Commentary

Over Half a Century of Burial of ρ, θ and ϕ in PDB

Wei Li[1]

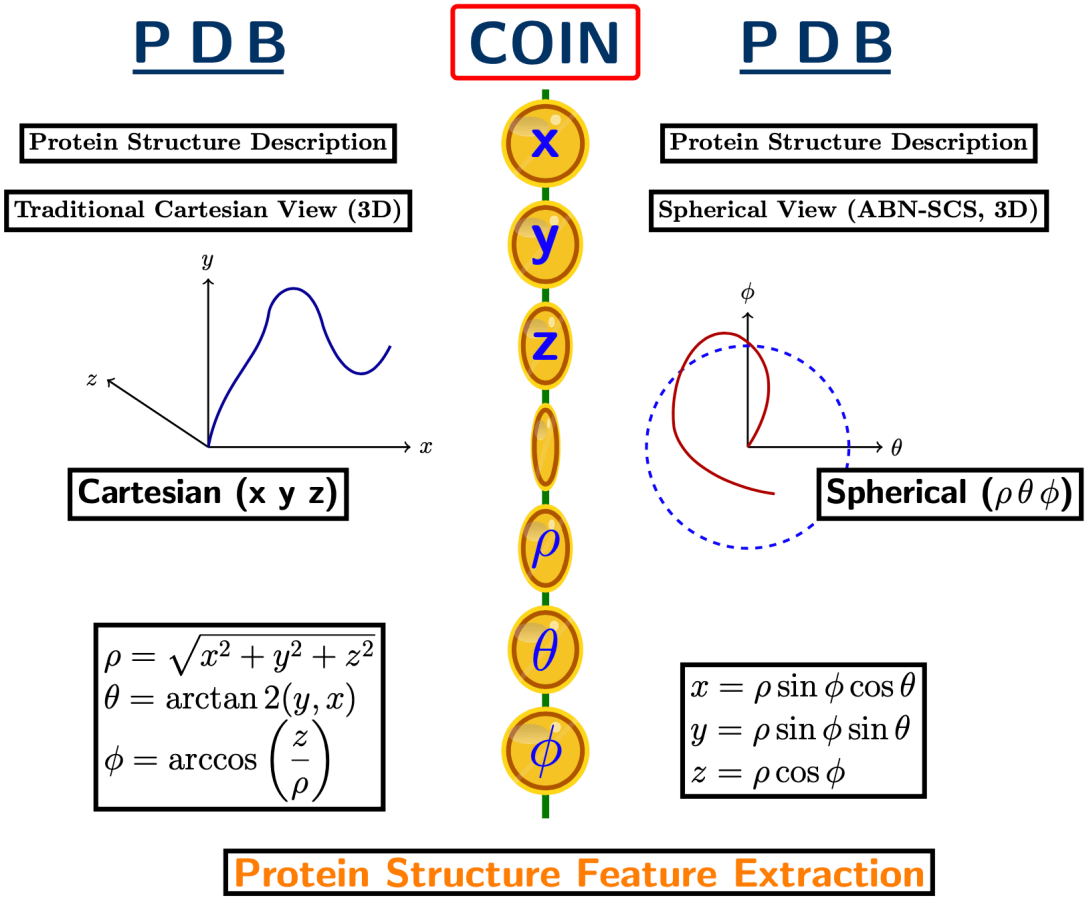
1. Independent researcher

Since its inception in 1971, the Protein Data Bank (PDB) has archived experimentally measured biomolecular structures in Cartesian coordinates $[x, y, z]$. While this convention aligns with the output of experimental techniques such as X-ray crystallography and NMR spectroscopy, it does not explicitly encode chemical connectivity or atomic bonding networks within protein structures. In light of the geometric equivalence of Cartesian (CCS) and spherical (SCS) coordinate systems, this manuscript proposes an alternative geometric framework—the Atomic Bonding Network-based Spherical Coordinate System (ABN-SCS)—in which atomic positions are expressed using spherical coordinates $[\rho, \theta, \phi]$ anchored to the covalent atomic bonding network within a protein structure. In ABN-SCS, ρ corresponds to the equilibrium covalent bond length, while θ and ϕ capture angular distributions, providing chemically meaningful descriptors that complement Cartesian-based representations. With Caenopore-5 as an example, this manuscript demonstrates that ABN-SCS enables the characterization of spherical bond-level geometries, expanding the feature space available for computational pipelines such as AlphaFold2. Finally, this work argues that integrating ABN-SCS features into protein structure prediction pipelines can enhance geometric fidelity and that the time is now ripe for the community to consider the other side of the coin (Ref: Graphical Abstract) and for the trapped spherical features [ρ, θ, ϕ] to be released from the PDB and integrated into algorithms such as AF2 to improve protein structure prediction performance.

