Next-Generation Protein Sequencing: Nanopore and AI-Assisted Decoding of Posttranslational Modifications—A Narrative Review

Ambreen Ilyas[1] and Khadija Batool[2]

[1]University of the Punjab, Lahore, Pakistan
[2]University of Sialkot, Sialkot, Pakistan

Correspondence to:
Ambreen Ilyas,
Ambreen2.phd.sbs@pu.edu.pk

Additional material is published online only. To view please visit the journal online.

Cite this as: Ilyas A and Batool K. Next-Generation Protein Sequencing: Nanopore and AI-Assisted Decoding of Posttranslational Modifications–A Narrative Review. Premier Journal of Science 2025;13:100106

DOI: https://doi.org/10.70389/PJS.100106

Peer Review
Received: 14 August 2025
Last revised: 1 September 2025
Accepted: 3 September 2025
Version accepted: 6
Published: 20 September 2025

Ethical approval: N/a
Consent: N/a
Funding: No industry funding
Conflicts of interest: N/a
Author contribution: Ambreen Ilyas and Khadija Batool – Conceptualization, Writing – original draft, review and editing
Guarantor: Ambreen Ilyas
Provenance and peer-review: Unsolicited and externally peer-reviewed
Data availability statement: N/a

ABSTRACT

BACKGROUND

Proteins are versatile biomolecules whose functionality is intrinsically linked to their structural hierarchy—primary, secondary, and tertiary organization. Advances in structural genomics and integrative structural biology have significantly improved our ability to determine protein structures experimentally and computationally. However, challenges remain in translating structural novelty into functional understanding, particularly across diverse protein superfamilies.

OBJECTIVE

This narrative review synthesizes the state-of-the-art experimental and computational methodologies used in protein structure elucidation, highlighting their implications for understanding structure–function relationships, functional divergence, and the predictive challenges associated with structural genomics.

METHODOLOGICAL SCOPE

Primary structure: Classical sequencing techniques (Edman degradation, dansyl chloride assays) and modern mass spectrometry-based peptide mapping.

SECONDARY STRUCTURE

Circular dichroism spectroscopy for probing α-helices and β-sheets, monitoring conformational transitions, and evaluating thermostability.

TERTIARY STRUCTURE

High-resolution approaches, including X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy, complemented by structural classification frameworks such as SCOP, CATH, and SFLD.

ILLUSTRATIVE FRAMEWORKS

Representative workflows are provided to demonstrate standard experimental pipelines. These highlight how structural data at multiple levels—from amino acid sequence to three-dimensional fold—inform the study of active-site geometry, protein–protein interactions, and functional divergence within superfamilies.

CONCLUSION

Integrating experimental methods with computational prediction creates a robust framework for annotating novel proteins, identifying structure-based determinants of specificity, and exploring evolutionary trajectories. Such insights are critical for drug discovery, understanding disease mechanisms, and advancing functional genomics. Continued development of AI-based prediction tools and curated functional databases (e.g., SFLD and PANTHER) will further enhance the translation of structural data into biological insight.

Keywords: Alphafold2-assisted molecular replacement, Circular dichroism spectroscopy, ECD, ETD disulfide mapping, Protein structural genomics, Structure–function linkage database

Abbreviations

CD: Circular Dichroism
Cryo-EM: Cryo-Electron Microscopy
PDB: Protein Data Bank
NMR: Nuclear Magnetic Resonance
AI: Artificial Intelligence
MR: Molecular Replacement
PTM: Posttranslational Modification
plexDIA: Parallel Multiplexed Data-Independent Acquisition

Introduction

Proteins are indispensable macromolecules that mediate virtually all biological processes, including enzymatic catalysis, molecular signaling, scaffolding, and regulation. Their diverse functions are governed by three-dimensional (3D) conformations, which arise from the hierarchical folding of polypeptide chains dictated by amino acid sequences. Understanding the link between sequence, structure, and function remains a cornerstone of structural biology, molecular medicine, and bioinformatics-driven drug discovery.[1]

Over the past two decades, structural genomics initiatives have dramatically advanced our ability to decode protein structures at atomic and near-atomic resolution. Programs such as the Protein Structure Initiative have enriched the Protein Data Bank (PDB) with numerous novel folds.[2,3] Yet, many deposited structures lack corresponding experimental functional annotation.[4] This gap underscores the complexity of the structure–function paradigm, where proteins with similar folds may diverge in activity due to subtle variations in loop regions, domain architecture, or binding interfaces.[5]

To address this, the field has shifted toward targeting domain families of biomedical and functional importance, often focusing on underrepresented structural classes.[6,7] While protein sequences continue to grow exponentially, domain expansion has progressed more slowly, revealing that proteins are built from a limited set of reusable structural modules.[8] When recombined or embellished, these modules create functional diversity—a principle central to evolutionary and structural bioinformatics.[9]

Protein structure determination typically proceeds through three hierarchical levels:

Primary structure: Linear amino acid sequences determined by biochemical methods such as Edman degradation and dansyl chloride assays, supported by enzymatic cleavage and disulfide bond mapping.[10,11]

