, see details in appendix A.

The duplicates in Fig. 5 were shown to the PDB validation team, who didn't know about the found coincidences (in coordinates) and differences (in amino acids) because

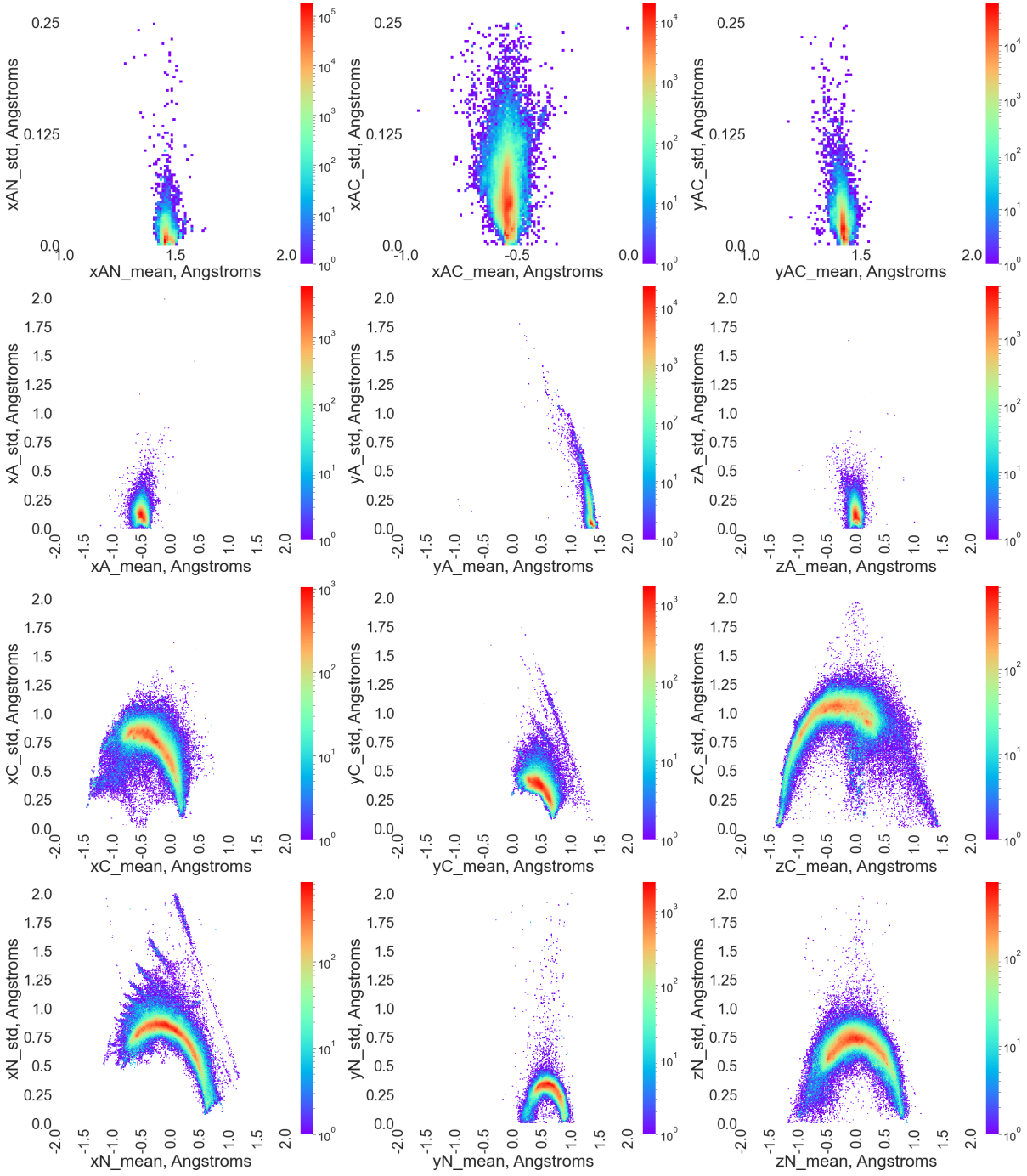

Figure 4. Heatmaps of average/standard deviation of the invariants across all 707K+ chains obtained by Protocol 3.2. The color indicates the number of chains whose pair of specified average/standard deviation is discretized to each pixel. **Row 2**: average/standard deviation pairs of the invariants $x(A_i), y(A_i), z(A_i)$ of the bonds N_iA_i from nitrogen to α-carbon in Definition 3.4. **Row 3**: average/standard deviation pairs of the invariants $x(C_i), y(C_i), z(C_i)$ of the bonds A_iC_i from α-carbon to the carbonyl carbon. **Row 4**: average/standard deviation pairs of the invariants $x(N_i), y(N_i), z(N_i)$ of the peptide bonds C_iN_{i+1}.

the validation is currently done for an individual protein only (checking atom clashes etc). The recently published method [39] didn't report any duplicates. Right now anyone can

PDB id1 & chain	PDB id2 & chain	all atoms have identical x, y, z	different residues
1a0t-Q	1oh2-Q	all 3×413	9
1ce7-A	2mll-A	all 3×241	1
1ruj-A	4rhv-A	all 3×237	1
1gli-B/D	3hhb-B/D	all 3×146	1
2hqe-A	2o4x-A	all 3×217	1
5adx-Z	5afu-T	all 3×165	1
5lj3-O	5lj5-P	all 3×252	1
8fdz-A	8fe0-A	all 3×200	1

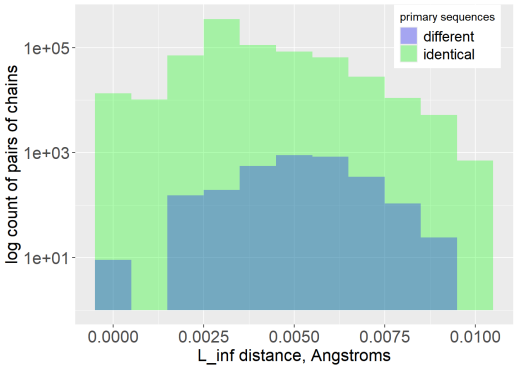

Figure 5. **Left**: pairs of identical backbones that differ in amino acids (corrections needed). **Right**: near-duplicates up to max deviation $L_\infty \leq 0.01$Å on the log scale: 13394+9 pairs with $L_\infty = 0$.

download the PDB files from Fig. 5 (left) and see the coincidences and differences with their own eyes without any computations. Here are the links to the identical files in the first row of Fig. 5 (left), where the 4-letter PDB id can be replaced with another one: https://files.rcsb.org/download/1A0T.cif, https://files.rcsb.org/download/1OH2.cif.