Corresponding author: Wei Li, wli148@aucklanduni.ac.nz

 doi.org/10.32388/O5RM3Y

Graphical Abstract



Graphical Abstract. Since 1971, experimentally measured protein structures have been deposited in the Protein Data Bank (PDB) with atomic positions expressed in Cartesian coordinates $[x, y, z]$. In light of the geometric interconvertibility of Cartesian and spherical (SCS, $[\rho, \theta, \phi]$) coordinate systems, this figure presents a conceptual shift toward a spherical view of protein structure. As an alternative to the Cartesian default since 1971, the ABN-SCS framework described in this manuscript offers a radial, angular description of protein structure with $[\rho, \theta, \phi]$, where ρ is defined as the equilibrium interatomic bond length (a physical constant), i.e., the internuclear distance at which the system energy minimum occurs. In the center of this figure are seven flipping golden coins, which constitute a metaphor for the geometric interconvertibility of Cartesian and spherical coordinate systems, advocating for the extraction of $[\rho, \theta, \phi]$ rooted in the atomic bonding network (ABN) within protein structure, and also for the integration of spherical structural features $[\rho, \theta, \phi]$ into protein structure prediction algorithms like AlphaFold.

doi.org/10.32388/O5RM3Y

Background

In 1971, the Protein Data Bank (PDB) was established as the central global repository for experimentally determined biomolecular structures [1][2][3][4]. Since then, the Cartesian coordinate system (CCS) has served as the default geometric framework to specify atomic positions as $[x, y, z]$ (Figure 1), primarily because it directly corresponds to experimental data from techniques such as X-ray crystallography and NMR spectroscopy [5][6], which inherently produce atomic positions as [x, y, z]. By definition, CCS treats each atom as an independent point in space and conceptualizes protein structure as a set of discrete points, without explicitly encoding their underlying chemical connectivity or bonding network in the PDB-format text. This conceptual aspect suggests that incorporating alternative coordinate systems or representations that explicitly capture the atomic bonding network and chemical connectivity—such as graph-based, network-based, or spherical coordinate systems tied to atomic bonding patterns—could complement or enhance protein structure description, feature extraction, and prediction beyond what is currently achievable with the traditional CCS approach. For example, the Cartesian coordinate system and spherical coordinate system (SCS)—$[\rho, \theta, \phi]$—are mathematically interconvertible [7], underscoring the technical and geometric feasibility of specifying atomic positions using SCS—$[\rho, \theta, \phi]$, in addition to the default [x, y, z] approach since 1971 [1].

