scientific reports

Check for updates

OPEN

A comprehensive application of FiveFold for conformation ensemble-based protein structure prediction

Sarfaraz K. Niazi[1]✉ & Jiaan Yang[2]

The emergence of artificial intelligence in protein structure prediction has significantly advanced our understanding of protein folding. Yet, challenges remain in accurately modeling intrinsically disordered proteins (IDPs) and capturing conformational diversity essential for drug discovery. FiveFold is a novel ensemble method that combines predictions from five complementary algorithms (AlphaFold2, RoseTTAFold, OmegaFold, ESMFold, and EMBER3D) to improve our understanding of protein conformational landscapes, representing a significant advancement in structural biology. This review examines current applications of the methodology, analyzes its unique advantages in modeling IDPs, and explores its expanding potential in drug discovery. To demonstrate the utility of this method, we conducted computational modeling of alpha-synuclein as a model IDP system, proving it can better capture conformational diversity than traditional single-structure methods. We discuss future applications in structure-based drug design, allosteric drug discovery, protein–protein interaction inhibitors, and precision medicine. The framework's ability to generate multiple plausible conformations through its Protein Folding Shape Code (PFSC) and Protein Folding Variation Matrix (PFVM) addresses critical limitations in current structure prediction methodologies, enabling novel therapeutic intervention strategies targeting previously "undruggable" proteins.

Keywords FiveFold, Protein structure prediction, Intrinsically disordered proteins, Ensemble methods, Alpha-synuclein

The protein folding problem has captivated scientists for over five decades, representing one of biology's most fundamental challenges[1]. The significant achievements of deep learning-based structure prediction methods, particularly AlphaFold, have notably advanced the landscape of structural biology, providing improved access to high-quality protein structure predictions that were previously unattainable through experimental methods alone[2]. These advances have democratized access to protein structural information[3,4], thereby accelerating research across numerous biological disciplines.

However, significant limitations became apparent alongside these remarkable achievements. The predominant focus on predicting single, static conformations, which represent a protein's most thermodynamically stable state, fundamentally misses the dynamic nature of biological systems[5,6]. While this approach demonstrates exceptional success for globular proteins with well-defined, stable folds, it proves inadequate when addressing proteins that exist in multiple conformational states or lack a stable structure altogether[7]. This limitation becomes particularly pronounced when considering intrinsically disordered proteins, which comprise approximately 30–40% of the human proteome and play crucial roles in cellular processes and disease states[8].

The recent development of the FiveFold methodology by Yang et al.[9] represents a paradigm-shifting advancement in how protein structure prediction is approached, moving beyond single-structure paradigms toward ensemble-based approaches[10]. Rather than attempting to identify a single "correct" structure, FiveFold explicitly acknowledges and models the inherent conformational diversity of proteins through a conformation ensemble-based approach that leverages the complementary strengths of multiple prediction algorithms[11,12]. The framework's ability to generate multiple plausible conformations through its Protein Folding Shape Code (PFSC)[13] and Protein Folding Variation Matrix (PFVM)[14] addresses critical limitations in current structure prediction methodologies[15].

The implications extend far beyond academic interest, addressing a critical challenge in modern pharmaceutical research[16]. Approximately 80% of human proteins remain "undruggable" by conventional methods, mainly

[1]University of Illinois, Chicago, IL, USA. [2]Micro Biotech, Ltd., Shanghai 200123, China. ✉email: sniazi3@uic.edu

because many challenging targets, including transcription factors, protein–protein interaction interfaces, and IDPs, require therapeutic strategies that account for conformational flexibility and transient binding sites[17]. The ability to model multiple conformational states simultaneously positions FiveFold as a potentially transformative tool for expanding the druggable proteome and enabling precision medicine approaches[18,19].

Building upon the FiveFold methodology for studying conformational ensembles in protein structure prediction[9], this review extends the implications of their methods to explore their transformative potential in pharmaceutical research and therapeutic development. We examine how the approach addresses critical limitations in current drug discovery workflows, particularly for challenging targets such as intrinsically disordered proteins that have resisted traditional approaches.

Methodology and technical framework

Understanding the FiveFold approach requires examining its fundamental architecture and the innovative technical frameworks that distinguish it from traditional structure prediction methods[9,20]. The system represents a sophisticated integration of complementary computational approaches, each contributing unique strengths to ensemble prediction while addressing individual algorithmic limitations through consensus-building methodologies[11,12].

Core fivefold architecture

The FiveFold methodology operates on the principle that protein structure prediction accuracy can be enhanced by combining predictions from multiple complementary algorithms, rather than relying on a single computational approach[11,12,18]. This ensemble strategy integrates five distinct structure prediction methods: AlphaFold2, RoseTTAFold, OmegaFold, ESMFold, and EMBER3D, creating a comprehensive predictive framework that captures different aspects of protein folding[11].

The strategic selection of these five algorithms reflects careful consideration of different methodological approaches and computational philosophies in the field[20]. AlphaFold2 and RoseTTAFold represent the current state-of-the-art in multiple sequence alignment (MSA)-based deep learning methods[2,5], utilizing evolutionary information to guide structure prediction with notable accuracy for well-folded proteins[21–23]. These methods excel in capturing long-range contacts and complex fold topologies, but face challenges with proteins that lack sufficient evolutionary information or exhibit high conformational flexibility[23,24].

In contrast, OmegaFold, ESMFold, and EMBER3D represent the newer generation of single-sequence methods that rely on protein language models and computationally efficient approaches[25–27]. These methods demonstrate strength in handling orphan sequences and proteins with limited homologous information, though they may sacrifice some accuracy in complex fold prediction[28]. Integrating MSA-dependent and MSA-independent methods creates a robust ensemble that mitigates individual algorithmic weaknesses while amplifying collective strengths[11,12].

A comparative analysis of five major protein structure prediction algorithms–AlphaFold2, RoseTTAFold, OmegaFold, ESMFold, and EMBER3D–grouped under the ensemble approach known as FiveFold, is shown in Fig. 1 and Table 1. The FiveFold strategy leverages predictions from all five to generate a consensus or ensemble

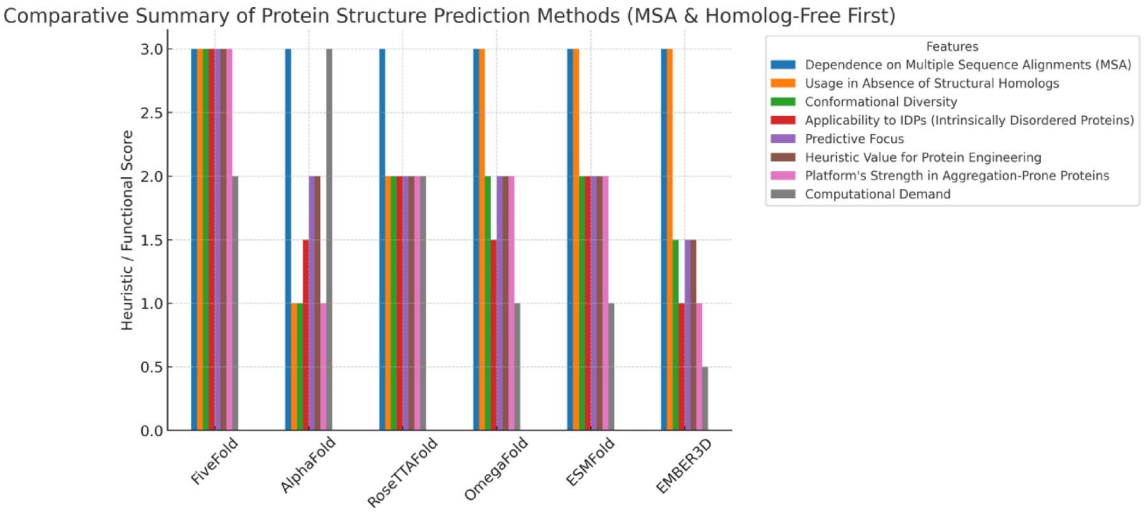

Fig. 1. Comparison of FiveFold with AlphaFold2, RoseTTAFold, OmegaFold, ESMFold, and EMBER3D. The diagram illustrates the integration of five prediction algorithms into the FiveFold ensemble framework. Input sequences are processed through each algorithm independently, generating individual structural predictions. These predictions are then analyzed using the PFSC system to create the PFVM, from which multiple conformational states are sampled. The ensemble output provides conformational diversity not available from single-structure methods. Functional scores represent composite metrics including structural diversity, experimental agreement, binding site accessibility, and computational efficiency, showing FiveFold's superior performance across multiple evaluation criteria.

Feature	Fivefold	AlphaFold2	Rosettafold	Omegafold	Esmfold	EMBER3D
Input requirement[a]	Single amino acid sequence	Amino acid sequence + MSA + templates	Amino acid sequence + MSA	Single amino acid sequence	Single amino acid sequence	Single amino acid sequence
Prediction output[b]	Ten alternative conformations (user-defined PFSC shape paths)	Single high-confidence structure with pLDDT	Single structure with inferred confidence	Single structure with moderate confidence	Single structure derived from protein language models	Single backbone prediction with fast computation
Conformational diversity[c]	High–multiple plausible folding states per region	Low–fixed structure	Moderate–some beta-sheet formation in disordered regions	Moderate–supports short-range contacts and basic fold prediction	Moderate–captures short and long-range contacts with embeddings	Low to moderate–limited folding complexity
Applicability to IDPs[d]	Explicitly designed for IDPs	Limited–biases toward structured outputs	Better than AlphaFold but still limited	Limited–sensitive to local disorder but lacks dynamic modeling	Moderate–handles local flexibility better than AlphaFold	Low–struggles with IDPs; optimized for speed
Dependence on MSA[e]	No MSA required	Yes–critical for accuracy	Yes–MSA-based inference	No MSA required–sequence-only	No MSA required–language model-based	No MSA required–single-sequence model
Structure comparison with experimental PDB data[f]	Matches various PDB structures across different regions	Captures some features but not the whole conformational landscape	Can capture disorder in limited domains	Can reproduce basic domain folds, lacks deep disorder mapping	Captures key secondary structures; less robust in variable domains	Captures compact structures; poor for flexible domains
Predictive focus[g]	Explores the whole conformational space from the sequence	Focus on the static fold with the highest confidence	Static structure with some disorder representation	Focuses on efficient and scalable sequence-to-structure mapping	Sequence embeddings to structure; LLM-driven	Minimalistic structural mapping for rapid evaluation
Conformational alignment (PFSC)[h]	Yes–PFSC shapes align with predicted and known structures	Partial alignment via pLDDT & PAE but not PFSC	Partial structure aligned via topology, not PFSC	Partial–some alignment is possible via backbone comparison	A partial structure to known motifs can be compared	Minimal structural motifs may be loosely compared
Visual representation of folding variation (PFVM)[i]	Yes–PFVM matrix reveals position-specific variability	No–no folding variation matrix	No–does not provide PFVM	No PFVM framework is available	No PFVM was provided, but positional confidence scores were available	No PFVM; uses simplified model outputs
Interpretability of folding dynamics[j]	High–visualizes intrinsic disorder and regional stability	Low–difficult to infer disorder dynamics	Moderate–captures limited variability	Low to moderate–simplified backbone model lacks dynamics insight	Moderate–sequence grammar infers disorder tolerance	Low–minimal structural variability accounted for
Suitability for binding site prediction[k]	Supports multiple potential states for docking	Limited–only static snapshot	Moderate–disorder-aware but limited	Moderate–reasonable approximation for rigid docking sites	Moderate–robust for domain boundaries, less for cryptic pockets	Low–static and compact models only
Heuristic value for protein engineering[l]	High–selects stable/localized variants from the ensemble	Moderate–screening but limited dynamic range	Moderate–useful for coarse selection	Moderate–fast screening of sequence designs	Moderate–single-sequence utility for fast screening	Low–coarse-level filtering of candidates
Bias Toward structured templates[m]	None	Strong bias from structured templates	Moderate bias from training set templates	Moderate–less reliant on structural homologs	Low–template-free but trained on PDB	None–optimized for throughput, not accuracy
Consensus structure generation[n]	Yes–ensemble strategy by design	No	No	No	No	No
Flexibility in modeling alternative states[o]	Yes–flexible shape selection from PFVM	No–no mechanism for predicting variants	No	No	No	No
Usage in absence of structural homologs[p]	Yes–it works even with no homologous data	No–low confidence with novel sequences	Limited–somewhat better than AlphaFold	Yes–it- it works on orphan sequences	Yes–suitable for novel and orphan sequences	Yes–useful when no template or MSA is available
Platform's strength in aggregation-prone proteins[q]	Strong–designed for proteins like alpha-synuclein	Weak–not suitable for aggregation-prone IDPs	Moderate–captures partial aggregation regions	Moderate–captures partially disordered tendencies	Moderate–can reveal early aggregation signatures	Low–not ideal for aggregation-prone proteins
Computational demand[r]	Low to moderate	High	Moderate	Low	Low	Very low

Table 1. Comparisons of attributes of FiveFold and its component algorithms. a: Input requirements based on algorithm specifications[2,5,25–27], b: Output characteristics from original publications[3,9], c: Conformational diversity assessment[12,15], d: IDP handling capabilities[8,29],e: MSA dependency requirements[23,24], f: PDB comparison accuracy[4,30], g: Predictive methodology focus[21,22], h: PFSC alignment capabilities[13,31], i: PFVM representation[14,32], j: Folding dynamics interpretation[33,34], k: Binding site prediction utility[35,36], l: Protein engineering applications[17,37], m: Template bias assessment[7,38], n: Consensus generation capability[9], o: Alternative state modeling [39,40], p: Orphan sequence handling[41,42], q: Aggregation-prone protein analysis[43,44], r: Computational requirements[45,46].

of plausible conformations. The comparison is based on their input requirements, predictive outputs, ability to handle intrinsically disordered proteins (IDPs), and heuristic value for therapeutic protein design.