One potential explanation of all coincidences is the use of the molecular replacement method [40], which copies data from a previously deposited entry in the PDB to a new structure that has the same (or similar) sequence but a low-quality electron density map. However, since a full protein often consisting of several chains is not expected to be identical to a past entry, all coordinates should be additionally refined [41, 42]. Fig. 5 (right) shows that such a refinement stage was missed for many thousands of PDB entries. See details of all duplicates in the supplementary materials.

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 the DALI server [43] also confirmed the found duplicates by the traditional Root Mean Square Deviation (RMSD) through optimal alignment. The DALI takes about 30 min on average to find a short list of nearest neighbors of one chain in the whole PDB. Extrapolating this time to 707K+ cleaned chains yields 40+ years, slower by orders of magnitude than 6 hours needed for all our comparisons of the new invariants BRI on the same desktop computer.

7 Discussion and scientific integrity in chemistry

Using protein structures as an important example, this paper advocates a justified approach to any real data. The first and often missed step is to define an equivalence relation for given data because real objects can be represented in (usually infinitely) many different ways. For example, a human can be recognized in a huge number of digital photos but science progressed by developing DNA codes 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 functional properties. Hence the input of all prediction algorithms should be invariant, ideally a complete invariant.

under rigid motion. The Lipschitz bi-continuity of invariants is also essential because adding a small noise should not lead to a drastically different output. Earlier versions of Problem 1.2 with weaker conditions were solved for 2D lattices [44], periodic crystals [18], and finite clouds of unordered points [45] in the new area of Geometric Data Science.