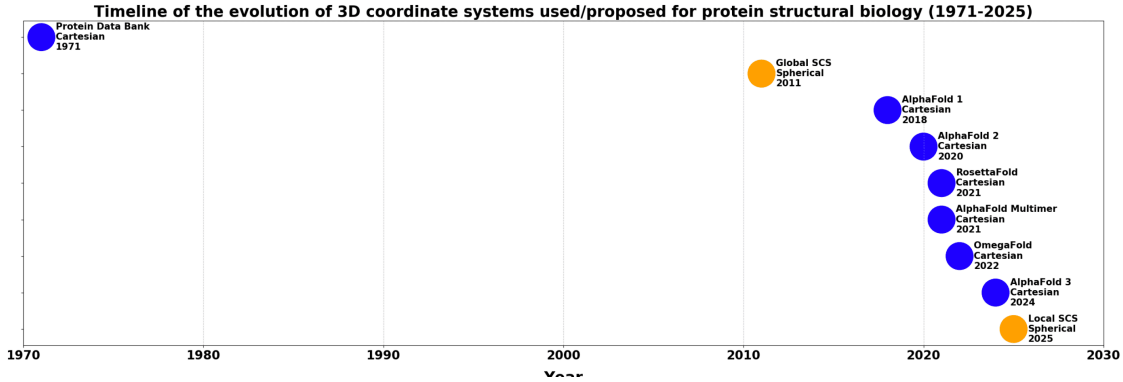


Figure 1. Over half a century of burial of ρ, θ, ϕ in the PDB [8][9]. In this figure, blue and orange solid circles represent Cartesian and spherical approaches for protein structure description, respectively, with **Local SCS** (the orange solid circle at the bottom) representing the ABN-SCS framework described in this manuscript, and **Global SCS** (the orange solid circle at the top) representing a global spherical coordinate system, which anchors atomic coordinates to an imaginary geometric centroid [10][11].

A Cartesian PDB for protein structure feature extraction

As of August 28, 2025, ~241,055 biomolecular structures have been deposited in the PDB [12], the majority of them resolved by experimental techniques such as X-ray crystallography [4], NMR spectroscopy [4], and cryo-electron microscopy [3]. This number, however, represents only a minute fraction of the billions of known protein sequences, highlighting a fundamental gap between sequence discovery and structural characterization [13]. In recent years, advances in computational methodologies—particularly deep learning—have addressed this disparity, enabling accurate large-scale protein structure prediction [14][15]. Notably, AlphaFold [16] and RoseTTAFold [14] (Figure 1) have leveraged neural networks to replace traditional energy models and sampling procedures, yielding substantial gains in predictive accuracy [17] [18][19][20][21][22]. Of further interest, geometric feature extraction is a common cornerstone for both experimental structure measurement and protein structure prediction, where experimentally measured raw data or computationally extracted structural features are transformed into geometric descriptors that underpin the construction of the final structural model or NMR ensemble of the molecule of interest. In other words, for neuralized protein structure prediction (e.g., by AF2 [16]), protein structural feature extraction is similar to what experimental measurements of chemical shifts or X-ray diffraction patterns are to protein structure determination using NMR spectroscopy or X-ray crystallography, respectively.

As the primary source of training and validation data, the PDB underpins the performance of most protein structure prediction tools [14][15][16][21][23]. For example, AlphaFold2 (AF2) both explicitly and implicitly extracts diverse geometric features from Cartesian-format PDB texts (see supplementary **supps.pdf**), including atomic coordinates, masks for valid atoms, and amino acid encodings. Subsequently, AF2 computes backbone atomic positions (N, Cα, C, O), torsion angles (ϕ, ψ, ω), and masks for valid residue pairs. These structural template-derived geometric features are supplemented by implicit geometric data generated by the **Evoformer** module during inference, such as inter-residue Euclidean distance maps from predicted C_β coordinates, backbone torsion angles (ϕ, ψ, ω), side-chain torsion χ angles, and pairwise orientation vectors. Afterward, AF2 constructs residue-specific rigid-body frames and frame-to-frame transformations to represent local coordinate systems, while geometric validity masks exclude physically implausible conformations, and attention biases based on distance and orientation restraints guide the generation of biologically plausible 3D coordinates in the format of $[x, y, z]$ [24][25]. Of note here, by definition, AF2's backbone torsions (ϕ, ψ) differ from the spherical coordinates $[\rho, \theta, \phi]$ as described in the Graphical Abstract (**grafic.pdf**) and in [26].

Collectively, these diverse geometric descriptors are critical ingredients for AF2's accurate structural predictions. While this Cartesian-based prediction pipeline has proven transformative in its performance [16], geometrically, it encodes only one viewpoint of protein 3D structure, representing only one side of the coin (Ref: Graphical Abstract). As mentioned above, the Cartesian coordinate system and spherical coordinate system (SCS) are mathematically interconvertible [7]. Hence, this article proposes an atomic bonding network–based spherical coordinate system (ABN-SCS) for protein structure representation and feature extraction, not just from Cartesian coordinates $[x, y, z]$ (Figure 1), but also from spherical coordinates $[\rho, \theta, \phi]$.

ABN-SCS Framework: definition, principles and methodology

Here in this manuscript, the ABN-SCS framework is defined as a chemically grounded geometric framework for the representation of protein structures by integrating covalent atomic bonding information with spherical coordinates $[\rho, \theta, \phi]$. Its core principles and methodology are outlined as follows:

1. **Foundation: Inter-atomic Bonding Network (ABN)**

   ABN-SCS anchors the coordinate system to the covalent bonding network of the protein rather than to an arbitrary geometric centroid [10][11]. Each atom's position is defined relative to its bonded neighbor, ensuring that chemical connectivity is inherently encoded in the ABN-SCS framework.