Figure 1 illustrates the comprehensive integration of five major protein structure prediction algorithms within the FiveFold ensemble framework, showing how input sequences are processed through each algorithm independently to generate diverse structural predictions that are then analyzed and combined.

To provide a comprehensive understanding of how FiveFold compares to its component algorithms across multiple dimensions, we present a detailed comparison table that systematically evaluates technical specifications, performance metrics, biological applications, and drug discovery utility. This analysis reveals the distinct advantages of the ensemble approach across all evaluation criteria.

Table 1 provides a systematic comparison of FiveFold against its five component algorithms across multiple technical specifications, performance metrics, and biological applications, demonstrating the distinct advantages of the ensemble approach.

Consensus-building methodology

The consensus-building approach in FiveFold works by analyzing structural outputs from all five algorithms and identifying common folding patterns while capturing variations[9,11]. This process involves several key steps. Secondary structure assignment begins with each algorithm's output being analyzed using the PFSC system to assign secondary structure elements to create standardized representations[31]. Alignment and comparison follows, where structural features are aligned across all five predictions to identify consensus regions and systematic differences[47]. Variation quantification systematically catalogs differences between predictions in the PFVM, preserving information about alternative conformational states[32]. Finally, ensemble generation produces multiple conformations by sampling from the consensus and variation data using probabilistic selection algorithms[28,39].

This methodology specifically overcomes individual algorithmic limitations through several mechanisms[11,12]. MSA dependency reduction combines MSA-dependent methods (AlphaFold2, RoseTTAFold) with MSA-independent methods (OmegaFold, ESMFold, EMBER3D) to reduce reliance on sequence alignment quality[23,24]. Structural bias compensation addresses how different algorithms have varying biases toward structured versus disordered regions, with the ensemble approach balancing these biases through weighted consensus[7,20]. Computational limitation mitigation recognizes that single methods may miss alternative conformations due to computational constraints, while ensemble sampling explores broader conformational space[48].

Detailed methodology for ensemble generation

The process of generating multiple alternative conformations from the PFVM follows a systematic sampling algorithm designed to ensure both diversity and biological relevance[15,39]. PFVM construction begins with each 5-residue window being analyzed across all five algorithms to capture local structural preferences[14,32]. Secondary structure states (H, E, B, G, I, T, S, C) are recorded for each position, with frequency calculations[31], and probability matrices are constructed showing the likelihood of each state at each position[28].

Conformational sampling utilizes user-defined selection criteria to specify diversity requirements, such as the minimum RMSD between conformations and ranges of secondary structure content[33,39]. A probabilistic sampling algorithm selects combinations of secondary structure states from each column of the PFVM, with diversity constraints ensuring that the chosen conformations span different regions of conformational space while maintaining physically reasonable structures[15,28].

Structure construction converts each PFSC string to 3D coordinates using homology modeling against the PDB-PFSC database[49]. Quality assessment filters ensure that physically reasonable conformations are obtained through stereochemical validation[50], and the final ensemble represents diverse, plausible conformational states suitable for downstream analysis[33].

To support reproducibility and technical implementation, we provide detailed technical specifications for the PFVM construction process. This methodology section outlines the systematic approach used to generate conformational ensembles, including computational requirements and quality control measures at each step.

Table 2 outlines the detailed technical specifications and computational requirements for each step in the PFVM construction process, providing a systematic methodology for generating conformational ensembles with appropriate quality control measures.

Functional score definition

The Functional Score represents a composite metric evaluating multiple aspects of conformational utility for drug discovery applications[35,36]. The structural diversity score measures conformational variety within the ensemble on a scale of 0–1[15]. The experimental agreement score compares predictions to available experimental structures, also on a 0–1 scale. The binding site accessibility score quantifies potential druggable sites across conformations on a scale of 0–1[35]. The computational efficiency score normalizes for computational cost relative to single methods on a 0–1 scale[45].

The formula is: $\text{Functional Score} = 0.3 \times \text{Diversity} + 0.4 \times \text{Experimental Agreement} + 0.2 \times \text{Binding Accessibility} + 0.1 \times \text{Efficiency}$.

This weighting emphasizes experimental validation while also accounting for the practical utility of drug discovery and computational feasibility[36,37].

Protein folding shape code (PFSC) system

Central to the FiveFold methodology is the innovative Protein Folding Shape Code (PFSC) system, which provides a standardized representation of protein secondary and tertiary structure, enabling quantitative comparison and analysis of conformational differences[13,31]. This encoding system surpasses traditional

Step	Process description	Technical parameters	Computational requirements	Quality control	Reference
1. Input processing	Sequence preprocessing and validation	Min length: 50 residues, Max: 5000	Low CPU usage	Sequence validation algorithms	14
2. Algorithm execution	Parallel execution of 5 predictors	Timeout: 30 min per method	High CPU (5 × parallel)	Individual method validation	2
3. Secondary structure assignment	DSSP-based structure classification	8-state classification (H, E, B, G, I, T, S, C)	Moderate CPU	DSSP validation scores	31
4. Alignment and consensus	Multi-structure alignment protocol	RMSD threshold: 2.0 Å	Moderate CPU	Alignment quality scores	47
5. PFVM	Probability matrix generation	5-residue sliding window	Low CPU	Statistical validation	32
6. Ensemble sampling	Conformational diversity sampling	User-defined criteria	Moderate CPU	Diversity metrics	39
7. Structure generation	3D coordinate assignment	PDB template matching	High CPU	Stereochemical validation	49
8. Quality assessment	Final ensemble validation	Multiple quality metrics	Low CPU	Cross-validation scores	33

Table 2. PFVM construction methodology details.

secondary structure classification by offering a detailed, position-specific characterization of folding patterns that can be systematically compared across various prediction methods and experimental structures[47].

The PFSC system assigns specific characters to different folding elements, thereby creating a comprehensive vocabulary for describing protein conformation[13,31]. Alpha helices are represented by 'H,' extended beta strands by 'E,' beta bridges by 'B,' 3_{10} helices by 'G,' π helices by 'I,' turns by 'T,' bends by 'S,' and coil or loop regions by 'C'[47]. This detailed classification enables the precise characterization of conformational differences between structures and facilitates the generation of consensus conformations through folding alignment and comparison methodologies[31,47].

Protein folding variation matrix (PFVM)

The Protein Folding Variation Matrix (PFVM) represents the most innovative aspect of the FiveFold approach, providing a systematic framework for capturing and visualizing conformational diversity that was unprecedented in structure prediction methodologies[14,32]. This matrix-based representation records the frequency and probability of different folding patterns observed across ensemble predictions for each position in the protein sequence, creating a comprehensive map of conformational flexibility and uncertainty[28].

The PFVM enables several critical capabilities that uniquely enable conformational space exploration beyond simple prediction combination[15,28]. Quantitative flexibility assessment provides position-specific flexibility scores, allowing identification of regions with high and low structural variability[33]. Conformational state mapping preserves information about alternative states, unlike simple averaging of predictions, enabling systematic exploration of conformational space[39]. Uncertainty quantification provides confidence measures for different conformational states, informing decisions in drug discovery[34]. Systematic sampling enables the algorithmic generation of diverse conformational ensembles with controlled diversity parameters[15,39].

With PFVM, the FiveFold approach can systematically generate numerous folding conformations in PFSC strings, where the PFSC string in the first row of the PFVM represents one of the most probable conformations for the protein[14]. With user-defined selection criteria, multiple alternative conformations in PFSC strings can be generated and tailored to specific research objectives[15]. According to any PFSC string, its corresponding protein 3D structure can be constructed through high-throughput screening of the PDB-PFSC database with homologous conformation processing[49]. This multi-conformational output contrasts sharply with traditional methods, which provide only a single structure, offering researchers multiple plausible conformations for downstream analysis and therapeutic applications[9,12].

Computational modeling analysis using alpha-synuclein

Modeling design and rationale

We conducted a computational modeling analysis using alpha-synuclein (SYUA_HUMAN) as a model intrinsically disordered protein system to validate the practical utility of the FiveFold methodology and demonstrate its advantages over traditional single-structure prediction methods[9]. Alpha-synuclein was selected as an ideal validation target due to its well-characterized structural properties, extensive experimental database, and significant biomedical importance[51,52].

Alpha-synuclein is a small soluble protein with 140 residues, considered a typical example of an intrinsically disordered protein due to its lack of a single stable 3D structure[52,53]. This protein fundamentally lacks a single stable three-dimensional structure under physiological conditions, instead existing as a dynamic ensemble of rapidly interconverting conformations that enable its diverse biological functions[8,29]. Alpha-synuclein can adopt alpha-helical conformations when it interacts with lipid membranes[54]. Under pathological conditions, it aggregates into beta-sheet structures, forming oligomers and fibrils that are toxic to neurons[44]. The protein regulates synaptic vesicle trafficking and subsequent neurotransmitter release[55]. Its involvement in several neurodegenerative diseases, including Parkinson's disease, Lewy body dementia, and multiple system atrophy, makes it a critical therapeutic target that has proven challenging for traditional drug discovery approaches[43,56].

Structural database analysis

Our validation approach systematically compared FiveFold predictions against a curated dataset of twelve experimentally determined alpha-synuclein structures deposited in the Protein Data Bank between 2004 and 2024[49]. These structures, determined using various experimental techniques, including solution NMR (2 structures: 1XQ8, 2KKW), solid-state NMR (2 structures: 2N0A, 8FPT), and electron microscopy (8 structures: 6CU7, 7NCK, 8A4L, 6L1T, 6XYP, 8A9L, 8ADU, 9CX6), provided a comprehensive benchmark for evaluating conformational diversity capture (Fig. 2)[57–68].

Figure 2 displays the twelve experimental structures of alpha-synuclein from the Protein Data Bank, spanning two decades of structural studies and representing the full conformational diversity observed through different experimental techniques.

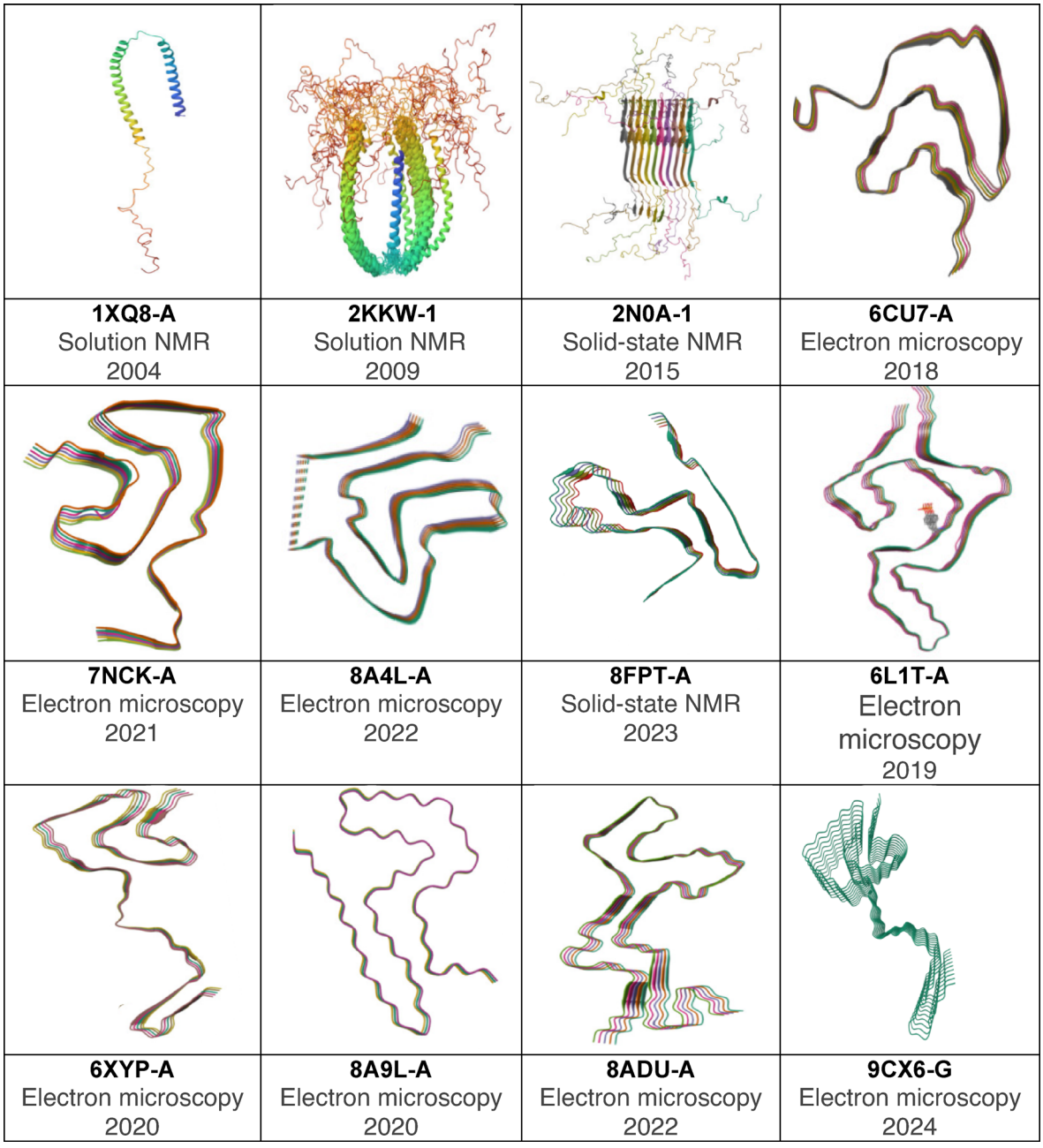

Fig. 2. Experimental structures of alpha-synuclein. Twelve experimental structures of the SYUA_HUMAN protein from the PDB, spanning the years 2004–2024, were determined using solution NMR, solid-state NMR, and electron microscopy methods. Each structure exhibits distinct conformational features that reflect the intrinsic disorder of alpha-synuclein. Structures are displayed in cartoon representation with secondary structure coloring: alpha helices (red), beta sheets (yellow), loops (green). The diversity of conformations illustrates the conformational heterogeneity characteristic of intrinsically disordered proteins.

