

Head Office:

Università degli Studi di Padova

Department:

Biomedical Sciences

Ph.D. Course in:

Biomedical Sciences

Series: 38

THESIS TITLE:

In-silico study of protein structural ensembles as potential novel drug targets

Thesis written with the financial contribution of the European Union - Next Generation EU - National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 - Investment 1.4 "Strengthening of research facilities and creation of "national R&D champions" on some Key Enabling Technologies".CN00000013. Project: "National Centre for HPC, Big Data and Quantum Computing" - CUP UNIPD C93C22002800006.

Coordinator: Prof. Ornella Rossetto

Supervisor: Prof. Damiano Piovesan

Co-Supervisor: Dr. Alexander Miguel Monzon

Ph.D. Student:

Hamidreza Ghafouri

To all my people in Iran who are fighting for their freedom.

Table of Contents

Summary	6
List of Publications	7
List of Abbreviations	8
1- Introduction	9
1.1 Proteins and Their Functional Diversity	9
1.2 Protein Folding and the Energy Landscape Framework	11
1.3 The Sequence-structure-function paradigm	14
1.4 Intrinsically Disordered Proteins and Regions (IDPs/IDRs)	17
1.5 Toward a Sequence-Ensemble-function Paradigm	19
1.6 Conformational Ensemble Properties and Sequence Determinants of IDPs	21
1.7 Databases Supporting IDP Research	24
2- Research objectives	29
2.1 – Infrastructure for Ensemble-Based Protein Structure Representation	29
2.2 – Standardised Workflows for the Determination of Conformational Ensembles	30
2.3 – Computational Tools for the Analysis of Conformational Ensembles	30
Chapter 1: Infrastructure for Ensemble-Based Protein Structure Representation	31
Chapter 2: Standardised Workflows for the Determination of Conformational Ensembles	49
Chapter 3: Computational Tools for the Analysis of Conformational Ensembles	76
3- Discussion	89
The Protein Ensemble Database: Development, Expansion, and Methodological Trends	89
Scalable Deposition and Integration of Predicted Conformational Ensembles:	90
Next steps and future perspective of PED:	91
Toward a Unified Framework for Determining Conformational Ensembles:	93
Summary of the major components of the ensemble determination framework:	94
The importance of benchmarking of conformational ensembles of IDPs as the next step:	97
The key important steps for future of modeling of conformational ensembles of IDPs: ..	98
Review of recent state-of-the-art methods for determining conformational ensembles: ...	99
Computational Methods for the Analysis of Conformational Ensembles:	99
4- Conclusion	101
Appendix A: DRMAAtic supplementary materials	104

Appendix B: Towards a unified framework for determination of conformational ensembles	108
Appendix C: Advances in the determination of conformational ensembles	125
Appendix D: IDPET supplementary materials	130
References	141

List of Figures

1	a) List of standard amino acids and b) the central dogma of the molecular biology	11
2	A schematic energy landscape for protein folding. A) The classical representation of the linear folding pathway. B) The modern view of multiple routes through a funneled landscape. This figure has been taken from[13]	13
3	A schematic energy landscape for protein folding. This figure has been taken from[17]	14
4	Sequence structure function paradigm and its challenges. The classical sequence–structure–function relationship is challenged by (a) allostery, (b) induced fit, (c) metamorphic proteins adopting multiple folds, and (d) intrinsically disordered proteins, highlighting the role of dynamics in protein function.	16
5	Key functional and pathological behaviours of intrinsically disordered proteins, including fuzzy binding, folding upon binding, biomolecular condensates, and disease-associated aggregation.	19
6	Overview of the shift from a sequence–structure–function view of proteins to a sequence–ensemble–function framework, illustrating how both folded and disordered conformational states give rise to diverse biological functions	21
7	a) IDRs exist as highly heterogeneous conformational ensembles rather than single structures, in contrast to folded domains, and their ensemble nature depends on molecular context. b–c), IDR ensembles are quantitatively described using ensemble properties (for example, radius of gyration, shape, and inter-residue distances), which are best represented as probability distributions. d–g), Sequence features and residue patterning encode IDR ensemble behavior, which in turn tunes functional outcomes such as interdomain communication and effective concentrations of folded domains. This figure has been taken from[31]	24
8	Deposition process and information content in DisProt. This figure has been taken from[74]	25
9	Information content in MobiDB. This figure has been taken from[75]	26
10	Overview of the paper-collection structure of this thesis, showing the three interconnected chapters and their objectives, which are addressed through published and submitted works on infrastructure development, standardised workflows for conformational ensemble determination, and computational tools for ensemble analysis within the sequence–ensemble–function framework.	31
11	Experimental SAXS profiles are compared with ensemble-averaged scattering curves calculated from five ensembles generated using different approaches: C36m molecular dynamics (PED00543), idpGAN (PED00477), IDPConformerGenerator with chemical-shift bias (PED00426), C22 molecular dynamics (PED00544), and a99SB-disp molecular dynamics (PED00545). Reduced χ^2 values and residuals show substantial differences in agreement with the experimental data, with the IDPConformerGenerator ensemble exhibiting the best fit and the C22 force field ensemble the poorest agreement.	93

Summary

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) play essential roles in cellular regulation, signaling, and molecular recognition. Unlike globular proteins, their biological function arises from dynamic conformational ensembles rather than from a single well-defined three-dimensional structure. Accurately representing and interpreting these ensembles therefore represents a central challenge in modern structural biology. This thesis addresses this challenge by contributing to the infrastructure, methodology, and computational analysis required for ensemble-based protein structure representation, with a particular focus on intrinsically disordered systems.

The first part of the thesis focuses on infrastructure for conformational ensembles of IDPs. A major contribution is the development and extension of the Protein Ensemble Database (PED), a database for conformational ensembles of flexible and disordered proteins. Through multiple updates, PED has evolved into a resource that supports experimentally derived ensembles, curated metadata, and automated deposition and validation workflows. This work also includes the integration of high-performance computing resources via DRMAAtic to automate validation and large-scale deposition pipelines, as well as the implementation of batch deposition strategies that enabled the inclusion of large predicted ensemble datasets.

The second part of the thesis addresses the lack of standardization in ensemble determination by proposing a unified conceptual framework. This framework decomposes ensemble determination into three core components: experimental methods, computational ensemble generation, and validation and comparison. Building on this framework, the thesis also contributes a comprehensive review of state-of-the-art experimental, computational, and integrative approaches for determining conformational ensembles of IDPs. Particular emphasis is placed on uncertainty quantification, avoidance of overfitting, and the need for systematic benchmarking. These considerations culminate in the proposal of the IDP Ensemble Benchmarking Challenge (IDP-Bench), a community-driven initiative designed to enable fair, transparent, and reproducible assessment of ensemble generation methods.

The final part of the thesis introduces a computational tool for the analysis and comparison of conformational ensembles. By developing and applying ensemble-level descriptors and comparison metrics, this work demonstrates how ensemble properties can be quantified and compared. This tool provides a bridge between raw ensemble representations and biological interpretation, enabling systematic assessment of structural heterogeneity, robustness, and physical plausibility.

Overall, this thesis advances ensemble-based modeling of IDPs by integrating infrastructure development, methodological standardization, and computational analysis. By enabling more reliable characterization of conformational ensembles and their dynamic properties, this work also provides a foundation for investigating IDPs as potential drug targets, where understanding ensemble modulation, transient binding sites, and context-dependent interactions is essential for rational therapeutic design.

List of Publications

The work presented in this thesis is based on the following publications:

1. **Ghafari H**, Lazar T, Del Conte A, Tenorio Ku LG, Tompa P, Tosatto SCE, Monzon AM. *PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins*. Nucleic Acids Research. 2024;52(D1):D536–D544.
2. Del Conte A, **Ghafari H** Clementel D, Mičetić I, Piovesan D, Tosatto SCE, Monzon AM. *DRMAAtic: dramatically improve your cluster potential*. Bioinformatics Advances. 2025;5(1):vbaf112.
3. **Ghafari H**, Kadeřávek P, Melo AM, Aspromonte MC, Bernadó P, Cortes J, Dosztányi Z, Erdős G, Feig M, Janson G, Lindorff-Larsen K, et al. *Towards a Unified Framework for Determining Conformational Ensembles of Disordered Proteins*. arXiv:2504.03590, 2025.(Accepted in Nature Methods)
4. **Ghafari H**, Tosatto SCE, Monzon AM. *Advances in the determination of disordered protein ensembles*. Current Opinion in Structural Biology. 2026;96:103198.
5. **Ghafari H**, Janson G, Tosatto SCE, Monzon AM. *IDPEnsembleTools: An open-source library for analysis of conformational ensembles of disordered proteins*. Protein Science. 2026;35(1):e70427.
6. Liu ZH, Zhang O, De Castro S, Sun K, **Ghafari H**, Attafi OA, Fawzi NL, Tosatto SC, Monzon AM, Moses AM, Head-Gordon T. *AlphaFlex:Ensembles of the human proteome representing disordered regions..* bioRxiv. 2025 Nov 25:2025-11.

Publications not included in this thesis

1. Nugnes MV, Bouhraoua KE, Zoubiri M, Pancsa R, Fichó E, Tompa P, Piovesan D, Tosatto SC, Aspromonte MC. *DisProt in 2026: enhancing intrinsically disordered proteins accessibility, deposition, and annotation*.Nucleic Acids Research. 2025 Nov 17:gkaf1175.
2. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A, Tosatto SC, Piovesan D. *DisProt in 2024: improving function annotation of intrinsically disordered proteins*. Nucleic Acids Research. 2024 Jan 5;52(D1):D434-41.

List of Abbreviations

Abbreviation	Meaning
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
PTM	Post-Translational Modification
LLPS	Liquid-Liquid Phase Separation
NMR	Nuclear Magnetic Resonance
SAXS	Small-Angle X-ray Scattering
CD	Circular Dichroism
HDX-MS	Hydrogen-Deuterium exchange Mass Spectroscopy
smFRET	Single-Molecule Förster Resonance Energy Transfer
MD	Molecular Dynamics
HPC	High-Performance Computing
PED	Protein Ensemble Database
PDB	Protein Data Bank
BMRB	Biological Magnetic Resonance Bank
ELM	Eukaryotic Linear Motif
SLiM	Short Linear Motif
AI	Artificial Intelligence
ML	Machine Learning
α -MoRE	α -Helical Molecular Recognition Element
CREB	Cyclic AMP Response Element-Binding protein
KIX	Kinase-Inducible domain-interacting domain
CBP	CREB-Binding Protein
CCRs	Composition-to-Conformation Relationships
FCR	Fraction of Charged Residues
NCPR	Net-Charge Per Residue
IDPET	Intrinsically Disordered Protein Ensemble Tools
PCA	Principle Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
LLPS	Liquid-Liquid Phase Separation
DRMAA	Distributed Resource Management Application API

1– Introduction

1.1 Proteins and Their Functional Diversity

Proteins have been recognised and studied for more than two centuries. In the late 1700s, Antoine Fourcroy and others described several “albuminous materials”, such as albumin, fibrin, gelatin, and gluten, as distinct substances found in animals and plants. In the 1830s, Gerardus Johannes Mulder analysed these materials and suggested that they shared a common formula, and Jöns Jacob Berzelius introduced the term *protein*, derived from the Greek word *proteios*, meaning “primary”, to emphasise their fundamental role in living organisms. Later contributions by nutritional scientists, including Carl von Voit, Thomas Burr Osborne, Lafayette Mendel, and William Cumming Rose, connected specific amino acids and dietary protein intake to growth and health. Progress in protein purification and structural methods during the 20th century, such as X-ray crystallography and protein sequencing, transformed proteins from vaguely defined “albumins” into well-characterised macromolecules [1]. Within the modern molecular framework, gene expression follows the central dogma (Fig 1-b), in which information stored in our genes, i.e DNA is transcribed into RNA and subsequently translated into proteins, generating polypeptide chains that fold and execute cellular functions.

Proteins are now understood as large biological macromolecules composed of one or more long chains of amino acids (Fig 1-a). Their extensive functional diversity underpins nearly all processes in living cells. Each polypeptide is formed by linking amino acids through peptide bonds into a specific sequence encoded by a gene, and this sequence normally folds into a defined three-dimensional structure that determines its functional properties. Even small sequence variations, or chemical modifications introduced after synthesis, can influence folding, stability, and activity, generating functional diversity from the same fundamental components. Proteins may also include non-peptide elements, such as metal ions or organic cofactors, and they frequently assemble into complexes, enabling multiple subunits to cooperate within a shared functional task. Because proteins are continuously synthesised and degraded through protein turnover, cells can adjust their proteome dynamically, replacing damaged or misfolded molecules and adapting functional capacity to environmental changes or developmental requirements.

This molecular versatility explains why proteins participate in almost every cellular process, including catalysis, signalling, and structural support. Many proteins act as highly specific enzymes, accelerating thousands of biochemical reactions across metabolism, DNA replication, repair, and transcription, often through a few key residues located in a precisely formed active site [2]. Other proteins play structural roles: fibrous proteins such as collagen and elastin provide rigidity or elasticity to tissues, whereas actin and myosin, although globular as monomers, assemble into cytoskeletal and contractile filaments that determine cell shape and drive muscle contraction [3]. Moreover, proteins also specialise in communication and defence. Hormones such as insulin controls the glucose metabolism through the dual feedback loop; linking glucose concentration with insulin in plasma [4], membrane receptors convert extracellular stimuli into intracellular responses [5], and antibodies recognise

foreign antigens with high specificity to coordinate appropriate immune reaction [6]. Transport proteins further expand functional diversity: haemoglobin binds and releases oxygen as it circulates through the bloodstream [7], membrane channels and carriers regulate ion and small-molecule transport across lipid bilayers [8], and numerous ligand-binding proteins deliver nutrients or signalling molecules to their appropriate destinations. Altogether, this wide range of catalytic, structural, signalling, transport, and regulatory functions shows how variations in amino acid sequence, structure, post translational modifications (PTMs), and interactions allow proteins to operate as a versatile toolkit that sustains life across different biological scales[9] .

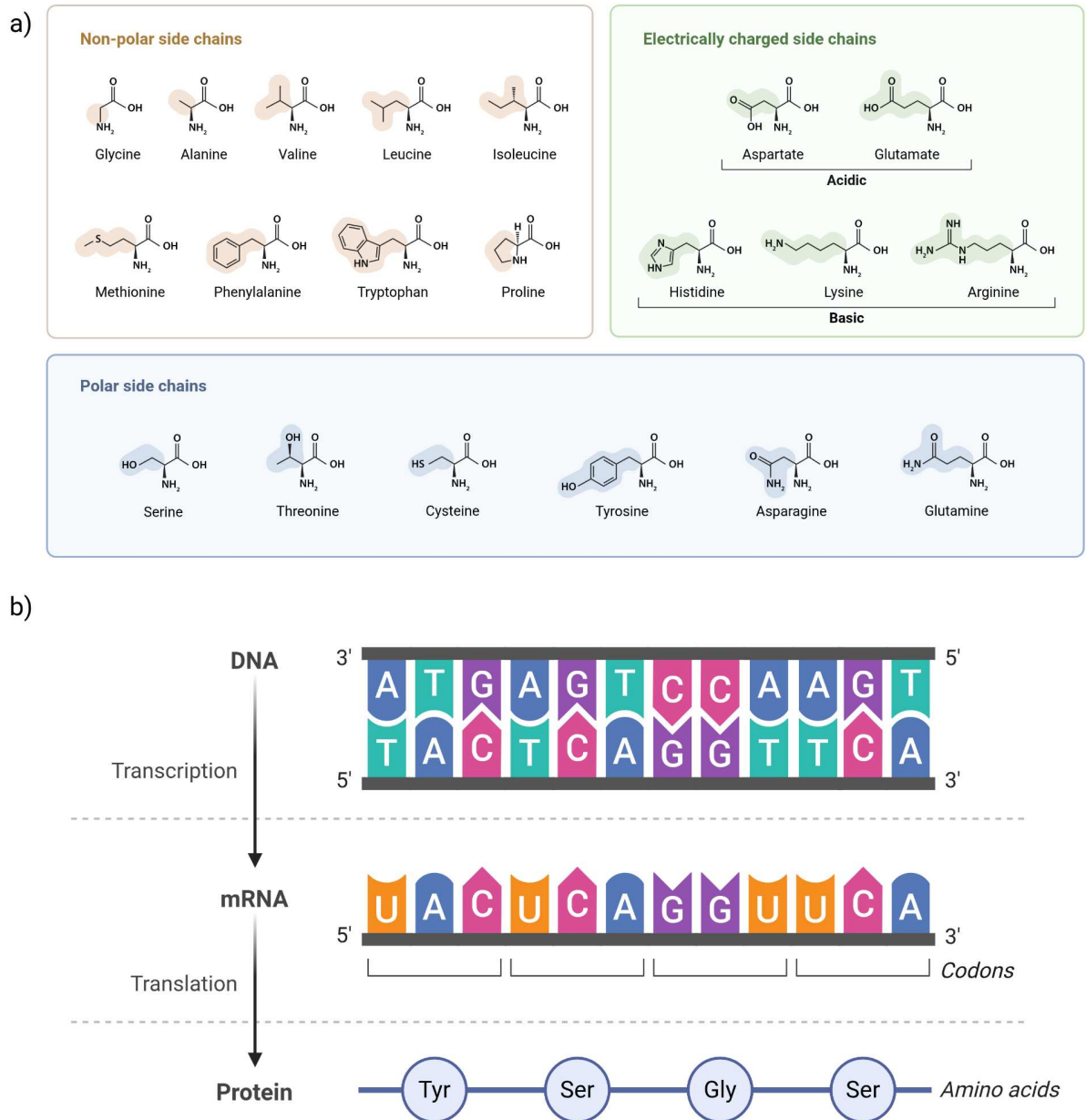


Figure 1: a) List of standard amino acids and b) the central dogma of the molecular biology

1.2 Protein Folding and the Energy Landscape Framework

As noted in the previous section, proteins are initially produced as linear polypeptide chains through the translation of RNA. In the conventional view, however, the central feature that confers functional diversity and specificity to proteins is the folding of these linear chains into a unique three-dimensional structure. According to this view, the native structure is dictated solely by the encoded amino-acid sequence of the polypeptide chain, a principle commonly referred to as Anfinsen's dogma[10, 11, 12]. This, however, raises a fundamental question: how do proteins reliably find their

native structures within the average timescales of milliseconds, despite the vastness of the available conformational space?

Numerous studies have attempted to address this question. In summary, early descriptions portrayed protein folding as a linear progression through discrete intermediates along a single dominant pathway (Fig-2A)[13]. In contrast, the modern energy-landscape framework views folding as a stochastic search across a high-dimensional conformational space, in which the folding dynamics are governed by the global architecture of the free-energy surface rather than by a single predefined route (Fig-2B)[3].

Statistical mechanical studies of random heteropolymers demonstrated that generic sequences generate rugged landscapes with numerous low-energy misfolded states, which produce slow, glass-like dynamics[14]. In contrast, natural proteins have evolved to be minimally frustrated: their energetic interactions favour native contacts and reduce the likelihood of kinetic trapping in misfolded conformations. This evolutionary optimisation creates a funneled energy landscape, where a broad ensemble of unfolded states progressively narrows toward a free-energy basin. Although the funnel contains rugged features and local minima, its overall downhill bias guides multiple microscopic trajectories toward the same native structure[15].

In this framework, folding does not require a single obligatory pathway. Instead, proteins move through ensembles of partially structured states, and many different routes descend toward the native conformation. This view explains why simple kinetic patterns, such as apparent two-state folding, can emerge despite considerable underlying complexity[16]. The transition state ensemble corresponds to a saddle region (Fig-3) of the landscape and consists of conformations that share essential native contacts. Φ -value analysis, in which the effects of point mutations on both thermodynamic stability and folding rates are measured, provides residue-level information about how native-like individual regions are in the transition state. Simulations and experimental studies show that transition states often involve delocalised nuclei, reflecting cooperative behaviour distributed across the chain rather than a single local folding trigger[17].

Energy-landscape theory also clarifies the origin of intermediates frequently observed in larger or topologically complex proteins. Depending on sequence and architecture, intermediates may support folding by stabilising partially organised regions, or they may behave as kinetic traps if they correspond to deep local minima. The balance between the funneled bias toward the native state and the residual ruggedness of the landscape determines whether folding proceeds rapidly and cooperatively or more slowly and heterogeneously[18]. Even when intermediates occur, the overall process is governed by the global topology and directional bias of the funneled landscape rather than by a unique linear mechanism. In this sense, the energy-landscape framework unifies thermodynamic, kinetic, and computational observations and describes protein folding as diffusive motion on a biased, rugged free-energy surface shaped by evolutionary pressures[15].

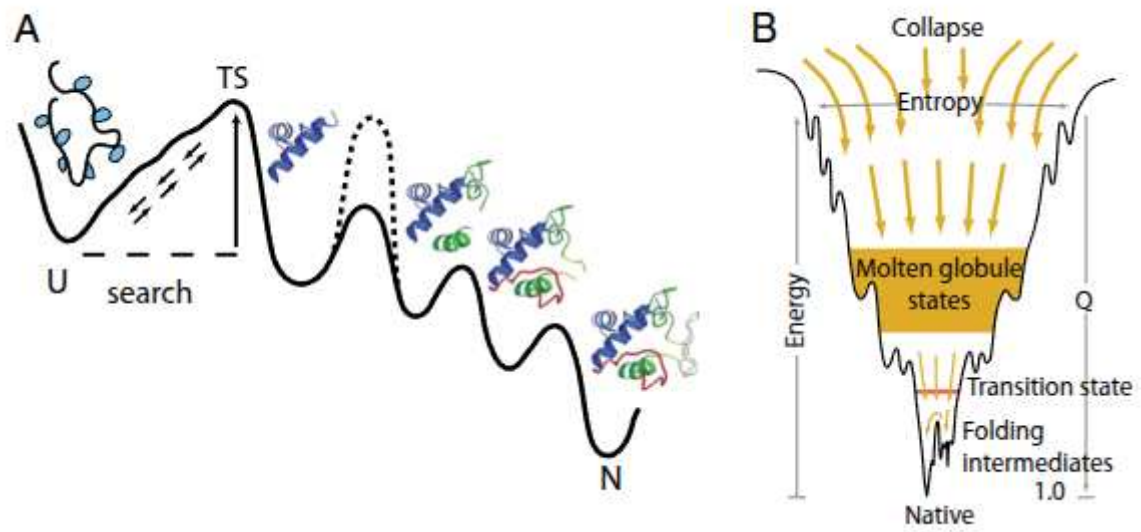


Figure 2: A schematic energy landscape for protein folding. A) The classical representation of the linear folding pathway. B) The modern view of multiple routes through a funneled landscape. This figure has been taken from[13]

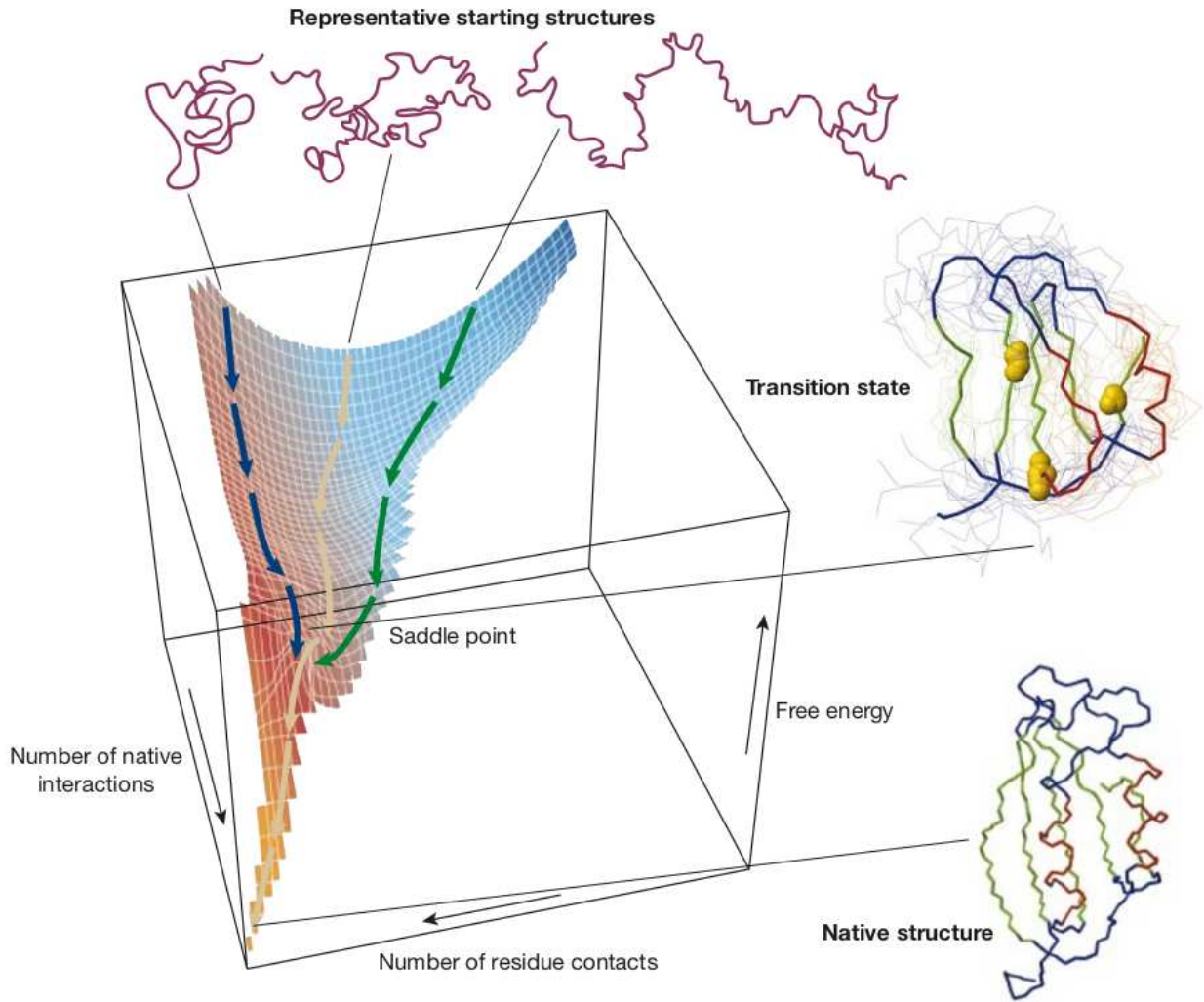


Figure 3: A schematic energy landscape for protein folding. This figure has been taken from[17]

1.3 The Sequence-structure-function paradigm

The idea that a protein’s amino-acid sequence determines its three-dimensional structure and, as a result, its biological function has been one of the central conceptual foundations of molecular biology. This sequence–structure–function paradigm emerged in the mid-20th century, supported by progress in protein chemistry and early structural studies[19]. Christian B. Anfinsen’s refolding experiments with ribonuclease showed that a fully denatured protein can regain its native conformation under suitable conditions, leading to the thermodynamic hypothesis that the native structure corresponds to the global minimum of the free-energy landscape[20]. This observation established that the information required for proper folding is encoded in the primary sequence and inspired the long-standing view that proteins behave as deterministic molecular machines defined by their linear amino-acid chains.