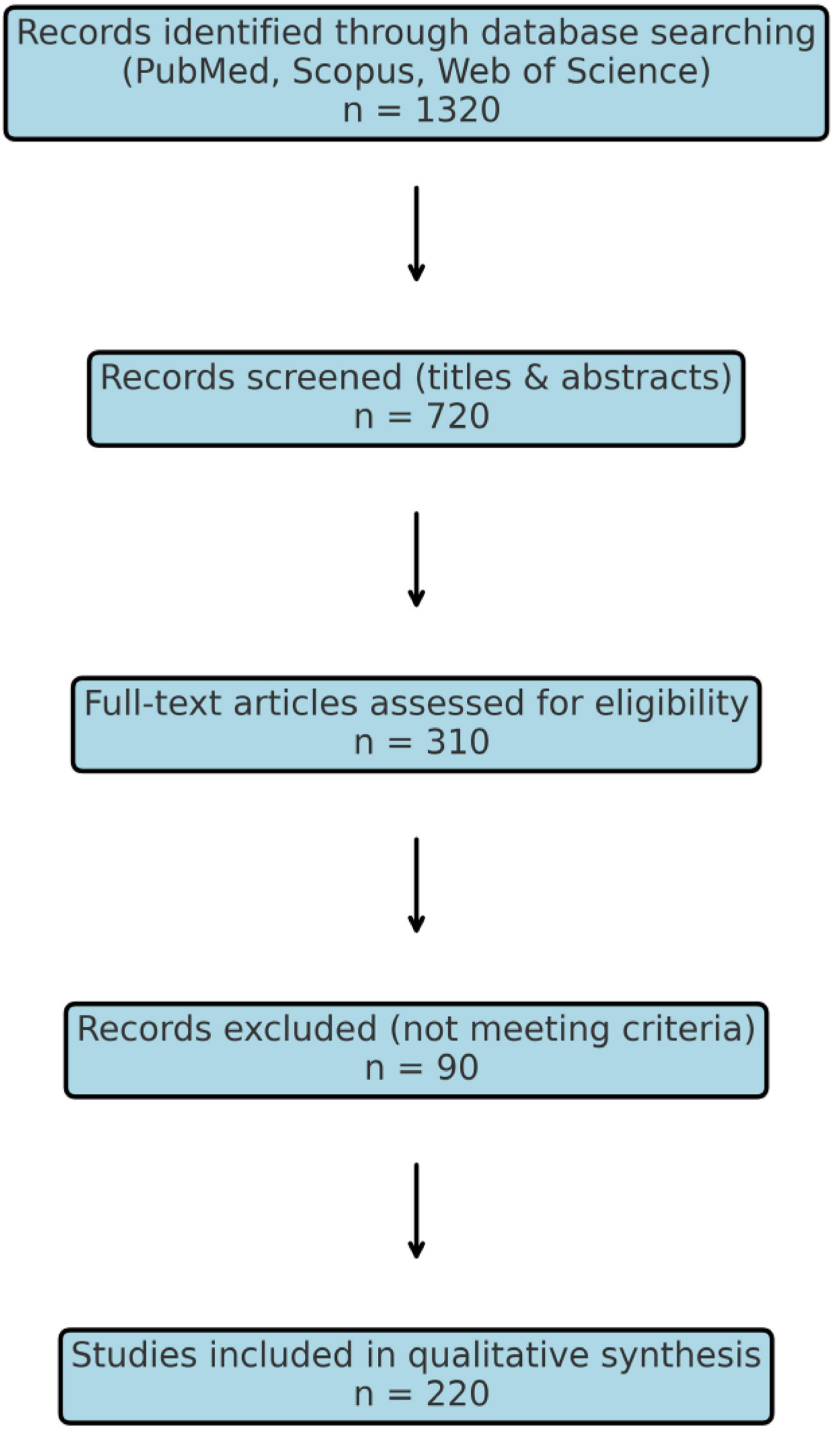
Records identified through database searching
(PubMed, Scopus, Web of Science)
n = 1320
Records screened (titles & abstracts)
n = 720
Full-text articles assessed for eligibility
n = 310
Records excluded (not meeting criteria)
n = 90
Studies included in qualitative synthesis
n = 220

Fig 1 | Literature retrieval and screening process. A PRISMA-style flow diagram illustrating the number of records identified (n = 1320), screened (n = 720), excluded (n = 90), and finally included (n = 220) in this review

Table 1 | Primary structure techniques

Method	Purpose	Reference
Ion-exchange/Affinity Chromatography	Protein purification	23
Disulfide Cleavage (DTT, β-ME)	Subunit dissociation	24
ECD/ETD Mass Spectrometry	Disulfide mapping	25,26
Enzymatic/Chemical Cleavage + HPLC-MS	Fragmentation and sequence reconstruction	27–29

Secondary structure: Local motifs such as α-helices and β-sheets, assessed using circular dichroism (CD) spectroscopy, which offers rapid solution-based evaluation of folding, thermostability, and conformational changes.[12,13]

Tertiary structure: Complete 3D folding patterns resolved by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (Cryo-EM), enabling visualization of catalytic sites, ligand-binding pockets, and protein–protein interfaces.[14,15]

However, structural data alone do not guarantee functional insight. Proteins sharing high structural similarity may perform divergent functions due to small sequence-specific insertions, deletions, or adaptive changes, particularly in catalytic or binding regions.[16–18] Consequently, emphasis has grown on subdividing superfamilies into functionally coherent subgroups using computational and evolutionary models.

Databases such as SFLD, SCOP, and CATH provide hierarchical frameworks for classifying domains and predicting function based on evolutionary lineage and biochemical context.[19] At the same time, computational methods—including entropy-based models, geometric potential mapping, and machine learning—are increasingly integrated with experimental pipelines to enhance structural annotation at the proteome scale.[2]

The convergence of computational predictions with experimental validation now represents the frontier of structural biology. Integrative approaches expand the scope, accuracy, and efficiency of structure determination, while global repositories like the PDB facilitate cross-comparison and predictive analysis.[20] For overall statistics on the distribution of structures by method in the PDB, see Figure 5 and the related discussion section.

Unlike earlier surveys that primarily emphasize either experimental or computational pipelines,[21,22] this review provides an integrative synthesis across all levels of protein structure determination. We highlight hybrid workflows that combine experimental precision with computational speed, incorporate very recent AI-driven advances such as AlphaFold2-guided molecular replacement and nanopore-based sequencing, and emphasize methodological controversies that remain unresolved. This combination offers a distinctive and timely perspective beyond descriptive catalogs of available methods.

This review consolidates current experimental and computational methodologies for protein structure determination, illustrates standard workflows (Figures 2–6), and emphasizes their importance in resolving the structure–function paradigm.