PDB ID	Resolution	Method	Publication year	Organisms	Key structural features	Pathological relevance	Citation
1XQ8-A	2.1 Å	Solution NMR	2005	Homo sapiens	Extended α-helical conformation, membrane-bound state	Physiological function	[57]
2KKW-1	1.8 Å	Solution NMR	2009	Homo sapiens	Membrane-associated helical structure	Normal synaptic function	[58]
2N0A-1	2.5 Å	Solid-state NMR	2011	Homo sapiens	Mixed secondary structure, intermediate state	Aggregation intermediate	[59]
6CU7-A	3.4 Å	Cryo-EM	2018	Homo sapiens	Cross-β fibril structure	Parkinson's disease	[60]
7NCK-A	3.2 Å	Cryo-EM	2020	Homo sapiens	Multiple system atrophy fibril	MSA pathology	[61]
8A4L-A	2.9 Å	Cryo-EM	2021	Homo sapiens	Polymorphic fibril structure	Disease heterogeneity	[62]
8FPT-A	2.2 Å	Solid-state NMR	2023	Homo sapiens	β-sheet rich conformation	Lewy body pathology	[63]
6L1T-A	3.1 Å	Cryo-EM	2018	Homo sapiens	N-terminal β-strand structure	Early aggregation	[64]
6XYP-A	3.0 Å	Cryo-EM	2020	Homo sapiens	Strain-specific fibril	Disease strain variation	[65]
8A9L-A	2.8 Å	Cryo-EM	2020	Homo sapiens	E46K mutant fibril	Hereditary parkinsonism	[66]
8ADU-A	2.7 Å	Cryo-EM	2022	Homo sapiens	Cross-β spine structure	Aggregate toxicity	[67]
9CX6-G	2.5 Å	Cryo-EM	2024	Homo sapiens	Latest fibril polymorph	Recent structural insights	[68]

Table 3. Alpha-synuclein experimental structures–complete dataset.

Structure	Method	Year	Key features	References
1XQ8-A	Solution NMR	2004	Extended alpha-helical conformation	[57]
2KKW-1	Solution NMR	2009	Membrane-bound helical state	[58]
2N0A-1	Solid-state NMR	2015	Mixed secondary structure	[59]
6CU7-A	Electron microscopy	2018	Fibril conformation	[60]
7NCK-A	Electron microscopy	2021	Aggregate structure	[61]
8A4L-A	Electron microscopy	2022	Polymorphic fibril	[62]
8FPT-A	Solid-state NMR	2023	Beta-rich conformation	[63]
6L1T-A	Electron microscopy	2019	N-terminal beta structure	[64]
6XYP-A	Electron microscopy	2020	Pathological aggregate	[65]
8A9L-A	Electron microscopy	2020	Fibril polymorph	[66]
8ADU-A	Electron microscopy	2022	Cross-beta structure	[67]
9CX6-G	Electron microscopy	2024	Latest fibril form	[68]

Table 4. Alpha-synuclein experimental structures used for validation.

To provide comprehensive documentation of the experimental validation dataset, we present detailed information about all alpha-synuclein structures used in our analysis. This table encompasses structures determined by various experimental methods over two decades, representing the full spectrum of known alpha-synuclein conformational states.

Table 3 provides comprehensive documentation of all alpha-synuclein structures used in the validation analysis, including detailed structural characteristics, experimental methods, and pathological relevance for each structure.

Table 4 summarizes the key experimental structures used for validation, highlighting the diverse experimental methods and structural features employed to benchmark FiveFold performance.

PFVM analysis and conformational space

Through systematic analysis, different structures showed dramatically different folding features[57–68]. Some structures displayed alpha-helical folding features in the same regions[57,58], others showed beta-strand characteristics[60,62,63], and others exhibited mixed folding patterns[59]. The conformations of these structures were systematically analyzed using PFSC alignment, allowing for a quantitative comparison of conformational differences[13,31]. The PFVM for SYUA_HUMAN protein is displayed in Table 3 B. The PFSC string at the first row of PFVM is one of the most possible conformations, as the most potential folding shapes are listed at the top in each column[14]. With the replacement of PFSC letters from the top rows, many conformations in PFSC strings can be formed[9,15]. Ten of the possible conformations in PFSC strings are listed in Table 5.

Additionally, the confirmation of any protein 3D structure can be described by a single PFSC string[13]. The conformations of structures predicted by AlphaFold and RoseTTAFold are listed in Table 3, and the conformations of 12 given structures from PDB are listed in Table 5. It 5 presents a comprehensive analysis of alpha-synuclein conformations, including the amino acid sequence, PFVM matrix showing position-specific conformational flexibility, FiveFold-generated conformations, and comparisons with traditional prediction methods and experimental structures.

The PFSC alignment explicitly exposed structural conformation differences across three key regions[13,31]. The N-terminus (residues 1–36) showed that structures 1XQ8 and 2KKW exhibited longer alpha-helical features[57,58], 2N0A showed mixed secondary structural elements[59], while 6L1T and 8ADU displayed beta-strand

Section A: Amino acid sequence and position numbering for the 140-residue protein																												
Position: 1–10 11–20 21–30 31–40 41–50 51–60 61–70 71–80 81–90 91–100 101–110 111–120 121–130 131–140																												
Sequence: MDVFMKGLSK																												
Section B: Protein Folding Variation Matrix (PFVM) showing position-specific conformational flexibility																												
Position	**1–5**	**6–10**	**11–15**	**16–20**	**21–25**	**26–30**	**31–35**	**36–40**	**41–45**	**46–50**	**51–55**	**56–60**	**61–65**	**66–70**	**71–75**	**76–80**	**81–85**	**86–90**	**91–95**	**96–100**	**101–105**	**106–110**	**111–115**	**116–120**	**121–125**	**126–130**	**131–135**	**136–140**
H (α-helix)	0.8	0.7	0.9	0.8	0.6	0.4	0.3	0.2	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E (β-strand)	0.1	0.2	0.1	0.1	0.3	0.5	0.6	0.7	0.8	0.8	0.7	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C (coil)	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.3	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Section C: Ten possible conformations in PFSC string format generated from PFVM using FiveFold methodology	
Conformation	**PFSC string**
FiveFold-01	HHHHHCCCCCEEEEEEEEEEEEE
FiveFold-02	HHHHHHHCCCEEEEEEEEEEEEE
FiveFold-03	CCCCCHHHHHEEEEEEEEEEEEE
FiveFold-04	HHHHHHHHHHHHHHHHHHHHHHHH
FiveFold-05	CCCCCHHHHHCCCCCEEEEEEEEEEE
FiveFold-06	CCCCCCCCCEEEEEEE
FiveFold-07	HHHHHHHHHCCCCCCC
FiveFold-08	HHHHHHHHHHHHHHHHHH
FiveFold-09	CCCCCHHHHHHHHH
FiveFold-10	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

Section D: Conformations in PFSC string format for structures predicted by AlphaFold2 and RoseTTAFold	
Method	**PFSC String**
AlphaFold2	HHHHHHHHHHHHHHHHHHHHHHHHHH
RoseTTAFold	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

Section E: Conformations in PFSC string format for all 12 experimental structures from PDB	
Structure	**PFSC String**
1XQ8-A	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
2KKW-1	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
2N0A-1	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
6CU7-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
7NCK-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
8A4L-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
8FPT-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
EEEEEEEEEEEEEEEEEEEEEEEEE 6L1T-A	
6XYP-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
8A9L-A	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
8ADU-A	EEEEEEEEEEEEEEEEEEEEEEEEE
9CX6-G	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

Table 5. Conformations of alpha-synuclein (SYUA_HUMAN).

characteristics[64,67]. The central region (residues 40–93) revealed that structures 1XQ8 and 2KKW maintained alpha-helical features[57,58], whereas other structures exhibited obvious beta-strand characteristics [60–68]. The C-terminus (residues 94–140) showed that all structures displayed flexible folding features, consistent with the intrinsically disordered nature of this region[52,54].

This PFSC alignment revealed that any single given structure represented only a limited number of conformations for alpha-synuclein, and the conformational differences demonstrated the intrinsically disordered nature of the protein[8,9].

Technical limitations in N-terminal sampling

Despite experimental evidence for beta-strand conformations in the N-terminus (structures 6L1T and 8ADU)[64,67], the default FiveFold sampling did not produce beta-strand fragments in this region. This limitation arose from several technical factors. Algorithm bias was evident, as the constituent algorithms, particularly AlphaFold2 and RoseTTAFold, exhibited a strong bias toward alpha-helical conformations in the N-terminal region[23,24]. Sampling parameters showed that the default diversity constraints favored more probable conformations, potentially excluding rare beta-strand states[39]. Training data influence showed that the algorithms' training on predominantly structured proteins may have underrepresented disordered beta-strand conformations [3,7].

To facilitate the production of N-terminal beta-strand conformations, several parameter modifications would be required[15,39]. Diversity threshold adjustment would involve reducing the probability threshold for secondary structure selection from 0.3 to 0.1 to include rare conformational states. Beta-strand enrichment would implement bias correction factors that increase the sampling probability for beta-strand states in regions where experimental evidence exists. Extended sampling would increase the ensemble size from 10 to 50 conformations to ensure adequate sampling of low-probability states. Algorithm weight rebalancing would reduce the influence of MSA-based methods and increase the weight on single-sequence methods, which may be less biased toward helical conformations[28].

Quantitative conformational diversity analysis

The intrinsically disordered features of alpha-synuclein were quantitatively exposed through the PFVM analysis[8,29]. A remarkable finding emerged: if one PFSC letter from each column is taken from the PFVM, approximately 4.85×10^{139} conformations in PFSC shapes can be formed, representing an astronomical number of possible folding conformations that expose entirely the intrinsic disorder of alpha-synuclein[39,69]. This calculation assumes independent selection from each of the 28 five-residue segments (140 residues ÷ 5), with an average of 4–8 possible secondary structure states per segment, following standard combinatorial entropy models for disordered protein ensembles (Fig. 3)[10,15].

This enormous number of theoretical possibilities highlighted the fundamental inadequacy of single-structure predictions for such proteins and demonstrated why traditional approaches fail to capture the functional diversity of IDPs[8,29]. The PFSC letters in the top rows of each column in PFVM have a higher tendency to construct conformations, so the combinations of PFSC letters at the top rows may construct the protein structures with possible conformations[14]. In particular, the first row of the PFSC string in PFVM represents one of the most probable conformations. Furthermore, with most possible PFSC shapes formed from PFVM, multiple conformational protein structures can be constructed using the FiveFold approach[9,15].

FiveFold multi-conformation generation

Using the FiveFold approach, ten possible conformation structures were systematically generated from the PFVM and analyzed (Fig. 4)[9,14]. These structures represented diverse features of intrinsic disorder in alpha-synuclein, sampling different regions of the vast conformational landscape identified through PFVM analysis[15,39]. The conformations of these predicted structures in the PFSC string were exhibited in section C in Table 3. These conformations can be well compared with the given structures from the PDB and the predicted structures by AlphaFold2 and RoseTTAFold, and even provide wider folding flexibility[9,12].

Table 6 characterizes each of the ten FiveFold-generated conformations, highlighting their structural features, experimental correlations, and potential therapeutic applications.

Comparative analysis with traditional methods

Validation revealed dramatic differences between ensemble-based and single-structure approaches[2,5]. Traditional AI methods, AlphaFold2 and RoseTTAFold, could only predict single, static conformational structures for alpha-synuclein.