In this classical model, the amino-acid sequence specifies a network of intramolecular

interactions, including hydrophobic packing, hydrogen bonding, electrostatic interactions, and steric constraints, which collectively guide the chain toward a single well-organised structure. Structural motifs such as α -helices, β -sheets, and conserved domains arise from recurring patterns of local interactions, and the overall fold reflects how these elements pack in three dimensions. Comparative studies of homologous proteins demonstrated that evolutionary variation in sequence often preserves the overall fold, suggesting that structural architectures are robust solutions optimised for biological tasks[21]. This principle underpins much of modern structural biology, where the fold is considered as the scaffold that shapes catalytic pockets, binding sites, regulatory elements, and other dynamic features essential for function[22].

The second major component of the paradigm is the link between structure and function. Specific three-dimensional arrangements of residues enable the precise chemistry performed by enzymes[23], the selectivity of ligand-binding proteins[24], the mechanical properties of cytoskeletal systems[25], and the signalling capabilities of receptors[26]. Even subtle structural perturbations can lead to major functional consequences, as illustrated by pathogenic mutations that destabilise a fold or alter an active site[27]. This structure-centred perspective has guided decades of work in biochemistry, enzymology, and protein engineering, and it continues to offer a strong framework for rational drug design focused on well-defined conformational states.

Despite its acceptance in the scientific community, the sequence–structure–function paradigm provides a conceptual description that does not capture the full diversity of protein behaviours, including protein dynamics. As shown in Section 1.2, energy-landscape theory demonstrates that folding is not a single linear pathway but a stochastic process occurring on a rugged, globally funneled free-energy surface. Even proteins traditionally regarded as well folded populate ensembles of conformations around the native basin and may sample higher-energy substates that influence function, regulation, or misfolding. This intrinsic conformational heterogeneity challenges the assumption that a unique sequence necessarily corresponds to a single biologically relevant structure[28].

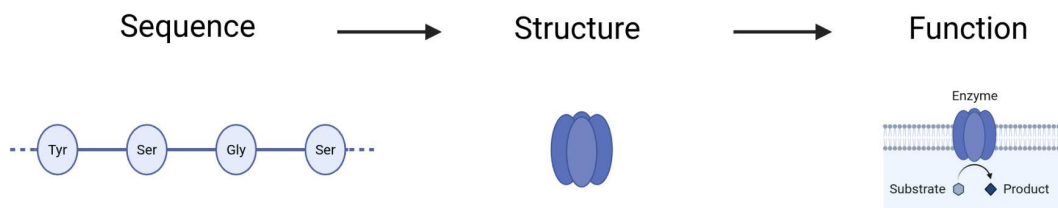
Several biological phenomena further highlight the limitations of the classical paradigm. For instance, allosteric regulation often arises from shifts in conformational equilibria rather than discrete structural transitions (Fig- 4a), showing that functional output may depend on ensemble redistribution rather than a single conformation[29]. Moreover, Induced fit and conformational selection mechanisms reveal that proteins interconvert among multiple states during ligand recognition (Fig- 4b). Metamorphic proteins present an even stronger challenge by adopting distinct, stable folds under different conditions, demonstrating that one sequence can support more than one functional structure (Fig- 4c)[30].

Even so, the most significant challenge comes from intrinsically disordered proteins and intrinsically disordered regions (collectively referred to hereafter as IDPs), which do not adopt stable three-dimensional structures under physiological conditions yet carry out essential cellular functions[31]. Their ability to remain conformationally heterogeneous enables interaction versatility, binding promiscuity, and rapid regulatory responses, features that are difficult to achieve with rigid structures. These proteins illustrate that functional mechanisms can be encoded in dynamic ensembles

rather than in fixed folds, exposing the limitations of the classical paradigm and motivating a more flexible conceptual framework (Fig- 4d)[32].

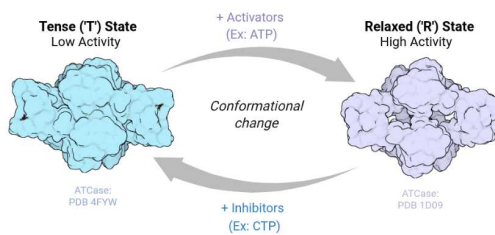
Taken together, these observations support an expanded view in which sequence still encodes the physicochemical properties that shape a protein's behaviour, but function may arise from an ensemble of accessible conformations instead of a single native state. This perspective is increasingly important for understanding disordered proteins, multistate systems, and conformationally dynamic complexes.

In this context, the sequence–structure–function paradigm remains valuable, but it must be complemented by a broader sequence–ensemble–function framework that captures the inherent dynamic natures of proteins.

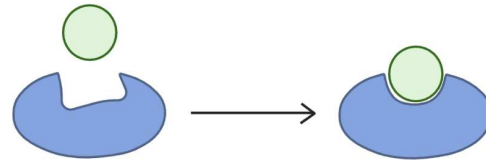


Challenges

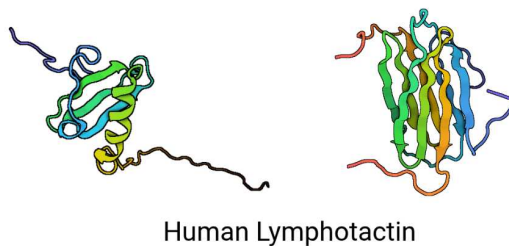
a) Allostery



b) Induced fit



c) Metamorphic Proteins



d) Intrinsically Disordered Proteins

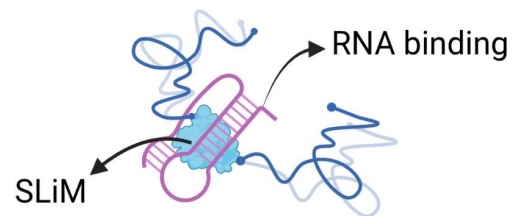


Figure 4: Sequence structure function paradigm and its challenges. The classical sequence–structure–function relationship is challenged by (a) allostery, (b) induced fit, (c) metamorphic proteins adopting multiple folds, and (d) intrinsically disordered proteins, highlighting the role of dynamics in protein function.

1.4 Intrinsically Disordered Proteins and Regions (IDPs/IDRs)

Protein function was long associated with a unique, sequence-encoded three-dimensional structure. However, beginning in the 1990s, accumulating evidence revealed a large class of proteins that do not conform to this principle. These proteins, now known as IDPs, do not maintain persistent tertiary structure under physiological conditions, yet they perform essential cellular functions[32, 33, 34]. Instead of adopting a single stable fold, IDPs populate broad conformational ensembles and sample multiple transient structures that interconvert on nanosecond to millisecond timescales[35, 36]. IDPs challenge the classical sequence–structure–function paradigm and shows that structural heterogeneity, rather than structural rigidity, can support biological activity.

IDPs have characteristic sequence features. In contrast to well-folded proteins, which typically contain a balanced mixture of hydrophobic, polar, and charged residues that enable compact packing of a hydrophobic core, disordered regions are enriched in polar and charged residues such as Lys, Glu, Ser, and Gln, and depleted in bulky hydrophobic residues[37, 38]. This composition reduces the tendency to form compact structures and favours expanded, dynamic conformations. Many IDPs also contain low-complexity sequences, repetitive motifs, or regions targeted by PTMs, which allow them to function as regulatory hubs[39]. Although IDPs lack stable tertiary structure, their ensembles are not random. They reflect sequence-encoded preferences, local secondary-structure propensities, long-range electrostatic interactions, and environmental factors such as salt concentration, pH, and binding partners[40, 41].

IDPs are widespread in biology. In eukaryotes, more than 30% of proteins contain long disordered segments, and many signalling, regulatory, and transcriptional proteins rely on disorder for their function. Disordered regions play central roles in processes including signal transduction, molecular recognition, cell-cycle control, gene expression, and stress response[31, 42]. Their functional versatility arises from their ability to bind multiple partners, adopt diverse binding modes, and undergo rapid conformational adjustments. Instead of a rigid lock-and-key mechanism, IDPs often operate through conformational selection or induced-fit mechanisms or a combination of both, where binding occurs either by recognising transient substates or through structural reorganisation triggered by partner interactions[43].

A distinctive feature of IDPs function is coupled folding and binding. Many IDPs remain disordered when isolated but fold upon interacting with a target, forming α -helices, β -strands, or more complex structures only in the bound state(Fig- 5b). This strategy supports high specificity with relatively low affinity, which is advantageous for dynamic regulatory systems that require rapid association and dissociation[44]. At the molecular level, folding upon binding can proceed through different mechanisms, including induced folding and conformational selection. Detailed insight into these processes has been obtained from atomistic studies of viral IDPs, such as the α -helical molecular recognition element (α -MoRE) of the intrinsically disordered C-terminal domain of the measles virus nucleoprotein (NTAIL) binding to the X domain of the viral phosphoprotein[44]. Unbiased molecular dynamics simulations capturing numerous binding and unbinding events revealed that complex formation

predominantly occurs through induced-folding pathways, in which intermolecular contacts are established before or concomitantly with secondary structure formation. The transition state ensemble is highly heterogeneous and characterised by only a small number of key intermolecular contacts, rather than by preformed secondary structure. Notably, conformations with extensive helical content early in the binding process were more likely to dissociate, whereas more disordered conformations exhibited a kinetic advantage in progressing toward the native complex. These observations illustrate how structural disorder can facilitate efficient target recognition and complex formation.

In many cases, however, binding does not result in complete ordering of the disordered region. Instead, IDPs often form fuzzy complexes (Fig- 5a), in which significant conformational flexibility is retained in the bound state. A well-characterised example is provided by transcriptional regulators such as the kinase-inducible domain of CREB interacting with the KIX domain of CBP, where multiple binding modes coexist and parts of the disordered chain remain dynamic while engaging the partner. As discussed in the referenced work, fuzziness allows dynamic redistribution of contacts, enabling tunable affinity, multivalency, and sensitivity to PTMs or competing interactions. Such partially ordered complexes highlight that functional binding does not require full structural stabilisation and that residual disorder can be an integral component of molecular recognition and regulation[45, 46].

IDPs also contribute to the formation of biomolecular condensates through liquid-liquid phase separation (LLPS) (Fig- 5c). Their low-complexity and multivalent interaction motifs enable them to form dynamic assemblies such as stress granules, nucleoli, and transcriptional hubs. In these contexts, disorder provides the flexibility and multivalency required for reversible assembly and disassembly, allowing cells to compartmentalize biochemical reactions without membrane-bound organelles[47, 48]. Although IDPs play many beneficial roles, dysregulation of IDPs is associated with numerous diseases. Misfolding, aberrant aggregation, or pathological phase transitions of disordered proteins contribute to neurodegenerative diseases such as Alzheimer’s disease, Parkinson’s disease, and amyotrophic lateral sclerosis[49](Fig- 5d). Mutations within IDRs often disrupt interaction networks or PTM patterns, leading to cancer, congenital disorders, or impaired regulation. These observations have stimulated major interest in understanding the structural and energetic principles governing IDPs and in designing therapeutic strategies that target disordered regions[50].

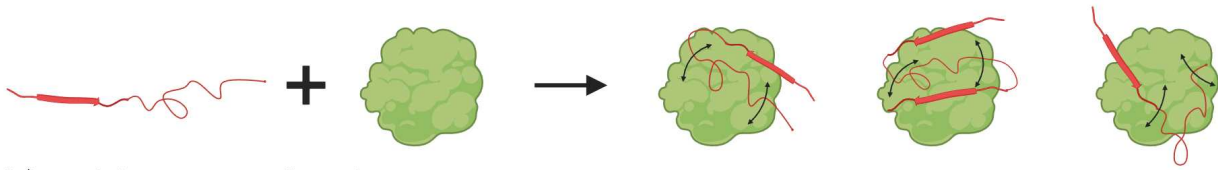
From a biophysical perspective, IDPs occupy a distinct protein energy landscape. Instead of a deep, narrow basin corresponding to a unique native structure, their landscapes are shallow and rugged, with many nearly isoenergetic conformations[51]. The absence of a dominant structural minimum allows IDPs to explore a broad conformational space, which is essential for their functional roles. This view aligns with the energy-landscape framework introduced in Section 1.2 and highlights a continuum between ordered and disordered proteins. Both are governed by landscapes shaped by sequence-encoded energetic biases, but IDPs exist in regions where entropy and frustration prevent collapse into a single fold[52].

The recognition of IDPs and IDRs has transformed our understanding of protein science. Disorder is now viewed not as a failure to form structure, but as an

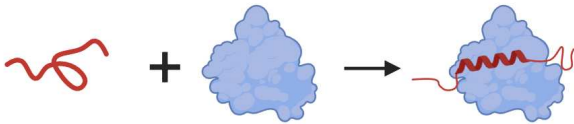
evolutionarily selected property that expands the functional repertoire of the proteome. IDPs exemplify the shift from the classical sequence–structure–function paradigm toward a broader sequence–ensemble–function paradigm, where dynamic structural heterogeneity is essential for biological activity.

In the following section, this ensemble-based perspective will be developed further with a discussion of conformational ensembles and the experimental and computational approaches used to characterise the structural heterogeneity of disordered proteins.

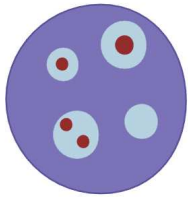
a) Fuzzy binding



b) Folding upon binding



c) Biomolecular condensates



d) IDP & diseases (Tau protein aggregation)

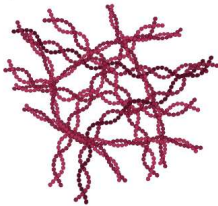


Figure 5: Key functional and pathological behaviours of intrinsically disordered proteins, including fuzzy binding, folding upon binding, biomolecular condensates, and disease-associated aggregation.

1.5 Toward a Sequence-Ensemble-function Paradigm

As protein science has progressed, it has become clear that many proteins cannot be described adequately by a single static conformation. Instead, their biological properties arise from the ensemble of conformations they populate under physiological conditions. This understanding has led to an expanded conceptual framework in which an amino-acid sequence encodes not one preferred structure but a distribution of states, and the relative populations of these states mediate function. The sequence–

ensemble–function paradigm offers a unified way to describe both ordered and disordered proteins by focusing on how sequence shapes conformational landscapes and how these landscapes give rise to biological activity[10, 53].

A central principle of this paradigm is that protein sequences define energy landscapes, which determine the accessibility, stability, and dynamics of conformational states. For well-folded proteins, evolutionary pressures create landscapes that funnel toward a narrow native basin. In contrast, disordered or partially disordered sequences produce shallow or broad landscapes that allow many conformations to remain populated at equilibrium. These ensembles represent the thermodynamic ground state for such proteins, with sequence composition, charge distribution, and residue patterning playing key roles in determining ensemble dimensions and properties[41, 54].

Within this ensemble-based view, function is understood as an outcome of population distributions and their modulation. Changes in sequence, such as charged or hydrophobic substitutions, altered patterning, or PTMs, can shift ensemble properties including compaction, flexibility, and long-range interactions. Such shifts can strongly influence binding affinity, recognition specificity, regulatory responses, or phase behaviour, even when no stable tertiary structure is present. Therefore, the functional effects of sequence variation are often better interpreted as changes in ensemble populations rather than alterations to discrete structural motifs[53].

The ensemble perspective also clarifies why proteins across the structural spectrum display diverse behaviours. Ordered proteins rely on fluctuations around the native basin, multi-domain and partially disordered proteins depend on interdomain motions and structural plasticity, and highly disordered proteins operate through broad, dynamic ensembles. The distinction among these cases lies not in the presence or absence of structure but in the scope and organisation of the ensemble and how it responds to cellular conditions.

The development of this paradigm has been supported by advances in methods that directly probe conformational heterogeneity. Experimental techniques such as nuclear magnetic resonance (NMR)[55, 56], single-molecule fluorescence resonance energy transfer (smFRET)[57], small-angle X-ray scattering (SAXS)[58], and mass spectrometry[59] provide observables that reflect distributions of states. Computational approaches, including molecular dynamics simulations (MD)[60] and integrative ensemble refinement[61], enable the construction of conformational ensembles that are consistent with sequence features and experimental data. These approaches shift the focus of structural characterisation from identifying a single representative conformation to mapping the ensemble that defines protein behaviour[62].

The sequence–ensemble–function paradigm also provides a coherent foundation for understanding emergent cellular phenomena such as multivalent interactions and biomolecular condensate formation. In these systems, ensemble properties, including flexibility and transient secondary structure, govern assembly and regulation more directly than any single structural form.

In summary, the sequence–ensemble–function paradigm recognises that protein sequences encode conformational landscapes and that biological function arises from the properties of these ensembles. This framework integrates ordered, flexible, and

fully disordered proteins and emphasises the essential role of structural dynamics in shaping molecular function (Fig- 6).

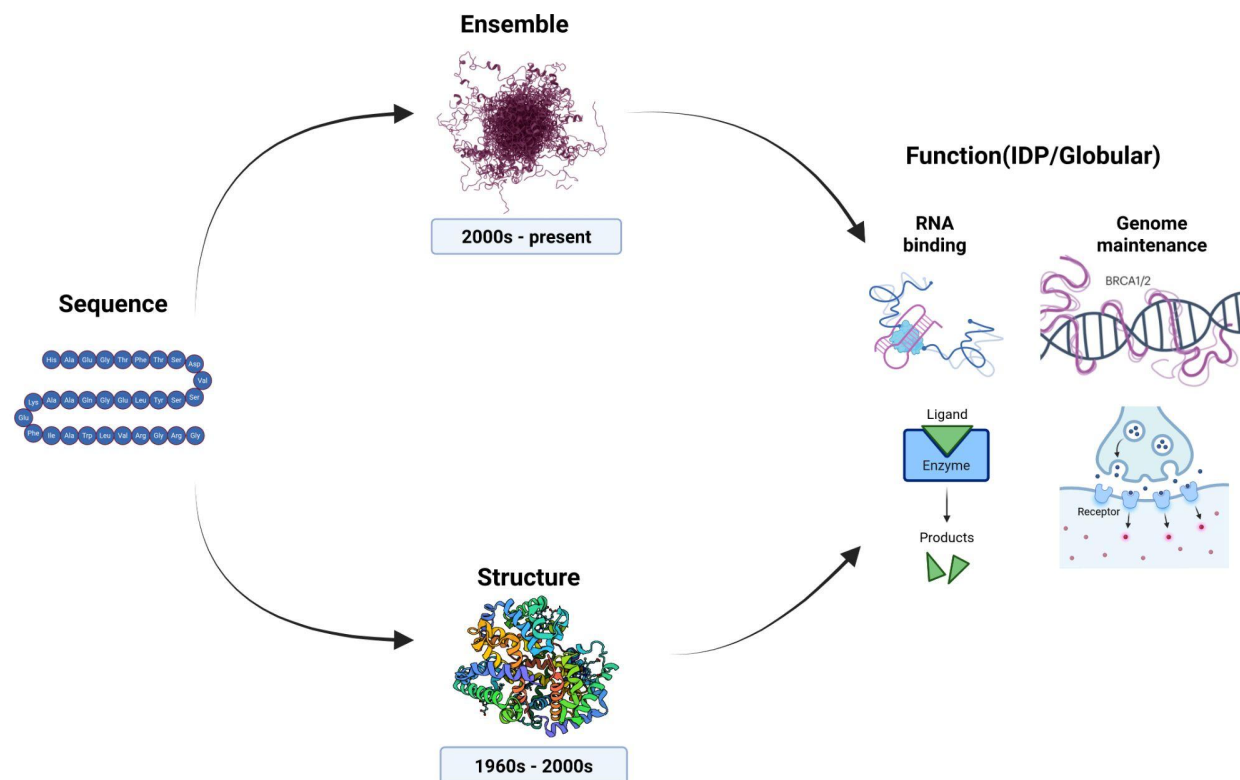


Figure 6: Overview of the shift from a sequence–structure–function view of proteins to a sequence–ensemble–function framework, illustrating how both folded and disordered conformational states give rise to diverse biological functions

1.6 Conformational Ensemble Properties and Sequence Determinants of IDPs

Because IDPs exist as dynamic ensembles (Fig- 7a) rather than single well-defined structures, their structural description necessarily differs from the classical view of a unique three-dimensional conformation. Instead, IDPs are characterized through ensemble properties, which quantify statistical features of the full set of conformations populated under given conditions (Fig- 7b-e). These properties provide a physically meaningful way to describe the size, shape, heterogeneity, and internal organization of disordered ensembles, and they can be measured experimentally, computed from simulations, or predicted from sequence[31, 63, 64].

Commonly used global ensemble descriptors include the radius of gyration and the end-to-end distance, which quantify the overall dimensions and degree of compaction of a disordered chain. The radius of gyration reports on how mass is distributed around the center of mass of the ensemble and is particularly useful for distinguishing between compact, semi-compact, and expanded ensembles. In contrast, the end-to-

end distance measures the separation between the N- and C-termini and is more sensitive to chain extension and large-scale fluctuations, making it informative for describing linker-like or spacer functions of IDPs. While these two metrics are often correlated, they capture complementary aspects of global ensemble dynamics[65, 66](Fig- 7b).

Shape-related metrics, such as asphericity, provide additional insight into the anisotropy of the ensemble. Asphericity distinguishes between roughly spherical, globular-like ensembles and more elongated or rod-like conformations, thereby reporting on whether conformations are isotropically compact or preferentially extended along one dimension. This property is particularly relevant for understanding how IDPs explore space and how they may sterically influence the relative positioning of interaction partners in multi-domain proteins or complexes[67](Fig- 7b).

Local ensemble descriptors capture sequence-dependent structural propensities within disordered regions. Transient secondary structure content, typically quantified as residue-resolved propensities for α -helices, β -strands, or turns, reflects local energetic biases encoded in the amino-acid sequence. Although these secondary structure elements are short-lived and sparsely populated, they can play important roles in molecular recognition, folding upon binding, and the formation of molecular recognition elements. Such local order often correlates with functional sites and interaction motifs embedded within IDPs[68, 69].

Inter-residue distance distributions provide a more detailed view of spatial organization within the ensemble by reporting on the probability of contacts or proximity between specific residue pairs. These distributions capture long-range interactions that are not apparent from global metrics alone and can reveal transient intramolecular contacts, sequence-encoded compaction, or preferences for specific topologies. Importantly, distance distributions retain information about heterogeneity and multimodality, allowing discrimination between ensembles that share similar average dimensions but differ in internal organization[70](Fig -7b).

A central insight from the last decade of IDP research is that these ensemble properties are not arbitrary but are strongly encoded by amino-acid sequence features(Fig- 7e). In particular, the amino-acid composition of IDPs differs systematically from that of folded proteins, being depleted in bulky hydrophobic residues and enriched in polar and charged residues. Building on this observation, a polymer-physics-based framework has been developed to relate sequence composition to conformational preferences through composition-to-conformation relationships (CCRs). In this view, IDP sequences can be classified into broad compositional categories, such as polar tracts, polyampholytes, and polyelectrolytes, that map onto distinct conformational states, ranging from compact globule-like ensembles to expanded coil-like states[71]. Quantitative parameters derived from sequence composition play a key role in this mapping. The fraction of charged residues (FCR) and the net charge per residue (NCPR) determine the balance between intra-chain electrostatic interactions and solvent-mediated effects, thereby influencing whether an ensemble is compact or expanded. Sequences with low FCR and low NCPR tend to favor more compact ensembles, whereas sequences with high absolute NCPR or high FCR typically populate expanded conformations due to electrostatic repulsion and favorable solvation

of charged side chains. Importantly, many IDPs occupy intermediate regimes in which attractions and repulsions are balanced, resulting in maximal conformational heterogeneity and ensembles that include mixtures of compact, semi-compact, and extended conformations[40, 41, 72].