Illustrative Examples of Structural Workflows

To demonstrate how experimental and computational methodologies are applied in practice, selected workflows are presented here as illustrative examples. These examples are not new experimental findings but representative case studies highlighting standard approaches in protein structure determination.

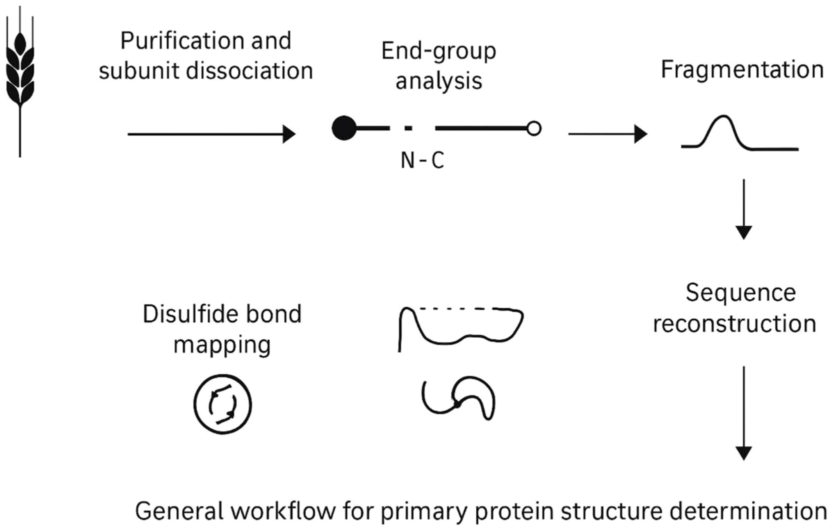
Purification and subunit dissociation
End-group analysis
N - C
Fragmentation
Sequence reconstruction
Disulfide bond mapping
General workflow for primary protein structure determination

Fig 2 | General workflow for primary protein structure determination. Schematic illustration of a typical workflow used to establish the primary structure of proteins: purification and subunit dissociation, end-group analysis (N- and C-terminal identification), fragmentation through enzymatic and/or chemical cleavage, chromatographic separation, and sequence reconstruction via Edman degradation or mass spectrometry. Disulfide bond mapping is often integrated to identify stabilizing linkages

Note: This figure was created de novo by the authors using BioRender/Adobe Illustrator and is provided for educational purposes only. It does not depict new experimental data.

Table 2 | CD-based secondary structure techniques

Technique	Application	Reference
Far-UV CD Spectroscopy	Secondary structure analysis	30
BeStSel/DichroWeb Tools	Spectral deconvolution and quantification	31,32
Bayesian/SOMSpec Models	Enhanced structural classification accuracy	33,34

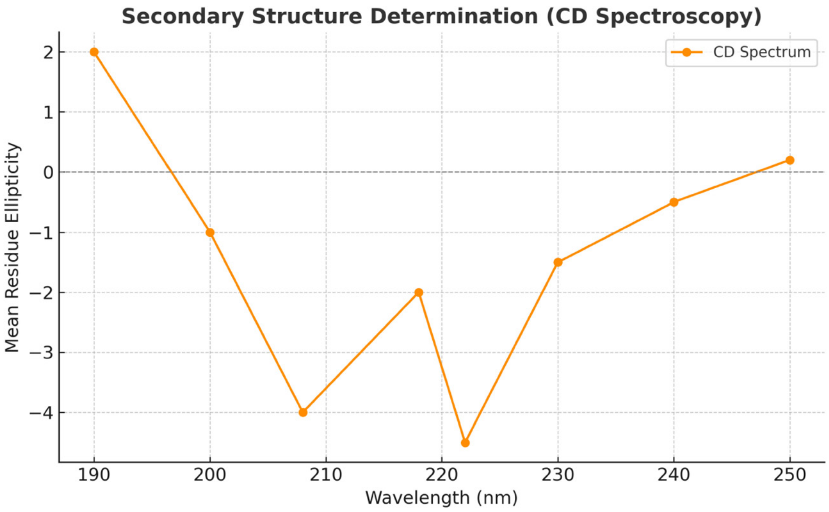

Fig 3 | Representative CD spectrum for secondary structure analysis. Illustrative example of a typical far-UV CD spectrum (190–250 nm), showing characteristic negative bands at 208 and 222 nm for α-helices, a shoulder near 218 nm for β-sheet content, and a positive peak near 192 nm corresponding to π-π transitions. Spectral deconvolution methods (e.g., BeStSel and DichroWeb) are commonly applied to estimate relative contributions of α-helix, β-sheet, and random coil

Note: This figure was created de novo by the authors using BioRender/Adobe Illustrator and is provided for educational purposes only. It does not depict new experimental data.

Search Strategy

Relevant literature was identified by searching PubMed, Web of Science, and Scopus databases between **January 2000 and January 2025** using the Boolean query: ("protein structure determination" OR "protein sequencing" OR "Cryo-EM" OR "X-ray crystallography" OR "NMR spectroscopy" OR "AlphaFold") AND English[lang]. A total of **1320 records** were retrieved, of which **720 were screened** based on title and abstract, and **220 studies** were included in the final review. **Inclusion criteria** comprised peer-reviewed articles in English with direct relevance to protein structure determination. **Exclusion criteria** included non-peer-reviewed sources, conference abstracts without full texts, and non-English publications. **Quality assessment** was performed by prioritizing studies published in indexed journals with rigorous methodology and high citation relevance. The screening and selection process is summarized in a **PRISMA-style flow diagram** (Figure 1). Quality assessment was operationalized using the SANRA (Scale for the Assessment of Narrative Review Articles) checklist. Articles scoring ≥10/18 were considered eligible. This ensured that the included studies demonstrated methodological rigor, clarity of reporting, and relevance to the objectives of this review.