2. **Spherical Coordinate Transformation**

   Atomic positions are expressed in spherical coordinates $[\rho, \theta, \phi]$:

$$\rho = \sqrt{x^2 + y^2 + z^2}, \tag{1}$$

$$\theta = \arctan 2(y, x), \tag{2}$$

$$\phi = \arccos \frac{z}{\rho} \tag{3}$$

   where ρ (Figure 2) denotes the equilibrium bond length, θ represents the azimuthal angle in the x-y plane from the x-axis, and ϕ represents the polar angle from the z-axis.

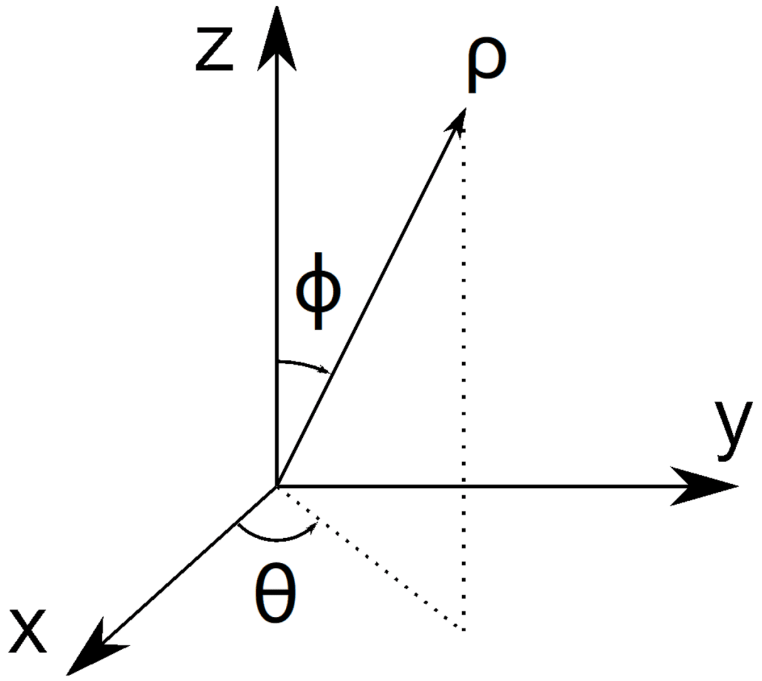


Figure 2. Cartesian and spherical coordinate systems are geometrically equivalent (Ref: Graphical Abstract **grafic.pdf**).

3. **Equilibrium Bond Lengths as Radial Distances**

   The radial coordinate ρ (Figure 2) is defined as the equilibrium inter-atomic bond length, i.e., the inter-nuclear distance at which the system energy minimum occurs [27][28]. This ensures that the ABN-SCS framework captures chemically relevant spacing between atoms and that the value of ρ is physically and structurally restrained [26], making it a more physically grounded paradigm compared to the default CCS in PDB.

4. **Chemical Significance of spherical coordinates** $[\theta, \phi]$

   As two spherical angles, θ and ϕ capture local spherical bond-level geometric features, enabling an ABN-SCS perspective for the extraction of spherical protein structure features, in addition to Cartesian features (see supplementary **supps.pdf**) as used by tools such as AF2 [16].

5. **Local SCS Frame Definition for Protein Structure Feature Extraction**

   For each covalently bonded atom pair, a local spherical coordinate frame can be established based on its covalently bonded neighbors, providing a consistent reference for measuring angles and

distances across the protein, where global rotations and translations do not affect protein structure feature extraction.