The crucial novelty in the proposed approach is treating (the rigid class of) any experimental structure (protein backbone), e.g. from the 'gold standard' 220K+ entries of the PDB, as *objective ground truth* instead of any manually assigned labels.

Problem 1.2 asked for an analytically defined invariant I whose explicit formula should remain unchanged for any new data without usual re-training in machine learning.

While traditional approaches explored more data within continuously infinite spaces in a 'horizontal' way, solutions to Problem 1.2 and its analogs offer 'vertical' breakthroughs by building geographic-style maps of data spaces as viewed from a satellite.

The continuous maps of the PDB in Fig. 2 can be zoomed in at any spot and mapped in further invariants. This is a huge advantage in comparison with any dimensionality reduction, which was proved [46] 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, 4.8, which solved Problem 1.2 for protein backbones and detected thousands of previously unknown (near-)duplicates in the PDB. Improving the 'gold standard' of the PDB is urgently needed to avoid predictions based on the currently skewed data. The supplementary materials (available by request) include the Python code and table of all 9366 pairs of exact duplicates whose corresponding coordinates coincide with all digits and hence need further refinement. We thank all reviewers in advance for their time and for supporting data integrity in science.

This research was supported by the Royal Academy Engineering Fellowship IF2122/186, EPSRC New Horizons EP/X018474/1, Royal Society APEX fellowship APX/R1/231152. The authors thank Mariusz Jaskolski and Alex Wlodawer for their helpful comments on the first drafts and any other reviewers for their valuable time and fruitful suggestions.

A Appendix: updates on duplicates in the PDB

This appendix includes the confirmations of several duplicates found here and confirmed by their authors, and also subsequent updates in the PDB. After finding the first duplicates in Fig. 5 (left), we contacted the authors of the underlying publications. Since these structures were quite old (up to 30 years), many of their authors could not be found online but we e-mailed everyone whose contact details were accessible. Only two authors replied with details. The common author of the PDB entries 1a0t and 1oh2, Kay Diederichs, has confirmed the duplication error, see the screenshot of his email in Fig. 6.

Prof John Helliwell studied our duplicates in the supplementary materials 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 reading the first draft of this paper, John confirmed on 13th October 2024 that "4yta supersedes 3unr", so we will pass this to the PDB. Though we received only two personal confirmations, after sending the e-mails with duplicates in December 2022, we noticed that five PDB entries in the initial list of duplicates were updated, see Table 2.

The first two updates for 1ruj and 4rhv appeared at the same time in January 2023. In February 2023, we talked to the PDB validation team, who confirmed that PDB entries

Re: structures 1a0t and 1oh2 in the PDB
1 message

Kay Diederichs <kay.diederichs@uni-konstanz.de> 9 December 2022 at 17:39
To: [redacted]
Cc: Wolfram Welte <wolfram.welte@uni-konstanz.de>

Dear [redacted]

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 duplication in the PDB entries 1a0t and 1oh2, see Fig. 5 (left).

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: [redacted]

Dear [redacted]
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. Author's confirmation of duplication for the PDB entries 3unr and 4yta, which can be found in the Excel file PDB_4872pairs_allCAs_equal_xyz_by_red_flags in the supplementary materials.

are updated only by authors' request or by their permission. After that, three more PDB entries were updated in March and April 2023, see the last three rows in Table 2.

Table 2. Five PDB entries from Fig. 5 (left) were modified after our initial contacts in December 2022. All original and updated files are still accessible online.

PDB entry	date of last modification	time of last modification
4rhv	Fri, 13 Jan 2023	13:55:14 GMT
1ruj	Fri, 13 Jan 2023	13:55:14 GMT
1gli	Fri, 10 Mar 2023	14:09:09 GMT
3hhb	Fri, 10 Mar 2023	14:09:09 GMT
1cov	Fri, 14 Apr 2023	13:14:08 GMT

The older versions of the PDB files are available via ftp://snapshots.rcsb.org/20230102. The other duplicates from the lower half of Fig. 5 (left) have not been publicly reported yet, so their current files show the existing duplication in geometry with unexplained differences in sequences of amino acids. If the sequences of two protein chains coincide, their geometries can be expected to be close, though the neighboring chains in different proteins should affect some geometric coordinates at least slightly. 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 is considered negligible, especially due to floating point errors.