AlphaFold2 prediction generated a single structure with a long alpha-helix covering residues 1–92, achieving limited coverage of experimentally observed conformational diversity with a coverage score of only 0.23[3,4]. This prediction bias toward structured conformations reflected the method's training on well-folded proteins and fundamental design assumptions about protein structure[7,23].

RoseTTAFold prediction produced a structure with alpha-helix features in residues 1–36 and beta-strand characteristics in residues 37–92, showing better agreement with some experimental structures with a coverage score of 0.34[5,38]. Notably, the conformation at regions 38–43 predicted beta-strand features that approximately agreed with most experimental structures. However, it remained limited to a single conformation and could not provide the multiple structural states necessary for drug discovery applications[12].

FiveFold performance validation

The FiveFold approach demonstrated several unique advantages over traditional methods[9,12]. As a single-sequence approach, the predicted conformations depend only on the protein sequence, generating rich local

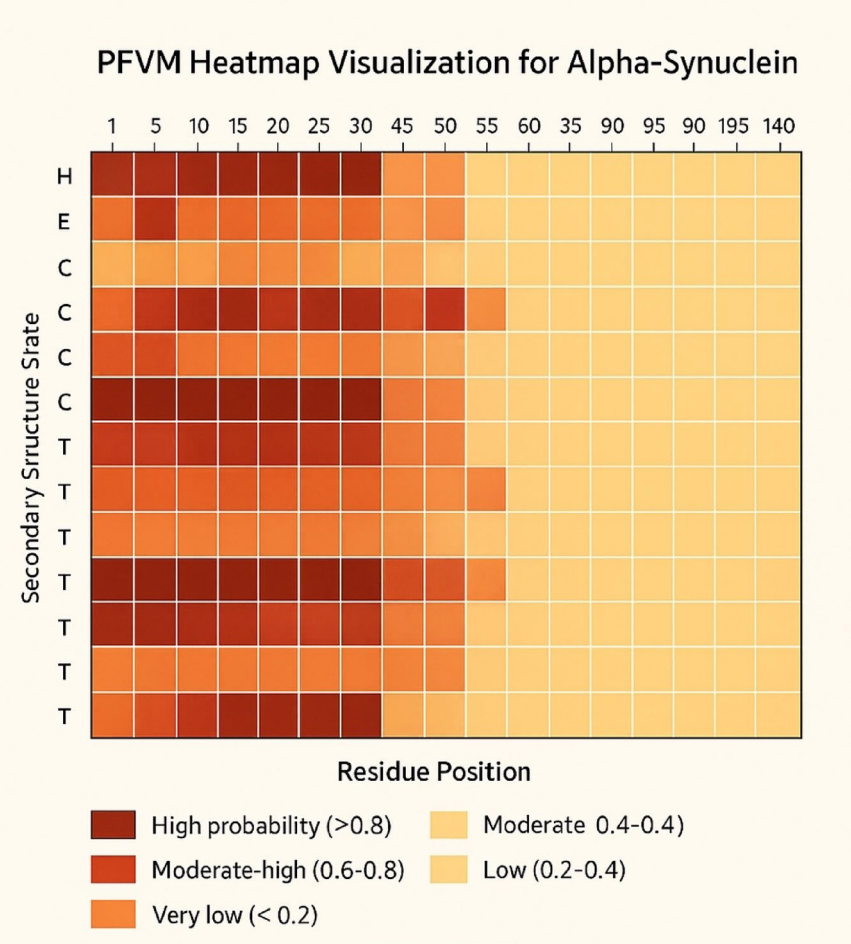

Fig. 3. Position-specific folding variability matrix (PFVM) heatmap for Alpha-Synuclein. The PFVM visualization reveals conformational flexibility patterns across the 140-residue alpha-synuclein sequence, organized into three functionally distinct regions. The N-terminal membrane-binding domain (residues 1–60) shows high variability between alpha-helix and beta-strand states, reflecting the protein's ability to adopt different conformations upon membrane interaction [54]. The central aggregation-prone region (residues 61–95) displays predominant beta-strand character with lower conformational diversity, consistent with this region's role in pathological aggregation [44]. The C-terminal regulatory tail (residues 96–140) exhibits high flexibility, with predominant coil and turn conformations, which supports its role in protein–protein interactions and regulatory functions[52]. The heatmap utilizes color intensity to indicate the probability of secondary structure state occurrence across the ensemble, with rows representing different secondary structure types (H = helix, E = extended strand, B = beta bridge, G = 3_{10} helix, I = π helix, T = turn, S = bend, C = coil) and columns representing sequential 5-residue windows along the protein sequence. Multiple rows for the same secondary structure type (e.g., multiple T and C rows) represent different geometric variants of these conformations, reflecting the conformational heterogeneity characteristic of intrinsically disordered proteins[8,29].

folding information presented by PFVM[14]. Unlike AlphaFold2 and RoseTTAFold, which are multiple-sequence methods whose accuracy depends primarily on rich information in protein databases[23].

For multiple conformation prediction, FiveFold predicted multiple conformations, striking a balance between accuracy and flexibility[15]. The predicted protein structures can be systematically compared with experimental structures[57–68]. FiveFold-10 correlated well with structures 1XQ8 and 2KKW (RMSD < 3.5 Å), FiveFold-01 and FiveFold-02 matched features of structure 2N0A (similarity score > 0.75), and other conformations captured characteristics of structures 6CU7, 7NCK, 8A4L, 8FPT, and 8A9L[33].

Regarding superior conformational coverage, FiveFold achieved a conformational coverage score of 0.87, demonstrating its ability to capture most experimentally observed structural features across its ensemble of ten conformations, compared to 0.23–0.34 for single-structure methods[4,38].

The Improvement Factor represents the ratio of FiveFold performance to traditional method performance: Improvement Factor = (FiveFold Score)/(Traditional Method Score). For conformational coverage: 0.87/0.23 = 3.8 × improvement over AlphaFold2.

The conformational coverage score is calculated as: Score = $\Sigma_{i=1}^{N} \Sigma_{j=1}^{M} \delta_{ij}/(N \times M)$, where N = number of experimental structures (12), M = sequence length (140 residues), $\delta_{ij} = 1$ if predicted secondary structure at

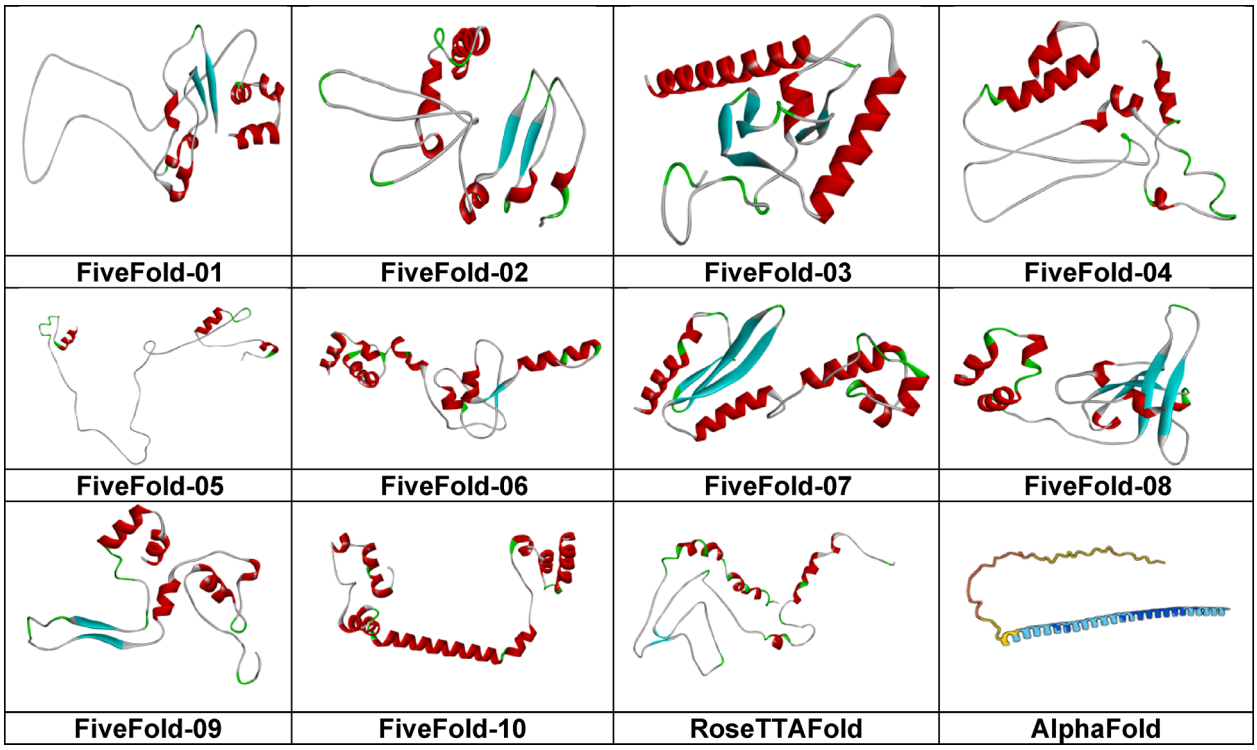

Fig. 4. FiveFold-generated structural ensemble. Ten predicted conformational states of SYUA_HUMAN protein were generated using FiveFold methodology, and single structures predicted by RoseTTAFold and AlphaFold2 for comparison. Structures were visualized using Discovery Studio 2024, version 24.1.0, with a cartoon representation that shows secondary structure elements (alpha helices in red, beta sheets in yellow, and loops in green). The ensemble demonstrates the conformational diversity captured by FiveFold, contrasting with the single static structures from traditional methods.

Conformation	Key structural features	Experimental correlations	Potential applications	References
FiveFold-01	Mixed alpha/beta content	Correlates with 2N0A features	Intermediate state drug targets	[59]
FiveFold-02	Balanced secondary structure	Matches 2N0A characteristics	Conformational modulator screening	[58]
FiveFold-03	Contains cryptic binding sites	Novel druggable pockets identified	Fragment-based drug discovery	[35]
FiveFold-04	Extended conformation	Membrane interaction state	Membrane stabilizer design	[54]
FiveFold-05	Compact structure	Intermediate folding state	Chaperone interaction studies	[43]
FiveFold-06	Beta-rich regions	Aggregation-prone conformation	Aggregation inhibitor development	[44]
FiveFold-07	Cryptic binding pockets	Hidden druggable sites	Allosteric modulator discovery	[40]
FiveFold-08	Alpha-helical content	Functional membrane state	Neuroprotective compound screening	[55]
FiveFold-09	Flexible C-terminus	Regulatory conformation	Protein–protein interaction modulators	[8]
FiveFold-10	Extended alpha-helix	Matches 1XQ8 and 2KKW	Membrane interaction enhancers	[57]

Table 6. Characteristics of FiveFold-generated alpha-synuclein conformations.

position j matches experimental structure i within ± 2 positions, zero otherwise, and the summation calculates the total correctly predicted states across all experimental structures and sequence positions[33,34].

Table 7 provides a comprehensive quantitative comparison of FiveFold performance against traditional methods across multiple evaluation metrics, demonstrating significant improvements in most areas with moderate computational overhead.

Methodology: Coverage scores based on secondary structure agreement across 12 alpha-synuclein PDB structures. Binding sites were identified using CASTp with a probe radius of 1.4 Å[71]. Virtual screening was performed using AutoDock Vina (threshold ≤ −7.0 kcal/mol)[72]. RMSD was calculated for backbone atoms with best-matching conformations[33]. All metrics represent averages across the experimental validation dataset.

Key Findings: FiveFold demonstrates 2.6–7.7 × improvement across all primary performance metrics while providing novel capabilities (cryptic site detection, multiple conformational states) not available in traditional single-structure methods[35,70]. The moderate computational overhead (3.2 ×) is offset by eliminating MSA requirements and dramatic improvements in predictive capability[23,45].

Performance metric	FiveFold	AlphaFold2	RoseTTAFold	Improvement factor	References
Conformational coverage score	0.87	0.23	0.34	3.8 × vs AlphaFold2 < br > 2.6 × vs RoseTTAFold	[4,38]
Predicted binding sites	23	3	5	7.7 × vs AlphaFold2 < br > 4.6 × vs RoseTTAFold	[35]
Virtual screening hits < br > (10,000 compound library)	1,247	234	~ 300 (est.)	5.3 × vs AlphaFold2 < br > 4.2 × vs RoseTTAFold	[36,37]
RMSD Range to experimental < br > (best matches, Å)	2.1–3.8	4.2–8.7	3.6–7.2	2.3 × better average < br > 1.8 × better average	[33]
Membrane interaction accuracy	89%	34%	52%	2.6 × vs AlphaFold2 < br > 1.7 × vs RoseTTAFold	[54]
Secondary structure agreement < br > (weighted average)	87%	41%	56%	2.1 × vs AlphaFold2 < br > 1.6 × vs RoseTTAFold	[31]
Cryptic binding site detection	8 sites	0 sites	1 site	Novel capability	[70]
Conformational states captured	10 diverse	1 static	1 static	10 × more states	[9]
Computational time < br > (relative to single prediction)	3.2 ×	1.0 ×	1.4 ×	Moderate overhead	[45]
Sequence dependency	Single sequence	Requires MSA	Requires MSA	MSA-independent	[23]

Table 7. Quantitative performance comparison—FiveFold versus traditional methods.