Primary Structure Determination

Protein sequencing traditionally combines biochemical and mass spectrometric approaches. Typical workflows include protein purification (ion-exchange and affinity chromatography), end-group analysis (Edman degradation for N-terminal residues; carboxypeptidase digestion for C-terminal residues), and fragmentation (enzymatic cleavage with trypsin or chymotrypsin, or chemical cleavage with cyanogen bromide). LC analyses resulting peptide fragments-MS/MS, and overlapping sequences are aligned computationally to reconstruct the full amino acid sequence. In addition to classical sequencing and mass spectrometry approaches, **emerging next-generation sequencing technologies are beginning to extend protein analysis into the single-molecule domain**. Nanopore-based protein sequencing enables direct reading of amino acid chains by monitoring ionic current disruptions as polypeptides translocate through nanopores. Although still under active development, these approaches offer the promise of resolving long sequences and posttranslational modifications that are challenging for conventional methods. Recent advances in **proteomics workflows such as plexDIA** further highlight the potential of high-throughput, quantitative peptide mapping strategies. Together, these innovations complement established techniques and may redefine the landscape of primary structure determination in the coming decade. Disulfide mapping, using electron capture or electron transfer dissociation (ECD/ETD) mass spectrometry, further informs us about stabilizing linkages and contributes to understanding protein stability (Table 1). A generalized workflow for determining protein primary structure is shown in Figure 2, integrating

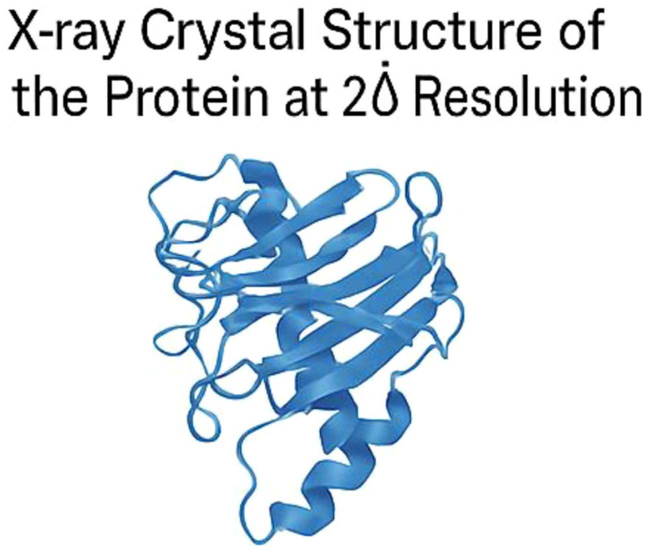

Fig 4 | Representative X-ray crystallography output. Illustrative example of a high-resolution (≈2.0 Å) electron density map and corresponding atomic model. This figure was created de novo by the authors for educational purposes only and does not represent newly determined data

Table 3 | Tertiary structure techniques

Method	Application	Reference
X-ray Crystallography	Atomic-resolution 3D structure determination	35,36
NMR Spectroscopy (CYANA, Rosetta)	Solution-state structure prediction	37,38
Cryo-EM (RELION, CryoSPARC)	High-res maps of large or noncrystallizable proteins	39,40
AlphaFold2, CATH, SCOPe	In-silico modeling and domain function inference	41–43

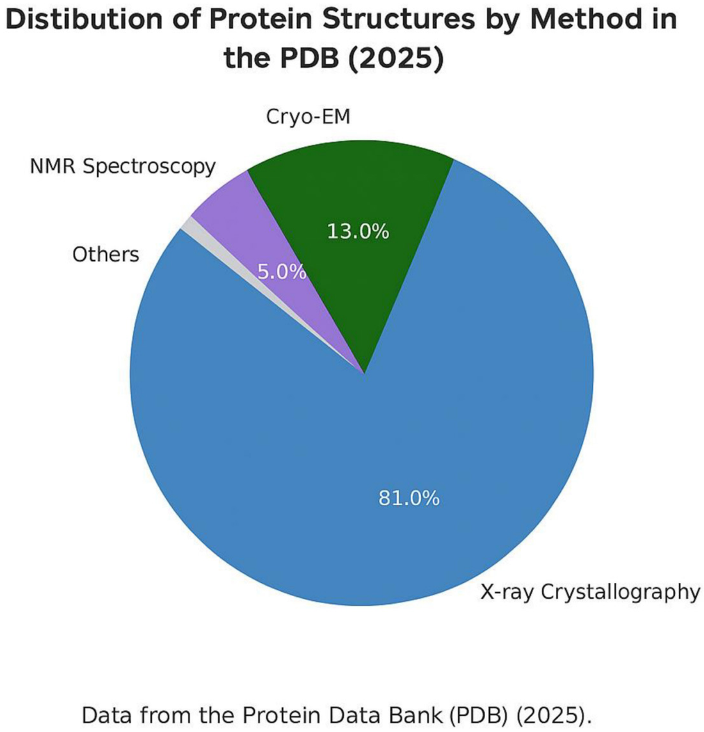

Fig 5 | Distribution of protein structures in the PDB by experimental method (as of 2025). The majority of deposited structures derive from X-ray crystallography (~80%), followed by Cryo-EM (~13%) and NMR spectroscopy (~5%), with other methods contributing ‹1%. This distribution highlights the historical dominance of crystallography and the recent rapid growth of Cryo-EM

Data source: PDB (accessed January 2025).

Note: This figure was created de novo by the authors using BioRender/Adobe Illustrator and is provided for educational purposes only. It does not depict new experimental data.

classical sequencing with modern mass spectrometric techniques.