Beyond overall composition, the linear patterning (Fig- 7d) of residues along the sequence has emerged as a crucial determinant of ensemble properties. For polyampholytic IDPs, the degree of mixing versus segregation of oppositely charged residues strongly influences conformational behavior. This effect can be quantified by parameters that describe charge patterning, which modulate the effective strength of intra-chain electrostatic attractions. Well-mixed charge distributions tend to favor expanded, coil-like ensembles, whereas segregated charge patterns promote compaction and, in some cases, the formation of locally collapsed domains(Fig- 7c). Thus, two sequences with identical amino-acid composition but different charge patterning can populate markedly different conformational ensembles, highlighting that sequence order, not only composition, is functionally relevant[41, 73].

These sequence–ensemble relationships provide a physical basis for understanding how IDPs achieve functional specificity without adopting stable tertiary structures. Many biological functions of IDPs depend sensitively on ensemble dimensions, flexibility, and heterogeneity rather than on precise atomic arrangements. For example, expanded ensembles can act as entropic spacers or flexible linkers, whereas more compact or chimeric ensembles can modulate interaction probabilities, regulate accessibility of short linear motifs, or facilitate context-dependent folding upon binding. Small sequence perturbations, such as point mutations, alternative splicing, or post-translational modifications, can shift ensemble properties by altering charge balance or patterning, thereby tuning function without requiring large structural rearrangements[31, 54].

Taken together, this ensemble-centric description establishes that IDP behavior is governed by a hierarchy of sequence-encoded features that shape conformational landscapes and ensemble properties. This perspective directly complements the sequence–ensemble–function paradigm introduced earlier and provides the conceptual foundation for both experimental ensemble determination and computational ensemble modeling.

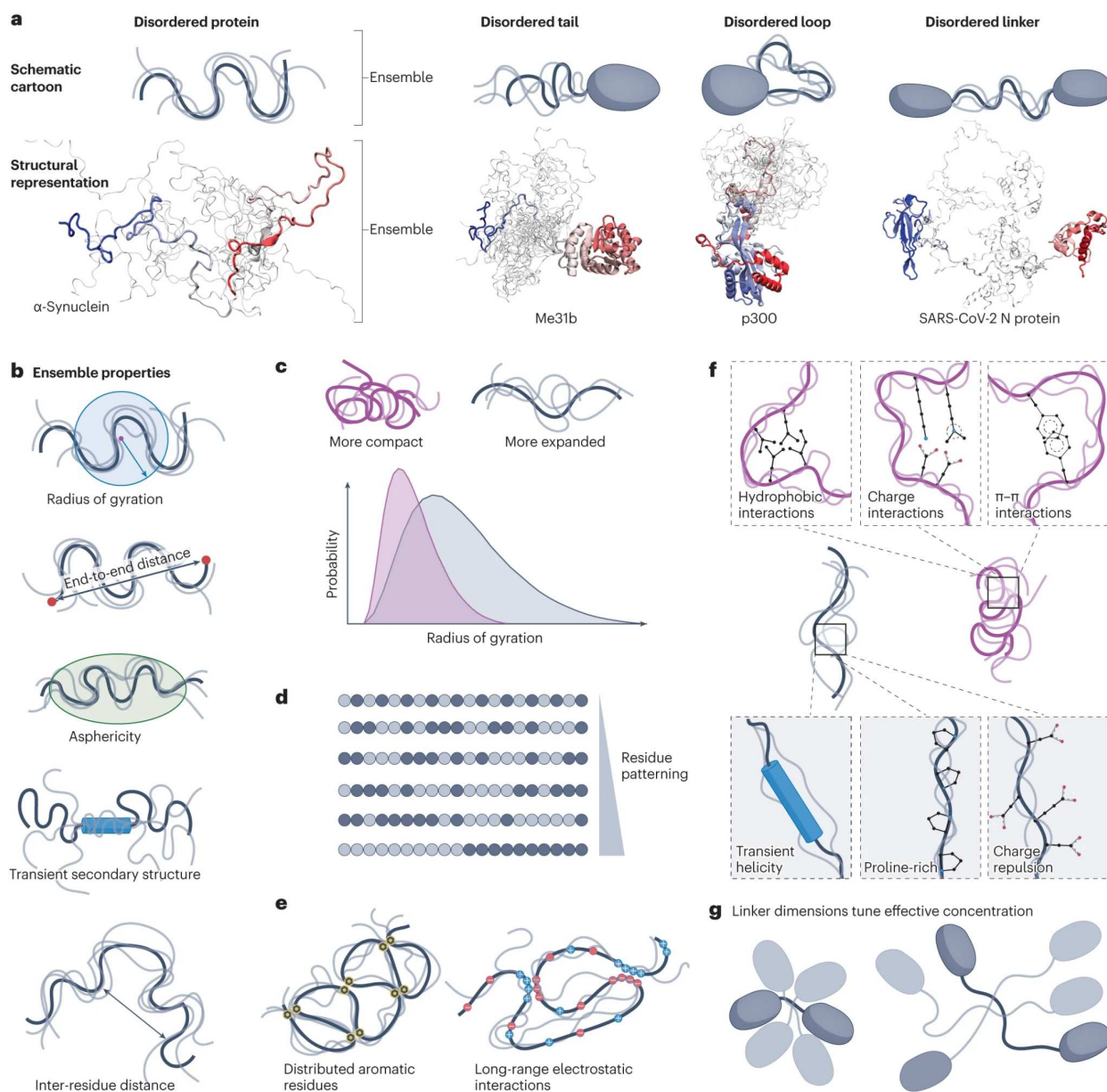


Figure 7: a) IDRs exist as highly heterogeneous conformational ensembles rather than single structures, in contrast to folded domains, and their ensemble nature depends on molecular context. b–c), IDR ensembles are quantitatively described using ensemble properties (for example, radius of gyration, shape, and inter-residue distances), which are best represented as probability distributions. d–g), Sequence features and residue patterning encode IDR ensemble behavior, which in turn tunes functional outcomes such as interdomain communication and effective concentrations of folded domains. This figure has been taken from [31]

1.7 Databases Supporting IDP Research

The rapid growth of IDP research has been accompanied by the development of specialised databases that collect, curate, and integrate information on the sequence, structure, function, and biophysical properties of disordered regions. These resources

are essential for large-scale analyses, benchmarking computational methods, and formulating new hypotheses. They also complement more general repositories such as UniProt, the Protein Data Bank (PDB), and the Biological Magnetic Resonance Bank (BMRB). Broadly, IDP-related databases (Table- 2) can be categorised into three groups: sequence- and annotation-centric resources, structural and ensemble databases, and functional or context-specific resources.

DisProt is the primary manually curated database dedicated to IDPs (Fig- 8). It provides experimentally supported annotations of disorder, order-disorder transitions, and functional roles of IDRs, using a controlled ontology that links each annotation to specific experimental evidence [74]. DisProt connects disordered segments to the underlying techniques used to identify them (such as NMR, SAXS, or missing electron density in X-ray structures) and to the corresponding literature, making it a gold-standard reference for training and evaluating disorder prediction algorithms. Recent updates have expanded the coverage and functional depth of the database, with improved evidence codes and curation strategies that enhance annotation interpretability.

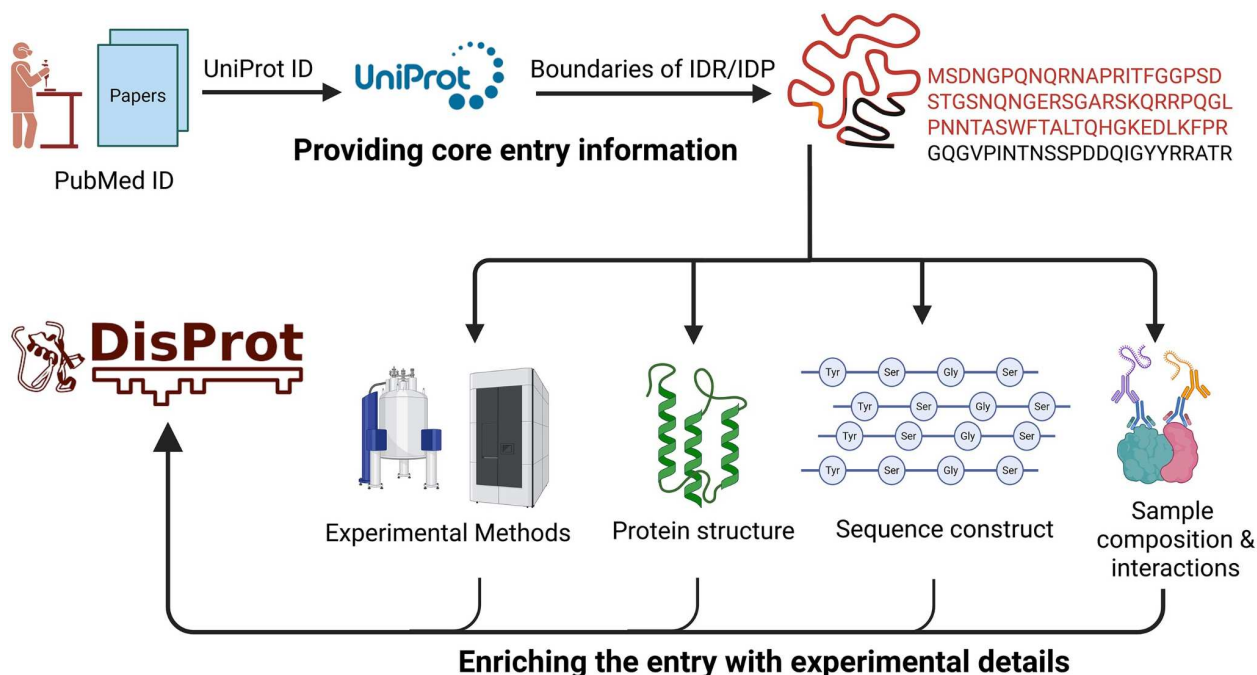


Figure 8: Deposition process and information content in DisProt. This figure has been taken from [74]

In parallel, MobiDB [75] provides a more integrative view of disorder and mobility across entire proteomes (Fig- 9). It aggregates information from curated sources such as DisProt and IDEAL [76], indirect experimental evidence (e.g. missing residues in PDB structures), and multiple computational predictors into a consensus per-residue annotation. MobiDB thus serves as a general-purpose resource for researchers who need disorder annotations for any UniProt sequence, providing both high-confidence curated regions and large-scale predictive coverage. Its consensus scheme, including an explicit “ambiguous” state when sources disagree, is particularly useful for mapping

the “disorder landscape” of proteins and for exploring relationships between disorder, mobility, and function[75].

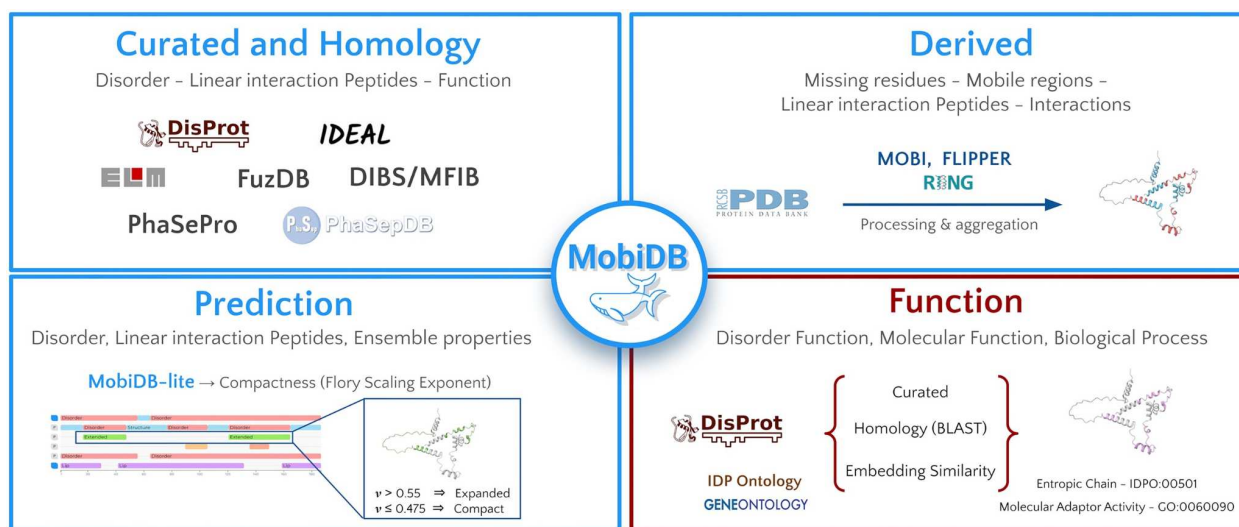


Figure 9: Information content in MobiDB. This figure has been taken from[75]

IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature) is another important curated resource focused on experimentally characterized IDPs[76].

Beyond simply labeling regions as ordered or disordered, IDEAL annotates interaction sites, post-translational modification sites, regions undergoing disorder-to-order transitions (Protean segments), structural domain assignments, and interaction networks involving disordered segments. This detailed functional and interaction information makes IDEAL particularly valuable for case studies and for understanding how specific disordered segments participate in regulatory and signaling pathways. A broader overview of such IDP-focused databases and their complementary roles is provided in recent reviews[77].

Whereas DisProt, MobiDB and IDEAL primarily provide sequence-level annotations, the Protein Ensemble Database (PED) addresses the need for storing and disseminating structural ensembles of IDPs and other flexible systems. PED is the major repository for conformational ensembles that have been derived by combining experimental data (e.g. NMR, SAXS, FRET) with computational modeling. Moreover, a recently added dataset of purely predicted conformational ensembles for the human proteome, enriched in disordered regions, has been deposited in PED[78]. It should be noted that these predicted ensembles are not associated with experimental data, and mechanistic interpretations based on them should therefore be made with caution.

Each PED entry includes the ensemble coordinates, when provided, weight assigned to each conformation, the associated experimental restraints, and rich metadata describing the system and modeling protocol. PED thus complements the Protein Data Bank[79], which primarily stores single conformers, by providing an infrastructure suited to the representation of conformational ensembles[80].

Beyond structural representation, understanding the functional roles of IDPs also requires resources that capture sequence-encoded interaction motifs and regulatory elements. IDPs often function through short linear motifs (SLiMs) embedded in disordered regions, which mediate transient and regulated interactions. The Eukaryotic Linear Motif (ELM) resource[81] collects manually curated, experimentally validated SLiMs and provides powerful tools to explore motif instances and their sequence context. ELM annotations, together with disorder information from MobiDB or DisProt, enable systematic mapping of functional sites[82] within IDRs and are widely used to study modular signaling proteins, viral hijacking of host machinery[83], and regulatory networks.

A number of additional resources capture specialized aspects of IDP behavior. Databases such as FuzDB collect examples of “fuzzy” complexes in which disordered regions remain partially disordered upon binding, emphasizing that functional complexes can be structurally heterogeneous[84]. Given the prominent role of IDRs in biomolecular condensates, several curated resources have been developed for proteins undergoing liquid–liquid phase separation (LLPS), including PhaSePro for LLPS drivers, PhaSepDB for phase-separation-associated proteins, and LLPSDB/LLPSDB v2.0 and DrLLPS for experimentally characterized LLPS systems across organisms. Although these databases are not limited to IDPs, many of their entries contain extensive IDRs, and they provide crucial contextual information on how disorder, multivalency and sequence features contribute to condensate formation[85, 86, 87]. Together, these databases form a rich ecosystem that supports IDP research from multiple angles: DisProt, IDEAL and MobiDB provide high-quality disorder annotations; PED captures conformational ensembles; ELM and related resources highlight functional motifs and interaction modules; and LLPS-focused databases reveal how disordered sequences participate in higher-order assemblies. Leveraging these complementary resources is essential for building realistic sequence–ensemble–function models and for designing future computational and experimental studies on intrinsically disordered proteins and regions.

Table 2: Summary of major databases supporting IDP research, including curated disorder annotations, structural ensembles, functional motifs, and resources related to phase separation.

Database	Type	Content & Key Features	Typical Use Cases
DisProt	Manually curated disorder annotations	Experimentally validated IDRs; evidence codes; disorder functions; curated literature links	Gold-standard dataset for benchmarking predictors; studying functional roles of disorder
MobiDB	Integrative disorder database	Consensus disorder from curated sources, PDB missing residues, and multiple predictors; per-residue annotations	Large-scale analyses; proteome-wide disorder annotation; generating consensus disorder profiles
IDEAL	Curated IDP/IDR functional annotation	Verified IDRs, protean segments (disorder-order transitions), interaction regions, PTMs	Case studies of IDP mechanisms; functional annotation of regulatory disordered segments
PED (Protein Ensemble Database)	Structural ensemble repository	Experimentally validated conformational ensembles from NMR, SAXS, sm-FRET, integrative modeling; full coordinate sets	Structural modeling of IDPs; training/testing ensemble inference methods; comparing modeling protocols
ELM (Eukaryotic Linear Motif Resource)	Functional motif database	Experimentally verified short linear motifs (SLiMs) in IDRs; motif classes (modification, docking, degradation, cleavage)	Identifying regulatory motifs in IDRs; mapping interaction sites; studying signaling networks
FuzDB	Fuzzy complex database	Cases where proteins retain disorder upon binding; structural heterogeneity in complexes	Studying fuzzy binding and dynamic complexes; characterizing partial ordering in IDPs
PhaSePro	Curated LLPS database	Experimentally validated phase separation drivers; disorder-rich proteins; molecular conditions of LLPS	Studying sequence determinants of phase separation; identifying LLPS drivers
PhaSepDB	Broad LLPS resource	Proteins associated with phase-separated assemblies; curated and computational evidence	Exploring relationships between disorder, multivalency, and condensates
LLPSDB / LLPSDB v2	Experimental LLPS dataset	In-vitro LLPS experiments with sequences, conditions, mutations, and droplet behavior	Mechanistic LLPS studies; training ML models for phase separation

2– Research objectives

The work described in this thesis is unified by the general goal of advancing the understanding and practical characterisation of protein conformational ensembles, with a particular focus on IDPs and conformationally dynamic systems (fig-6). Motivated by the shift from static structure-based descriptions toward ensemble-based representations of protein behaviour, this thesis aims to address key conceptual, methodological, and infrastructural challenges that currently limit the widespread and reproducible use of conformational ensembles.

To achieve this goal, the thesis combines infrastructure development, methodological standardisation, and computational tool design. First, dedicated computational infrastructure is developed and extended to support the storage, processing, and dissemination of conformational ensembles at scale. Second, principled workflows are defined for the determination of conformational ensembles by integrating experimental data and computational modelling in a transparent and reproducible manner. Third, computational tools are developed to analyse and compare ensembles quantitatively, enabling extraction of biologically meaningful information from heterogeneous and dynamic structural data.

In order to organise the content of this thesis coherently, the work is divided into three connected chapters, each addressing a specific aspect of ensemble-based protein science. The objectives of these chapters are summarised below.

2.1 – Infrastructure for Ensemble-Based Protein Structure Representation

This chapter focuses on the development and consolidation of computational and data infrastructure required to support ensemble-based descriptions of protein structure. The aim is to address the limitations of traditional structure repositories and workflow systems when applied to intrinsically disordered and conformationally flexible proteins.

The specific objectives of this chapter are:

- a) To contribute to the development and maintenance of the Protein Ensemble Database (PED) as a dedicated repository for experimentally supported conformational ensembles, including the definition of data models, deposition workflows, and metadata standards.
- b) To develop and deploy DRMAAtic, a workflow and job-management framework that enables reproducible, scalable execution of ensemble generation and validation pipelines across heterogeneous computing environments.
- c) To design and implement backend workflows for bulk deposition of conformational ensembles, enabling large-scale ensemble integration into PED, including proteome-wide ensemble datasets such as those generated in the AlphaFlex project.
- d) To evaluate infrastructure scalability and robustness by applying these tools to large ensemble collections, thereby identifying practical constraints and best practices for ensemble-centric data management.

2.2 – Standardised Workflows for the Determination of Conformational Ensembles

This chapter addresses the methodological challenges associated with determining protein conformational ensembles from experimental data and computational sampling. The aim is to define clear, reproducible workflows that support ensemble generation while explicitly accounting for uncertainty, heterogeneity, and methodological assumptions.

The specific objectives of this chapter are:

- a) To formalise ensemble determination as a multi-step workflow, including experimental methods, computational generation techniques, comparison methods and validation process.
- b) To analyse and compare existing ensemble determination strategies, such as Bayesian inference, maximum entropy and maximum parsimony refinement, and reweighting approaches, highlighting their shared principles and methodological differences.
- c) To propose best practices for ensemble validation and robustness assessment, including cross-validation, sensitivity analysis, and the use of independent experimental observables.
- d) To align ensemble determination workflows with existing infrastructure, ensuring compatibility with ensemble repositories and enabling automated and reproducible ensemble deposition.
- e) Suggesting a benchmarking pipeline to quantitatively validate conformational ensembles of disordered proteins.
- f) Review the state-of-art methods in both experimental approaches and computational techniques in order to determine conformational ensembles.

2.3 – Computational Tools for the Analysis of Conformational Ensembles

This chapter focuses on my work on the development of a computational tool for the quantitative analysis of conformational ensembles. The objective is to move beyond single-structure metrics and provide ensemble-aware descriptors that capture heterogeneity and population distributions.

The specific objectives of this chapter are:

- a) To develop a computational method for describing ensemble-level properties, including global dimensions, contact patterns, secondary-structure propensities, and measures of conformational diversity.
- b) To design analysis tools that are robust to ensemble heterogeneity, enabling comparison between ensembles generated using different sampling strategies or experimental restraints.
- c) To integrate ensemble analysis tools into standardised workflows and infrastructure, facilitating large-scale and automated analysis of deposited ensembles.
- d) To demonstrate the applicability of ensemble analysis methods to intrin-

sically disordered proteins, flexible linkers, and other conformationally dynamic systems, highlighting their relevance within the sequence–ensemble–function paradigm.

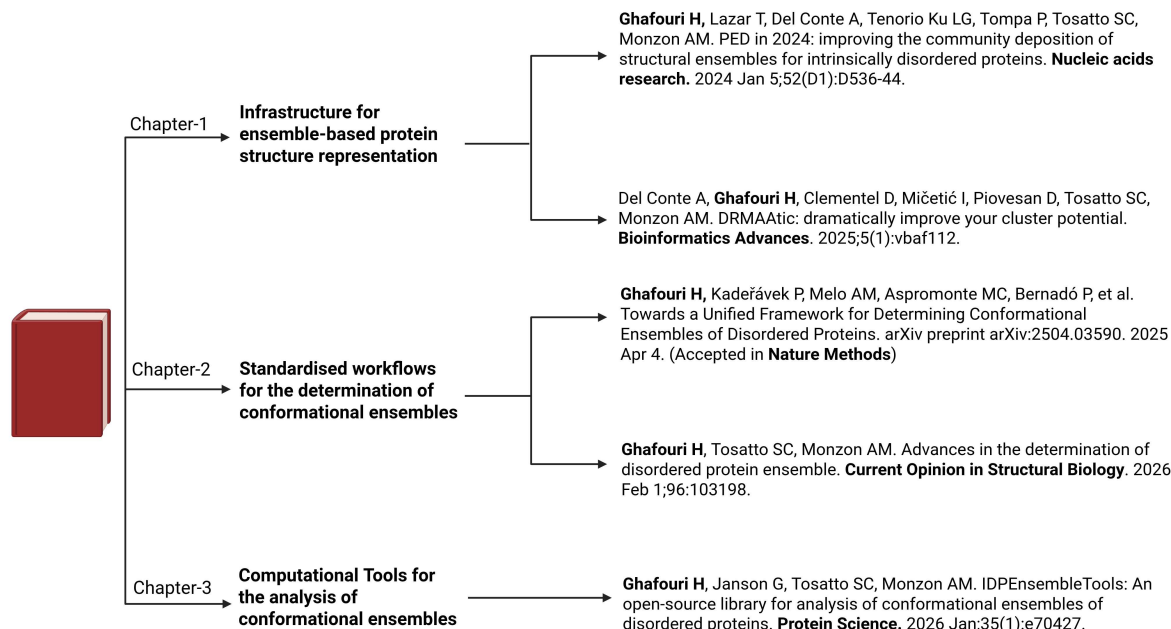


Figure 10: Overview of the paper-collection structure of this thesis, showing the three interconnected chapters and their objectives, which are addressed through published and submitted works on infrastructure development, standardised workflows for conformational ensemble determination, and computational tools for ensemble analysis within the sequence–ensemble–function framework.