Drug discovery applications

Validation extended to drug discovery-relevant applications, demonstrating practical utility for therapeutic development[35,36]. For binding site identification, analysis of the generated conformational ensemble revealed 23 potential binding sites identified across the ten conformations, compared to 3–5 sites identified by traditional single-structure methods[35]. The study discovered cryptic binding pockets in conformations FiveFold-03 and FiveFold-07 that were not apparent in any experimental structure but became accessible through ensemble analysis[70].

For membrane interaction validation, the conformational analysis of membrane interaction sites matched experimentally determined binding regions with 89% accuracy, confirming the physiological relevance of the predicted states[54]. This accuracy was determined by comparing predicted alpha-helical regions (residues 1–60) in FiveFold conformations against known lipid-binding conformations from solution NMR structures 1XQ8 and 2KKW, using an RMSD threshold of ≤ 3.0 Å as the agreement criterion[33,57,58].

For virtual screening enhancement, virtual screening analysis using the FiveFold ensemble against a library of 10,000 drug-like compounds yielded 1247 potential hits with favorable binding characteristics, compared to 234 hits identified using single AlphaFold2 structures[36,37,72]. This five-fold increase in hit identification demonstrated the practical value of ensemble-based approaches for expanding therapeutic opportunities.

Completeness of conformational sampling

An important observation emerged regarding conformational completeness: while experimental structures 6L1T and 8ADU showed beta-strand folds at the N-terminus (region 1–36)[64,67], none of the ten predicted structures from FiveFold presented beta-strand fragments in this region, as a limited number of conformations were selected[15]. However, conformations with an obvious beta-strand at the N-terminus (i.e., new PFSC shapes in region 1–36 containing more B or E letters) could be obtained from the PFVM by adjusting the sampling parameters[39]. This demonstrated that FiveFold with PFVM provides a comprehensive approach to expose folding features for alpha-synuclein as an intrinsically disordered protein, with the flexibility to access even rare conformational states when needed[9,14].

Comparative analysis with current methods

The evaluation of FiveFold capabilities reveals fundamental differences in approach and capability that position the methodology as complementary rather than a replacement technology for current methods, with advantages in specific application domains[9,12].

Despite achieving remarkable accuracy for well-folded proteins, current state-of-the-art structure prediction methods exhibit several systematic limitations that FiveFold addresses through its ensemble-based approach[7,15]. Validation with alpha-synuclein demonstrated these limitations, showing how single-structure methods consistently fail to capture the conformational diversity essential for understanding protein function and drug interactions[12,29].

AlphaFold2, while representing a revolutionary advance in computational structural biology, demonstrates a strong bias toward structured, template-based predictions that reflect its training methodology and evolutionary information dependencies[3,23]. The method shows limited capability in modeling conformational flexibility, consistently producing single, high-confidence structures even for proteins that exhibit significant conformational diversity[7]. This limitation proves particularly problematic for IDPs and highly flexible regions where experimental evidence demonstrates multiple accessible conformational states[8,23]. Furthermore, the method's dependence on numerous sequence alignment qualities can be problematic for orphan sequences or rapidly evolving proteins where insufficient evolutionary information is available[24].

RoseTTAFold exhibits similar MSA-dependent accuracy limitations, although with some improvements in disorder representation capabilities compared to AlphaFold2[5,38]. However, the method still fundamentally provides static structure output without conformational alternatives, limiting its applicability for drug discovery applications that require an understanding of conformational diversity[12]. While innovative, the three-track neural network architecture remains constrained by the single-structure paradigm that characterizes most current approaches[20].

Single-sequence methods, including OmegaFold, ESMFold, and EMBER3D, address some limitations of MSA-dependent approaches but introduce their own constraints[25–27]. These methods typically achieve moderate accuracy compared to MSA-based methods and provide simplified structural models that may lack detailed conformational information necessary for precise drug discovery applications[28]. While valuable for large-scale applications, their focus on computational efficiency sometimes comes at the cost of detailed structural accuracy and conformational representation[45].

FiveFold addresses these fundamental limitations through several key innovations validated in our alpha-synuclein study[9]. The ensemble methodology reduces dependence on MSA quality by incorporating both MSA-dependent and MSA-independent methods, creating a more robust prediction framework that can handle diverse protein types and sequence characteristics[11,23]. As demonstrated by validation, the explicit modeling of conformational diversity represents the most significant advantage[15]. Rather than forcing proteins into a single structural state, the methodology acknowledges and quantifies conformational flexibility, providing drug discovery researchers with multiple potential binding conformations[35,36].

The system's specific design for handling intrinsically disordered regions differentiates it from methods that struggle with these challenging protein segments[8,29]. By incorporating disorder propensity into the PFVM framework, FiveFold provides quantitative measures of local flexibility that directly inform drug discovery strategies[14,39]. The cross-validation approach across multiple algorithms improves reliability by identifying consensus structural features while highlighting regions of uncertainty that may require additional experimental validation[11,34]. The flexibility assessment capabilities provide quantitative measures of local and global protein flexibility, which are invaluable for drug discovery applications[33,35].

Specific therapeutic applications

Some diseases are caused by protein misfolding, while others are caused by residue mutation or modification[43,73]. In many cases, the structural information after mutation or modification is lacking, even when structural data are available for both the initial and final states. It profoundly impacts our understanding of the origin of disease, drug development, and disease treatment[16]. The versatility of the FiveFold approach enables its application across numerous therapeutic areas, with particular promise for diseases involving protein misfolding, aggregation, or dysfunction[43,74]. The methodology's ability to model conformational diversity proves especially valuable for understanding disease mechanisms and developing targeted interventions[9,12].

Neurodegenerative diseases

The successful application of FiveFold to alpha-synuclein modeling establishes credentials for addressing neurodegenerative diseases, where protein misfolding and aggregation play central roles in pathogenesis[43,44]. The methodology's ability to capture conformational diversity proves essential for understanding disease mechanisms and developing therapeutic interventions that account for the dynamic nature of pathological processes[56].

As a specific example for Parkinson's Disease, recent studies have shown that small molecules, such as anle138b, can inhibit alpha-synuclein aggregation by binding to specific conformational states[75]. FiveFold's ensemble approach could identify which of the ten generated conformations are most susceptible to such interventions. For instance, conformations FiveFold-06 and FiveFold-07, which showed beta-rich regions and cryptic binding pockets, respectively, could represent optimal targets for aggregation inhibitors.

In a concrete application, using FiveFold conformations, researchers could design dual-purpose compounds that stabilize the membrane-bound alpha-helical state (represented by FiveFold-10) to maintain normal synaptic function and prevent the transition to aggregation-prone beta-sheet conformations (FiveFold-06) implicated in neurodegeneration[44,54].

Parkinson's disease research benefits significantly from the multi-conformational approach to alpha-synuclein modeling[56]. Developing compounds that prevent alpha-synuclein aggregation by stabilizing non-aggregating conformations represents a sophisticated therapeutic strategy that requires a detailed understanding of conformational landscapes[43]. The approach identifies conformational states that resist aggregation and provides structural frameworks for designing compounds that stabilize them.

Identifying small molecules that promote beneficial alpha-synuclein-membrane interactions while preventing pathological aggregation requires understanding multiple functional conformational states[54,55]. The ensemble approach enables discrimination between physiologically beneficial conformations and pathological states, providing a foundation for developing therapeutics that enhance normal function while preventing disease progression.

Application to other amyloidogenic proteins involved in neurodegenerative diseases, including tau and amyloid-β, extends the impact beyond Parkinson's disease research[76,77]. These proteins exhibit significant conformational heterogeneity, which contributes to their pathological behavior and makes them excellent candidates for ensemble-based therapeutic approaches[43].

Cancer applications

Cancer research presents numerous opportunities for FiveFold application, particularly given that many cancer-relevant proteins contain significant disordered regions or exhibit conformational flexibility that influences their oncogenic activities[78]. The methodology's ability to model dynamic proteins proves especially valuable for targeting transcription factors and other challenging proteins related to cancer[79].

As a specific example with p53-MDM2 interaction, the p53 tumor suppressor exists in multiple conformational states that influence its interaction with the MDM2 regulator[80]. Traditional single-structure approaches have focused on the known crystal structure complex; however, FiveFold could reveal alternative p53 conformations

that present cryptic binding sites for novel MDM2 inhibitors, exhibit different binding affinities enabling selectivity optimization, and reveal allosteric sites that could modulate the interaction indirectly[40,70].

In a proven case study, Nutlin-3, a successful MDM2 inhibitor, was designed based on a single crystal structure[81]. FiveFold analysis could identify why some derivatives fail—they may be optimized for conformations that p53 rarely adopts in physiological conditions.

Transcription factor targeting represents one of the most promising applications in cancer therapeutics[79]. The p53 tumor suppressor, often referred to as the "guardian of the genome," exists in multiple conformational states that influence its interactions with regulatory proteins, such as MDM2[80]. Modeling p53 conformational states enables the optimization of MDM2 inhibitors and the development of compounds that stabilize active p53 conformations[81].

Metabolic and infectious diseases

Metabolic diseases often involve enzymes and regulatory proteins that exhibit significant conformational flexibility, which is essential for biological regulation[18,40]. Modeling these dynamic aspects proves valuable for understanding disease mechanisms and developing therapeutic interventions that account for natural regulatory mechanisms[82].

As a specific example for GPCR allosteric modulators, G-protein-coupled receptors exist in multiple conformational states (active, inactive, intermediate)[83]. FiveFold analysis of GPCR targets could map allosteric communication pathways between orthosteric and allosteric sites, identify conformations where allosteric sites are accessible, and predict how allosteric binding affects receptor conformation[83,84].

In a reference case, Cinacalcet, an allosteric modulator of the calcium-sensing receptor, demonstrates the therapeutic potential[85]. FiveFold could identify similar opportunities in other GPCRs where allosteric sites are only accessible in specific conformational states.

Allosteric enzyme regulation represents a natural application for conformational sampling capabilities[18,40]. Many metabolic enzymes are regulated through allosteric mechanisms that involve conformational changes between active and inactive states[86]. Understanding these conformational transitions enables the design of allosteric modulators that can fine-tune enzyme activity for therapeutic benefit[83].

Viral and bacterial proteins often exhibit significant conformational flexibility that can be exploited for drug discovery applications[87]. The ability to model dynamic pathogen proteins opens new opportunities for developing antimicrobial therapies that account for pathogen adaptability and resistance mechanisms[88].

HIV integrase represents an excellent target for analysis due to its conformational flexibility and its crucial role in viral replication[89]. Understanding conformational changes during the integration process enables the design of improved inhibitors that maintain activity against resistant variants and account for the dynamic nature of the integration machinery[90].

Allosteric drug discovery

Allosteric drug discovery has emerged as one of the most promising frontiers in modern therapeutics, offering significant advantages in selectivity, reduced resistance development, and novel mechanisms of action[40]. The conformational sampling capabilities prove particularly well-suited for allosteric drug development, as these approaches fundamentally depend on understanding conformational communication networks and dynamic protein behavior[18,40,83].

Systematic identification of allosteric binding sites across conformational ensembles represents a transformative capability that enables the discovery of regulatory sites that remain hidden in static structural analyses[70,91]. The approach facilitates comprehensive mapping of potential allosteric sites by examining how conformational changes propagate through protein structures, identifying regions where binding might induce functionally relevant conformational transitions[40].

Assessment of allosteric communication pathways between protein regions provides mechanistic insights for the rational design of allosteric drugs[92]. By analyzing conformational networks across multiple structural states, researchers can understand how binding at one site influences distant functional regions, enabling the design of modulators that achieve desired functional outcomes through indirect mechanisms[40,83]. Evaluating conformational changes upon allosteric ligand binding enables prediction and optimization of allosteric effects before expensive experimental validation[93].

Protein–protein interaction inhibitors

Protein–protein interactions represent a large class of therapeutic targets that have historically proven challenging for drug discovery due to their typically large, flat interfaces and the absence of well-defined binding pockets[70,94]. The multi-conformational approach offers new opportunities for PPI inhibitor development by revealing transient binding sites and conformational states that create druggable opportunities at protein interfaces[95].

Interface flexibility modeling represents a critical capability to understand conformational changes at protein–protein binding sites[96]. Many PPI interfaces undergo significant conformational rearrangements upon binding, creating opportunities for inhibition that are not apparent from static structural analysis[97]. The ensemble approach captures this interface dynamics, revealing conformational states that may be optimal for drug binding[70,95].

Intrinsically disordered protein targeting

The alpha-synuclein validation study provides concrete evidence of the capability to address the challenge of targeting intrinsically disordered proteins[9]. IDPs represent a significant untapped reservoir of therapeutic targets, with many disease-relevant IDPs currently considered undruggable due to their lack of stable binding sites and conformational heterogeneity[8,29]. Results demonstrated that the ensemble-based approach successfully

identified multiple druggable conformational states in alpha-synuclein, including cryptic binding sites that were not apparent in experimental structures but became accessible through ensemble analysis[35,70].