To sum up, ABN-SCS encodes both chemical connectivity and three-dimensional geometry, offering a self-contained and reproducible representation of protein structures, which have been described using CCS since 1971 [1][2]. Furthermore, given that Cartesian (x, y, z) and spherical (ρ, θ, ϕ) coordinates are mathematically interconvertible and both applicable for atomic position specification [7][29], transforming from Cartesian to spherical coordinates (Ref: Graphical Abstract **grafic.pdf**) decouples spatial information into chemically interpretable parameters as below and in Figure 2:

- ρ: radial distance (equilibrium covalent bond length),
- θ: azimuthal angle from the positive x-axis in the xy-plane.
- ϕ: polar angle from the positive z-axis,

Example: ABN-SCS Representation of a Model Pentapeptide

To begin with, an ABN-SCS description of the chemical bonding network of twenty stand-alone natural amino acids is provided in the supplementary file **supps.pdf**.

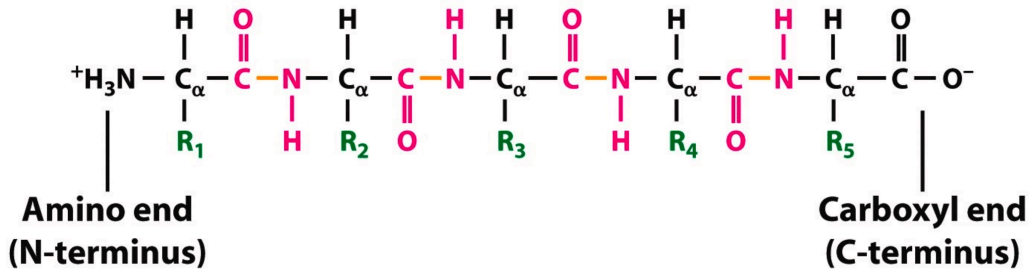


Figure 3. Chemical structure of the model pentapeptide (**peptideX**).

Here, a short peptide with five amino acids (referred to below as **peptideX**) (Figure 3) is employed as an example to illustrate how the ABN-SCS framework describes protein structure with ρ, θ, and ϕ [29]. As shown in Figure 4, a peptide bond is formed by dehydration, which is the removal of the carboxyl hydroxyl on one amino acid and one hydrogen atom on the amine end of another amino acid. Thus, the entire peptide bonding network of the backbone of peptideX (Figure 3) is defined sequentially as below:

1. a peptide bond formed between R_1 and R_2 of **peptideX** (Figure 3);
2. a peptide bond formed between R_2 and R_3 of **peptideX** (Figure 3);
3. a peptide bond formed between R_3 and R_4 of **peptideX** (Figure 3);
4. a peptide bond formed between R_4 and R_5 of peptideX (Figure 3);

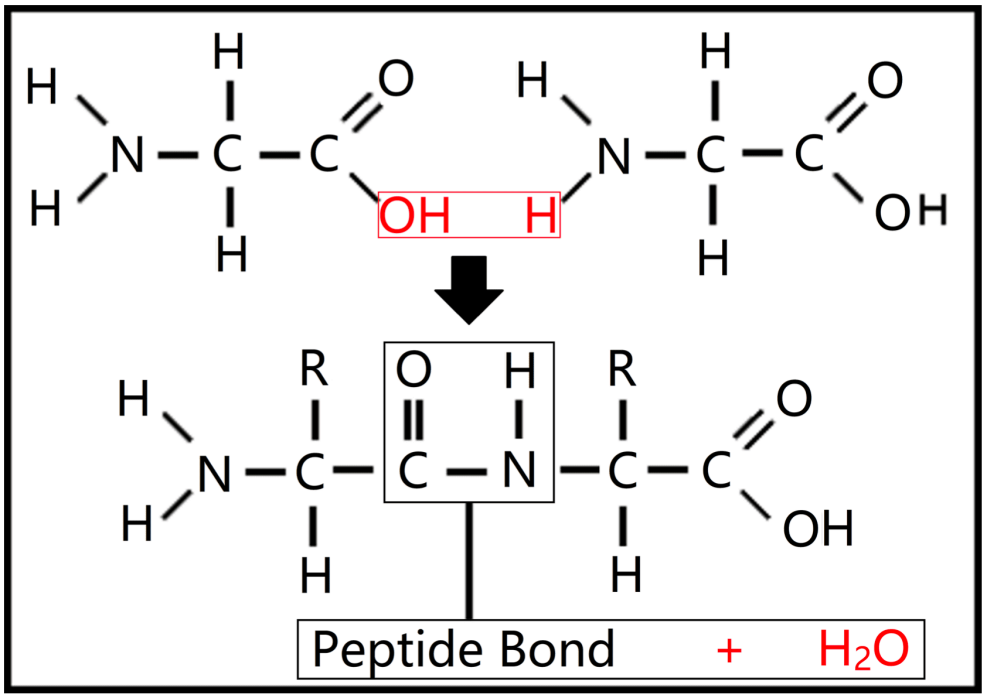


Figure 4. Peptide bond formation through the covalent linkage between the carbonyl carbon and the amide nitrogen via dehydration.