Chapter 1: Infrastructure for Ensemble-Based Protein Structure Representation

This chapter contains two articles, one published in Nucleic Acid Research (NAR), one in Bioinformatics Advances.

1. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. **Ghafari H**, Lazar T, Del Conte A, Tenorio Ku LG, Tompa P, Tosatto SC, Monzon AM. *Nucleic acids research*. 2024 Jan 5;52(D1):D536-44. (<https://doi.org/10.1093/nar/gkad947>)
2. DRMAAatic: dramatically improve your cluster potential. Del Conte A, **Ghafari H**, Clementel D, Mičetić I, Piovesan D, Tosatto SC, Monzon AM. *Bioinformatics Advances*. 2025;5(1):vbaf112. (<https://doi.org/10.1093/bioadv/vbaf112>) (The appendix associated with this paper is provided in **Appendix A** at the end of the thesis.)

PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins

Hamidreza Ghafouri^{1,†}, Tamas Lazar^{2,3,†}, Alessio Del Conte¹, Luigi G Tenorio Ku¹, PED Consortium, Peter Tompa^{2,3,4}, Silvio C.E. Tosatto^{1,*} and Alexander Miguel Monzon^{5,*}

¹Department of Biomedical Sciences, University of Padova, Padova, Italy

²VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie (VIB), Brussels, Belgium

³Structural Biology Brussels, Department of Bioengineering, Vrije Universiteit Brussel (VUB), Brussels, Belgium

⁴Institute of Enzymology, Research Centre for Natural Sciences (RCNS), Budapest, Hungary

⁵Department of Information Engineering, University of Padova, Padova, Italy

*To whom correspondence should be addressed. Tel: +39 049 827 6269; Email: alexander.monzon@unipd.it

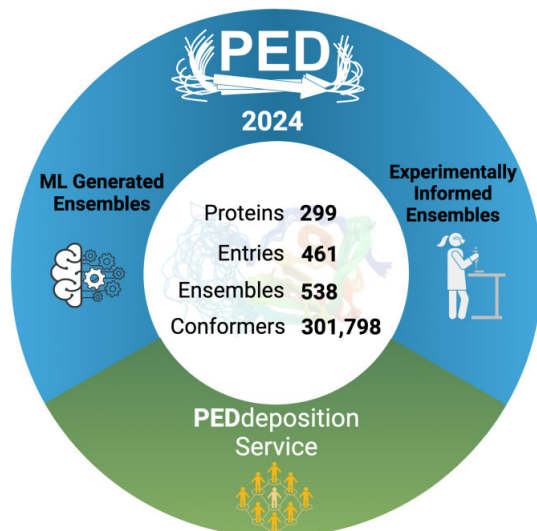
Correspondence may also be addressed to Silvio C.E. Tosatto. Email: silvio.tosatto@unipd.it

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

The Protein Ensemble Database (PED) (URL: <https://proteinensemble.org>) is the primary resource for depositing structural ensembles of intrinsically disordered proteins. This updated version of PED reflects advancements in the field, denoting a continual expansion with a total of 461 entries and 538 ensembles, including those generated without explicit experimental data through novel machine learning (ML) techniques. With this significant increment in the number of ensembles, a few yet-unprecedented new entries entered the database, including those also determined or refined by electron paramagnetic resonance or circular dichroism data. In addition, PED was enriched with several new features, including a novel deposition service, improved user interface, new database cross-referencing options and integration with the 3D-Beacons network—all representing efforts to improve the FAIRness of the database. Foreseeably, PED will keep growing in size and expanding with new types of ensembles generated by accurate and fast ML-based generative models and coarse-grained simulations. Therefore, among future efforts, priority will be given to further develop the database to be compatible with ensembles modeled at a coarse-grained level.

Graphical abstract



Received: September 15, 2023. Revised: October 10, 2023. Editorial Decision: October 11, 2023. Accepted: October 13, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Intrinsically disordered proteins or regions (IDPs/IDRs) lack a specific, stable structure and instead exist as rapidly interconverting conformers. This arises from their relatively uniform free-energy landscape, resulting in their highly dynamic and heterogeneous nature (1). IDPs/IDRs play significant roles in various essential functions such as cell signaling, regulation and recognition. Furthermore, their involvement in numerous human diseases renders them highly attractive targets for therapeutic drug discovery (2). While binding modes of some IDPs/IDRs that fold upon interaction offer valuable structural insights (3), gaining a thorough comprehension of the complex mechanisms governing the function of IDPs also requires knowledge of their structural dynamics in the unbound state, and many IDPs/IDRs form fuzzy complexes (3). Given their extreme conformational dynamics, modeling IDPs/IDRs in terms of ensembles is the only valid strategy for structurally studying IDP function. By definition, a conformational ensemble consists of multiple structures, each with their statistical weights representing their relative populations and transition rates that quantify their dynamics (4). Despite the steady expansion of experimentally determined protein structures in the Protein Data Bank (5) and the recent AlphaFold Protein Structure Database (6), which contains accurate structural models of millions of proteins, the information they offer about the dynamic nature of proteins remains limited, especially in the context of ensemble representation of IDPs. In 2014, the Protein Ensemble Database (PED) (7) was established to bridge this gap, and over time, it has consistently evolved, enhancing the quantity and quality of deposited ensembles.

Generally, conformational ensembles are determined by integrating experimental and computational methods. This involves a diverse range of experimental techniques, including nuclear magnetic resonance (NMR) spectroscopy, small angle X-ray scattering (SAXS), single-molecule Förster resonance energy transfer (smFRET), electron paramagnetic resonance (EPR) and circular dichroism (CD) (4,8). These experimental measurements serve as global and/or local constraints, enabling the resampling and reweighting of a pool of conformers generated through statistical conformer generators or molecular dynamics (MD)/Monte Carlo (MC) simulations. Moreover, the advent of AlphaFold2 (9), RoseTTAFold (10) and the advancements in machine learning approaches have fostered the development of various pipelines aimed at effectively modeling multiple conformational states or predicting conformational ensembles (11–13). Nevertheless, despite recent progress in the field, modeling conformational ensembles, especially for IDRs/IDPs, remains challenging. On the computational front, a significant obstacle arises from the lack of a precise energy function to guide MD or MC simulations (14,15), coupled with limited computational resources for thorough sampling of the conformational space (16). On the other hand, from an experimental standpoint, a major challenge is accurately quantifying all sources of errors and uncertainties in both the experimental data and the predictors (forward models). Additionally, the observable data are averaged over all members of the ensemble, leading to a reduction in information content. Because of these limitations, resolving structural ensembles has persisted as an ‘underdetermined’ challenge. This viewpoint arises from the fact that the number of degrees of freedom in the ensembles significantly surpasses the

available experimental restraints, leading to multiple potential solutions for the problem without a distinct ‘best’ option. In such a context, having comprehensive and manually curated IDP-related databases, e.g. PED (17), DisProt (18), MobiDB (19), FuzDB (3) and IDEAL (20), can serve multiple purposes. First and foremost, they can serve as a foundational reference and a valuable resource for establishing a validation pipeline to assess the reliability of IDP conformational ensembles. Furthermore, they function as extensive training datasets for upcoming machine learning (ML) models (21). Since the latest PED publication in 2021 (17), a strong emphasis has been given by the scientific community to predict IDP conformational ensembles from sequence by combining ML approaches and MD simulations (22,23), as well as to compare conformational ensembles of flexible proteins (24,25). In this article, we present the new version of the PED (Protein Ensemble Database, <https://proteinensemble.org>), aimed at addressing the evolving challenges and advancements in the field of IDPs/IDRs. Our primary goal has consistently been to enhance the size of our database. In this updated release of PED, we have now accumulated a total of 461 entries and 538 ensembles. This time, we also included IDP ensembles generated without experimental data by novel ML and sampling methods from sequences. A new restyled website with an improved user interface and novel features is presented, as well as a dedicated web-server for the ensemble’s deposition and curation.

Progress and new features

Database growth

PED aims to be the gold-standard primary deposition database for conformational ensembles of non-globular proteins (NGPs) or regions. Therefore, the main goal of PED is to provide an ever-growing platform of structural ensemble entries with a user-friendly deposition pipeline while maintaining high standards for data quality and the FAIR data principles. PED is cross-linked with the main resources to deposit ensembles’ primary experimental data, including BRMB (26), SASBDB (27) and PCDDDB (28).

In this new release, the number of PED entries has increased almost three times compared to the version presented in the last publication (461 versus 162) (17). The source of this data increment comes from depositions from data owners (42 entries in this release), ensembles generated without experimental data (61 entries in this release) and ensemble identification by the PED biocurator team from databases and publications. As detailed below, a larger number of NMR ensemble entries were identified by an automated computational pipeline applied to BMRB (1409 protein structures), which were then subsequently revised, filtered and published by the biocurators (totalling 189 entries).

New entries

Novel ensembles

As in previous releases, new PED ensembles were predominantly modeled using SAXS, NMR spectroscopy, FRET data and their combinations (Figure 1A). NMR data included chemical shifts (CSs), nuclear Overhauser effects (NOEs), J-couplings, residual dipolar couplings (RDCs), relaxation data and paramagnetic relaxation enhancements (PREs) (29). On top of these, for a few new entries, methods such as electron paramagnetic resonance (EPR) spec-

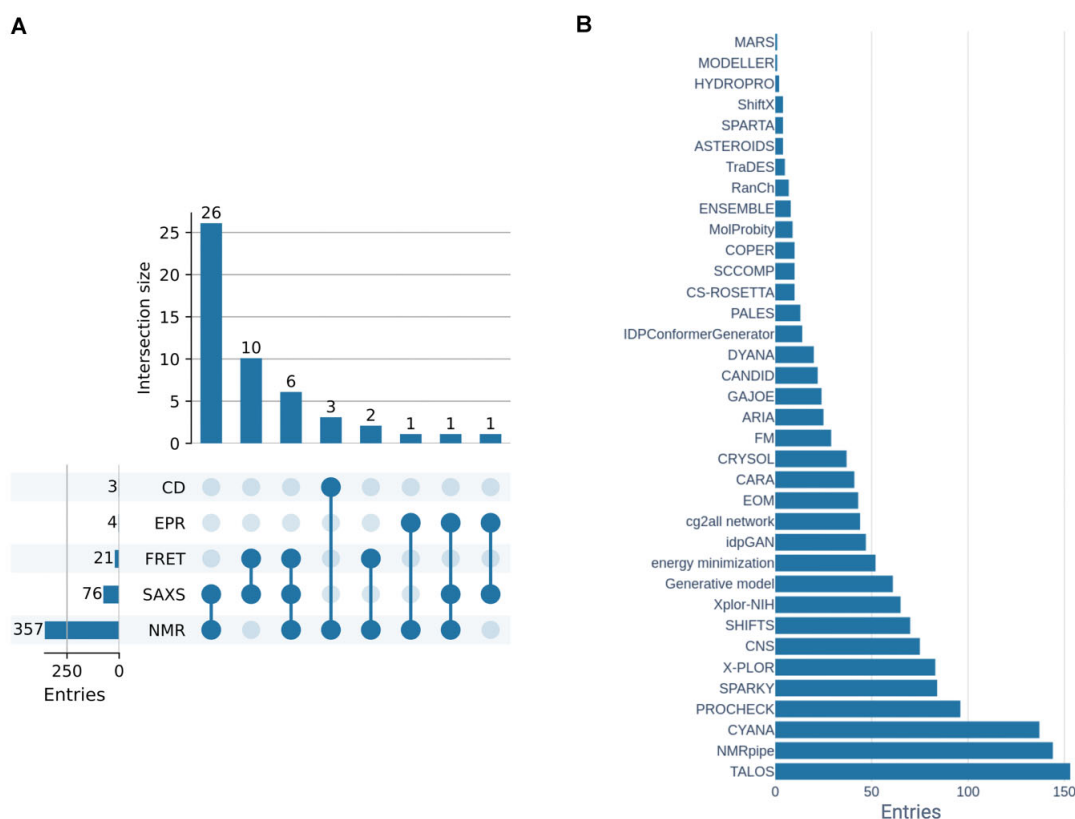


Figure 1. Experimental techniques and ensemble generation methods. **(A)** Matrix layout quantifies the combinations of experimental techniques for PED entries, sorted by size. Filled circles in the matrix indicate which experimental measurement is part of the intersection. **(B)** Distribution of ensemble generation methods and auxiliary software applied in PED. The X axis represents the number of PED entries.

troscopy techniques (e.g. double electron–electron resonance (DEER)) and CD were also used to characterize the protein ensembles (30,31); often in combination with other techniques. These combinations included EPR + NMR, EPR + SAXS, EPR + NMR + SAXS, CD + NMR (30–32). NMR data have already been cross-referenced from BMRB (26) and SAXS data from SASDB (27), but now CD data can also be cross-referenced from PCDDb (28), which will enable PED depositors during submission to reference their CD data already deposited in its primary resource.

Besides these experimental datasets, several PED depositions also used computationally expensive MD simulations to perform integrative structural modeling by reweighting the ensembles. These MD simulations comprised among others trajectories generated by a CHARMM force field and the EEf1 implicit solvent model in a replica-exchange MD setup (33), or AMBER03w force field with TIP4P/2005s water model (34), or replica-exchange Discrete MD (DMD) using the MEDUSA force field in implicit water (35), or coarse-grained Langevin MD in multiple replicas (36), or AMBER99SB-disp force field with its own water model using replica exchange with solute tempering (37). Furthermore, the repertoire of ensemble generation methods and auxiliary software is continuously expanding to encompass state-of-the-art techniques in the field (Figure 1B).

It is often emphasized that IDPs have a high-degree of conformational heterogeneity, which is harder to capture by a single technique. Therefore the integration of simulations and various experiments can better characterize the highly dynamic nature of IDPs (38). Now, there is an increasing number of IDP ensembles in PED determined by different combinations of techniques under the same or slightly different conditions, e.g. hnRNPA1, alpha-synuclein, Tau. We envisage that these ensemble data will reveal not only how sensitive IDPs are to environmental conditions but also the strengths and weaknesses of methods in capturing certain structural aspects.

NMR structural ensembles

A significant upgrade in the new PED version involves the inclusion of a large number of NMR structural ensembles comprising IDRs of IDPs sourced from the PDB. These NMR ensembles represent collections of different conformations (models) that individually satisfy the experimentally derived constraints (8).

To achieve this, we systematically searched the MobiDB (19) to identify NMR ensembles containing IDRs/IDPs. As a starting point, we identified a subset of 2064 proteins containing large RMSD regions defined as ‘mobile’ in MobiDB. Mobile regions are calculated for all NMR ensembles using the Mobi software (39); it is an analogous definition to the

presence of missing residues in X-ray structures. This feature represents highly flexible regions based on structural superposition that change their local conformation in the NMR ensembles.

To further refine our dataset, we also consider the disorder content percentage of the proteins based on two criteria: AlphaFold-disorder (40) and MobiDB-lite predictions (41). Initially, we focused on proteins for which both predictors indicated a disorder content percentage exceeding 50%. In the subsequent phase, we expanded our inclusion criteria to cover proteins where at least one of these two predictors indicated a disorder content percentage above 50%. During this stage, we also verified the availability of experimental NMR data for each protein in the BMRB database (26).

We then established three key criteria to determine the eligibility of NMR ensembles for inclusion in PED: (i) publication availability: we confirmed the existence of a corresponding publication; (ii) consistency in disorder prediction: we ensured that a minimum of ten consecutive residues within the mobile region were classified as disordered by AlphaFold-disorder and/or MobiDB-lite, the cutoff representing the minimum length of IDRs in DisProt; (iii) sufficient conformational coverage: the NMR ensemble had to consist of at least ten distinct structures.

Ensembles without explicit experimental data

Given the recent advancements in ML algorithms for modeling protein structural dynamics (42) and in new methods for sampling IDP conformational ensembles (43,44), we expanded PED and its controlled vocabulary (CV) (<https://proteinensemble.org/about>) to accommodate ensembles calculated without incorporating specific experimental data constraints.

The idpGAN generative model (22) was trained on coarse-grained molecular dynamics (MD) simulations (45) of IDRs from the DisProt database. It is capable of rapidly generating ensembles for arbitrary IDR sequences. IdpGAN does not incorporate experimental data in the ensemble-generation process and, for this update, we did not adopt any reweighting scheme (4) to improve compatibility with the experimental data of the entries. idpGAN was applied to a specific set of sequences from the PED database, involving the careful selection of 47 entries meeting both idpGAN's technical prerequisites and exhibiting a significant fraction of disorder. In this context, idpGAN generated 1000 C α -only conformers for each selected entry, which were then converted into full all-atom structures using the cg2all neural network (46). These resulting structures underwent an energy minimization relaxation process similar to the one in AF predictions. For reproducibility, the entire pipeline was made accessible at https://github.com/feiglab/idpgan_ped.

We have also included fourteen ensembles generated with the new IDPConformerGenerator software suite, which allows statistical or experimentally (chemical shift) biased sampling of torsion angles from the PDB to create all-atom IDPs and IDRs (tails, linkers and loops) in the context of full-length proteins containing folded domains (43,44). IDPConformerGenerator allows exploration of multiple torsion-angle sampling methods that enrich the ensembles' conformational diversity and account for post-translational modifications, multi-chain protein complexes, non-protein ligands such as nucleic acids and lipid bilayers around membrane-bound proteins containing IDRs. The ensembles deposited

were assessed in the original publications (43,44). IDPConformerGenerator is open-source, fully documented with examples, and is accessible at <https://github.com/julie-forman-kay-lab/IDPConformerGenerator>.

PEDdeposition service

PED introduces a dedicated deposition user interface accessible to everyone. This service allows depositors to upload ensembles and metadata, calculate and visualize structural features, and assess ensemble quality through automated validation. Deposition of new ensembles into the PED can be described in three main stages: ensemble deposition by the user, calculation of ensemble properties, and finally, manual curation by PED expert curators (Figure 2).

The initial step in submitting an ensemble involves user authentication through ORCID ID credentials. Within the deposition service, users encounter two primary sections: one for creating a new ensemble draft and another for managing existing drafts. Additionally, the service offers an example ensemble draft to help users become acquainted with the required deposition information. After creating a new draft, the user can begin depositing information, which is organized into three main tabs: description, ensemble upload and construct definition.

Ensemble description

In the 'Experimental procedure' section, users can provide a brief overview of the experimental techniques employed to determine the protein's structural characteristics. The 'Structural ensemble calculation' field captures computational methods, including software for pool generation, forward models and tools for fitting experimental observables with back-calculated measurements from predicted models, as well as validation efforts on the ensemble. For ensembles generated through MD/MC simulations, there is a specific section to detail simulation parameters like software, force field and water model, simulation duration, enhanced sampling, clustering of frames, etc. Additionally, the database offers a controlled vocabulary (CV) organized into an ontology to enhance searchability and standardize keywords describing experimental methodologies, ensemble generation and MD/MC simulations. The last two sections in the ensemble description focus on specifying the NCBI taxonomy ID of the expression organism and providing cross-references to other databases, including the BMRB, SASBDB, PCDDb, DisProt and IntAct (47).

Upload

The upload section of the PEDdeposition service facilitates efficient submission of ensembles. Users can upload multiple-model PDB files that contain the ensemble. Additionally, if available, they can upload a tab-separated file containing weights. These weights indicate the percentage contribution of each conformer to the ensemble. The PED deposition service initiates a validation pipeline to ensure accurate data formatting. External tools like DSSP (48) and MolProbity (49) are employed to compute essential parameters, including secondary structure propensity, accessible surface area, radius of gyration, Ramachandran outliers and steric clash analysis. All resultant data are made available for download.

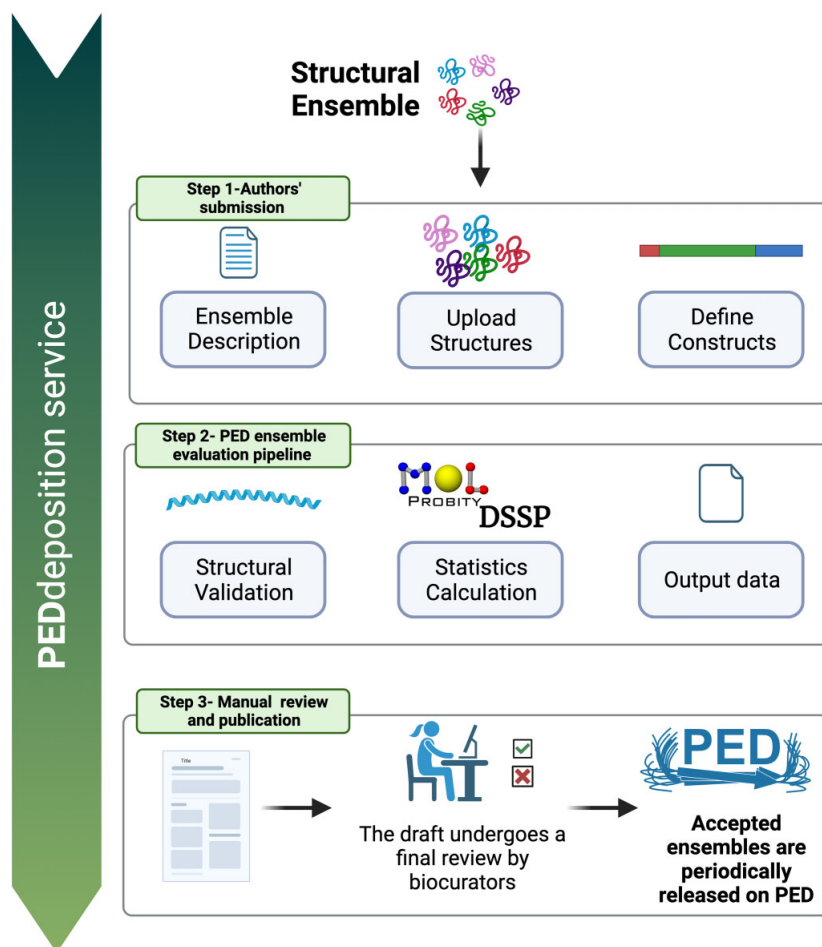


Figure 2. PEDdeposition service workflow. The workflow begins with the submission of an ensemble, which includes the description of both experimental and computational components, the deposition of conformers and the specification of the protein construct via UniProt accessions and/or protein sequence. The next step involves running the validation pipeline to evaluate the uploaded structures and generating insightful statistics through tools such as MolProbity, DSSP and calculating the radius of gyration. At the final stage, the submitted ensemble entry undergoes a final review by biocurators who determine whether to accept or reject it. Ultimately, approved ensembles are subsequently published on PED for public access.

Construct

Here, users define constructs corresponding to the deposited protein or region. Constructs are assembled from 'fragments' which can be defined using UniProt accession numbers, isoform identifiers and regions. For engineered constructs, manual input of the sequence is also an option. The feature viewer highlights deviations and modifications in the sequence, aiding in accurate definition.

Manual curation and validation

The final stage involves expert review and validation. The PED deposition service distinguishes between general depositors and expert biocurators. Biocurators have access to a dashboard where all deposited ensembles are organized based on their review status. Upon submission, deposited information undergoes thorough review and validation. If accepted, the ensemble draft is prepared for release in the PED database; if not,

depositors are promptly informed of the reasons for rejection. This automated process significantly reduces the time between ensemble deposition and publication, streamlining the entire workflow.

Implementation

The newly re-designed user interface (UI) provides a more enriched user experience and notable features. A prominent new addition to this UI version is a feature allowing users to access supplementary data from the ensemble's deposition phase. This represents a departure from the traditional report in PDF format, as users can now leverage a multitude of data assets available in CSV or JSON formats. This transition empowers users to have more flexibility to access and work with the ensemble data according to their particular needs. Furthermore, these data assets are also available on an im-

proved REST API for programmatic access. Constructed using the Django REST framework in Python, this REST API is meticulously documented according to the OpenAPI 3.0 standard. This documentation is presented using Swagger UI, enabling users to interact with the API on the website (<https://proteinsenemble.org/api>). The API allows users to selectively download specific data of interest, providing programmatic access to all PED data. PED also introduces a minor redesign of the browse page by grouping proteins for each entry, mitigating redundancy during the exploration of the database content. Another feature in this new version is the integration of PED in the 3D-Beacons (50), a network that provides programmatic access to macromolecular data from different data resources, therefore PED data is now also available through the 3D-Beacons API.

PED introduces a dedicated deposition interface that is accessible to all users. This new service enables depositors to measure the quality of their ensembles through an automated validation process utilizing calculations provided by tools such as MolProbity and DSSP. These calculations are facilitated by our distributed system, efficiently managing the computational workload through SLURM, enabling parallel calculation of multiple ensembles. The results are delivered in CSV and JSON formats, similar to the main user interface mentioned earlier. The service is integrated with the ORCID authentication service, utilizing the OpenID standard to verify user identity. This authentication allows users to track the status of their uploaded ensembles, which will undergo manual validation by curators before being published in the main database for public access. During this process, depositors must provide an accurate description of their ensembles and cross-reference them with other databases to enhance findability.