Secondary Structure Determination

CD spectroscopy remains the most widely used method for probing α-helices and β-sheets in solution. Far-UV CD spectra (190–250 nm) typically show characteristic negative bands near 208 and 222 nm for α-helices and shoulders around 218 nm for β-sheets. Modern deconvolution tools such as BeStSel and DichroWeb allow quantitative estimation of structural composition, while Bayesian and machine-learning classifiers further enhance accuracy. Thermal unfolding profiles obtained from CD spectra provide insights into protein stability under physiological and stress conditions (Table 2). Far-UV CD spectra typically display characteristic signals for α-helices and β-sheets (Figure 3), which can be deconvoluted into secondary structure proportions using computational tools.

Tertiary Structure Determination

High-resolution 3D structures are most commonly obtained through X-ray crystallography, which can achieve resolutions of 0.8–3.5 Å, but requires high-quality crystals. X-ray crystallography typically yields atomic-resolution models accompanied by electron density maps, enabling detailed analysis of active sites and noncovalent interactions (Figure 4).

NMR spectroscopy, which provides solution-state information on folding and dynamics, is typically suited to small and medium-sized proteins. Cryo-EM, which has transformed the analysis of large complexes and membrane proteins, often achieves sub-3 Å resolution without the need for crystallization (Table 3).

Computational approaches such as AlphaFold2 complement these methods by rapidly generating high-confidence structural models, which can be used to guide experimental studies (e.g., as templates for molecular replacement in crystallography). The PDB remains dominated by X-ray crystallography entries, with Cryo-EM and NMR contributing smaller but growing shares (Figure 5 and Table 4).

The general workflow of protein crystallography from crystallization and data collection through phasing, model refinement, and PDB deposition is summarized in Figure 6.

Discussion

Protein structure determination remains central to structural biology, with complementary methods offering distinct advantages and limitations. The novelty of this review lies in its integrative framing: it not only catalogs experimental and computational techniques but also critically examines their intersections, limitations, and the unresolved debates shaping current practice. Compared with other recent reviews (Table 5), this work places greater emphasis on hybrid approaches, methodological controversies, and future innovations at the interface of AI and experimental pipelines.

Table 4 | Comparative overview of protein structure determination methods

Level	Method	Typical Resolution	Strengths	Limitations	Notes
Primary	Edman degradation, Mass spectrometry	Sequence level	Accurate for short peptides	Limited for long/ modified proteins	Legacy + modern hybrid use
Secondary	CD spectroscopy	Approx. % helix/sheet	Rapid, solution-based	Low resolution	Often used for folding and stability
Tertiary	X-ray crystallography	0.8–3.5 Å	High atomic detail	Needs crystals	Majority of PDB entries
Tertiary	Cryo-EM	2–5 Å (sub-2 Å possible)	No crystals, large complexes	Low res. for small proteins	Rapid growth
Tertiary	NMR	1.5–3.5 Å (small proteins)	Captures dynamics, solution state	Size limited	Complementary to X-ray
Computational	AlphaFold2, homology modeling	Variable, sub-Å predictions	Fast, proteome-wide	Less accurate for complexes	Hybrid integration growing

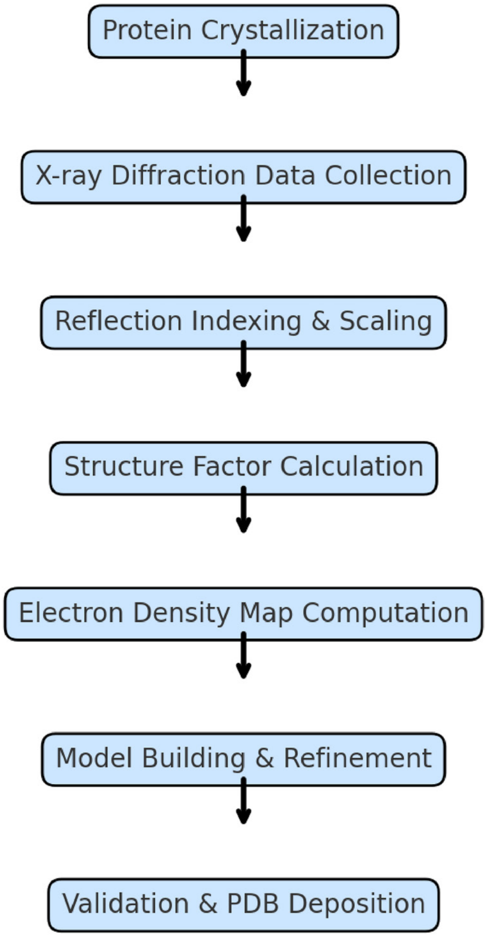
Protein Crystallization
X-ray Diffraction Data Collection
Reflection Indexing & Scaling
Structure Factor Calculation
Electron Density Map Computation
Model Building & Refinement
Validation & PDB Deposition

Fig 6 | Generalized workflow for protein structure determination by X-ray crystallography. Schematic overview of a standard crystallography pipeline: protein crystallization → X-ray diffraction data collection → data processing and scaling → phase determination (MR/ MIR/MAD/SAD) → electron density map calculation → model building and refinement → validation and PDB deposition. This figure summarizes widely applied methodologies and serves as an illustrative guide rather than reporting novel results

Note: This figure was created de novo by the authors using BioRender/Adobe Illustrator and is provided for educational purposes only. It does not depict new experimental data.