Afterwards, the entire atomic bonding network (ABN) of peptideX, or its chemical connectivity, is defined sequentially in a two-step loop, as below:

1. an ABN-SCS definition of the chemical connectivity of standalone amino acids (see supplementary **supps.pdf**), focusing on the covalent bonds within each residue's backbone: N–H, N–Cα, Cα–Hα, Cα–C=O, and C=O–N (to the next residue), and also on the covalent bonds within each residue's side chain;
2. formation of peptide bonds between adjacent residues, as described above and in Fig. 4.

With this generalizable ABN-SCS framework, the complete covalent chemical connectivity of any protein structure—including both backbone and side-chain atoms—can be fully and reversibly mapped into

spherical coordinates (ρ, θ, ϕ), thereby embedding intrinsic chemical topology directly into the protein structure description with the ABN-SCS framework. Looking ahead, a natural and transformative next step would be the creation of a spherical Protein Data Bank that stores structural information in $[\rho, \theta, \phi]$, serving as a complement to the traditional Cartesian PDB since 1971 [1][2]. Such a resource could provide an alternative perspective for us to understand and describe protein structure and also provide new opportunities for protein structure feature extraction, which is inextricably linked to the performance of protein structure prediction pipelines such as AF2 [16].

Towards a Spherical PDB for Protein Structure Feature Extraction

In recent years, the application of artificial intelligence (AI) has expanded rapidly across many scientific disciplines [30], including neural network–based protein structure prediction by tools such as AF2 [16]. Nevertheless, the adoption of AI in science is not without limitations; epistemological concerns have been raised regarding its widespread use [30]. Take astronomy as an analogy [30]. The model of the Universe with Earth at its center was extremely accurate at predicting planetary motions because of tricks such as 'epicycles'—the assumption that planets move in circles whose centers revolve around Earth along a larger circular path. This analogy suggests that extremely accurate predictions do not necessarily equate to genuine scientific understanding in an unbiased manner, which is critical for the continued improvement of the performance of technologies and computational tools (e.g., AF2) built upon the way protein structure is conceptualized and described in the central global repository [1].

Today, AI excels at producing the equivalent of 'epicycles' [30]. All else being equal, being able to squeeze more predictive juice [31] out of flawed theories or inadequate paradigms will help them stick around for longer, impeding true scientific progress [30] and the continued improvement of the performance of protein structure prediction pipelines such as AF2 [16]. To this end, for protein structure description with $[x, y, z]$, CCS is an inadequate paradigm in the sense that CCS treats each atom as an independent point in three-dimensional space, representing a protein as a collection of discrete (x, y, z) points without explicitly encoding its underlying covalent bonding network or chemical connectivity in the PDB text format. In other words, Cartesian coordinates have been the default "language" of protein structure description since 1971 (Figure 1) [1][2], reflecting a geometric rather than a chemically and physically grounded perspective of how protein structure is conceptualized and described in the central global repository [2].

Yet, this by no means suggests that CCS is wrong or flawed. As one of the foundational paradigms of structural biology, CCS has been not only geometrically correct but also proven indispensable and immensely useful in its real-world applications. Nevertheless, it is conceptually inadequate, as it omits explicit representation of chemical connectivity within protein structure. From a mathematical standpoint, CCS and spherical coordinate systems (SCS; ρ, θ, ϕ) are fully interconvertible [7]. Thus, specifying atomic positions using spherical coordinates alongside the conventional (x, y, z) representation has been both technically and geometrically feasible and enforceable. Building on this geometric equivalence (Ref: **Graphical Abstract**), this manuscript puts forward an ABN-SCS framework for protein structure representation and feature extraction. Unlike earlier global SCS approaches (e.g., the first orange circle at the top of Figure 1) [10][11], which anchor coordinates to an imaginary geometric centroid, ABN-SCS is intrinsically rooted in the covalent bonding network of the target protein. This allows the derivation of chemically meaningful geometric descriptors that can augment conventional Cartesian-based features. For example, the NMR ensemble of Caenopore-5 [32] [33] was recently used to demonstrate how the ABN-SCS framework transforms its backbone structure into a 3×477 matrix [26]. The three dimensions correspond to bond lengths (ρ) and angular distributions (ϕ versus θ, Figure 5), yielding geometric descriptors that complement the Cartesian-based features (see supplementary file supps.pdf) as extracted and used by tools such as AF2 [16].