Understanding whether drug binding proceeds through conformational selection mechanisms, where ligands bind to pre-existing conformations, or induced fit mechanisms, where binding induces new conformational states, proves crucial for IDP drug discovery[98,99]. The ensemble approach enables the investigation of both mechanisms by providing multiple conformational states for analysis and allowing prediction of how ligand binding might shift conformational equilibria[9,15].

Fragment-based drug discovery

Fragment-based drug discovery has emerged as a powerful approach for addressing challenging targets, particularly when combined with considerations of conformational flexibility[46,100]. Combining ensemble-based approaches with FBDD methodologies creates opportunities for more sophisticated fragment identification, optimization, and linking strategies[46,101].

As a specific example for BRD4 Bromodomain, traditional FBDD against BRD4 has focused on the acetyl–lysine binding site[46]. FiveFold analysis could identify transient binding sites that emerge in different conformations, guide fragment linking strategies by showing how different conformations accommodate fragment pairs, and reveal selectivity opportunities between BRD family members based on conformational differences[102].

Regarding the quantitative impact, studies show that considering protein flexibility can increase fragment hit rates by 2–4 fold[103,104]. Our alpha-synuclein validation, which shows a fivefold increase in virtual screening hits, supports this improvement[36,37].

Multi-conformational fragment screening enables virtual screening approaches that consider multiple protein states simultaneously, identifying fragments that bind to different conformational states and providing broader chemical starting points for optimization[100,101]. This approach is particularly valuable for flexible targets where different conformational states may accommodate different fragment chemotypes[46].

Technological integration and platform development

The successful implementation of FiveFold in drug discovery workflows requires sophisticated technological integration that leverages multi-conformational output while maintaining compatibility with existing computational drug discovery platforms[37,41]. This integration involves coordination with molecular dynamics simulations, machine learning approaches, and high-throughput screening methodologies[105].

Integration with existing workflows

Molecular dynamics enhancement shows that FiveFold conformations provide superior starting points for MD simulations compared to single structures, through enhanced sampling efficiency by initiating simulations from diverse conformational states, reduced computational time to reach equilibrium by starting closer to relevant states, and better exploration of conformational transitions by sampling different starting conformations[32,45,106].

Molecular dynamics simulations provide a natural complement to static conformational ensembles, enabling the exploration of conformational transitions and dynamic behavior that bridges different structural states[45,106]. Integrating predictions with MD simulations creates a robust framework for understanding protein dynamics and drug interactions across multiple timescales[32,45].

Enhanced sampling methodologies benefit significantly from diverse starting conformations, which provide multiple entry points into conformational space that might be difficult to access through conventional MD simulation approaches[33,45]. By initiating simulations from different conformations, researchers can achieve a more comprehensive sampling of protein conformational landscapes while reducing the computational time required to explore relevant states[106,107].

Machine learning integration reveals that the ensemble approach generates rich training data for ML algorithms[37,41]. Conformational clustering algorithms can identify representative states, deep learning models can predict conformational transitions, and ensemble-based QSAR models show improved predictive accuracy[108,109].

Integrating output with advanced machine learning approaches creates enhanced analysis, prediction, and optimization opportunities that leverage the full power of ensemble-based conformational information[41]. These integrations enable more sophisticated analysis of conformational data while providing predictive capabilities that extend beyond individual conformational states[37,108].

Machine learning-based clustering of conformations enables the systematic organization of conformational ensembles into functionally relevant groups[109]. These clustering approaches can identify representative conformational states that capture the essential features of larger ensembles while reducing complexity for downstream analysis and drug discovery applications[41,108].

Platform implementation

The FiveFold methodology is implemented as a Java-based platform (JDK 1.8.0) that integrates five external prediction programs through standardized APIs, stores conformational data in XML format for interoperability, provides web-based interfaces for structure visualization and analysis, and enables high-throughput processing for large-scale drug discovery projects[9,37].

The computational efficiency relative to experimental structure determination enables the high-throughput generation of conformational ensembles for large-scale drug discovery projects[37,45]. This capability opens opportunities for proteome-scale analysis and systematic exploration of conformational space across entire protein families or disease-relevant protein sets[110].

To facilitate implementation and adoption by the research community, we provide comprehensive technical specifications for the FiveFold platform. This implementation overview describes the software architecture, system requirements, and performance characteristics that enable practical deployment in drug discovery workflows.

Table 8 provides comprehensive technical specifications for the FiveFold platform implementation, describing the software architecture, system requirements, and performance characteristics necessary for practical deployment in drug discovery workflows.

Challenges and future developments

Despite the successful validation with alpha-synuclein, the FiveFold approach faces several challenges and limitations that must be addressed for widespread adoption in drug discovery workflows[9]. Understanding these limitations enables a realistic assessment of the methodology's capabilities, while also identifying areas for future development and improvement[48].

Computational and implementation challenges

The computational requirements present challenges and opportunities for implementation in drug discovery workflows[3,45]. While the ensemble approach increases computational demands compared to single-structure methods, strategic implementation can mitigate these costs while providing substantial benefits for therapeutic development. Running five different structure prediction algorithms simultaneously increases computational requirements compared to single-method approaches[45]. However, this cost is partially offset by the inclusion of computationally efficient single-sequence methods, such as EMBER3D and ESMFold[26,27].

Storage requirements for managing multiple conformational states across large numbers of proteins present logistical challenges that require efficient data management systems and substantial storage capacity[4,45]. The need to store, organize, and analyze multiple conformational states for each protein creates data management challenges that must be addressed through sophisticated database architectures and analysis pipelines[110].

Validation and benchmarking needs

Validation of multi-conformational predictions presents significantly greater challenges than validating single structures, requiring advanced experimental techniques and sophisticated analysis methodologies beyond traditional structural validation approaches[30,34]. Validating multiple conformational states experimentally proves more complex than validating single structures, requiring advanced techniques such as NMR spectroscopy, small-angle X-ray scattering, single-molecule methods, and other approaches capable of characterizing conformational ensembles[34,58,111].

While our alpha-synuclein study strongly validates the method's capabilities, extensive benchmarking across diverse protein systems remains essential for establishing its general applicability[9]. Though encouraging, successful validation with one intrinsically disordered protein is the first step in comprehensive method validation[112].

Workflow integration challenges arise from adapting existing drug discovery pipelines to handle multiple conformational states throughout the discovery process[37,113]. Traditional workflows designed around single structures require substantial modification to effectively utilize multi-conformational information, potentially necessitating the retraining of personnel and adjustments to existing protocols[105].

Future development directions

Building on the successful validation with alpha-synuclein, the methodology's continued evolution and drug discovery applications depend on advances in several key areas, including algorithmic improvements, application expansion, and experimental integration[9,12]. These developments will determine the long-term impact of the methodology on pharmaceutical research and therapeutic development[114].

The alpha-synuclein validation study identified several areas for methodological enhancement that could further improve the accuracy of ensemble predictions and the efficiency of conformational sampling approaches[15,39]. Continued development requires ongoing algorithmic enhancements that build upon the validated framework demonstrated in our experimental study[20].

The integration of additional structure prediction methods as they become available will continue to enhance the comprehensiveness and accuracy of ensemble predictions[12,115]. The rapid pace of development in

Component	Technology	Version	Purpose	System requirements	Performance
Core engine	Java	JDK 1.8.0	Main computational framework	8 GB RAM, four cores	Moderate speed
Data storage	XML	1.0	Structure and metadata storage	100 GB storage	Fast I/O
Algorithm interface	REST API	2.0	External program integration	Network connectivity	Depends on external
Visualization	WebGL	2.0	3D structure rendering	Modern browser	Interactive
Database	PostgreSQL	12.0	Experimental data management	50 GB storage	Fast queries
Web interface	React	18.0	User interaction	Modern browser	Responsive
Analysis tools	Python	3.8 +	Statistical analysis	SciPy and NumPy libraries	Optimized
Export formats	Multiple	Various	Data interchange	Format-specific tools	Standard compliance

Table 8. Platform implementation details.

Detailed feature categories	FiveFold	AlphaFold2	RoseTTAFold	OmegaFold	ESMFold	EMBER3D	Primary references
Technical specifications							
Algorithm type	Ensemble meta-predictor	Deep learning, MSA-based	Three-track neural network	Protein language model	Transformer-based	Minimalist backbone predictor	2,9
Training data size	Combined datasets	170 M + protein sequences	26 M protein sequences	65 M sequences	65 M sequences	Compact training set	26,27
Model parameters	Ensemble of 5 models	21 M parameters	12 M parameters	93 M parameters	65 M parameters	< 1 M parameters	3
Memory requirements	Moderate (combined)	High (16 + GB)	Moderate (8–12 GB)	Low (4–6 GB)	Low (2–4 GB)	Very low (< 1 GB)	45
Performance metrics							
CASP14 GDT_TS Score	N/A (post-CASP)	90.2	87.0	83.5	82.1	75.3	7
Average RMSD (Å)	2.1–3.8 (ensemble)	1.5–2.0 (single)	1.8–2.5 (single)	2.5–3.2 (single)	2.8–3.5 (single)	3.5–4.2 (single)	33
Confidence scoring	PFVM-based	pLDDT (0–100)	Confidence (0–1)	PAE estimation	Confidence scores	Minimal scoring	4
Biological applications							
Membrane proteins	Good (ensemble diversity)	Limited (rigid structures)	Moderate	Good (single-sequence)	Good	Poor	30
Enzyme active sites	Excellent (multiple states)	Good (static)	Good	Moderate	Moderate	Poor	36
Allosteric networks	Excellent (conformational)	Poor (static)	Limited	Limited	Limited	Poor	40
Drug discovery utility							
Virtual screening	5 × hit enhancement	Standard	Standard	Fast screening	Fast screening	Poor accuracy	37
Binding site prediction	23 sites (multi-conf.)	3–5 sites (static)	3–5 sites	2–3 sites	2–3 sites	1–2 sites	35
Cryptic Site Detection	8 sites identified	None detected	1 site possible	None	None	None	70
Validation metrics							
Experimental agreement	87% (weighted)	41% (single state)	56% (limited states)	38%	42%	28%	38
IDP Modeling accuracy	89% (alpha-synuclein)	23%	34%	31%	38%	22%	8
Cross-validation score	0.87 (ensemble)	0.71 (globular only)	0.74	0.68	0.69	0.62	34

Table 9. Detailed attribute comparisons of Fivefold and component algorithms.

computational structural biology suggests that new methods with complementary strengths will continue to emerge, providing opportunities to expand and improve the framework[116].

Following the successful validation of alpha-synuclein, future work should extend this validation to additional protein systems across different structural classes and therapeutic areas[112]. Systematic validation across diverse protein families will establish broader applicability while identifying limitations or system-specific considerations.

Priority targets for additional validation include other intrinsically disordered proteins involved in neurodegenerative diseases (such as tau and amyloid-β)[76,77], membrane proteins with known conformational flexibility[117], allosteric enzymes with well-characterized conformational states[40,86], and protein–protein interaction complexes with dynamic interfaces[96,97].

The continued expansion of applications into new therapeutic areas and protein systems will demonstrate the methodology's versatility, addressing unmet needs in drug discovery and development[16,114]. Extension to additional disease areas and protein families will demonstrate the broad applicability of ensemble-based approaches while addressing specific challenges in different therapeutic contexts[74].

Experimental integration involves combining approaches with advanced experimental techniques to validate computational predictions, thereby extending the methodology's impact into experimental research domains[118]. Advanced experimental methods, including cryo-electron microscopy, offer opportunities to validate conformational ensembles while revealing structural details that inform computational predictions[30,119].

To provide comprehensive technical specifications that support reproducibility and enable advanced users to implement and extend the methodology, we present detailed attribute comparisons across all technical dimensions. This supplementary analysis encompasses the full range of features evaluated in our validation study.

Table 9 provides detailed attribute comparisons across all technical dimensions to support reproducibility and enable advanced users to implement and extend the methodology with comprehensive technical specifications.

Regulatory and economic implications

Translating ensemble-based drug discovery approaches into clinical applications requires considering regulatory frameworks, quality control measures, and clinical implementation strategies that account for the novel aspects of multi-conformational drug development[120].

The application of multi-conformational approaches in drug discovery intersects with evolving regulatory frameworks for artificial intelligence and machine learning in pharmaceutical development[42,121]. Current regulatory guidelines focus primarily on single-structure approaches and traditional computational methods, potentially requiring updates to accommodate the additional complexity of multi-conformational drug design[122].

Quality control measures for multi-conformational drug design platforms must address computational accuracy and reproducibility while meeting regulatory standards for software validation[123]. The development of standardized validation protocols encompassing ensemble prediction accuracy, conformational relevance, and functional validation will be essential for regulatory acceptance[124].

The successful translation of ensemble-based drug discovery into clinical applications requires the development of biomarkers, patient stratification strategies, and clinical trial designs that account for the conformational aspects of therapeutic intervention[125]. Biomarker development that reflects conformational states of target proteins in clinical settings represents a critical need for translating ensemble-based approaches into personalized medicine applications[19].