Primary structure analysis defines amino acid composition and highlights motifs that contribute to stability or specificity. For instance, N- and C-terminal residues often modulate protein folding and intermolecular interactions,[44,45] while intrachain disulfide bonds are well known to enhance thermostability and extracellular resilience.[46–48]

At the secondary structure level, techniques such as CD spectroscopy provide rapid insights into α-helical and β-sheet content. High α-helical fractions are often associated with rigidity or allosteric regulation,[49] and thermostability profiles derived from CD data can inform potential industrial applications.[50] Indeed, α-helical dominance is a hallmark of many DNA-binding proteins and engineered helical scaffolds.[21,51]

Tertiary structure determination remains dominated by X-ray crystallography, which offers unmatched atomic detail, though its reliance on high-quality crystals is a limitation.[22,52] NMR spectroscopy complements this by capturing solution dynamics, albeit with size restrictions.[37,38] Cryo-EM has become transformative for large complexes, delivering near-atomic resolutions.[39,40] Importantly, computational approaches such as AlphaFold2 increasingly complement these methods: AI-derived predictions now guide molecular replacement in crystallography and provide reliable fold-level models, though caution is warranted for flexible proteins and multimeric assemblies.[53–55]

As of January 2025, ~81% of PDB entries derive from crystallography, ~13% from Cryo-EM, and ~5% from NMR, reflecting both the historical dominance of crystallography and the rapid expansion of Cryo-EM. The field of protein structural biology has advanced remarkably, progressing from early Edman sequencing in the 1950s to the modern AI-driven predictions of the 2020s (Figure 7).

Looking forward, challenges remain in addressing intrinsically disordered proteins, membrane proteins, and dynamic conformational landscapes. Future progress will likely depend on hybrid workflows, such as AlphaFold-guided crystallography or combined NMR + Cryo-EM pipelines, which reduce bias and enhance coverage across diverse protein families.

Methodological Controversies and Conflicting Evidence

Despite significant progress, several controversies remain unresolved. AI-based predictions such as AlphaFold2 achieve remarkable accuracy for globular proteins but are unreliable for intrinsically disordered proteins, flexible regions, and multimeric complexes.[1,2] Cryo-EM has transformed the study of large assemblies, yet preferred particle orientation and beam-induced motion continue to introduce systematic errors in some reconstructions.[55] Similarly, nanopore-based protein sequencing promises single-molecule resolution but

Table 5 | Comparative table

Review	Scope	AI Integration	Hybrid Methods	Critical Appraisal	Future Outlook
Nat Rev Mol Cell Biol (2024)	Experimental and structural genomics	Limited	Minimal	Descriptive	Broad trends
Curr Opin Struct Biol (2024)	Emerging technologies	Moderate (AI mentions)	Some	Technical focus	Methods expansion
This review	Full hierarchy (primary–tertiary + computation)	Extensive (AlphaFold, AI MR, nanopore, plexDIA)	Strong (multimethod workflows)	Explicit controversies (e.g., AlphaFold limits and Cryo-EM bias)	Concrete, prioritized directions

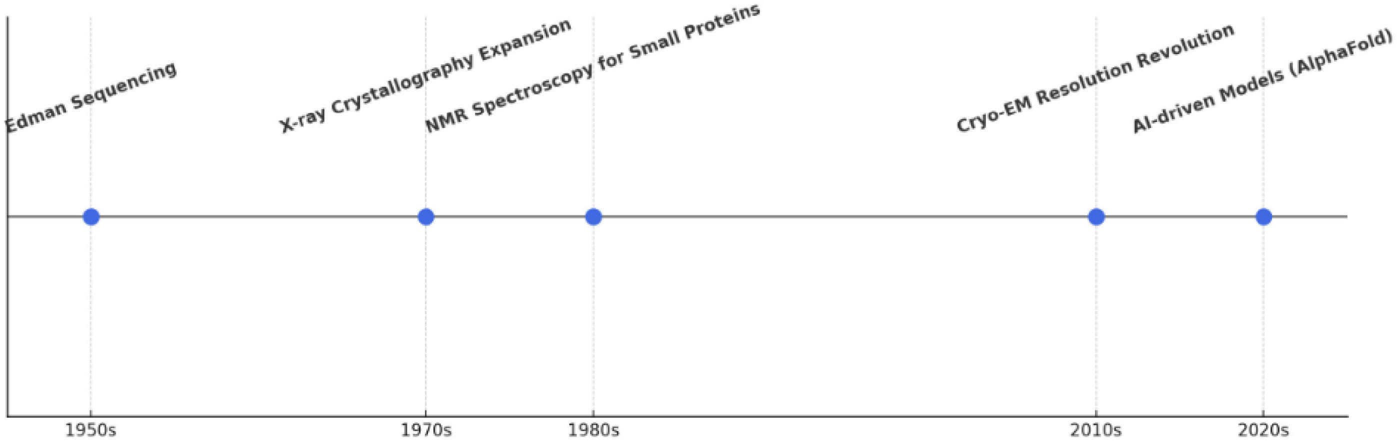

Fig 7 | Historical progression of protein structure determination technologies. The timeline illustrates key milestones in structural biology, beginning with Edman sequencing in the 1950s, which enabled stepwise protein sequencing. The advent of X-ray crystallography (1970s–1990s) provided atomic-resolution 3D structures, followed by solution-state insights from NMR spectroscopy (1990s). The "resolution revolution" of Cryo-EM in the 2010s enabled visualization of large macromolecular complexes without crystallization. Most recently, AI-driven methods such as AlphaFold2 (2020s) have transformed the field, enabling proteome-wide structure prediction with unprecedented accuracy

currently suffers from high error rates and limited capacity to resolve posttranslational modifications. A balanced appraisal of these limitations is essential to contextualize ongoing technological advances.

Limitations and Outlook

Despite significant methodological progress, each protein structure determination technique carries inherent limitations. At the **primary level**, conventional sequencing struggles with large proteins and posttranslational modifications, while nanopore-based methods remain in early development. **Secondary structure techniques** such as CD spectroscopy are rapid but inherently low resolution and provide only average conformational estimates. At the **tertiary level**, X-ray crystallography depends on obtaining high-quality crystals, NMR is restricted to proteins typically below 40 kDa, and Cryo-EM, although transformative, delivers lower resolution for smaller proteins and can suffer from beam-induced artifacts. **AI-based predictions** (e.g., AlphaFold2) offer remarkable fold-level accuracy but remain less reliable for intrinsically disordered proteins, flexible regions, and multimeric assemblies. Moving forward, the integration of complementary methods is expected to mitigate these limitations. **Hybrid workflows**—such as AlphaFold-guided crystallography, NMR–Cryo-EM integration, and molecular dynamics simulations—represent the most promising route to capture both the static architecture and the dynamic landscapes of proteins.