Conclusions and future work

Over the past 3 years, research on IDPs/IDRs has made significant progress, marked by the introduction of a diverse range of novel computational, experimental and ML-derived techniques for resolving structural ensembles. In alignment with the most recent advancements in this field, we remain devoted to customizing the database to the community's needs. Through a large community effort, the PED has experienced a substantial increase in its repertoire, with a noteworthy rise in the number of ensembles, entries and conformers. The repertoire has also expanded with an ever-growing range of different methods and their diverse combinations, recently enriched in EPR spectroscopy. Furthermore, we have integrated NMR-derived ensembles of IDPs/IDRs into PED and generated structural ensembles using advanced ML and sampling techniques without biasing them with experimental data.

Another key improvement in this release is the complete reimplementation and redesign of the PED deposition service. This tool has evolved beyond its previous capabilities, and now offers depositors a user-friendly, step-by-step workflow for retaining their structural ensembles. Furthermore, it includes a fully automated validation pipeline that comprehensively assesses the structure file format and generates insightful statistics. Additionally, the validation pipeline is now accessible as a standalone resource, enabling anyone interested to assess the quality of structural ensembles independently.

Further developments must be made in the future to address several key areas. To begin with, there is a need to smoothly

integrate coarse-grained (CG) ensembles into PED. Recently, a pair of IDP force fields have emerged that efficiently generate CG models of random coil-like IDPs/IDRs and capture the global characteristics of disordered proteins, such as the radius of gyration (23,51–53). The integration of such models into PED is long-awaited and has the potential to significantly expand the number of available ensembles for IDPs/IDRs.

The need for a conceptual categorization of entries within PED becomes increasingly important as we progress towards incorporating *ab-initio* and NMR ensembles, and in the near future, CG ensembles. Furthermore, by including predicted ensembles by diverse algorithms, we can facilitate the benchmarking and comparison of various ensemble generation methods for IDPs. Recently, the rapid growth of DisProt enabled the design of two community benchmark efforts, termed Critical Assessment of Protein Intrinsic Disorder prediction (CAID) challenge (website: <https://caid.idpcentral.org/challenge>) (54–56). We envision that the consistent growth of PED will also facilitate organizing a similar benchmark using withheld high-quality ensemble data and promote the development of predictors for IDP structural ensembles.

The long-term sustainability of PED is ensured by its central role in various initiatives involving large communities of bioinformaticians and structural biologists working in the disordered proteins field. Such communities include the 'ML4NGP' COST Action and the ELIXIR IDP Community, both of which foster collaboration and knowledge exchange among experts.

Data availability

The data that support the findings of this study are openly available in PED at <https://proteinsenemble.org/>.

Funding

European Union's Horizon 2020 research and innovation programme [778247 MSCA-RISE 'IDPfun' and 823886 MSCA-RISE 'REFRACT']; ML4NGP CA21160 project supported by COST (European Cooperation in Science and Technology) under Horizon Europe; H.G. and M.C.A. are funded by the European Union—NextGenerationEU through 'Italiadomani—PNRR' projects 'National Centre for HPC, Big Data and Quantum Computing' [codice identificativo MUR CN00000013, C93C22002800006]; 'National Center for Gene Therapy and Drugs based on RNA Technology' [codice fiscale 92315700283, codice identificativo CN00000041]; ELIXIR, the research infrastructure for life-science data, and by the European Union—NextGenerationEU through 'Italiadomani—PNRR project IR000010 'ELIXIR × NextGenerationIT: Consolidamento dell'Infrastruttura Italiana per i Dati Omici e la Bioinformatica—ElixirxNextGenIT' and project ECS00000017 'THE—Tuscany Health Ecosystem'. T.L. is holder of a postdoctoral innovation mandate [HBC.2022.0194] by the Flanders Innovation & Entrepreneurship Agency (VLAIO). Funding for open access charge: European Union's Horizon 2020 research and innovation programme [778247 MSCA-RISE 'IDPfun' and 823886 MSCA-RISE 'REFRACT'].

Conflict of interest statement

None declared.

References

1. Tompa,P. and Fersht,A. (2009) *Structure and Function of Intrinsically Disordered Proteins*. CRC Press.
2. Wang,H., Xiong,R. and Lai,L. (2016) Rational drug design targeting intrinsically disordered proteins. *WIREs Comput. Mol. Sci.*, **11**, 65–77.
3. Hatos,A., Monzon,A.M., Tosatto,S.C.E., Piovesan,D. and Fuxreiter,M. (2021) FuzDB: a new phase in understanding fuzzy interactions. *Nucleic Acids Res.*, **50**, D509–D517.
4. Bonomi,M., Heller,G.T., Camilloni,C. and Vendruscolo,M. (2017) Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.*, **42**, 106–116.
5. PDBe-KB consortium (2022) PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.*, **50**, D534–D542.
6. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
7. Varadi,M., Kosol,S., Lebrun,P., Valentini,E., Blackledge,M., Dunker,A.K., Felli,I.C., Forman-Kay,J.D., Kriwacki,R.W., Pierattelli,R., *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.
8. Sormanni,P., Piovesan,D., Heller,G.T., Bonomi,M., Kukic,P., Camilloni,C., Fuxreiter,M., Dosztanyi,Z., Pappu,R.V., Babu,M.M., *et al.* (2017) Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.*, **13**, 339–342.
9. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
10. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
11. Sala,D., Engelberger,F., Mchaourab,H.S. and Meiler,J. (2023) Modeling conformational states of proteins with AlphaFold. *Curr. Opin. Struct. Biol.*, **81**, 102645.
12. Del Alamo,D., Sala,D., Mchaourab,H.S. and Meiler,J. (2022) Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, **11**, e75751.
13. Stein,R.A. and Mchaourab,H.S. (2022) SPEACH_AF: sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput. Biol.*, **18**, e1010483.
14. Henriques,J., Craggell,C. and Skepö,M. (2015) Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theory Comput.*, **11**, 3420–3431.
15. Rauscher,S., Gapsys,V., Gajda,M.J., Zweckstetter,M., de Groot,B.L. and Grubmüller,H. (2015) Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.*, **11**, 5513–5524.
16. Abrams,C. and Bussi,G. (2014) Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, **16**, 163–199.
17. Lazar,T., Martínez-Pérez,E., Quaglia,F., Hatos,A., Chemes,L.B., Iserte,J.A., Méndez,N.A., Garrone,N.A., Saldaña,T.E., Marchetti,J., *et al.* (2021) PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, **49**, D404–D411.
18. Hatos,A., Hajdu-Soltész,B., Monzon,A.M., Palopoli,N., Álvarez,L., Aykac-Fas,B., Bassot,C., Benítez,G.I., Bevilacqua,M., Chasapi,A., *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.*, **48**, D269–D276.
19. Piovesan,D., Del Conte,A., Clementel,D., Monzon,A.M., Bevilacqua,M., Aspromonte,M.C., Iserte,J.A., Orti,F.E., Marino-Buslje,C. and Tosatto,S.C.E. (2023) MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res.*, **51**, D438–D444.
20. Fukuchi,S., Amemiya,T., Sakamoto,S., Nobe,Y., Hosoda,K., Kado,Y., Murakami,S.D., Koike,R., Hiroaki,H. and Ota,M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320–D325.
21. Lindorff-Larsen,K. and Kragelund,B.B. (2021) On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.*, **433**, 167196.
22. Janson,G., Valdes-Garcia,G., Heo,L. and Feig,M. (2023) Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.*, **14**, 774.
23. Tesei,G., Trolle,A.I., Jonsson,N., Betz,J., Pesce,F., Johansson,K.E. and Lindorff-Larsen,K. (2023) Conformational ensembles of the human intrinsically disordered proteome: bridging chain compaction with function and sequence conservation.
24. González-Delgado,J., Sagar,A., Zanon,C., Lindorff-Larsen,K., Bernadó,P., Neuvial,P. and Cortés,J. (2023) WASCO: a Wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. *J. Mol. Biol.*, **435**, 168053.
25. Lazar,T., Guharoy,M., Vranken,W., Rauscher,S., Wodak,S.J. and Tompa,P. (2020) Distance-based metrics for comparing conformational ensembles of intrinsically disordered proteins. *Biophys. J.*, **118**, 2952–2965.
26. Romero,P.R., Kobayashi,N., Wedell,J.R., Baskaran,K., Iwata,T., Yokochi,M., Maziuk,D., Yao,H., Fujiwara,T., Kurusu,G., *et al.* (2020) BioMagResBank (BMRB) as a Resource for Structural Biology. *Methods Mol. Biol. Clifton NJ*, **2112**, 187–218.
27. Kikhney,A.G., Borges,C.R., Molodenskiy,D.S., Jeffries,C.M. and Svergun,D.I. (2020) SASBDB: towards an automatically curated and validated repository for biological scattering data. *Protein Sci.*, **29**, 66–75.
28. Ramalli,S.G., Miles,A.J., Janes,R.W. and Wallace,B.A. (2022) The PCDDb (Protein Circular Dichroism Data Bank): a Bioinformatics Resource for Protein Characterisations and Methods Development. *J. Mol. Biol.*, **434**, 167441.
29. Felli,I.C. and Pierattelli,R. (2015) *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*. Springer, Cham.
30. Ritsch,I., Lehmann,E., Emmanouilidis,L., Yulikov,M., Allain,F. and Jeschke,G. (2022) Phase separation of heterogeneous nuclear ribonucleoprotein A1 upon specific RNA-binding observed by magnetic resonance. *Angew. Chem. Int. Ed. Engl.*, **61**, e202204311.
31. Galano-Frutos,J.J., Torreblanca,R., García-Cebollada,H. and Sancho,J. (2022) A look at the face of the molten globule: structural model of the *Helicobacter pylori* apoflavodoxin ensemble at acidic pH. *Protein Sci. Publ. Protein Soc.*, **31**, e4445.
32. Rao,J.N., Jao,C.C., Hegde,B.G., Langen,R. and Ulmer,T.S. (2010) A combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. *J. Am. Chem. Soc.*, **132**, 8657–8668.
33. Fisher,C.K., Huang,A. and Stultz,C.M. (2010) Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.*, **132**, 14919–14927.
34. Chan-Yao-Chong,M., Marsin,S., Quevillon-Cheruel,S., Durand,D. and Ha-Duong,T. (2020) Structural ensemble and biological activity of DciA intrinsically disordered region. *J. Struct. Biol.*, **212**, 107573.

35. Chen, J., Zaer, S., Drori, P., Zamel, J., Joron, K., Kalisman, N., Lerner, E. and Dokholyan, N.V. (2021) The structural heterogeneity of α -synuclein is governed by several distinct subpopulations with interconversion times slower than milliseconds. *Structure*, **29**, 1048–1064.
36. Bjarnason, S., McIvor, J.A.P., Prestel, A., Demény, K.S., Bullerjahn, J.T., Kragelund, B.B., Mercadante, D. and Heidarsson, P.O. (2023) DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2. bioRxiv doi: <https://doi.org/10.1101/2023.06.16.545083>, 16 June 2023, preprint: not peer reviewed.
37. Zhu, J., Salvatella, X. and Robustelli, P. (2022) Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states. *Nat. Commun.*, **13**, 6390.
38. Gomes, G.-N.W., Krzeminski, M., Namini, A., Martin, E.W., Mittag, T., Head-Gordon, T., Forman-Kay, J.D. and Gradinaru, C.C. (2020) Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.*, **142**, 15697–15710.
39. Piovesan, D. and Tosatto, S.C.E. (2018) Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinforma. Oxf. Engl.*, **34**, 122–123.
40. Piovesan, D., Monzon, A.M. and Tosatto, S.C.E. (2022) Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci. Publ. Protein Soc.*, **31**, e4466.
41. Necci, M., Piovesan, D., Clementel, D., Dosztányi, Z. and Tosatto, S.C.E. (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinforma. Oxf. Engl.*, **36**, 5533–5534.
42. Zheng, L.-E., Barethiya, S., Nordquist, E. and Chen, J. (2023) Machine learning generation of dynamic protein conformational ensembles. *Mol. Basel Switz.*, **28**, 4047.
43. Teixeira, J.M.C., Liu, Z.H., Namini, A., Li, J., Vernon, R.M., Krzeminski, M., Shamandy, A.A., Zhang, O., Haghighatdari, M., Yu, L., et al. (2022) IDPConformerGenerator: a flexible software suite for sampling the conformational space of disordered protein states. *J. Phys. Chem. A*, **126**, 5985–6003.
44. Liu, Z.H., Teixeira, J.M.C., Zhang, O., Tsangaris, T.E., Li, J., Gradinaru, C.C., Head-Gordon, T. and Forman-Kay, J.D. (2023) Local disordered region sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments. bioRxiv doi: <https://doi.org/10.1101/2023.07.25.550520>, 27 July 2023, preprint: not peer reviewed.
45. Valdes-Garcia, G., Heo, L., Lapidus, L.J. and Feig, M. (2023) Modeling concentration-dependent phase separation processes involving peptides and RNA via residue-based coarse-graining. *J. Chem. Theory Comput.*, **19**, 669–678.
46. Heo, L. and Feig, M. (2023) One particle per residue is sufficient to describe all-atom protein structures. bioRxiv doi: <https://doi.org/10.1101/2023.05.22.541652>, 23 May 2023, preprint: not peer reviewed.
47. del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., Perfetto, L., How, K., Ratan, P., Shirodkar, G., et al. (2021) The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.*, **50**, D648–D653.
48. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
49. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018) MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.*, **27**, 293–315.
50. Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S., Bienert, S., Borges, C., Deshpande, M., Green, T., Hassabis, D., et al. (2022) 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. *GigaScience*, **11**, giac118.
51. Klein, F., Barrera, E.E. and Pantano, S. (2021) Assessing SIRAH's capability to simulate intrinsically disordered proteins and peptides. *J. Chem. Theory Comput.*, **17**, 599–604.
52. Thomasen, F.E., Pesce, F., Roesgaard, M.A., Tesi, G. and Lindorff-Larsen, K. (2022) Improving Martini 3 for disordered and multidomain proteins. *J. Chem. Theory Comput.*, **18**, 2033–2041.
53. Fagerberg, E. and Skepö, M. (2023) Comparative performance of computer simulation models of intrinsically disordered proteins at different levels of coarse-graining. *J. Chem. Inf. Model.*, **63**, 4079–4087.
54. Necci, M., Piovesan, D., Predictors, C.A.I.D., Curators, D.P. and Tosatto, S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*, **18**, 472–481.
55. Conte, A.D., Mehdiabadi, M., Bouhraoua, A., Miguel Monzon, A., Tosatto, S.C.E. and Piovesan, D. (2023) Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins Struct. Funct. Bioinforma.*, <https://doi.org/10.1002/prot.26582>.
56. Del Conte, A., Bouhraoua, A., Mehdiabadi, M., Clementel, D., Monzon, A.M. and CAID predictors CAID predictors, Tosatto, S.C.E. and Piovesan, D. (2023) CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. *Nucleic Acids Res.*, **51**, W62–W69.

Appendix

PED Consortium

Maria C. Aspromonte¹, Pau Bernadó⁶, Belén Chaves-Arquero⁷, Lucia Beatriz Chemes⁸, Damiano Clementel¹, Tiago N. Cordeiro⁹, Carlos A. Elena-Real⁶, Michael Feig¹⁰, Isabella C. Felli¹¹, Carlo Ferrarini⁵, Julie D. Forman-Kay^{12,13}, Tiago Gomes^{9,26}, Frank Gondelaud¹⁴, Claudiu C. Gradinaru^{15,16}, Tâp Ha-Duong¹⁷, Teresa Head-Gordon^{18,19,20,21}, Pétur O. Heidarsson²², Giacomo Janson¹⁰, Gunnar Jeschke²³, Emanuela Leonardi¹, Zi Hao Liu^{12,13}, Sonia Longhi¹⁴, Xamuel L. Lund^{6,24}, Maria J. Macias^{25,26}, Pau Martin-Malpartida²⁶, Davide Mercadante²⁷, Assia Mouhand⁶, Gabor Nagy²⁸, María Victoria Nugnes¹, José Manuel Pérez-Cañadillas²⁹, Giulia Pesce¹⁴, Roberta Pierattelli¹¹, Damiano Piovesan¹, Federica Quaglia^{1,30}, Sylvie Ricard-Blum³¹, Paul Robustelli³², Amin Sagar⁶, Edoardo Salladini³³, Lucile Sénicourt⁶, Nathalie Sibille⁶, João M. C. Teixeira¹², Thomas E. Tsangaris^{15,16}, Mihaly Varadi³⁴

⁶ Centre de Biologie Structurale (CBS), Université de Montpellier, INSERM, CNRS

⁷ Centro de Investigaciones Biológicas Margarita Salas (CIB), CSIC, 28040 Madrid, Spain

⁸ Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín (UNSAM) – Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), San Martín, Argentina

⁹ Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal

¹⁰ Department of Biochemistry and Molecular Biology, Michigan State University, USA

¹¹ Department of Chemistry ‘Ugo Schiff’ and Magnetic Resonance Center (CERM), University of Florence, Florence, Italy

¹² Molecular Medicine Program, The Hospital for Sick Children, Toronto, Ontario Canada

¹³ Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada

¹⁴ Lab. Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, Aix Marseille University and

Centre National de la Recherche Scientifique (CNRS), 163 Avenue de Luminy, Case 932, 13288, Marseille, FRANCE

¹⁵ Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

¹⁶ Department of Chemical & Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada

¹⁷ BioCIS, CNRS, Université Paris-Saclay, France

¹⁸ Kenneth S. Pitzer Center for Theoretical Chemistry, University of California, Berkeley, California, USA

¹⁹ Department of Chemistry, University of California, Berkeley, California, USA

²⁰ Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California, USA

²¹ Department of Bioengineering, University of California, Berkeley, California, USA

²² Department of Biochemistry, Science Institute, University of Iceland, Reykjavík, Iceland

²³ Department of Chemistry and Applied Biosciences, ETH Zürich, Zürich, Switzerland

²⁴ Institut Laue-Langevin, 71 avenue de Martyrs, Grenoble 38042, France

²⁵ Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain

²⁶ Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona 08028, Spain

²⁷ School of Chemical Sciences, The University of Auckland, Auckland, New Zealand

²⁸ Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, D-37077 Göttingen, Germany

²⁹ Instituto de Química Física 'Blas Cabrera', Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

³⁰ Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy

³¹ University Lyon 1, ICBMS, UMR 5246 CNRS, Villeurbanne, France

³² Department of Chemistry, Dartmouth College, New Hampshire, USA

³³ Department of Drug Science and Technology, Università degli Studi di Torino, Torino, Italy

³⁴ Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome

Received: September 15, 2023. Revised: October 10, 2023. Editorial Decision: October 11, 2023. Accepted: October 13, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Software

DRMAAatic: dramatically improve your cluster potential

Alessio Del Conte¹, Hamidreza Ghafouri¹, Damiano Clementel¹, Ivan Mičetić¹,
Damiano Piovesan¹ , Silvio C.E. Tosatto^{1,2,*} , Alexander Miguel Monzon^{1,*} 

¹Department of Biomedical Sciences, University of Padova, Padova 35131, Italy

²Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari 70126, Italy

*Corresponding authors. Silvio C.E. Tosatto, Department of Biomedical Sciences, University of Padova, Viale G. Colombo, 3, Padova 35131, Italy.

E-mail: silvio.tosatto@unipd.it; Alexander Miguel Monzon, Department of Biomedical Sciences, University of Padova, Viale G. Colombo, 3, Padova 35131, Italy.

E-mail: alexander.monzon@unipd.it.

Associate Editor: Toma Tebaldi

Abstract

Motivation: The accessibility and usability of high-performance computing (HPC) resources remain significant challenges in bioinformatics, particularly for researchers lacking extensive technical expertise. While Distributed Resource Managers (DRMs) optimize resource utilization, the complexities of interfacing with these systems often hinder broader adoption. DRMAAatic addresses these challenges by integrating the Distributed Resource Management Application API (DRMAA) with a user-friendly RESTful interface, simplifying job management across diverse HPC environments. This framework empowers researchers to submit, monitor, and retrieve computational jobs securely and efficiently, without requiring deep knowledge of underlying cluster configurations.

Results: We present DRMAAatic, a flexible and scalable tool that bridges the gap between web interfaces and HPC infrastructures. Built on the Django REST Framework, DRMAAatic supports seamless job submission and management via HTTP calls. Its modular architecture enables integration with any DRM supporting DRMAA APIs and offers robust features such as role-based access control, throttling mechanisms, and dependency management. Successful applications of DRMAAatic include the RING web server for protein structure analysis, the CAID Prediction Portal for disorder and binding predictions, and the Protein Ensemble Database deposition server. These deployments demonstrate DRMAAatic's potential to enhance computational workflows, improve resource efficiency, and facilitate open science in life sciences.

Availability and implementation: <https://github.com/BioComputingUP/DRMAAatic>, <https://drmaatic.biocomputingup.it/>.

1 Introduction

In contemporary scientific research, the dissemination of computational tools and algorithms through intuitive web interfaces has become essential for engaging a broad spectrum of users. In life sciences, particularly in bioinformatics and computational biology, web interfaces not only enhance accessibility but also facilitate reproducibility, enabling researchers worldwide to utilize advanced computational methods without needing extensive technical expertise or dedicated local infrastructure. The web server issue of the *Nucleic Acids Research* journal underscores the growing significance of web-based computational solutions for the biological community. For example, 89 articles were published in 2021 (Rigden and Fernández 2021), 102 in 2022 (Rigden and Fernández 2022), 81 in 2023 (Rigden and Fernández 2023), and 74 in 2024 (Rigden and Fernández 2024), highlighting the continued demand for streamlined job creation and execution in web-based environments. Moreover, the Galaxy Project (The Galaxy Community 2024) is an open-source, web-based platform designed to make computational biology accessible to researchers without programming expertise. It provides a user-friendly interface for constructing and executing complex bioinformatics workflows,

emphasizing accessibility, reproducibility, and transparency in data analysis.

However, facilitating secure and efficient access for external applications to internal high-performance computing (HPC) clusters presents significant challenges. Traditional HTTP protocols are not inherently designed to manage the demands of long-running computational tasks that require substantial resources and often necessitate distribution across multiple nodes within a cluster. This limitation arises because HTTP operates on a request–response model, which is typically synchronous and does not accommodate prolonged operations well. For example, consider a scenario where a user submits a request to perform a resource-intensive data analysis or machine learning task. In such cases, implementing a solution often involves introducing mechanisms to handle the asynchronous nature of these tasks.

On the server side, the application must manage the job, spawning a new process, and monitor its progress and completion, all without waiting or blocking the main application process. This could involve checking if the results are available on the file system, querying a database or intra-process communication. These additional layers of logic introduce challenges related to state management, consistency, and

Received: January 17, 2025; Revised: April 16, 2025; Editorial Decision: May 5, 2025; Accepted: May 14, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

fault tolerance, which traditional HTTP protocols do not natively address. This gap is compounded by existing web server solutions, which are typically tailored to specific use cases and lack the flexibility for general-purpose, customizable job execution.

Distributed Resource Managers (DRMs) are key in managing computational tasks across networked nodes. These systems optimize resource utilization through sophisticated scheduling and management policies, distributing tasks to maximize throughput and efficiency. The Distributed Resource Management Application API (DRMAA) (Troger *et al.* 2007) (<https://drmaa.org/>) complements these systems by offering a standardized interface for interacting with DRMs, abstracting their complexities to facilitate seamless job submission, control, and monitoring. Despite its potential, the adoption of DRMAA has been limited, with only a subset of DRMs fully implementing the latest standards. This disconnect hinders the development of interoperable tools that can bridge the gap between web-based interfaces and distributed computing environments.

Nowadays, the development of modern Web APIs is predominantly centered around the Representational State Transfer (REST) architecture (Fielding and Taylor 2000). A RESTful API is an implementation of an API that strictly adheres to the principles and constraints of REST architecture, following a standardized design pattern that defines a consistent and stateless set of operations. This architecture is particularly well-suited for web-based services, offering key advantages such as enhanced performance, scalability, and flexibility. RESTful Web APIs facilitate interoperability between systems and applications, making them a cornerstone of web software development (Cruz *et al.* 2020).

Several frameworks have been developed to expose HPC resources through RESTful APIs, notably NEWT (Cholia and Sun 2015), FirecREST (Cruz *et al.* 2020), and SCEAPI (Rongqiang *et al.* 2017). NEWT provides a web API for remote job submission, monitoring, and file system access on HPC systems. FirecREST takes a microservice-based approach to HPC middleware, handling job management and data transfer with tight integration into existing supercomputing infrastructure. SCEAPI offers a unified RESTful interface for job execution, file transfer, and authentication across multiple supercomputing centers. While these solutions significantly simplify HPC interactions, they are often tied to specific configurations and require users to understand cluster details, which can hinder accessibility for non-experts (see Table 1 for a more detailed comparison). In this work, we introduce DRMAAtic that fits into this ecosystem as a more flexible, user-centric alternative. It leverages the standardized DRMAA interface within a RESTful Django-based framework (DRF) to abstract away cluster-specific intricacies, enabling users to submit predefined computational tasks, monitor progress, and retrieve results via simple HTTP calls. By adopting a job-centric approach (with task templates and built-in result handling) and supporting robust authentication, DRMAAtic extends HPC capabilities to web applications and services with minimal user-side complexity, effectively bridging the gap between complex HPC environments and user-friendly web interfaces. Here, we present the architecture and functionalities of DRMAAtic, highlighting its potential as a key tool for organizations, research groups, and individual users seeking to enhance the accessibility and usability of computational resources.