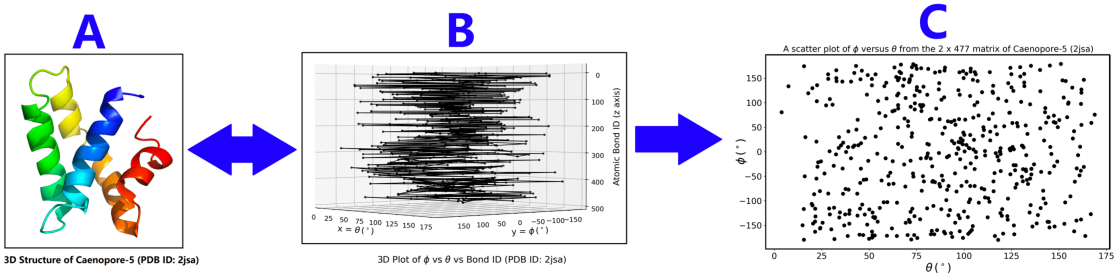


Figure 5. Protein structure feature extraction using the ABN-SCS framework for Caenopore-5 [26]. (A) NMR structure of Caenopore-5 (PDB ID: 2jsa) [32]; (B) 3D plot of ϕ versus θ versus bond ID for Caenopore-5 [26]; (C) 2D projection of panel B onto the ϕ–θ plane [26].

Taken together, this manuscript advocates an alternative structural biology perspective (Ref: **Graphical Abstract**), and with this manuscript, I argue that the time is now ripe for the computational structural biology and bioinformatics community to take a look at the other side of the coin (Ref: **Graphical**

Abstract) and for the trapped spherical features [ρ, θ, and ϕ] to be released from the PDB and integrated into algorithms such as AF2 toward protein structure prediction with improved performance, as this dual-coordinate system strategy is not just geometrically and technically feasible and enforceable but also paves a way to further narrow the gap between structure predictions and the physicochemical reality not just of proteins but also of biomolecular architectures in general.

Statements and Declarations

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used OpenAI's ChatGPT to improve the readability of the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. ^a, b, c, d, e, f *(1971). "Crystallography: Protein Data Bank." Nat New Biol. **233**(42):223–223.*
2. ^a, b, c, d, e *Berman H, Henrick K, Nakamura H (2003). "Announcing the worldwide Protein Data Bank." Nat St ruct Mol Biol. **10**(12):980–980.*
3. ^a, b *García-Nafría J, Tate CG (2021). "Structure determination of GPCRs: cryo-EM compared with X-ray cryst allography." Biochem Soc Trans. **49**(5):2345–2355. doi:10.1042/bst20210431.*
4. ^a, b, c *Steven AC, Baumeister W (2008). "The future is hybrid." J Struct Biol. **163**(3):186–195.*
5. ^ *Diamond R (1971). "A real-space refinement procedure for proteins." Acta Crystallogr A. **27**(5):436–452. doi: 10.1107/s0567739471000986.*
6. ^ *Diamond R (1974). "Real-space refinement of the structure of hen egg-white lysozyme." J Mol Biol. **82**(3):3 71–391. doi:10.1016/0022-2836(74)90598-1.*
7. ^a, b, c, d *Li W (2024). "A Reversible Spherical Geometric Conversion of Protein Backbone Structure Coordinate Matrix to Three Independent Vectors of ρ, θ and φ." doi:10.20944/preprints202404.0576.v1.*
8. ^ *Bonvin AMJJ (2021). "50 years of PDB: a catalyst in structural biology." Nat Methods. **18**(5):448–449.*
9. ^ *Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S (2017). "Protein Data Bank (PD B): The Single Global Macromolecular Structure Archive." New York: Springer. pp. 627–641.*