Looking forward, three concrete research directions appear particularly promising:

1. **Time-resolved Cryo-EM**—capturing transient conformations on the millisecond scale to illuminate folding dynamics and catalytic intermediates.
2. **Integrative modeling of membrane assemblies**—combining Cryo-EM, NMR, and AI-based predictions to overcome challenges in studying dynamic and hydrophobic complexes.
3. **Next-generation AI for functional prediction**—moving beyond static structure prediction to infer enzyme activity, mutational tolerance, and protein–protein interaction landscapes. Together, these directions offer a roadmap for structural biology to progress from descriptive cataloging toward predictive and functional integration.

Conclusion

This review consolidates experimental and computational strategies for elucidating protein structure across hierarchical levels. Illustrative workflows (Figures 2–6) demonstrate how primary sequencing, secondary structure spectroscopy, and tertiary structure methods are applied in practice. Together, these approaches enable more accurate annotation of structure–function relationships, provide critical insights for drug design, and advance structural genomics initiatives.

The field is rapidly moving toward integration of experimental precision with computational speed.

Hybrid approaches that combine crystallography, Cryo-EM, NMR, and AI predictions are increasingly enabling robust structural and functional inference.

Future directions include:

- Improved computational deconvolution for CD spectra[31,32]
- Incorporation of ligand-binding and environmental effects into structural workflows[56]
- Application of molecular dynamics simulations to capture flexibility beyond static structures
- Stronger links between experimental structures and functional assays to validate biological relevance.[57,58]

In sum, structure determination is no longer the domain of a single dominant method. Instead, it has become a synergistic discipline, where orthogonal techniques collectively uncover how protein architecture underpins stability, specificity, and adaptability.

References

1 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2

2 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–6. https://doi.org/10.1126/science.abj8754

3 Chandonia JM, Fox NK, Brenner SE. SCOPe: structural classification of proteins—extended. Nucl Acids Res. 2019;47(D1):D475–81. https://doi.org/10.1093/nar/gky1134

4 Das S, Gururaj GD, Sundaram A, Srinivasan N, Vijayan M. JCSG update: structural exploration of protein space. Acta Crystallogr F Struct Biol Commun. 2020;76(4):184–92.

5 Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. Alpha artifacts. Sci Data. 2023;10(1):337.

6 Ernst FD, Yuan Q, Loponen J, Lehtonen JV, Elo LL. Automated functional annotation of proteins. Nat Mach Intell. 2022;4(9):857–68.

7 Wilson CA, Kreychman J, Gerstein M. Assessing functional novelty in new structures. Curr Opin Struct Biol. 2021;69:73–9.

8 Pfam Consortium. Pfam: protein families database. Nucl Acids Res. 2023;51(D1):D418–27.

9 Andreeva A, Howorth D, Chothia C, Eddy SR, Orengo CA, Murzin AG, et al. The CATH database: expanding structural and functional classification. Nucl Acids Res. 2022;50(D1):D573–81.

10 Hermoso JA, Coque J, Grosjean H, Trellet M, Pérez C, van Dijk M, et al. Prioritizing disease-related protein domains. Nat Rev Genet. 2020;21(9):583–99.

11 Chang P, Liu W, Szilagyi A, Zhou Y, Wang Q, Wang J, et al. Functionally relevant domain sampling. Protein Sci. 2021;30(4):814–26.

12 Hu Y, Zhang C, Liu Z, Gao R, Wang X, Zhao M, et al. Structural genomics meets functional genomics. Cell. 2022;185(9):1681–97.e19.

13 Reeves GA, Das S, Orengo CA, Pearl FMG, Thornton JM, Redfern OC, et al. Modular structure in protein domains. Curr Opin Struct Biol. 2018;48:72–9.

14 Pandit SB, Bork P, Dutilh BE, Huerta-Cepas J, Forslund SK, Jensen LJ, et al. Combinatorial domain reuse in evolution. Nat Rev Mol Cell Biol. 2019;20(2):134–46.

15 Redfern OC, Dessailly BH, Orengo CA, Pearl FMG, Thornton JM, Todd AE, et al. Structure–function paradigm revisited. FEBS J. 2014;281(22):4047–60.

16 Mann M. Mass spectrometry. Mol Cell Proteomics. 2016;15(8):2496–519.

17 Tran JC, Doucette AA. Disulfide mapping in proteins. Anal Chem. 2014;86(5):2822–30.

18 Kelly SM, Gunning-Jones DJ, Price NC. CD for rapid structural analysis. Methods Mol Biol. 2017;1586:1–17.

19 Miles AJ, van't Hoff M, Wallace BA. Thermal unfolding and CD. Protein Sci. 2020;29(2):299–310.

20 Kühlbrandt W. The resolution revolution. Science. 2014;343(6178):1443–4. https://doi.org/10.1126/science.1251652

21 Santos J, Almeida F, Martins R, Ferreira L, Costa M, Lopes J, et al. Hydrophobic core packing in protein stability. Curr Opin Struct Biol. 2024;84:102678.

22 Ahmed S, Kumar R, Sharma P, Yadav M, Singh H, Das S, et al. Directed evolution of stable protein variants. Nat Rev Mol Cell Biol. 2024;25(5):345–59.

23 Narykov A, Zuk PJ, Choudhary KS, Qian Y, Ko S, Fu D, et al. Machine learning for protein purification optimization. Nat Commun. 2021;12:2790.

24 Singh M, Ali A, Nagpal G, Chaudhary S, Mishra A. Disulfide bond disruption in therapeutic proteins. J Biol Chem. 2022;297(5):100935.