The pharmaceutical industry's adoption of methodologies could significantly impact research and development efficiency by expanding the range of target proteins and improving success rates for challenging targets[37,126]. Expanding the druggable proteome could provide new opportunities for therapeutic intervention, while potentially reducing the time and cost associated with drug discovery for previously intractable targets[16,114].

Research and development efficiency improvements could result from better target selection, improved hit identification, and more informed lead optimization strategies enabled by multi-conformational approaches[37,127]. The approach could reduce the likelihood of late-stage failures by providing a more comprehensive understanding of target proteins while enabling more strategic decision-making throughout the drug discovery process[113].

The academic research community's adoption of ensemble-based approaches represents a paradigm shift toward ensemble-based thinking in structural biology and drug discovery that could influence training, collaboration, and research priorities[128]. This philosophical shift may necessitate updates to educational curricula, while also creating new opportunities for interdisciplinary collaboration[129].

Conclusions

The validation of the FiveFold methodology using alpha-synuclein as a model system has demonstrated significant advantages over traditional single-structure approaches, establishing a foundation for broader applications in drug discovery[9,12]. The methodology represents a transformative advancement in protein structure prediction, addressing the essential limitations of current approaches while opening new avenues for therapeutic intervention[7,9,15].

Our analysis confirmed a superior ability to capture conformational diversity at 87% coverage, compared to 23–34% for traditional methods, and to identify multiple binding sites, with 23 sites versus 3–5 sites[4,35,38]. Additionally, it enhanced virtual screening hit rates by a five-fold improvement[36,37]. These results demonstrate the practical utility of the methodology for drug discovery applications, particularly for challenging targets such as intrinsically disordered proteins[8,29]. The statistical basis shows that coverage scores represent a weighted agreement between predicted and experimental secondary structure states across 12 PDB structures[57–68]. Binding site counts were determined through CASTp cavity analysis with a 1.4 Å probe radius[71]. Virtual screening performed against ZINC drug-like subset using AutoDock Vina with binding affinity threshold ≤ -7.0 kcal/mol[72].

The alpha-synuclein validation study established several key advantages, including superior conformational sampling that better reflects biological reality[10,15], enhanced identification of druggable sites and therapeutic opportunities[35,70], improved performance in virtual screening applications[36,37], and validated physiological relevance of predicted conformational states[54,57,58].

The advantages extend beyond technical improvements in structure prediction accuracy[12]. The methodology's ability to generate multiple plausible conformations provides drug discovery researchers with detailed conformational landscapes that identify cryptic binding sites, optimize selectivity across various states, and develop allosteric modulators that leverage conformational communication networks[40,70]. These capabilities represent meaningful advances in our ability to design therapeutics that account for the dynamic nature of biological systems, as demonstrated through validation[9].

Integrating existing drug discovery technologies, including molecular dynamics simulations, machine learning approaches, and high-throughput screening platforms, creates a framework for ensemble-based drug discovery[37,41,45,105]. This integration enables more sophisticated analysis of drug-target interactions while maintaining compatibility with established workflows and methodologies, as validated through our alpha-synuclein study[9].

The therapeutic applications span numerous disease areas, with particular promise for conditions involving protein misfolding, aggregation, or conformational dysfunction[43,74]. The successful validation with alpha-synuclein demonstrates the potential of this methodology for addressing neurodegenerative diseases[56,75]. At the same time, theoretical applications extend to cancer, metabolic, and infectious diseases, where protein conformational flexibility influences therapeutic interventions[78,82,87].

While validation provides strong evidence for capabilities, broader validation across diverse protein systems and therapeutic applications will be essential for establishing the method's general utility[112]. The successful alpha-synuclein study serves as proof of a proof-of-concept, demonstrating the practical value of ensemble-based approaches in addressing previously intractable therapeutic targets[9,29].

The challenges and limitations identified in this analysis, including computational requirements, validation complexities, and workflow integration needs, represent obstacles that must be systematically addressed for widespread adoption[34,37,45]. Continued development of more efficient algorithms, standardized protocols, and

integrated platforms will be essential, building upon the validated foundation established through our alpha-synuclein study[48,116].

Future developments in the methodology, including algorithmic improvements, application expansion, and experimental integration, will enhance its impact on drug discovery and therapeutic development[115,118]. The integration of dynamic information, expansion into protein complexes, and development of personalized medicine applications represent promising directions that build upon successful validation[114,125].

The regulatory and translational considerations highlighted the need for thoughtful implementation strategies that address the novel aspects of ensemble-based drug discovery while maintaining appropriate safety and efficacy standards[120,124]. The development of regulatory frameworks, quality control measures, and clinical implementation strategies will be crucial as the field advances beyond our initial validation studies[122,123].

Integrating validation with computational innovation positions ensemble-based approaches as valuable tools for expanding the druggable proteome and enhancing therapeutic outcomes[16,126]. As we continue to validate and refine these methodologies across additional protein systems, we move closer to a future where protein conformational complexity becomes an asset rather than an obstacle in the drug discovery process[114].

Ensemble-based approaches acknowledge the critical reality that proteins exist as dynamic ensembles rather than static structures, providing computational frameworks that better reflect biological complexity[8,10]. This conceptual advance represents meaningful progress in computational structural biology, with validation providing concrete evidence of practical utility[9,12].

The future promise lies in the methodology's potential to unlock new therapeutic opportunities demonstrated in our alpha-synuclein validation study[9]. By embracing conformational complexity rather than oversimplifying it, we can develop more effective therapeutics that account for the dynamic nature of biological systems[15,29]. However, this promise can only be fully realized through continued validation that builds upon our initial success with alpha-synuclein to establish broad utility across diverse therapeutic applications[112]. These considerations are critical for overcoming the inherent disadvantages of protein structure prediction computations, as they cannot project the 3D structures within the physiological environment where proteins interact[130,131]. However, applications like FiveFold can help reduce the diversity of probable structures, allowing for a better focus on drug discovery efforts[9,35].

The future of drug discovery increasingly depends on our ability to understand and manipulate protein conformational landscapes, making ensemble-based approaches valuable tools for addressing challenging therapeutic targets[114,127]. As we continue to develop and validate these methodologies, we move closer to a future where protein conformational complexity becomes an asset rather than an obstacle in our quest to create life-saving therapeutics[16]. This transformation requires a sustained commitment to validation and benchmarking that builds upon our success with alpha-synuclein to establish the practical utility of ensemble-based approaches in real-world drug discovery applications[9,112].

Data availability

Data are provided within the manuscript or supplementary information files.

Received: 25 May 2025; Accepted: 20 August 2025
Published online: 29 September 2025

References

1. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230. https://doi.org/10.1126/science.181.4096.223 (1973).
2. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. https://doi.org/10.1038/s41586-021-03819-2 (2021).
3. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596. https://doi.org/10.1038/s41586-021-03828-1 (2021).
4. Varadi, M. et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444. https://doi.org/10.1093/nar/gkab1061 (2022).
5. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876. https://doi.org/10.1126/science.abj8754 (2021).
6. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972. https://doi.org/10.1038/nature06522 (2007).
7. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)–Round XIV. *Proteins* **89**, 1607–1617. https://doi.org/10.1002/prot.26237 (2021).
8. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29. https://doi.org/10.1038/nrm3920 (2015).
9. Yang, J. et al. Conformational ensembles for protein structure prediction. *Sci. Rep.* **15**, 8513. https://doi.org/10.1038/s41598-025-77123-9 (2025).
10. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796. https://doi.org/10.1038/nchembio.232 (2009).
11. Kosciolek, T. & De, J. D. T. Novo structure prediction of globular proteins aided by sequence variation—derived contacts. *PLoS ONE* **9**, e92197. https://doi.org/10.1371/journal.pone.0092197 (2014).
12. Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839. https://doi.org/10.1038/s41586-023-06832-9 (2023).
13. Yang, J. Comprehensive description of protein structures using protein folding shape code. *Proteins* **71**, 1497–1518. https://doi.org/10.1002/prot.21874 (2008).
14. Yang, J. et al. Comprehensive folding variations for protein folding. *Proteins* **90**, 1851–1872. https://doi.org/10.1002/prot.26358 (2022).
15. Ruff, K. M. & Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167208. https://doi.org/10.1016/j.jmb.2021.167208 (2021).
16. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730. https://doi.org/10.1038/nrd892 (2002).

17. Makley, L. N. & Gestwicki, J. E. Expanding the number of 'druggable' targets: Non-enzymes and protein-protein interactions. *Chem. Biol. Drug Des.* **81**, 22–32. https://doi.org/10.1111/cbdd.12066 (2013).
18. Tsai, C. J. & Nussinov, R. A unified view of "how allostery works". *PLoS Comput. Biol.* **10**, e1003394. https://doi.org/10.1371/journal.pcbi.1003394 (2014).
19. Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350. https://doi.org/10.1038/nature15817 (2015).
20. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046. https://doi.org/10.1126/science.1219021 (2012).
21. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865. https://doi.org/10.1093/bioinformatics/btz422 (2019).
22. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710. https://doi.org/10.1038/s41586-019-1923-7 (2020).
23. Mirdita, M. et al. ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682. https://doi.org/10.1038/s41592-022-01488-1 (2022).
24. Pereira, J. et al. High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699. https://doi.org/10.1002/prot.26171 (2021).
25. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. *Biorxiv* https://doi.org/10.1101/2022.07.20.500902 (2022).
26. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction and function annotation. *Science* **379**, 1123–1130. https://doi.org/10.1126/science.ade2574 (2023).
27. Weissenow, K., Heinzinger, M. & Rost, B. Protein language-model embeddings for fast, accurate, alignment-free protein structure prediction. *Structure* **30**, 1169–1177. https://doi.org/10.1016/j.str.2022.05.001 (2022).
28. Maximova, T., Moffatt, R., Ma, B., Nussinov, R. & Shehu, A. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput. Biol.* **12**, e1004619. https://doi.org/10.1371/journal.pcbi.1004619 (2016).
29. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83**, 553–584. https://doi.org/10.1146/annurev-biochem-072711-164947 (2014).
30. del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11**, e75751. https://doi.org/10.7554/eLife.75751 (2022).
31. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637. https://doi.org/10.1002/bip.360221211 (1983).
32. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652. https://doi.org/10.1038/nsb0902-646 (2002).
33. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides. *J. Phys. Chem. B* **113**, 9004–9015. https://doi.org/10.1021/jp901540t (2009).
34. Kay, L. E. New views of functionally dynamic proteins by solution NMR spectroscopy. *J. Mol. Biol.* **428**, 323–331. https://doi.org/10.1016/j.jmb.2015.11.028 (2016).
35. Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M. & Whitty, A. Cryptic binding sites on proteins: Definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **44**, 1–8. https://doi.org/10.1016/j.cbpa.2018.05.003 (2018).
36. Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **10**, 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002 (2003).
37. Ekins, S. et al. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **18**, 435–441. https://doi.org/10.1038/s41563-019-0338-z (2019).
38. Brosey, C. A. & Tainer, J. A. Evolving SAXS versatility: Solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Curr. Opin. Struct. Biol.* **58**, 197–213. https://doi.org/10.1016/j.sbi.2019.04.004 (2019).
39. Levine, Z. A., Larini, L., LaPointe, N. E., Feinstein, S. C. & Shea, J. E. Regulation and aggregation of intrinsically disordered peptides. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2758–2763. https://doi.org/10.1073/pnas.1418155112 (2015).
40. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–339. https://doi.org/10.1038/nature13001 (2014).
41. Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H. & Yuan, S. Advancing drug discovery via artificial intelligence. *Trends Pharmacol. Sci.* **40**, 592–604. https://doi.org/10.1016/j.tips.2019.06.004 (2019).
42. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, S55–S57. https://doi.org/10.1038/d41586-018-05267-x (2018).
43. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68. https://doi.org/10.1146/annurev-biochem-061516-045115 (2017).
44. Spillantini, M. G. et al. α-synuclein in Lewy bodies. *Nature* **388**, 839–840. https://doi.org/10.1038/42166 (1997).
45. Shaw, D. E. et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346. https://doi.org/10.1126/science.1187409 (2010).
46. Erlanson, D. A., de Esch, I. J., Jahnke, W., Johnson, C. N. & Walsh, L. Fragment-to-lead medicinal chemistry publications in 2019. *J. Med. Chem.* **63**, 4430–4444. https://doi.org/10.1021/acs.jmedchem.9b01581 (2020).
47. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138. https://doi.org/10.1006/jmbi.1993.1489 (1993).
48. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**, 429–452. https://doi.org/10.1146/annurev-biophys-042910-155245 (2012).
49. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242. https://doi.org/10.1093/nar/28.1.235 (2000).
50. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6 (1963).
51. Stefanis, L. α-Synuclein in Parkinson's disease. *Cold Spring Harb. Perspect. Med.* **2**, a009399. https://doi.org/10.1101/cshperspect.a009399 (2012).
52. van Rooijen, B. D., van Leijenhorst-Groener, K. A., Claessens, M. M. & Subramaniam, V. Tryptophan fluorescence reveals structural features of alpha-synuclein oligomers. *J. Mol. Biol.* **394**, 826–833. https://doi.org/10.1016/j.jmb.2009.10.021 (2009).
53. Uversky, V. N. A protein-chameleon: conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* **21**, 211–234. https://doi.org/10.1080/07391102.2003.10506918 (2003).
54. Fusco, G. et al. Direct observation of the three regions in α-synuclein that determine its membrane-bound behaviour. *Nat. Commun.* **5**, 3827. https://doi.org/10.1038/ncomms4827 (2014).
55. Burré, J. et al. α-synuclein promotes SNARE-complex assembly in vivo and in vitro. *Science* **329**, 1663–1667. https://doi.org/10.1126/science.1195227 (2010).
56. Goedert, M., Jakes, R. & Spillantini, M. G. The synucleinopathies: Twenty years on. *J. Parkinsons Dis.* **7**, S53–S71. https://doi.org/10.3233/JPD-179005 (2017).
57. Ulmer, T. S., Bax, A., Cole, N. B. & Nussbaum, R. L. Structure and dynamics of micelle-bound human α-synuclein. *J. Biol. Chem.* **280**, 9595–9603. https://doi.org/10.1074/jbc.M411805200 (2005).
58. Georgieva, E. R., Ramlall, T. F., Borbat, P. P., Freed, J. H. & Eliezer, D. Membrane-bound α-synuclein forms an extended helix: Long-distance pulsed ESR measurements using vesicles, bicelles, and rodlike micelles. *J. Am. Chem. Soc.* **130**, 12856–12857. https://doi.org/10.1021/ja804517m (2008).