2 Implementation

The core of the DRMAAtic architecture is the execution of a job in a DRM. Each job must be defined as an executable task with pre-configured parameters specifying how and where the task will be executed. These parameters also include a set of user-defined options provided at execution time. Users initiate job execution by selecting a task, providing necessary inputs, and submitting the job. The job is subsequently created and forwarded to the DRM for execution. A universally unique identifier (UUID) is generated for the job, allowing users to track its state and interact with it as needed.

2.1 Data modeling and management

DRMAAtic leverages the Django REST Framework (DRF) and its Object-Relational Mapping (ORM) capabilities to model the underlying database structure. The database schema (Fig. 1, available as supplementary data at *Bioinformatics Advances* online) centers around the job entity, which includes critical attributes like a UUID, description, and creation timestamp. Jobs are interconnected with auxiliary entities, such as tasks to be executed, through relational mappings. Additionally, jobs may inherit execution directories by referencing parent jobs, and the “job_dependencies” table enables the specification of complex inter-job relationships essential for constructing dependency-driven workflows.

Tasks are defined by attributes such as a unique identifier, execution command, queue assignment, CPU, and memory requirements. Each task can include parameters with a variety of types (e.g. Boolean, integer, float, string, or file) and features such as default values, administrator-only visibility, and required status. User groups further refine execution permissions, facilitating resource allocation via throttling rates and execution tokens.

2.2 API endpoints and user interaction

API endpoints in DRMAAtic are implemented using DRF views, enabling easy interactions through HTTP requests. These endpoints cover the full spectrum of job lifecycle management, including submission, monitoring, and retrieval of results (Table 2). For instance, a user submits a job by issuing a POST request to the `/job` endpoint. This triggers server-side validation of input parameters, job creation, and DRM submission, with a structured response containing job metadata. Users may also leverage endpoints for querying job states, downloading outputs, and managing dependencies. An example (Fig. 2, available as supplementary data at *Bioinformatics Advances* online): a user submitting a job can specify a list of parameters in the request body, along with the task name to be executed and optional fields such as `parent_job`, `dependencies`, and `dependency_type`.

Detailed API documentation, enriched with Swagger-based descriptions, is available at URL: <https://drmaatic.biocomputingup.it> to facilitate user adoption.

2.3 Authentication and authorization

DRMAAtic employs external authentication services, such as ORCID (URL: <https://orcid.org/>), for identity verification, while handling resource authorization internally. Authentication is facilitated via OAuth 2.0 implicit flow, where access tokens are converted into stateless JSON Web Tokens (JWTs). Upon receipt of a valid JWT in an HTTP header, DRMAAtic authenticates the user, retrieves metadata from the authentication

Table 1. Comparative summary of DRMAAatic and other RESTful HPC API solutions.^a

Framework	Ease of integration	Task control model	File handling	User management	Extensibility	Deployment complexity
DRMAAatic	(***)—standalone REST API (Django) easily integrates into web apps; minimal HPC knowledge needed for end-users.	Predefined tasks, simplifying submission but not permitting arbitrary jobs.	Built-in file staging and result retrieval via job context (outputs downloadable through API).	External auth support (e.g. OAuth/ORCID) with role-based access; users need not have local HPC accounts.	Designed for extension: new task types and parameters can be added; portable to any DRM via DRMAA.	(**)—runs as a single web service (Docker deployment available) on top of the cluster's scheduler (requires DRMAA library).
NEWT	(**)—provides REST endpoints; integration requires users to have HPC accounts and knowledge of the target system.	General job submission (no preset tasks; users submit scripts or commands), offering flexibility with more user responsibility.	Direct file system access and transfer endpoints for upload/download and directory listing.	Tied to HPC center authentication (user accounts or API keys); uses existing user credentials for job execution.	Extensible via plugin framework to support new commands or resources.	(**)—a monolithic service that must be set up within the HPC environment; less complex than microservices but site-specific.
FirecREST	Variable-rich API for HPC operations, but primarily for portal developers; requires token-based access and HPC-center setup.	General job submission and system operations through microservices (no built-in task templates).	Dedicated microservices for data transfer (optimized for large files and HPC storage systems).	Integrates with HPC site identity management (tokens/OAuth with HPC accounts); enforces per-user permissions.	Loosely coupled microservices allow adding or upgrading components (e.g. new services) at the cost of greater system complexity.	(***)—complex microservice architecture (API gateway + multiple services); intended for deployment by large HPC centers.
SCEAPI	(**)—offers a unified API across sites; integration is easier for users once deployed, but setting up across multiple HPC centers is involved.	General job submission across federated resources (standard CRUD for jobs; no specific task templates).	File transfer included (endpoints for moving data to/from each site's storage).	Centralized SSO and OAuth2 for multiple clusters; maps users to site accounts, ensuring security across centers.	Extensible to additional HPC centers and services within its federation model; not focused on user-defined task abstraction.	(***)—requires SCE middleware and coordination between HPC sites; suitable for national or multi-institutional HPC infrastructure.

^a Asterisks indicate relative levels among frameworks as follows: (*) = Low, (**) = Moderate, and (***) = High.

provider, and ensures secure interaction (Fig. 3, available as supplementary data at *Bioinformatics Advances* online).

Sensitive resources and endpoints are protected using role-based access control, ensuring that only authorized users can access specific functionalities. This dual-layered approach enhances security while maintaining a robust user experience.

2.4 Throttling and resource management

DRMAAatic employs a dual-throttling strategy to ensure efficient and fair use of HPC resources. API request throttling limits the rate of user interactions to prevent unnecessary strain on the system. More critically, job submission is governed by a CPU token-based mechanism. Each user is assigned a fixed-size CPU token bucket (e.g. 100 tokens), which is replenished at a defined rate (e.g. 10 tokens per hour). Jobs can be configured to consume tokens on submission proportional to their CPU needs, for example, a task requiring 4 CPUs will consume 4 tokens. If a user lacks sufficient tokens, the job submission is deferred with a message indicating the wait time until enough tokens are available (Fig. 4, available as supplementary data at

Bioinformatics Advances online). The system identifies users either through the authentication mechanism or, if they are not authenticated, by their IP address.

In high-demand scenarios, this mechanism effectively smooths submission spikes, preventing overloads on the DRM. The throttling system guarantees that users cannot monopolize cluster resources, maintaining responsiveness for all users. Internal tests showed that even under concurrent multi-user submissions (e.g. 50+ jobs across 5 users), DRMAAatic maintained consistent job handling without system crashes or degraded service. When the cluster reaches full load, new job requests are queued by the DRM, while DRMAAatic provides informative status updates without blocking the interface.

2.5 Administration and configuration interface

DRMAAatic provides an administration web interface accessible via a dedicated “/admin” route (Fig. 5, available as supplementary data at *Bioinformatics Advances* online). Administrators can configure execution tasks, queues, and

Table 2. DRMAAAtic API endpoints.^a

Method	Route	Description
GET	/[provider]/token/	Allows users to obtain JWT tokens for authenticating with DRMAAAtic, using access tokens from external providers.
GET	/cpu-credit/	Provides the amount of available CPU credits for the user.
POST	/job/	Enables users to create new jobs, defining job parameters based on specific tasks.
GET	/job/	Allows users to list and query jobs, including filtering by state, job name, and other criteria.
GET	/job/{uuid}/	Retrieves all the metadata of the job using its UUID, updating its state if it is not in a final state.
DELETE	/job/{uuid}/	Permits users to delete specific jobs, stopping them in DRM while preserving the execution directory and the metadata.
GET	/job/{uuid}/download/	Allows users to download a zip file containing all the output files of a job.
GET	/job/{uuid}/file/	Lists all the output files of a job, including their paths.
GET	/job/{uuid}/file/{path}	Enables users to download a specific output file of a job, given its path in the output directory.
GET	/job/{uuid}/status/	Provides the state of a job, updating it if necessary.
PUT	/job/{uuid}/stop/	Stops a job that is executing in the DRM, returns an error if the job was already executed.
GET	/task/	Lists all available tasks in the DRMAAAtic system, including their parameters and descriptions.
GET	/task/{name}/	Retrieve a specific task.

^a The first column specifies the HTTP method allowed for the endpoint. The second column, the route, provides the address of the endpoint where the client should send the request (the root indicates the targeted resource). Lastly, the description column explains the purpose and functionality of the endpoint.

user groups using this interface. Through the web-based administration interface, an administrator can define new tasks by specifying their name, execution command, associated computational requirements (e.g. memory, CPUs, queue), and any input parameters required at runtime. Parameters can be customized with type validation (e.g. integer, float, file) and can include optional flags to modify the execution behavior. The interface also enables monitoring of submitted jobs, parameter configurations, and debugging processes (Fig. 6, available as supplementary data at *Bioinformatics Advances* online).

Additionally, administrators can manage system performance and security by configuring throttling parameters, execution tokens, and the longevity of access tokens. This low-code configuration process enables seamless integration of existing scripts or tools with minimal technical overhead, making DRMAAAtic accessible even for non-expert system administrators.

3 Results

In this section, the general job execution workflow is described, along with the integration of DRMAAAtic with the SLURM workload manager. Finally, three application examples utilizing DRMAAAtic as a core component for job execution are presented and discussed.

3.1 Job execution workflow

The job lifecycle in DRMAAAtic encompasses multiple states, ensuring systematic management from submission to completion (Fig. 1). Each state is uniquely defined and linked to its management context, either by the web server or the underlying DRM. Upon job receipt, the server validates input parameters to prevent computational errors and resource misuse. Jobs failing to meet predefined constraints, such as incorrect configurations or insufficient CPU credits, are rejected to protect cluster resources and maintain availability.

Validated jobs are forwarded to the DRM, where the scheduler assigns them to the cluster queue and marks them as “queued active.” Jobs transition to the “running” state

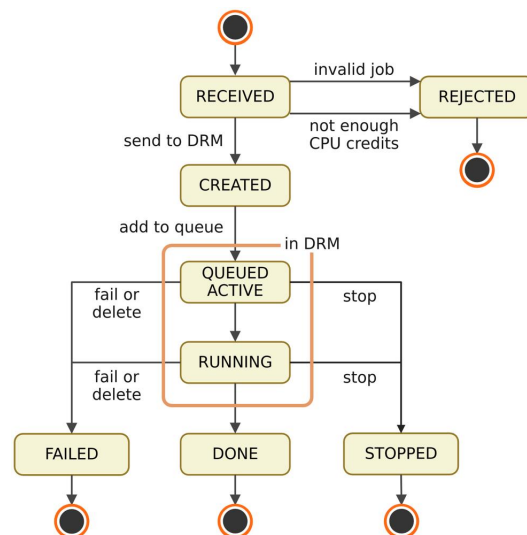


Figure 1. Execution flow of a job with its relative states. A job can have different states in its lifetime, depending on the user input or the result of its execution.

when execution begins. On successful completion (exit code zero), jobs are marked as “done.” Errors or disruptions during execution result in a “failed” state, while users can issue stop commands to halt active jobs, transitioning them to the “stopped” state.

DRMAAAtic dynamically tracks job states through direct queries to the DRM during user requests, avoiding resource-intensive polling. The server updates job states in its database and communicates them to users in real time. Final states (done, failed, stopped, rejected) are preserved in the DRMAAAtic database without further DRM updates, minimizing unnecessary system interactions. This structured workflow ensures efficient job management and optimal

resource utilization while maintaining efficient user interaction with the DRM.

3.2 DRMAAAtic and SLURM integration

As previously highlighted, DRMAAAtic is designed to integrate smoothly with any DRM system that supports or exposes DRMAA APIs (Fig. 2C and D). For this study, our focus was on integrating DRMAAAtic with SLURM, a widely adopted workload manager in HPC environments. DRMAAAtic's Python-based implementation utilizes the DRMAA-Python library (available at <https://github.com/pygridtools/drmaa-python>) to access DRMAA bindings directly, enabling efficient job submission and management.

To ensure cross-compatibility, we developed a job structure that abstracts job configurations from DRM-specific implementations. This abstraction allows DRMAAAtic to incorporate additional SLURM-specific parameters, referred to as “native specifications,” without deviating from the DRMAA v1 standard. Key native specifications included in DRMAAAtic are CPU-per-task allocation, minimum RAM requirements, and partition or queue designations. These enhancements ensure robust job configuration while maintaining the flexibility to integrate other DRMs in the future. Notably, the nature and extent of native specifications vary across DRMs, determined by their respective implementations of DRMAA bindings.

3.2.1 Deploying DRMAAAtic with Docker

Deploying DRMAAAtic in a production environment requires proper configuration of the cluster and its dependencies. For simplified local testing, a Docker Compose configuration was developed to provide all necessary services:

- MariaDB: The primary relational database management system utilized by both SLURM and DRMAAAtic.
- phpMyAdmin: A web-based interface that simplifies MariaDB administration.

- slurmdbd: SLURM's database daemon, responsible for job accounting and tracking.
- slurmctld: The central management component in SLURM, overseeing job scheduling, resource allocation, and cluster monitoring.
- Worker Nodes (c1 and c2): Nodes configured to execute computational tasks within the cluster.
- DRMAAAtic: Serves as the interface to SLURM via the DRMAA API, using MariaDB for data storage.

During the initial execution of the Docker Compose setup, a preconfigured test database is deployed, which includes configurations for a sample task and an administrator account. This enables immediate functionality testing of DRMAAAtic. For production environments, DRMAAAtic can be deployed as a Docker container interfacing with an existing SLURM control daemon. In such cases, careful configuration of SLURM is essential to ensure robust communication with the control machine. Additionally, deploying DRMAAAtic in production requires a HTTP server (e.g. Apache) paired with a Web Server Gateway Interface (WSGI) (Fig. 2C). This setup provides the stability, scalability, and security required for production, including features such as SSL/TLS termination and advanced access control.

File and Directory Management: Docker-based deployments of DRMAAAtic require proper handling of job-specific files and directories within mounted volumes. Ensuring that the user executing jobs has appropriate write permissions is critical. The internal user in the Docker container must correspond to an external user to prevent permission conflicts, particularly when Docker user remapping is enabled. Addressing this challenge involves creating a dedicated user whose UID is synchronized with external users across all worker nodes.

3.3 Application examples

DRMAAAtic supports client-side mashup integration, where REST API calls originate from the web interface front-end (e.g.

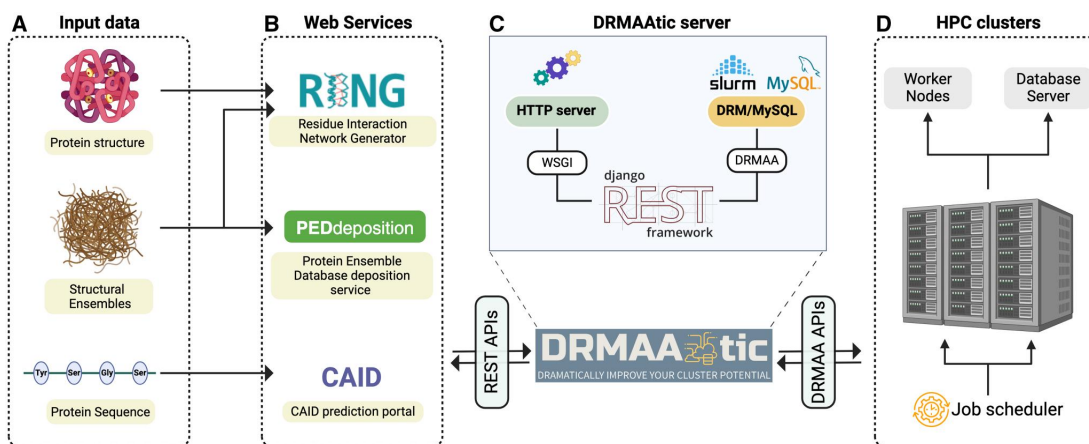


Figure 2. DRMAAAtic workflow and architecture. (A) Input data from user-facing web interfaces. Examples of web services utilizing these inputs are RING, which accepts protein structures and conformational ensembles, PEDdeposition for structural ensemble submission, and CAID for sequence-based predictions. (B) Web services act as clients making REST API calls to DRMAAAtic (from front-end, not back-end), facilitating input submission and delivering results. (C) DRMAAAtic translates requests into job submissions to the HPC cluster, leveraging the functionality of DRMAA with RESTful APIs, implemented through the Django REST Framework. (D) Job execution and tracking occur on worker nodes, with metadata stored in the backend database.

JavaScript or React apps) rather than backend middleware. This architecture enables seamless integration into web portals, aligning with modern Web 2.0 principles.

3.3.1 RING web-server

The first operative web application interacting with the DRMAAtic web server is the Residue Interaction Network Generator (RING) (URL: <https://ring.biocomputingup.it/>). RING web server allows users to execute the RING software (Del Conte *et al.* 2024) remotely on a cluster, eliminating the need for local installation. Moreover, the server presents results interactively, enabling users to analyze Residue Interaction Networks (RINs), representing the interactions between residues within a protein structure.

The UI allows users to upload a protein structure or structural ensemble for analysis or retrieve it from Protein Data Bank (Berman *et al.* 2003) and AlphaFoldDB (Varadi *et al.* 2024). Additionally, users can define various parameters for the RING software, customizing the analysis and setting thresholds for identifying interactions. Upon job submission, the user-defined parameters are packaged and sent to DRMAAtic for validation. Once validated in the back-end, a RING task is executed within the DRM using the specified inputs and parameters. The server then identifies the job with a UUID, enabling users to monitor its status and retrieve output files upon completion.

Upon successful execution, the outputs are downloaded and processed for integration into interactive components of the UI such as graphs, 3D protein structure representations, and contact heatmaps. DRMAAtic's architecture ensures efficient user and job management, cluster execution of RING tasks, and easy retrieval of prior jobs and their outputs.

3.3.2 CAID Prediction Portal

The CAID Prediction Portal (Del Conte *et al.* 2023; <https://caid.idpcentral.org/portal>) exemplifies the advanced use of DRMAAtic for bioinformatics workflows. This portal accepts a protein sequence, job description, and optional email address as inputs, running disorder and binding prediction algorithms to identify disordered regions and binding sites. The portal integrates 40 distinct predictors from the literature, each executable against the query sequence. To enhance user experience, completed predictions are made available for download or visualization on the web page, managing different execution times across predictors.

The implementation strategy leverages DRMAAtic to manage dependencies systematically. A “main job” validates and organizes input data within the execution directory, spawning individual prediction tasks as dependent jobs. This approach ensures predictions are executed sequentially, with dependencies handled via DRMAAtic's “afterok” mechanism for predictors requiring supplementary inputs. Each predictor generates outputs in its designated directory, populating the main job's execution directory with comprehensive results.

The frontend employs UUIDs to track jobs, enabling an easy retrieval of related tasks. A polling mechanism monitors job states, automatically downloading results upon completion. If an email address is provided, a notification job is triggered to alert the user when all tasks have concluded, even in cases of partial failure, utilizing DRMAAtic's “afterany” dependency setting.

This systematic and modular approach underscores the flexibility of DRMAAtic, ensuring efficient resource utilization and

user-centric results presentation for complex bioinformatics tasks.

3.3.3 Deposition web-server of Protein Ensemble Database

The Protein Ensemble Database (PED) (Ghafouri *et al.* 2024) serves as the main repository for structural ensembles of intrinsically disordered proteins (IDPs). Ensemble deposition is facilitated through a dedicated web server (URL: <https://deposition.proteinensemble.org/>), allowing users to submit structural ensemble data and execute the ensemble validation pipeline.

The user interface (UI) supports input of ensemble meta-data, uploading structural ensemble files, and, when available, the weights of individual conformations. The backend efficiently manages this data and integrates with DRMAAtic via its API, enabling automated job submissions for ensemble validation and statistical analysis. Tools such as MolProbity (Williams *et al.* 2018), DSSP (Kabsch and Sander 1983), and relative solvent accessibility (RSA) calculations, among other global parameters derived from structural data, are used to process and evaluate the ensembles.

The backend facilitates interaction with DRMAAtic and updates the deposition database with job statuses in real time. Upon completion, result files are securely stored in protected storage, ensuring long-term preservation even after the associated DRMAAtic jobs are deleted. The integration between the backend and DRMAAtic improves the deposition process, keeping calculations and data processing nearly invisible to users until results are finalized and made available through the UI.

4 Discussion

The expanding role of web applications in life sciences, particularly in bioinformatics, highlights the importance of tools that simplify access to complex computational infrastructures. One of the main challenges in this domain is the heterogeneity of cluster architectures and resource managers, which makes web-based integration with HPC systems harder.

DRMAAtic addresses this challenge through its job-centric design, which emphasizes predefined task templates, simplified job submission, secure authentication, and intuitive result retrieval. Built on the DRF, DRMAAtic offers a modular, scalable platform with native support for the DRMAA standard, enabling compatibility with multiple DRM such as SLURM and PBS. Its architecture supports external authentication (e.g. ORCID) and role-based access control, making it suitable for public-facing services that do not rely on local Unix user accounts. Throttling and token-based resource allocation mechanisms further ensure fair usage in multi-user environments.

When compared to existing HPC REST APIs such as NEWT, FirecREST, and SCEAPI, DRMAAtic offers distinct advantages and trade-offs. It is ideal for web portals or science gateways where users interact with predefined workflows, requiring minimal technical knowledge. In contrast, NEWT and FirecREST support arbitrary job submissions and broader system access but are more complex to use and deploy, particularly for non-experts. DRMAAtic also simplifies integration into web mashups, thanks to its lightweight, self-contained architecture. While FirecREST employs a microservice model best suited for large HPC centers, and SCEAPI supports cross-institutional federation, DRMAAtic caters to individual labs or institutes needing a focused and easy-to-deploy solution.

From a security and usability standpoint, DRMAAAtic allows users to submit and monitor jobs without requiring local accounts, while still maintaining control over access policies. File management is tightly linked to each job, allowing users to retrieve results without needing to navigate complex storage systems. While this limits flexibility in some advanced data workflows, it greatly enhances usability for predefined pipelines.

The versatility of DRMAAAtic is demonstrated through its integration into real-world applications such as the RING web server, the CAID Prediction Portal, and the PED deposition system. These deployments showcase its ability to manage complex and interdependent workflows efficiently, all while providing an accessible user experience. DRMAAAtic's integration with SLURM, combined with its Docker-based deployment option, further emphasizes its adaptability and ease of maintenance.

In summary, DRMAAAtic is most suitable when ease of integration, controlled job execution, and broad accessibility are top priorities, for instance, in biomedical or life sciences research portals that require HPC-backed computation through user-friendly web interfaces. It may be less appropriate in scenarios demanding unrestricted HPC access, deep system integration, or federation across HPC centers, where more comprehensive solutions like FirecREST, NEWT, or SCEAPI may be more aligned with user needs. DRMAAAtic's capacity to foster collaboration, improve accessibility, and optimize resource efficiency aligns with the overarching goals of open science and reproducibility in life sciences. As the demand for user-friendly and versatile web-based tools continues to grow, DRMAAAtic represents a significant step toward addressing these evolving needs.

Author contributions

Alessio Del Conte (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing—original draft [lead]), Hamidreza Ghafouri (Formal analysis [equal], Visualization [equal], Writing—review & editing [equal]), Damiano Clementel (Investigation [supporting], Methodology [supporting], Software [supporting], Visualization [supporting], Writing—original draft [supporting]), Ivan Mićetić (Resources [supporting], Software [supporting]), Damiano Piovesan (Conceptualization [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Silvio C. E. Tosatto (Funding acquisition [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]), and Alexander Miguel Monzon (Formal analysis [equal], Supervision [lead], Writing—original draft [lead], Writing—review & editing [lead])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by: ELIXIR, the research infrastructure for life-science data; COST Action ML4NGP [CA21160], supported by COST (European Cooperation in Science and Technology); European Union through NextGenerationEU; PNRR project ELIXIRxNextGenIT [IR0000010]; National Center for Gene Therapy and Drugs based on RNA Technology [CN00000041]; Italian Ministry of Education and Research through the NextGenerationEU fund PRIN 2022 project: PLANS [2022W93FTW]; co-funded by the European Union [101160233] (HORIZON-Twinning project IDP2Biomed) and under grant agreement no. 101182949 (HORIZON-MSCA-SE project IDPfun2); views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Funding for open access charge: University of Padua.

Data availability

The data underlying this article are available in *GitHub* at <https://github.com/BioComputingUP/DRMAAAtic>.