References

[1] V. R. Somnath, C. Bunne, A. Krause, Multi-scale representation learning on proteins, *Advances in Neural Information Processing Systems* **34** (2021) 25244–25255.

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

[3] J. Jumper, et al., Highly accurate protein structure prediction with alphafold, *Nature* **596** (2021) 583–589.

[4] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al., Accurate prediction of protein structures and interactions using a three-track neural network, *Science* **373** (2021) 871–876.

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

[6] 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 Biotechnology* **42** (2024) 243–246.

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

[8] 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 crystallography: methods and protocols* (2017) 627–641.

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

[10] S. Rass, S. König, S. Ahmad, M. Goman, Metricizing the euclidean space towards desired distance relations in point clouds, *IEEE Transactions on Information Forensics and Security* (2024).

[11] 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.

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

[13] P. Kumar, M. Bansal, Identification of local variations within secondary structures of proteins, *Acta Crystallographica Section D: Biological Crystallography* **71** (2015) 1077–1086.

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

[15] E. Parthé, L. Gelato, B. Chabot, M. Penzo, K. Cenzual, R. Gladyshevskii, *TYPIX standardized data and crystal chemical characterization of inorganic structure types*, Springer Science & Business Media, 2013.

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

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

[18] D. Widdowson, V. Kurlin, Resolving the data ambiguity for periodic crystals, *Advances in Neural Information Processing Systems (Proceedings of NeurIPS 2022)* **35** (2022).

[19] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, *Proteins: Structure, Function, and Bioinformatics* **57** (2004) 702–710.

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

[21] R. H. Lathrop, The protein threading problem with sequence amino acid interaction preferences is np-complete, *Protein Engineering, Design and Selection* **7** (1994) 1059–1068.

[22] C. Mura, E. Draizen, S. Veretnik, P. Bourne, Deep generative models of protein structure: A method to uncover distant relationships across a continuous fold space, *https://www.researchsquare.com/article/rs-2834307/latest.pdf* (2023).

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

[24] J. K. Leman, et al., Macromolecular modeling and design in rosetta: recent methods and frameworks, *Nature methods* **17** (2020) 665–680.

[25] 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, et al., Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination, *Nature Methods* **21** (2024) 110–116.

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

[27] G. t. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins, *Advances in protein chemistry* **23** (1968) 283–437.

[28] J. Yim, B. L. Trippe, V. De Bortoli, E. Mathieu, A. Doucet, R. Barzilay, T. Jaakkola, Se(3) diffusion model with application to protein backbone generation, in: *ICML*, 2023.

[29] 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, *Annals of Mathematics* (1935) 724–732.

[30] B. V. Dekster, J. B. Wilker, Edge lengths guaranteed to form a simplex, *Archiv der Mathematik* **49** (1987) 351–366.

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

[32] 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, *Journal of computer-aided molecular design* **27** (2013) 511–524.

[33] 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 bioinformatics* **16** (2015) 1–15.

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

[35] D. G. Kendall, The diffusion of shape, *Advances in applied probability* **9** (1977) 428–430.

[36] D. G. Kendall, D. Barden, T. K. Carne, H. Le, *Shape and shape theory*, John Wiley & Sons, 2009.

[37] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers: Original Research on Biomolecules* **22** (1983) 2577–2637.

[38] D. S. Chawla, Crystallography databases hunt for fraudulent structures, `https://cen.acs.org/research-integrity/Crystallography-databases-hunt-fraudulent-structures/102/i8`, 2024.

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

[40] M. G. Rossmann, The molecular replacement method, *Acta Crystallographica A: Foundations* **46** (1990) 73–82.

[41] 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 Crystallographica Section D: Biological Crystallography* **67** (2011) 355–367.

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

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

[44] V. A. Kurlin, Mathematics of 2-dimensional lattices, *Foundations of Computational Mathematics* (2022).

[45] 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*, 2023, pp. 1275–1284.

[46] P. S. Landweber, E. A. Lazar, N. Patel, On fiber diameters of continuous maps, *The American Mathematical Monthly* **123** (2016) 392–397.