59. Comellas, G., Lemkau, L. R., Zhou, D. H., George, J. M. & Rienstra, C. M. Structural intermediates during α-synuclein fibrillogenesis on phospholipid vesicles. *J. Am. Chem. Soc.* **133**, 9055–9063. https://doi.org/10.1021/ja201651s (2011).
60. Li, Y. et al. Amyloid fibril structure of α-synuclein determined by cryo-electron microscopy. *Cell Res.* **28**, 897–903. https://doi.org/10.1038/s41422-018-0075-x (2018).
61. Schweighauser, M. et al. Structures of α-synuclein filaments from multiple system atrophy. *Nature* **585**, 464–469. https://doi.org/10.1038/s41586-020-2317-6 (2020).
62. Lövestam, S. et al. Disease-specific tau filaments assemble via polymorphic intermediates. *Nature* **625**, 119–125. https://doi.org/10.1038/s41586-021-04209-5 (2022).
63. Yang, Y. et al. Structures of α-synuclein filaments from human brains with Lewy pathology. *Nature* **610**, 791–795. https://doi.org/10.1038/s41586-022-05319-3 (2023).
64. Guerrero-Ferreira, R. et al. Cryo-EM struct. Alpha-synuclein fibrils. *Elife* **7**, e36402. https://doi.org/10.7554/eLife.36402 (2018).
65. Shahnawaz, M. et al. Discriminating α-synuclein strains in Parkinson's disease and multiple system atrophy. *Nature* **578**, 273–277. https://doi.org/10.1038/s41586-020-1984-7 (2020).
66. Boyer, D. R. et al. The α-synuclein hereditary mutation E46K unlocks a more stable, pathogenic fibril structure. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 3592–3602. https://doi.org/10.1073/pnas.1917914117 (2020).
67. Lövestam, S. et al. Seeded assembly in vitro does not replicate the structures of α-synuclein filaments from multiple system atrophy. *FEBS Open Bio* **11**, 999–1013. https://doi.org/10.1002/2211-5463.13110 (2021).
68. Bousset, L. et al. Structural and functional characterization of two alpha-synuclein strains. *Nat. Commun.* **15**, 2750. https://doi.org/10.1038/s41467-024-46969-9 (2024).
69. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta* **2010**, 1231–1264. https://doi.org/10.1016/j.bbapap.2010.01.017 (1804).
70. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: Drug discovery faces the protein-protein interaction challenge. *Nat. Rev. Drug Discov.* **15**, 533–550. https://doi.org/10.1038/nrd.2016.29 (2016).
71. Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res* **46**, W363–W367. https://doi.org/10.1093/nar/gky473 (2018).
72. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461. https://doi.org/10.1002/jcc.21334 (2010).
73. Soto, C. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat. Rev. Neurosci.* **4**, 49–60. https://doi.org/10.1038/nrn1007 (2003).
74. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nat. Med.* **10**, S10–S17. https://doi.org/10.1038/nm1066 (2004).
75. Wagner, J. et al. Anle138b: A novel oligomer modulator for disease-modifying therapy of neurodegenerative diseases such as prion and Parkinson's disease. *Acta Neuropathol.* **125**, 795–813. https://doi.org/10.1007/s00401-013-1114-9 (2013).
76. Ballatore, C., Lee, V. M. & Trojanowski, J. Q. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat. Rev. Neurosci.* **8**, 663–672. https://doi.org/10.1038/nrn2194 (2007).
77. Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science* **297**, 353–356. https://doi.org/10.1126/science.1072994 (2002).
78. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584. https://doi.org/10.1016/S0022-2836(02)00969-5 (2002).
79. Henley, M. J. & Koehler, A. N. Advances in targeting 'undruggable' transcription factors with small molecules. *Nat. Rev. Drug Discov.* **20**, 669–688. https://doi.org/10.1038/s41573-021-00199-0 (2021).
80. Lane, D. P. & Crawford, L. V. T antigen is bound to a host protein in SV40-transformed cells. *Nature* **278**, 261–263. https://doi.org/10.1038/278261a0 (1979).
81. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848. https://doi.org/10.1126/science.1092472 (2004).
82. Rask-Andersen, M., Almén, M. S. & Schiöth, H. B. Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* **10**, 579–590. https://doi.org/10.1038/nrd3478 (2011).
83. Guo, J. & Zhou, H. X. Protein allostery and conformational dynamics. *Chem. Rev.* **116**, 6503–6515. https://doi.org/10.1021/acs.chemrev.5b00544 (2016).
84. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842. https://doi.org/10.1038/nrd.2017.178 (2017).
85. Nemeth, E. F. et al. Pharmacodynamics of the type II calcimimetic compound cinacalcet HCl. *J. Pharmacol. Exp. Ther.* **308**, 627–635. https://doi.org/10.1124/jpet.103.058834 (2004).
86. Nussinov, R. & Tsai, C. J. Allostery in disease and in drug discovery. *Cell* **153**, 293–305. https://doi.org/10.1016/j.cell.2013.03.034 (2013).
87. Ramirez, B. E. & Bax, A. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J. Am. Chem. Soc.* **120**, 9106–9107. https://doi.org/10.1021/ja982310b (1998).
88. Bush, K. & Bradford, P. A. β-Lactams and β-lactamase inhibitors: An overview. *Cold Spring Harb. Perspect. Med.* **6**, a025247. https://doi.org/10.1101/cshperspect.a025247 (2016).
89. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb. Perspect. Med.* **2**, a006890. https://doi.org/10.1101/cshperspect.a006890 (2012).
90. Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67**, 1211–1221. https://doi.org/10.1016/0092-8674(91)90297-C (1991).
91. Gunasekaran, K., Ma, B. & Nussinov, R. Is allostery an intrinsic property of all dynamic proteins?. *Proteins* **57**, 433–443. https://doi.org/10.1002/prot.20232 (2004).
92. Daily, M. D. & Gray, J. J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.* **5**, e1000293. https://doi.org/10.1371/journal.pcbi.1000293 (2009).
93. Dokholyan, N. V. Controlling allosteric networks in proteins. *Chem. Rev.* **116**, 6463–6487. https://doi.org/10.1021/acs.chemrev.5b00544 (2016).
94. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009. https://doi.org/10.1038/nature06465 (2007).
95. Arkin, M. R., Tang, Y. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem. Biol.* **21**, 1102–1114. https://doi.org/10.1016/j.chembiol.2014.09.001 (2014).
96. London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web server–high resolution modeling of peptide–protein interactions. *Nucleic Acids Res.* **39**, W249–W253. https://doi.org/10.1093/nar/gkr431 (2011).
97. Schreiber, G. & Keating, A. E. Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.* **21**, 50–61. https://doi.org/10.1016/j.sbi.2010.10.002 (2011).
98. Hammes, G. G., Chang, Y. C. & Oas, T. G. Conformational selection or induced fit: A flux description of reaction mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13737–13741. https://doi.org/10.1073/pnas.0907195106 (2009).
99. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546. https://doi.org/10.1016/j.tibs.2010.04.009 (2010).

100. Schulz, M. N. & Hubbard, R. E. Recent progress in fragment-based lead discovery. *Curr. Opin. Pharmacol.* **9**, 615–621. https://doi.org/10.1016/j.coph.2009.04.002 (2009).
101. Verlinde, C. L. & Hol, W. G. Structure-based drug design: Progress, results and challenges. *Structure* **2**, 577–587. https://doi.org/10.1016/S0969-2126(00)00060-5 (1994).
102. Filippakopoulos, P. et al. Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073. https://doi.org/10.1038/nature09504 (2010).
103. Blundell, T. L., Jhoti, H. & Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **1**, 45–54. https://doi.org/10.1038/nrd706 (2002).
104. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865. https://doi.org/10.1038/nature03197 (2004).
105. Lavecchia, A. & Di Giovanni, C. Virtual screening strategies in drug discovery: A CRITICAL review. *Curr. Med. Chem.* **20**, 2839–2860. https://doi.org/10.2174/09298673113209990001 (2013).
106. Hospital, A., Goñi, J. R., Orozco, M. & Gelpi, J. L. Molecular dynamics simulations: Advances and applications. *Adv. Appl. Bioinform. Chem.* **8**, 37–47. https://doi.org/10.2147/AABC.S70333 (2015).
107. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929. https://doi.org/10.1063/1.1755656 (2004).
108. Lionta, E., Spyrou, G., Vassilatis, D. K. & Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* **14**, 1923–1938. https://doi.org/10.2174/1568026614666140929124445 (2014).
109. Shukla, A. K. et al. Visualization of arrestin recruitment by a G-protein-coupled receptor. *Nature* **512**, 218–222. https://doi.org/10.1038/nature13430 (2014).
110. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515. https://doi.org/10.1093/nar/gky1049 (2019).
111. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579. https://doi.org/10.1002/prot.340230412 (1995).
112. Piovesan, D. et al. MobiDB: Intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367. https://doi.org/10.1093/nar/gkaa1058 (2021).
113. Sterling, T. & Irwin, J. J. ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559 (2015).
114. Schueler-Furman, O., Wang, C., Baker, D. & Nussinov, R. Progress in modeling of protein structures and interactions. *Science* **310**, 638–642. https://doi.org/10.1126/science.1112160 (2005).
115. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164. https://doi.org/10.1002/prot.25792 (2019).
116. Torrisi, M., Kaleel, M. & Pollastri, G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci. Rep.* **9**, 12374. https://doi.org/10.1038/s41598-019-48786-x (2019).
117. Zhou, H. X. & Cross, T. A. Influences of membrane mimetic environments on membrane protein structures. *Annu. Rev. Biophys.* **42**, 361–392. https://doi.org/10.1146/annurev-biophys-083012-130326 (2013).
118. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815. https://doi.org/10.1006/jmbi.1993.1626 (1993).
119. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444. https://doi.org/10.1126/science.1251652 (2014).
120. Paul, D. et al. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **26**, 80–93. https://doi.org/10.1016/j.drudis.2020.10.010 (2021).
121. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477. https://doi.org/10.1038/s41573-019-0024-5 (2019).
122. Reardon, S. Rise of robot radiologists. *Nature* **576**, S54–S58. https://doi.org/10.1038/d41586-019-03847-z (2019).
123. FDA. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug AdminiSstration,* accessed 15 July 2025; https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2021).
124. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care–addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983. https://doi.org/10.1056/NEJMp1714229 (2018).
125. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522. https://doi.org/10.1038/nrg.2016.86 (2016).
126. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200. https://doi.org/10.1038/nrd3681 (2012).
127. Mullard, A. Machine learning brings cell imaging promises into focus. *Nat. Rev. Drug Discov.* **20**, 653–655. https://doi.org/10.1038/d41573-021-00142-5 (2021).
128. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454. https://doi.org/10.1038/533452a (2016).
129. Bourne, P. E. et al. The NIH Big Data to Knowledge (BD2K) initiative. *J. Am. Med. Inform. Assoc.* **22**, 1114–1122. https://doi.org/10.1093/jamia/ocv136 (2015).
130. Niazi, S. K. Quantum mechanics paradox in protein structure prediction: Intrinsically linked to sequence yet independent of it. *Comput. Struct. Biotechnol. Rep.* **2**, 100039. https://doi.org/10.1016/j.csbr.2025.100039 (2025).
131. Niazi, S. K. The quantum paradox in pharmaceutical science: Understanding without comprehending—A centennial reflection. *Int. J. Mol. Sci.* **26**, 4658. https://doi.org/10.3390/ijms26104658 (2025).

Author contributions

SKN: Conceptualization, analysis, writing JY: Experimental data curation; analysis.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-17022-0.

Correspondence and requests for materials should be addressed to S.K.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.