10. ^ a, b, c *Reyes VM (2015). "Implementation of the Spherical Coordinate Representation of Protein 3D Structures and its Applications Using FORTRAN 77/90 Language."*

11. ^ a, b, c *Reyes VM (2011). "Representation of protein 3D structures in spherical (ρ, φ, θ) coordinates and two of its potential applications." Interdiscip Sci Comput Life Sci. **3**(3).*

12. ^ *Berman HM (2000). "The Protein Data Bank." Nucleic Acids Res. **28**(1):235–242. doi:10.1093/nar/28.1.235.*

13. ^ *Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature. **596**(7873):583–589.*

14. ^ a, b, c *Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. (2024). "Generalized biomolecular modeling and design with RoseTTAFold All-Atom." Science. **384**(6693). doi:10.1126/science.adl2528.*

15. ^ a, b *Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. (2022). "High-resolutionde novostructure prediction from primary sequence." doi:10.1101/2022.07.21.500999.*

16. ^ a, b, c, d, e, f, g, h, i *Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature. **596**(7873):583–589.*

17. ^ *AlQuraishi M (2020). "A watershed moment for protein structure prediction." Nature. **577**(7792):627–628.*

18. ^ *Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. (2020). "Improved protein structure prediction using potentials from deep learning." Nature. **577**(7792):706–710.*

19. ^ *Hey T, Butler K, Jackson S, Thiyagalingam J (2020). "Machine learning and big scientific data." Philos Trans R Soc A Math Phys Eng Sci. **378**(2166):20190054.*

20. ^ *Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. (2019). "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)." Proteins. **87**(12):1141–1148.*

21. ^ a, b *Zweckstetter M (2021). "NMR hawk-eyed view of AlphaFold2 structures." Protein Sci.*

22. ^ *Strodel B (2021). "Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins." J Mol Biol. 167182.*

23. ^ *Ahdritz G, Bouatta N, Floristean C, Kadyan S, Xia Q, Gerecke W, et al. (2024). "OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization." Nat Methods. **21**(8):1514–1524. doi:10.1038/s41592-024-02272-z.*

24. ^ *DeLano WL (2002). "Pymol: An open-source molecular graphics tool." CCP4 Newsl Protein Crystallogr. **40**:82–92.*

25. ^Xu W (2024). "Current Status of Computational Approaches for Small Molecule Drug Discovery." J Med Chem. **67**(21):18633–18636. doi:10.1021/acs.jmedchem.4c02462.

26. a, b, c, d, e, fLi W (2025). "Protein Structure Description with ρ, θ and φ: A Case Study with Caenopore-5." bioRxiv. doi:10.1101/2025.01.08.631847.

27. ^Batsanov SS (1999). "Calculation of van der Waals radii of atoms from bond distances." J Mol Struct THEOCHEM. **468**(1-2):151–159.

28. ^Caine BA, Bronzato M, Popelier PLA (2019). "Experiment stands corrected: accurate prediction of the aqueous pKa values of sulfonamide drugs using equilibrium bond lengths." Chem Sci. **10**(25):6368–6381.

29. a, bLi W (2020). "A Local Spherical Coordinate System Approach to Protein 3D Structure Description." doi:10.20944/preprints202001.0133.v1.

30. a, b, c, d, eNarayanan A, Kapoor S (2025). "Why an overreliance on AI-driven modelling is bad for science." Nature. **640**(8058):312–314. doi:10.1038/d41586-025-01067-2.

31. ^Callaway E (2025). "AlphaFold is running out of data — so drug firms are building their own version." Nature. **640**(8058):297–298. doi:10.1038/d41586-025-00868-9.

32. a, bMysliwy J, Dingley AJ, Stanisak M, Jung S, Lorenzen I, Roeder T, et al. (2010). "Caenopore-5: The three-dimensional structure of an antimicrobial protein from Caenorhabditis elegans." Dev Comp Immunol. **34**:323–330.

33. ^Roeder T, Stanisak M, Gelhaus C, Bruchhaus I, Grötzinger J, Leippe M (2010). "Caenopores are antimicrobial peptides in the nematode Caenorhabditis elegans instrumental in nutrition and immunity." Dev Comp Immunol. **34**:203–209.

Supplementary data: available at https://doi.org/10.32388/O5RM3Y

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.