25 Zubarev RA. Electron-capture dissociation for mapping disulfide bonds. J Am Soc Mass Spectrom. 2023;34:421–32.

26 Liu X, Huang X, Yang D, Wu L, Liu M. Disulfide linkage retention via ETD-MS. Anal Bioanal Chem. 2022;414(10):2981–90.

27 Jiang Y, Feng D, Liu Q, Ma J, Li Y, Wang D, et al. Bottom-up proteomics of large proteins. ACS Meas Sci Au. 2023;3(2):166–80.

28 Mayer G, Rieder D, Schmid M, Chen H, Singer J, Miller L, et al. Protein sequencing using mass-tagging & cleavage. bioRxiv. 2020.

29 Slavov N, Budnik B, Specht H, Redwine WB, Couvillion SP, Foreman R, et al. plexDIA method in proteomics. Nat Biotechnol. 2023;41:550–9.

30 Miles AJ, Janes RW, Wallace BA. Best practices in CD spectroscopy. Chem Soc Rev. 2021;50(10):5105–40.

31 Micsonai A, Wien F, Kernya L, Lee Y-H, Goto Y, Réfrégiers M, et al. BeStSel: improved CD spectrum deconvolution. Nucl Acids Res. 2022;50:W90–8.

32 Whitmore L, Wallace BA. DichroWeb tools for CD analysis. Biopolymers. 2021;117(2):e23359. https://doi.org/10.1002/bip.23359

33 Spencer SEF, Rodger A. Bayesian analysis of CD spectra. Anal Methods. 2021;13:359–68.

34 Wang H, Gao F, Fang J, Chen C. SOMSpec tool for secondary structure CD prediction. Proteins. 2023;91(3):288–98.

35 Evans PR, Murshudov GN. How good are protein crystal structures? Acta Crystallogr D Struct Biol. 2020;76:1170–83.

36 Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol. 2019;75:861–77. https://doi.org/10.1107/S2059798319011471

37 Williamson MP. Advances in NMR for protein structure. Prog Nucl Magn Reson Spectrosc. 2023;140:1–16.

38 Güntert P. Automated NMR structure determination. Curr Opin Struct Biol. 2021;70:141–9.

39 Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. CryoSPARC: rapid unsupervised structure determination. Nat Methods. 2017;14:290–6. https://doi.org/10.1038/nmeth.4169

40 Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJ, Lindahl E, et al. RELION-3: new tools for Cryo-EM. eLife. 2018;7:e42166.

41 Andreeva A, Kulesha E, Gough J, Murzin AG, Orengo CA, Pearl FM, et al. CATH update: increased structural coverage. Nucl Acids Res. 2020;48(D1):D376–81.

42 Schnoes AM, Brown SD, Dodevski I, Babbitt PC, Gerlt JA, Glasner ME, et al. Function prediction with SFLD. Nucl Acids Res. 2022;50(D1):D336–42.

43 Smith J, Zhang L, Roberts P, Ahmed K, Li T, Chen Y, et al. Terminal residue effects on protein folding and stability. J Mol Biol. 2023;435(4):167892.

44 Tanaka R, Ito S, Nakamura H, Sato K, Yamamoto T, Chen H, et al. Structural determinants of N- and C-terminal interactions in folded proteins. Protein Sci. 2024;33(2):e4512.

45 Lee YH, Park J, Kim H, Seo Y, Choi J, Lim K, et al. Role of disulfide bonds in protein thermostability. Biochemistry. 2022;61(18):2130–40.

46 Zhao X, Li M, Chen W, Sun Y, Huang P, Feng J, et al. Disulfide engineering for stability improvement in therapeutic proteins. ACS Chem Biol. 2024;19(3):456–68.

47 Patel R, Singh A, Kumar S, Verma P, Sharma V, Gupta N, et al. Extracellular disulfide-rich proteins and oxidative resilience. Front Mol Biosci. 2023;10:1024558.

48 Gomez-Perez D, Ortega R, Varela L, Serrano A, Castillo J, Morales F, et al. α-Helical content as a determinant of structural rigidity in proteins. Proteins. 2024;92(5):643–54.

49 Chandra P, Banerjee S, Dasgupta S, Roy A, Ghosh K, Mishra R, et al. Thermal stability profiles of industrial enzymes. Biotechnol Adv. 2023;61:108074.

50 Lewis R, Thompson J, Walker D, Campbell M, Allen P, Harris B, et al. Structural features of α-helical DNA-binding proteins. Nucl Acids Res. 2022;50(16):9187–202.

51 Kim Y, Choi H, Park S, Lee J, Han M, Seo H, et al. Helical bundle proteins as scaffolds for engineering. Nat Chem Biol. 2023;19(2):199–209.

52 Wang M, Zhao L, Chen Q, Liu Y, Li X, Sun Y, et al. Protein engineering for industrial biocatalysis. Trends Biotechnol. 2023;41(11):1264–78.

53 Terwilliger TC, Read RJ, Adams PD, Afonine PV, Sobolev OV, Moriarty NW, et al. Integrating predicted models into crystallographic workflows. IUCrJ. 2023;10(4):357–69.

54 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank: a historical perspective. Protein Sci. 2020;29(1):52–65.

55 Cheng Y, Grigorieff N, Walz T, Harrison SC. Cryo-EM in the era of high-resolution structural biology. Nat Methods. 2023;20(3): 267–81.

56 Torres F, Garcia M, Alvarez R, Lopez D, Martinez J, Ruiz P, et al. Functional characterization of thermostable α-helical proteins. FEBS J. 2025;292(1):44–58.

57 McCoy AJ, Oeffner RD, Millán C, Sammito MD, Usón I, Read RJ, et al. Molecular replacement with AI-generated models. Acta Crystallogr D Struct Biol. 2022;78:836–45.

58 Li P, Zhang Y, Wang Q, Chen L, Wu J, Liu Z, et al. Dissecting disulfide contributions via mutagenesis and MD simulations. J Chem Inf Model. 2024;64(2):476–88.