References

- Berman H, Henrick K, Nakamura H *et al.* Announcing the worldwide protein data bank. *Nat Struct Biol* 2003;10:980.
- Cholia S, Sun T. The NEWT platform: an extensible plugin framework for creating ReSTful HPC APIs. *Concurr Comput Pract E* 2015; 27:4304–17.
- Cruz FA, Dabin AJ, Dorsch JP. *et al.* FirecREST: a RESTful API to HPC systems. In: *2020 IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud)*, Atlanta, GA, USA, 2020, 21–6.
- Del Conte A, Bouhraoua A, Mehdiabadi M *et al.* CAID prediction portal: A comprehensive service for predicting intrinsic disorder and binding regions in proteins. *Nucleic Acids Res* 2023;51:W62–9.
- Del Conte A, Camagni GF, Clementel D *et al.* RING 4.0: faster residue interaction networks with novel interaction types across over 35,000 different chemical structures. *Nucleic Acids Res* 2024; 52:W306–12.
- Fielding RT, Taylor RN. Architectural styles and the design of network-based software architectures. Ph.D Dissertation. University of California, Irvine, 2000.
- Ghafouri H, Lazar T, Del Conte A *et al.*; PED Consortium. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res* 2024; 52:D536–44.
- Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–637.
- Rigden DJ, Fernández XM. The 2021 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2021;49:D1–9.
- Rigden DJ, Fernández XM. The 2022 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2022;50:D1–10.
- Rigden DJ, Fernández XM. The 2023 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2023;51:D1–8.
- Rigden DJ, Fernández XM. The 2024 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2024;52:D1–9.

- Rongqiang C, Haili X, Shasha L *et al.* SCEAPI: a unified restful web API for high-performance computing. *J Phys Conf Ser* 2017; **898**:092022.
- The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res* 2024;**52**:W83–94.
- Troger P, Rajic H, Hass A. *et al.* Standardization of an API for Distributed Resource Management Systems. In: *Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07)*, Rio De Janeiro, 2007. USA: IEEE Computer Society, 2007, 619–26.
- Varadi M, Bertoni D, Magana P *et al.* AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024; **52**:D368–75.
- Williams CJ, Headd JJ, Moriarty NW *et al.* MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 2018;**27**:293–315.

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
 Bioinformatics Advances, 2025, 00, 1–8
<https://doi.org/10.1093/bioadv/vbaf112>
 Original Article

Chapter 2: Standardised Workflows for the Determination of Conformational Ensembles

This chapter contains two articles, one published in Current Opinion in Structural Biology (COSB) and the other accepted in Nature Methods and available in bioRxiv.

1. **Ghafouri H**, Kadeřávek, P., Melo, A.M. et al. Toward a unified framework for determining conformational ensembles of disordered proteins. Nat Methods (2026). (<https://doi.org/10.1038/s41592-026-03003-2>) (The appendix associated with this paper is provided in **Appendix B** at the end of the thesis.)
2. **Ghafouri H**, Tosatto SC, Monzon AM. Advances in the determination of disordered protein ensemble. Current Opinion in Structural Biology. 2026 Feb 1;96:103198.(<https://doi.org/10.1016/j.sbi.2025.103198>)(The appendix associated with this paper is provided in **Appendix C** at the end of the thesis.)

Toward a unified framework for determining conformational ensembles of disordered proteins

Received: 30 March 2025

Accepted: 15 December 2025

Published online: 09 March 2026



Hamidreza Ghafouri¹, Pavel Kadeřávek^{2,3}, Ana M. Melo^{4,5}, Maria Cristina Aspromonte¹, Pau Bernadó⁶, Juan Cortés⁷, Zsuzsanna Dosztányi⁸, Gábor Erdős⁸, Michael Feig⁹, Giacomo Janson⁹, Kresten Lindorff-Larsen¹⁰, Frans A. A. Mulder¹¹, Peter Nagy^{12,13,14,15}, Richard Pestell^{14,15,16,17,18}, Damiano Piovesan¹, Marco Schiavina¹⁹, Benjamin Schuler^{20,21}, Nathalie Sibille²², Giulio Tesei¹⁰, Peter Tompa^{18,23}, Michele Vendruscolo²⁴, Jiri Vondrasek²⁵, Wim Vranken^{26,27,28,29,30}, Lukas Zidek^{2,3}, Silvio C. E. Tosatto¹✉ & Alexander Miguel Monzon¹✉

Disordered proteins play essential roles in myriad cellular processes, yet their structural characterization remains a major challenge due to their dynamic and heterogeneous nature. Here we present a community-driven initiative to address this problem by advocating a unified framework for determining conformational ensembles of disordered proteins. Our aim is to integrate state-of-the-art experimental techniques with advanced computational methods, including knowledge-based sampling, enhanced molecular dynamics and machine learning models. The modular framework comprises three interconnected components: experimental data acquisition, computational ensemble generation and validation. The systematic development of this framework will ensure the accurate and reproducible determination of conformational ensembles of disordered proteins. We highlight the open challenges necessary to achieve this goal, including force-field accuracy, efficient sampling, and environmental dependence, advocating for collaborative benchmarking and standardized protocols.

Intrinsically disordered proteins and regions (hereafter collectively termed IDPs) are prevalent in eukaryotic proteomes, making up an estimated one-third of the human proteome¹. These proteins exist as conformational ensembles of rapidly interconverting conformations (Box 1)². IDPs are implicated in a broad range of signaling and regulatory functions, such as signal transduction, gene expression, genome organization, cell-cycle control and the formation of biomolecular condensates³. Consequently, IDPs are central targets for biological and medical research. A defining feature of IDPs is the link between the features of their conformational ensembles and their biological functions^{4–6}. Understanding this relationship requires a robust

framework for generating and analyzing conformational ensembles by combining computational and experimental techniques. In this Perspective, we advocate a unified framework for generating, analyzing and validating conformational ensembles of IDPs. This framework, developed through extensive discussions initiated during a workshop held in Prague in May 2023 (<https://ml4ngp.eu/workshop-prague/>), is structured around three key modules: (1) experimental observations informative of conformational ensembles, (2) computational generation of conformational ensembles consistent with the experimental information and (3) validation and comparison of the resulting conformational ensembles. By integrating insights from different

A full list of affiliations appears at the end of the paper. ✉ e-mail: silvio.tosatto@unipd.it; alexander.monzon@unipd.it

Nature Methods

BOX 1**Glossary of terms used within the scope of this paper**

Term	Definition
IDPs and disordered protein regions	Full-length proteins or protein regions that lack a stable ordered three-dimensional structure under physiological conditions
Conformers	Structures able to interconvert without making or breaking covalent bonds
Conformational collection	A set of conformers
Statistical ensemble	A conformational collection endowed with statistical weights
Conformational state	A thermodynamic state of a protein described by a statistical ensemble
Forward model	A computational framework that predicts experimental observables based on protein structures

methodologies in structural and computational biology, our goal is to progress toward establishing standardized processes that support integrative studies of IDPs. This modular framework (Fig. 1) underscores the importance of interdisciplinary collaboration in addressing the complexities of IDPs. By standardizing methodologies and establishing consensus within the scientific community, this framework seeks to enhance our understanding of the functions of IDPs, advance methodological rigor and accelerate therapeutic discovery.

Experimental techniques for studying IDP structural ensembles**NMR spectroscopy**

Nuclear magnetic resonance (NMR) spectroscopy is uniquely suited for studying IDPs in solution, offering atomic-level information. This technique can capture the dynamic and heterogeneous nature of IDPs by exploiting the sensitivity of nuclear magnetic moments to their local magnetic environment. This sensitivity enables precise, selective and minimally invasive atomic-resolution assessment of structure, dynamics and interactions of IDPs under near-physiological conditions⁷, typically following isotope labeling. Because IDPs have shallow free-energy landscapes, NMR observables are time averages that represent the statistical ensemble. However, the lack of information on individual states can be viewed as a fundamental limitation. The ensemble-averaged NMR parameters do not enable the specific reconstruction of individual conformers of the ensemble. Instead, the NMR parameters can be compared to those derived by forward models from simulated structural ensembles and thus benchmark the latter. Also, there are cases when some states of IDPs can be distinguished (for example, the presence of *cis* peptide bonds, which give rise to sets of additional signals^{8–10}), or when the NMR spectrum clearly carries a mark of their distribution (for example, dynamics of molten globules manifested by line broadening¹¹). NMR can also unveil sparsely populated NMR-invisible structural states by exploiting the transfer of magnetization between these states and NMR-visible states, through techniques like saturation transfer difference NMR and relaxation dispersion NMR. By analyzing the transferred magnetization, these techniques provide structural and dynamic information about minor states populated during structural transitions upon binding¹². Moreover, NMR methods of selective detection of certain states have been designed (for example, paramagnetic relaxation enhancement (PRE) discussed below).

Key NMR observables in IDP studies. Several NMR parameters can be directly correlated to conformational properties of disordered proteins. Chemical shifts reflect local electronic environments and provide insights into secondary-structure propensities (SSPs)^{13–15}. Computational tools such as ShiftX¹⁶, SPARTA+¹⁷, PPM¹⁸ and CamShift¹⁹ include forward models to predict chemical shifts from conformational ensembles (Table 1).

J-couplings, particularly three-bond couplings, directly report on backbone torsion angles through the Karplus equation²⁰. These measurements are limited by the accuracy of the corresponding forward models and the availability of measurable *J*-couplings for specific torsion angles. For example, ϕ angles can reliably be probed²¹. Furthermore, ψ angles often require complementary approaches such as cross-correlated relaxation^{22,23}.

Residual dipolar couplings (RDCs) provide highly sensitive insights into IDP conformational preferences by reporting on the average orientation of bond vectors across all coexisting structures in solution^{24,25}. RDCs reveal local secondary-structure biases, having an opposite sign in α -helices and extended structures. However, their detailed structural interpretation requires calculating or fitting the alignment tensor, which is particularly challenging for IDPs. To avoid this calculation, forward models that use the laboratory reference frame have been proposed²⁶. Various alignment media exist, yet careful selection is crucial to avoid unwanted interactions that could alter measurements. Another use of RDCs is for the cross-validation of conformational ensembles derived from other NMR and small-angle X-ray scattering (SAXS) data, ensuring consistency in structural interpretations²⁷.

Relaxation rates offer insights into IDP dynamics across time-scales. Through IDP-specific model-free formalism, the analysis of several backbone ¹⁵N relaxation experiments in multiple magnetic fields probes the local and segmental motional amplitudes and their associated correlation times^{28,29}. These data are also valuable for benchmarking molecular dynamics simulations, in which ensemble reweighting may be required to match experimental relaxation rates³⁰.

PRE is particularly powerful for disordered proteins, revealing transient long-range contacts and compact conformations within heterogeneous ensembles. For instance, PRE measurements have revealed transiently folded states in different IDPs^{31–33}, complementing global techniques like SAXS in characterizing conformational distributions^{34–36}. Paramagnetic relaxation interference methods further expand this analysis, identifying correlated or anticorrelated motions in transiently compacted regions³⁷.

Because IDPs are highly sensitive to their environments, in-cell NMR is particularly indicated to study their context-dependent behavior under native-like conditions, such as in bacterial and eukaryotic cells. These studies, while technically challenging due to low sensitivity and high background noise, highlight the potential of NMR to bridge in vitro and in vivo IDP investigations^{38–41}.

smFRET

Single-molecule Förster resonance energy transfer (smFRET) is a powerful tool for investigating the conformational dynamics and structural heterogeneity of IDPs. Förster resonance energy transfer (FRET), a non-radiative energy transfer from an excited donor fluorophore to an acceptor fluorophore, is highly sensitive to donor–acceptor distances, following an inverse sixth-power relationship. This makes smFRET a spectroscopic ruler capable of measuring intermolecular and intramolecular distances within a 2–10-nm range⁴². Unlike ensemble-averaging methods, smFRET provides a unique ability to resolve subpopulations, detect rare states and capture dynamic transitions of individual molecules, making it particularly suited for IDP studies^{43,44}.

smFRET applications to IDPs. smFRET has been instrumental in studying IDPs under diverse conditions. It has been used to elucidate the conformational ensembles of IDPs in dilute solution^{43,45,46} and

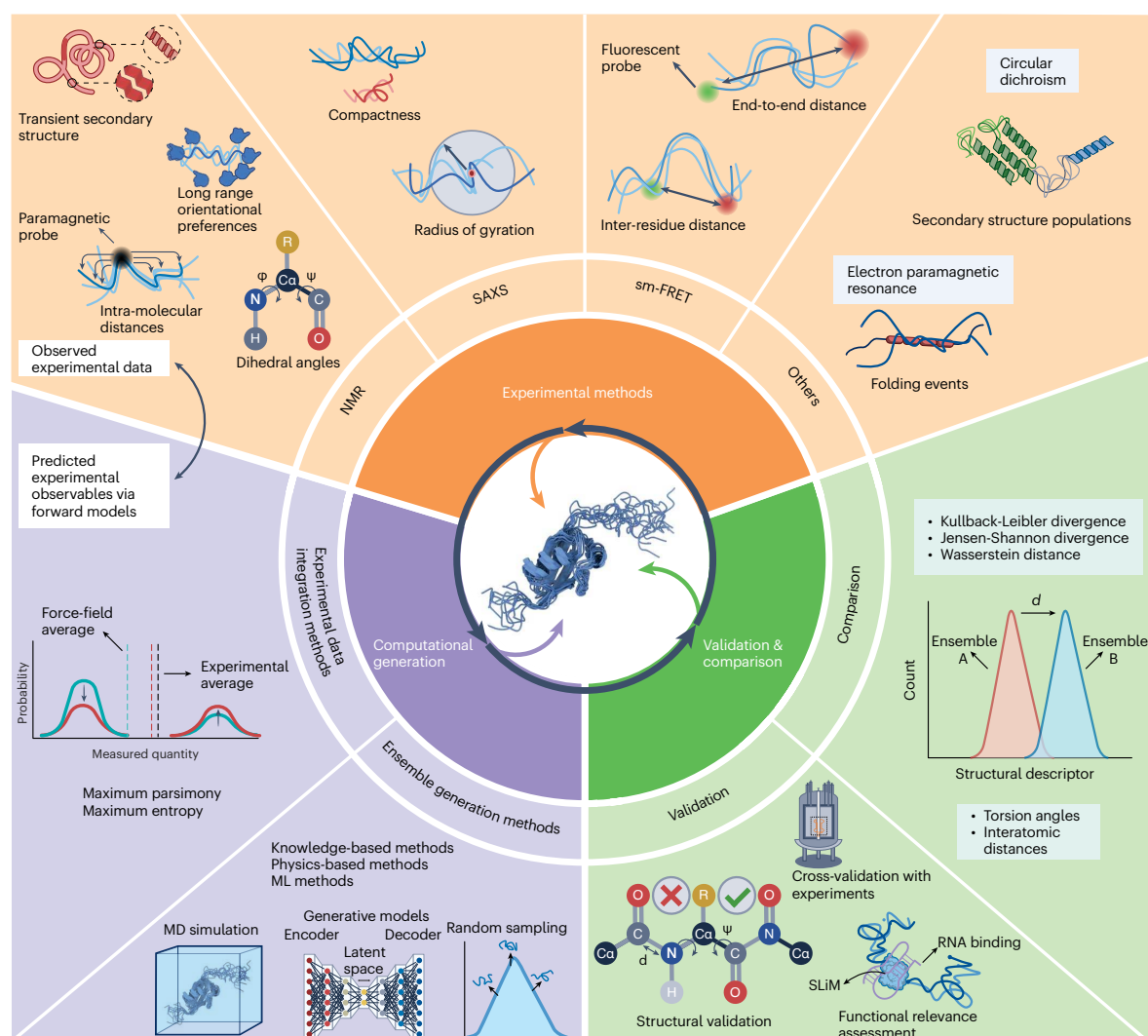


Fig. 1 | Schematic representation of the modular framework for the determination of conformational ensembles of IDPs. The illustration presents the modular framework for determining conformational ensembles of IDPs, comprising three interconnected modules: experimental methods, computational ensemble generation and validation and comparison. The first module focuses on experimental techniques such as NMR spectroscopy, single-molecule FRET, SAXS, circular dichroism and electron paramagnetic resonance spectroscopy. These methods provide local and global structural information for characterizing the heterogeneous and dynamic nature of IDPs. The second module of the framework addresses the computational determination of conformational ensembles of IDPs. This module involves the generation of conformational ensembles and the integration of experimental data obtained in the first module. Ensemble generation methods can be categorized into physics-based (molecular dynamics simulations), knowledge-based and machine

learning methods. In addition, maximum parsimony and maximum entropy are the main approaches for the integration of experimental data. The role of forward models is crucial in this step since they link the experimental observables to the generated models to ensure the goodness of fit to observed experimental data. Lastly, the validation and comparison tools complete the circular process of determining IDP conformational ensembles. Validation can be achieved through compliance of structures in the ensemble with basic physicochemical rules (structural validation), cross-validation using complementary experimental data and functional relevance assessment. Comparison tools particularly involve statistical methods for analyzing distributions, such as Kullback–Leibler, Jensen–Shannon divergence scores and Wasserstein distances. MD, molecular dynamics; ML, machine learning. Figure created in BioRender; Ghafouri, H. <https://biorender.com/aq5pa86> (2025).

crowded environments⁴⁷, including phase-separated biomolecular condensates^{48,49} and live cells⁵⁰. Furthermore, this technique has shed light on dynamic IDP complexes, revealing the interplay of transient structural states during functional interactions⁴⁴. Single-molecule studies can be performed down to picomolar concentrations⁴²,

avoiding the complications of aggregation and allowing functional/dysfunctional states of aggregation-prone IDPs to be probed^{51–53}.

smFRET experimental setup and protein labeling. smFRET experiments typically use either total internal reflection fluorescence

Table 1 | Main experimental methods to determine structural ensembles of IDPs

Technique	Experimental measurement	Forward models/ computational tools/ theoretical models	Information content	Structural interpretation (for example, residue level versus global)	System perturbation	Environment (in vivo/in vitro/ in cell)	Reference database
NMR	Chemical shifts (traditional, ^{13}C -NMR, ...)	ShiftX, SHIFTS, Sparta, PPM1, CamShift	Local SSP; ensemble-averaged deviations from random coil values. Limited	Dihedral angles, secondary structure population, dynamics	Isotope labeling		
	PRE, paramagnetic relaxation interference	Dipole-dipole equation + correlation function, DEER-PREDict, MMM (Multiscale Modeling of Macromolecules)	Ensemble-averaged distances between spin labels and nuclei; provides insight into transient, weak or low-population interactions and structural heterogeneity.	Distances/regions with correlated folding/dynamics	Chemical attachment of a paramagnetic label		
	Relaxation rates	Model to data: molecular dynamics simulation + correlation function analysis + weighting Data to model: spectral density mapping, model-free analysis, IMPACT	Time constants for intramolecular and global motions; order parameters reflecting flexibility; reveals timescales of molecular motions.	Correlation times + weights/order parameters		In cell/in vitro	https://bmrdb.io/
	DOSY	Data to model: Stokes-Einstein equation Model to data: Kirkwood-Riseman equation, HYDROPRO	Global hydrodynamic radius; informs on chain expansion or compaction	Translation diffusion coefficient	Isotope labeling		
	J-couplings	Karplus equation	Backbone dihedral angles (ϕ); detects transient secondary structure.	Dihedral angles			
	Cross-correlated cross-relaxation	Data to model: correlation functions and Karplus-like equation	Dihedral angles based on correlated relaxation processes between two couplings.	Dihedral angles			
	RDCs	Modeling alignment and calculation of the coupling	Bond vector orientation; indicates transient structural elements and long-range order	Mutual orientation	External alignment (mechanically stressed gels, alignment media)		
SAXS	Scattering intensity as a function of the angle	CRY SOL/WAXSiS Pepsi-SAXS/FoXS	Provides global structural information	Global shapes and sizes, low-resolution structures	None required	In vitro	https://www.sasbdb.org/
SANS	Scattering intensity as a function of the angle	CRYSON/ pepsi-SANS	Complementary to SAXS	Global shapes and sizes, low-resolution structures	Deuteration	In vitro	https://www.sasbdb.org/
(sm)FRET	Transfer efficiency (fluorescence bursts or trajectories) Fluorescence lifetime decay Time correlation (nanosecond fluorescence correlation spectroscopy)	Förster equation + polymer models, molecular simulations>+ dye model (for example, FRET positioning and screening, Seidel), rotamer-library approaches (for example, FRETpredict)	Number of populations, FRET efficiencies, fluorescence lifetimes, fluorescence anisotropies	Intramolecular or intermolecular distance/distance distribution, conformational changes Intramolecular or intermolecular distance distribution Distance relaxation time	Dye labeling	In vitro/in cell	Not applicable

microscopy to observe immobilized molecules, or confocal spectroscopy to probe both freely diffusing and immobilized molecules. For IDPs, confocal-based smFRET avoids perturbations from immobilization and enables multiparameter detection (wavelength, lifetime, anisotropy, correlation)⁵⁴.

Proteins are subjected to site-specific labeling with donor–acceptor fluorophore pairs via cysteine residues or more advanced labeling strategies, such as nonnatural amino acids or intein-based ligation methods^{55,56}. Heterogeneously labeled populations and different stoichiometries of intermolecular FRET can be resolved by alternating laser excitation or pulsed interleaved excitation^{44,57}. The functional impact of labeling must be carefully assessed through control experiments, such as comparing the labeled and unlabeled IDP behavior in functional assays. Fluorophore mobility can be evaluated through time-resolved fluorescence anisotropy to ensure rotational freedom^{43,44,54}.

Quantitative analysis of smFRET data for IDPs. The FRET efficiency, E , for IDPs cannot be interpreted in terms of a fixed donor–acceptor distance. Instead, IDPs require analytical approaches that can account for the broad distributions, $P(r)$, of the inter-fluorophore distances r , and the dynamic timescales of protein and dye motions. Suitable models from polymer physics have commonly been used for describing global IDP features and to approximate $P(r)$ ⁴⁵. Fluorescence lifetime measurements complement smFRET for accurate $P(r)$ determination. Two-dimensional plots of donor and acceptor fluorescence lifetime versus E can reveal whether a linear relationship exists (indicating fixed distances) or if deviations suggest broad and rapidly sampled distributions, offering a more stringent test of polymer models^{43,44}. The integration with advanced computational approaches, including molecular simulations and Bayesian reweighting, can provide more detailed descriptions of $P(r)$ and deviations from simple polymer models, for instance in the presence of folded parts^{46,58}. Nanosecond fluorescence correlation spectroscopy, combined with smFRET, enables the measurement of fast dynamics, including the reconfiguration times of disordered chains, which can be directly compared to simulations⁴⁵.

Important aspects for smFRET are to assess potential perturbations of the protein by fluorescent labeling, the analysis of complex distance distributions and the role of the relevant timescales of the measurement for signal averaging^{45,59}. Advances in labeling technologies and integration with computational modeling, including an explicit representation of fluorophores, and complementary biophysical techniques make smFRET invaluable in integrative approaches of IDP ensemble determination^{60,61}.

SAS of X-rays and neutrons

Small-angle scattering (SAS) of X-rays (SAXS) and neutrons (SANS) have emerged as powerful techniques for the structural characterization of biomolecular systems in solution⁶². Despite being a low-resolution (1–2 nm) technique, SAS offers valuable structural information on the overall size and shape of biomolecules and biomolecular assemblies. In the context of IDPs, this information is key to constrain the ensemble description, and it is complementary to residue-level information derived from other techniques^{63–65}. Major advances in beamline instrumentation, automatization, sample preparation and computational methods in the last decade have led to a tremendous increase in the application of SAS to structural biology⁶⁶. The fast readout capabilities of current detectors enable kinetic monitoring of reactions across a broad timescale, from microseconds to hours, triggered by stopped-flow⁶⁷ or temperature jumps⁶⁸.

Although the conceptual framework of X-ray and neutron scattering as well as the strategies applied for their analysis are similar, these techniques differ in practical aspects, especially related to their sensitivity, which is much higher for SAXS^{69,70}. The advantage of SANS arises from the different scattering power of nuclei, which is particularly large between the proton and the deuteron. Exploiting this isotope

effect through contrast variation experiments of partially deuterated samples has been a key concept to enrich the structural content of SANS experiments^{69,70}.

Analysis of SAS data. The scattering profile of an IDP is the average of those arising from all coexisting conformations sampled in solution. As a consequence, SAS profiles measured for IDPs do not display features along the momentum transfer (q) range⁶³. Traditionally, Kratky plots ($I(q) \times q^2$ as a function of q), especially in their dimensionless form⁶⁴, have been used to qualitatively identify (partly) disordered states and distinguish them from globular particles. The Kratky representation has the capacity to enhance particular features of scattering profiles, allowing an easier identification of different degrees of compactness⁷¹.

The radius of gyration, R_g , which can be obtained from the smallest angles of a SAS curve using the Guinier approximation, is the most common descriptor to quantify the overall size of particles in solution. The experimental R_g is a single-value representation of the size of the molecule, which for disordered states represents an ensemble average over all accessible conformations of the IDP and its hydration layer. The most common quantitative interpretation of R_g for IDPs is based on Flory's equation, which relates it to the length of the protein chain through a power law ($R_g \propto N^\nu$), where N is the number of residues and ν is an exponential scaling factor. For chemically denatured proteins, a ν value of ≈ 0.6 is expected⁷². However, according to theoretical models, IDPs are slightly more compact, with a ν value of 0.522 ± 0.01 (ref. 73). Recently, an approach has been developed to simultaneously derive the R_g and the ν value from a single SAXS profile⁷⁴, which has been used to connect the compactness of IDPs with their capacity for phase separation^{36,75}.

The pairwise distance distribution, $P(r)$, derived from the Fourier transformation of the SAS profile, provides a one-dimensional representation of the molecular structure, and it can also be used to qualitatively identify protein flexibility. The $P(r)$ also provides the maximum intramolecular distance, $D_{\max}$, among all coexisting conformers. However, in the context of disordered proteins, $D_{\max}$ is not a robust parameter, and its quantitative interpretation is not recommended^{76,77}.

For IDPs, ensemble approaches are required, relying on two key ingredients: (i) accurate models of disordered states, and (ii) forward models with the capacity to robustly predict scattering properties including contributions from the solvent to the measured SAS data⁷⁸.

Other biophysical techniques

Other complementary experimental methods have also gained attention in the studying of IDP structural properties (Supplementary Table 1). Hydrogen deuterium exchange (HDX) relies on measuring the rate at which amide protons in the protein backbone exchange with deuterium in a heavy water solution, a process highly sensitive to structural fluctuations and hydrogen bonding. For IDPs, HDX by NMR spectroscopy or in particular by mass spectrometry (HDX-MS) provides valuable insights into disorder-to-order transitions, transient secondary-structure formation and conformational flexibility, making it a complementary technique to other structural methods such as NMR, smFRET or SAS. However, given the high exchange rates of IDPs, careful control of experimental conditions, such as pH, temperature and labeling times, becomes critical to ensure meaningful interpretation of the data⁷⁹. Mass spectrometry-based techniques beyond HDX-MS, for example covalent labeling approaches such as hydroxyl radical footprinting or DEPC, and carboxyl or lysine-specific covalent modifications, provide complementary insights into IDP dynamics by irreversibly labeling solvent-exposed side chains on rapid timescales⁸⁰. These methods capture transient and heterogeneous conformations that HDX may miss, offering residue-level resolution of side-chain exposure. Applications include detecting structural transitions during amyloid formation⁸¹, mapping interaction interfaces⁸² and identifying microenvironment-sensitive regions^{83,84}. When combined

with modeling, these approaches enhance our ability to characterize the dynamic ensembles and functional states of IDPs.

Circular dichroism spectroscopy is a common technique for characterizing protein secondary structures by measuring the differential absorption of left- and right-circularly polarized light. Traditionally calibrated for globular proteins, circular dichroism spectroscopy has faced challenges in analyzing IDPs due to their weak and featureless spectra. Recent advancements have introduced new IDP reference datasets^{85,86}, enhancing the ability of this technique to detect conformational changes under various conditions. Despite these improvements, precise secondary-structure quantification of IDPs still remains a challenge.

Electron paramagnetic resonance spectroscopy, combined with site-directed spin labeling, is used to detect unpaired electrons and measure distance distributions between specific residues. It has emerged as a powerful technique for studying IDPs⁸⁷. Electron paramagnetic resonance spectroscopy coupled with site-directed spin labeling is particularly useful for monitoring folding events and structural behavior, as it is not limited by protein size or system complexity⁸⁸. The technique has been successfully applied to study disease-associated IDPs such as α -synuclein and tau, providing valuable insights into their conformational ensembles and interactions with the environment^{87,89}.

Integrative experimental modeling

Integrative experimental approaches rely on combining complementary experimental data to determine conformational ensembles. An analysis of the Protein Ensemble Database⁹⁰ shows that the most common hybrid approaches combine SAXS with NMR and/or smFRET. Notably, Gomes et al.⁹¹ used experimental NMR and SAXS data to construct the conformational ensembles for Sic1 and phosphorylated Sic1, and subsequently used smFRET data as an independent validation criterion. In another recent study, Borthakur et al.⁹² demonstrated that reweighted ensembles obtained from molecular dynamics simulations using different force fields converge to highly similar ensembles when incorporating extensive experimental data from NMR and SAXS. In addition, other combinations can be useful, such as SAXS, smFRET, fluorescence correlation spectroscopy and dynamic light scattering, as shown by Borgia et al.⁹³. Such integrative studies illustrate how the information from different techniques is complementary.

Bridging experiments and theory: forward models

A critical element in translating experimental data into conformational ensembles is to define procedures, referred to as forward models, to calculate experimental observables from protein structures^{94,95}. A major problem of forward models for IDPs is that most experimental methods provide data describing the average properties of proteins in solution, often over different timescales, rather than features of individual protein molecules, thus requiring a corresponding conformational averaging in the calculations^{94,95}. Forward models estimating the experimental observables from generated models enable fitting experimental data obtained from various sources with a representative collection of conformations, as outlined in Table 1. Because the use of forward models influences the fitting procedure, selecting the appropriate model requires detailed understanding and considerations⁹⁶. For instance, parameters of forward models need to be adjusted, ideally through self-consistent refinement against experimental datasets of both forward models and conformational ensembles^{96–98}. Ultimately, experimental observations are influenced by both systematic and random errors, which must be accounted for to produce accurate and precise models⁹⁵.

To consider errors and the inherent variability of experimental observables, different Bayesian inference frameworks have been suggested^{95,96}. Nonetheless, the development of more accurate and computationally efficient forward models and machine learning

tools to account for different sources of error remains an active area of research.

Computational generation of conformational ensembles

Generating conformational ensembles through computational techniques is a critical step in integrative modeling of IDPs, providing atomic-level insight that complements experimental measurements. Recent progress has been driven by tighter integration between experimental and modeling efforts, where experimental data refine models and models aid in the interpretation of complex experiments⁹⁴. Three main strategies are commonly used to generate conformational ensembles: knowledge-based methods, which draw on empirical patterns or structural fragments; physics-based simulations, which are based on biophysical principles to sample the conformational landscape; and machine learning-driven approaches, which infer structural features from data without explicit physical modeling. Each strategy offers distinct advantages, and their selection depends on the level of detail required, available data and modeling goals.

Knowledge-based methods

Knowledge-based methods (Table 2) generate conformational ensembles using heuristics or statistical patterns derived from known protein structures. These approaches often rely on fragment libraries, rotamer statistics or backbone angle distributions informed by structural databases. Examples of such methods include Flexible-Meccano⁹⁹, TraDES¹⁰⁰ and IDPConformerGenerator¹⁰¹, which use fragments from nonredundant structural databases and statistical coil models to generate ensembles. Because they do not require energy calculations, they are computationally efficient and suitable for generating large conformational pools. These conformational collections can then be refined or reweighted based on experimental constraints to approximate statistical ensembles. Importantly, knowledge-based approaches are particularly useful when only limited experimental data are available, or when rapid conformational collection generation is needed for large-scale applications. A summary of representative methods and typical use cases is provided in Supplementary Table 2.

Physics-based methods

Physics-based approaches (Table 2) use molecular simulations to generate IDP ensembles. Since force fields originally developed for folded proteins often underestimate the flexibility and expanded conformations typical of disordered chains, IDP-specific force fields such as a99SB-disp¹⁰² and Amber ff99SBws/ff03ws¹⁰³ have been developed to improve the balance between intramolecular and solvent interactions to better reflect IDP conformational preferences. Another challenge is sampling efficiency, as IDPs explore vast and shallow free-energy landscapes. Techniques like replica-exchange molecular dynamics¹⁰⁴, which run simulations in parallel at different temperatures to enhance conformational exploration, are commonly used to improve sampling. More recently, hybrid strategies that combine simulation with machine learning—for example, using machine learning-derived potentials to guide sampling—have emerged as promising tools^{105,106}. Alternative frameworks such as hydrophobicity scale models¹⁰⁷ and CALVADOS^{108,109}, coarse-grained models parameterized against experimental data on IDPs and multidomain proteins, reduce computational cost while preserving some of the essential sequence-dependent behaviors. Beyond ensemble generation, physics-based methods increasingly contribute to design-oriented frameworks—from modeling the thermodynamic basis of multiphase condensates¹¹⁰, to uncovering sequence–ensemble–function relationships in live cells¹¹¹, and enabling inverse design of IDRs with target ensemble properties^{112–114}. Representative methods and applications are summarized in Supplementary Table 2.

Table 2 | Comparison of computational strategies for the generation of conformational ensembles of IDPs

	Knowledge based	Physics based	Machine learning
Input	Protein sequence Structural fragments and SSPs (optional)	Three-dimensional structural model Force fields adapted for IDPs	Protein sequence
Core principle	Sampling backbone dihedrals using Protein Data Bank (PDB)-derived statistics or fragments	Simulates atomic motions over time using force fields	Learns statistical patterns from structural and/or experimental data to generate IDP ensembles from sequence
Resource/speed	Low; runs quickly on CPU with minimal setup	High; computationally intensive, often needs GPUs, HPC and force-field optimization	Moderate; fast inference after training, may require GPUs for large models
Output	Diverse ensembles reflecting backbone variability without including thermodynamic/kinetic information	Time-resolved or equilibrium ensembles with atomistic or coarse-grained resolution	Diverse coarse-grained/all-atom conformers sampled from a learned distribution
Typical use cases	Rapid conformer generation, SAXS ensemble preparation, fragment-based modeling	High-resolution structural refinement, validating experimental data (SAXS, NMR)	Sequence-to-ensemble modeling, large-scale IDP studies and synthetic ensemble generation

Machine learning methods

Machine learning approaches (Table 2) have recently emerged as an alternative strategy for generating conformational ensembles of IDPs. These methods aim to learn sequence–ensemble relationships directly from structural data, providing a way to overcome some of the limitations of traditional simulation-based techniques. For example, idpGAN¹¹⁵ is a generative adversarial network trained on ensembles generated by coarse-grained molecular dynamics simulations, capable of producing novel ensembles with realistic R_g . While still in early stages, machine learning-based models offer a promising route for efficient ensemble generation, especially when integrated with simulation-derived^{116,117} and/or experimental data¹¹⁸. In addition to generating ensembles, machine learning and related statistical parameter-learning methods can be used to tune force-field parameters against experimental data^{114,119}, predicting IDP properties directly from sequence^{4,6}, designing novel IDP variants with desired properties^{112,120} and predicting phase-separation propensities of IDPs¹²¹. Representative machine learning-based methods for computational ensemble generation are summarized in Supplementary Table 2.

Integration of computational models with experimental data

Experimental techniques such as NMR, SAXS and smFRET provide important insights into the behavior of IDPs, but these measurements typically report on averages over a wide range of rapidly fluctuating conformations. By contrast, computational methods—such as molecular simulations—can generate detailed conformational ensembles from a protein sequence. However, discrepancies often arise between simulations and experiments, reflecting inaccuracies in force fields, limited sampling or the nature of the experimental observable itself¹²².

To address this aspect, an integrative modeling strategy is needed—one in which experimental data and computational ensembles inform and refine each other. The basic workflow (Fig. 2) involves generating an ensemble using a physics-based, knowledge-based or machine learning model, computing experimental observables from that ensemble using forward models, and comparing them with actual experimental measurements. When agreement is poor, the ensemble can be refined or reweighted so that its average properties better match the data.

Two main classes of approaches are available for this purpose. The first is based on maximum entropy¹²³, a foundational concept in statistical mechanics for generating the most unbiased ensemble consistent with known constraints. In its simplest form, and in the absence of experimental data, structural ensembles follow the Boltzmann distribution: $P_i = \frac{1}{Z} e^{-\beta E_i}$ where E_i is the energy of conformation i , β is the inverse thermal energy, and Z is the partition function. However, when

experimental observables are introduced, such as NMR chemical shifts or SAXS profiles, this framework is generalized to a Gibbs distribution:

$P_i = \frac{1}{Z} e^{-\sum_k \beta_k E_{ik}}$. Here, E_{ik} represents an effective energy term for the deviation of conformation i from experimental observable k , and β_k (a Lagrange multiplier) adjusts the weight of that constraint. This formulation ensures the structural ensemble remains as close as possible to the original distribution while still fulfilling the experimental data. In practice, this statistical foundation, together with the principles from Bayesian statistics, underlies a number of maximum-entropy-based methods such as Bayesian/maximum entropy (BME)¹²⁴, BioEn¹²⁵ and metainference⁹⁵. These approaches implement the principle either by reweighting existing conformers post hoc, or by applying biasing forces during simulations to guide sampling. Such techniques are especially effective for IDPs, whose native states are characterized by extensive structural heterogeneity that must be preserved during model refinement.

The second category, known as maximum parsimony methods, seeks to identify the smallest subset of conformers from a large pool that can adequately explain the experimental data, minimizing structural complexity. In contrast to maximum entropy, maximum parsimony methods aim to simplify the conformational collection by selecting or reweighting structures to maximize agreement with experimental observations while minimizing the size of the conformational collection. These approaches, such as EOM¹²⁶ and ASTER-OIDS¹²⁷, are particularly useful when only a few conformational states are dominant and are thus most suitable for low-entropy systems with limited conformational heterogeneity, where a description in terms of statistical ensemble is not required.

These integrative strategies yield experimentally validated, physically realistic ensembles. Representative techniques for integration of experimental data are provided in Supplementary Table 3.

Hybrid computational techniques

Some computational methods for modeling IDPs leverage complementary modeling strategies. These hybrid approaches (Supplementary Table 2) integrate complementary ensemble generation techniques, or couple ensemble generation directly with experimental data integration strategies. One strategy involves combining physics-based and knowledge-based approaches to enhance sampling efficiency and structural diversity. For example, FastFloppyTail¹²⁸ uses fragment-based backbone sampling alongside energy-based refinement. Other techniques couple machine learning with physics-based modeling, such as AlphaFold–Metainference¹⁰⁶ and bAIES¹⁰⁵, which refines deep learning predictions from AlphaFold via Bayesian ensemble inference and molecular simulation. A different class of hybrid

methods links ensemble generation with experimentally data-driven refinement. For instance, DynamICE¹¹⁸ integrates machine learning with maximum-entropy principles to iteratively optimize ensemble structures based on NMR data. These flexible and composable pipelines offer practical advantages when addressing the complexity of IDP ensembles, particularly in cases where no single modeling paradigm is sufficient.

Open challenges in conformational ensemble generation

Despite recent advances, several challenges remain in generating conformational ensembles¹²². First, force-field accuracy continues to be a critical limitation. While parametrizations using experimental data and quantum chemistry calculations are continuously improving their accuracy, current force fields still rely on rather simple functional forms established decades ago, and fail to describe subtle interactions, such as those arising from polarization effects, especially in complex or highly flexible biomolecules. Progress in developing more accurate force fields may require closer collaboration between experimental and computational researchers and the development of top-down approaches, including the adoption of machine learning methods^{108,119,129}.

Second, the optimal integration of experimental data into simulations is still an open problem^{95,122}. Many forward models used to translate experimental data into simulation restraints are limited by their assumptions and approximations. Accounting for the ensemble-averaged nature of most experimental observables on IDPs, such as those derived from NMR or smFRET, or the effect of the hydration for SAS, is critical for accurate modeling. Robust frameworks that address uncertainties in both experimental data and forward models are essential to bridge this gap^{95,122}. In the absence of ground-truth conformational ensembles, such frameworks require experimental data from orthogonal techniques and for different proteins to self-consistently select, refine and validate forward models^{96,98,130}. These experimental data are typically insufficient to describe a unique conformational ensemble, with the final result dependent on the computational model used to generate the ensemble.

Third, the efficient sampling of conformational space also remains a critical bottleneck. Insufficient sampling can lead to incomplete or biased conclusions, making it crucial to assess convergence rigorously. Advancements in some techniques, such as accelerated molecular dynamics and Markov state models^{131,132}, offer potential solutions but require widespread adoption and careful validation.

Conformational ensemble validation and comparison

Comparison of conformational ensembles

The determination of an IDP ensemble from sparse experimental data is a mathematically underdetermined problem, leading to multiple solutions. Hence, the comparison of these ensembles may be critical for validating them (see next section), and/or for deriving mechanistic insights by identifying functional correlates (for example, binding regions, dynamic linkers), comparing homologous IDPs and assessing the influence of solution conditions. As conformational ensemble comparison methods for folded proteins¹³³ are difficult to generalize to IDPs, various approaches have been developed in recent years. One common strategy is to represent a conformational ensemble by a single scalar that quantifies its conformational heterogeneity¹³⁴, and differences in this value can be used to compare two conformational ensembles. While this metric is easy to interpret, it offers only a limited view of the differences between conformational ensembles and lacks the resolution to capture more local structural or dynamical variations. A more sophisticated approach involves analyzing the differences in distance distributions of individual residue pairs and assessing their statistical significance between two conformational ensembles¹³⁵. However, given the vast conformational variability of IDPs, it is essential to consider both average properties and their distributions. Indeed, reducing IDP

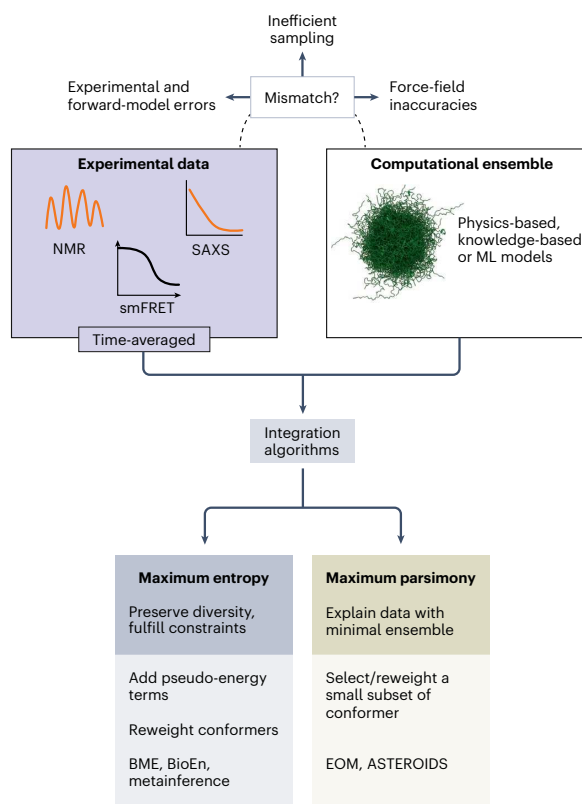


Fig. 2 | Overview of integrative modeling strategies for IDPs. Experimental methods such as NMR, SAXS and smFRET provide averaged data across many conformations, while computational models generate detailed structural ensembles. When the two disagree, for example due to limited sampling, force-field inaccuracies or experimental noise, different strategies can be used to improve agreement. These include maximum-entropy methods that adjust structural ensemble weights and maximum parsimony methods that generate conformational collections by selecting a minimal subset of structures. These strategies help align computational models with experimental data. Figure created in BioRender; Ghafouri, H. <https://biorender.com/194xtx1> (2025).

conformational descriptors to their mean can lead to major information loss. Therefore, several methods have been proposed for extending comparative structural analyses to a distributional framework, using various similarity measures to compare probability distribution functions, such as the Kullback–Leibler and the Jensen–Shannon divergence^{133,135,136}, Hellinger distance¹³⁷ or the Wasserstein distance¹³⁸. Among these, the Wasserstein distance is particularly appealing for its incorporation of the geometry of the underlying space, although it is more computationally demanding than the others. In this framework, different descriptors can be used to characterize conformations inside each ensemble and to compute pairwise distances. The most natural one is to use atomic coordinates and RMSD¹³³, however, the required structural superposition hinders its application to IDPs. To circumvent this problem, other descriptors such as torsion angles¹³⁹ or other local structural properties¹³⁶ can be considered. For a more comprehensive comparison, local and global structural descriptors can be taken into account in a unified way¹³⁸. Other approaches for the comparative analysis of conformational ensembles, such as methods incorporating topology concepts¹⁴⁰ or the identification of local energy traps at the residue level¹⁴¹, have been explored. Supplementary Table 4 summarizes conformational ensemble comparison methods.

Validation of conformational ensembles

Validating conformational ensembles of IDPs is challenging, as their accuracy and precision are difficult to establish. Nonetheless, objective criteria can be applied, generally relying on orthogonal experimental data and emphasizing the importance of integrative approaches.

Validation typically begins with basic stereochemical checks to ensure that each conformation follows physicochemical rules. These include proper bond lengths, torsion angles, Ramachandran distributions and the absence of atomic clashes. Such conditions are not always satisfied in Protein Ensemble Database ensembles⁹⁰. The next step is to verify that ensembles reproduce global structural parameters derived from experiments like SAS and smFRET, or estimated from sequence using polymer-physical principles^{4,45,142}. These include R_g , end-to-end distance, asphericity and scaling exponents (ν)^{143,144}. Because global parameters can be affected by secondary structure and long-range contacts, validation should proceed to more local descriptors, including smFRET-derived and PRE-derived distances, inter-residue contact probabilities^{145,146}, secondary-structure content from circular dichroism or NMR and residue-specific conformational preferences obtained from NMR observables such as chemical shifts, RDCs and nuclear Overhauser effects.

Robustness is another essential criterion. Ensembles that change substantially when subsets of input data are removed indicate limited accuracy^{122,147}. By contrast, predictive power, that is, the ability to anticipate independent experimental observables, provides strong validation¹⁴⁸.

Functional relevance is equally important. Ensembles should capture biologically meaningful features such as disorder-to-order transitions and the presence of bound-like states in the unbound ensemble¹⁴⁹. They may also reflect evolutionary conservation of global dimensions or correlations between compaction and Gene Ontology functional descriptors^{4,5}. In disease contexts, ensembles can help explain mechanisms or guide therapeutic design, for example by predicting small-molecule binding and supporting structure–activity relationship studies^{150–152}.

Finally, most ensembles are generated under *in vitro* conditions, which may not fully represent *in-cell* conformations. The ultimate benchmark for validation will, therefore, come from cellular structural data, such as *in-cell* NMR or smFRET^{50,153}, which remain in early stages of development^{40,154}.

Roadmap for conformational ensemble determination

The determination of conformational ensembles of IDPs is at a critical moment: methodologies are expanding rapidly, but standardized practices for validation, comparison and application remain elusive. To address these challenges, we propose a four-pronged roadmap: establishing a benchmarking infrastructure, developing robust approaches for uncertainty quantification, incorporating nonequilibrium and kinetic aspects, and offering practical guidelines for method selection.

IDP Ensemble Benchmarking Challenge

A major step toward advancing ensemble modeling is the creation of a transparent and standardized benchmarking initiative. Inspired by the success of CASP in protein structure prediction¹⁵⁵ and CAID in disorder prediction¹⁵⁶, we propose the IDP Ensemble Benchmarking Challenge (IDP-Bench), a recurring, community-driven assessment platform (Supplementary Fig. 1). Its biennial cycle will include selection of diverse experimental datasets by an independent committee, submission of ensemble predictions by participating groups, automated scoring against withheld validation data and continuous dissemination of results through a dedicated website.

The first edition will focus on well-characterized IDPs such as α -synuclein, Sic1, the K18 repeat and full-length tau, p27Kip1 and the measles-virus NTAIL domain, which together cover diverse sequence

features, biological functions and experimental accessibility. These proteins have extensive datasets, including SAXS, PRE, RDC, smFRET and HDX-MS. Scoring will combine global metrics such as χ^2 for SAXS profiles and Kullback–Leibler divergence for distance distributions with local measures such as PRE RMSE, RDC Q-factors and Wasserstein distances for smFRET (Supplementary Table 5). These will be combined into a weighted composite score, with scripts and error models released as open-source tools. To test predictive capability more rigorously, blind datasets such as SAXS or NMR under defined conditions will be introduced in a second phase. IDP-Bench will provide a sustainable benchmarking ecosystem that promotes transparency, drives innovation and enables long-term progress tracking.

Uncertainty in conformational ensemble determination

Benchmarking requires not only standardized data and metrics but also systematic treatment of uncertainty. We propose that uncertainty quantification be integrated directly into the IDP-Bench framework. Supplementary Fig. 2 summarizes the main error sources across experimental observables and serves as a guide for reproducibility and model fitting. Forward-model errors must also be included. For example, SPARTA+ predictions of backbone C α chemical shifts have a standard deviation of about 1 ppm on its training dataset¹⁵⁷, and SAXS predictions with CRY SOL can deviate systematically due to hydration shell modeling and background corrections^{158,159}. Where empirical error estimates are lacking, bootstrapping over modeling parameters such as fluorophore grids or rotamer libraries can provide confidence intervals.

The scoring pipeline will report both standard metrics and statistical descriptors such as reduced χ^2 , residual distributions and Bayes factors, allowing discrimination between genuine fits and overfitting. Transparency will be enforced by requiring participants to submit a checklist detailing experimental uncertainties, sources of forward-model error, bootstrap or posterior distributions of fitted parameters, and treatment of outliers. By embedding uncertainty quantification at every stage, ensemble quality can be judged not only in absolute terms but also relative to experimental noise and modeling limitations.

Dynamic properties and nonequilibrium behavior of IDPs

Many IDPs act through transient interactions and conformational switching that are not captured by equilibrium models. Progress in time-resolved techniques, including X-ray crystallography with free-electron lasers¹⁶⁰, real-time NMR¹⁶¹, time-resolved and single-molecule FRET¹⁶², time-resolved HDX-MS¹⁶³ and SAXS¹⁶⁴ and liquid-state electron microscopy¹⁶⁵ now enables access to kinetic processes. These measurements can be used as restraints for simulations that reconstruct the time evolution of ensembles.

Theoretical frameworks such as maximum entropy and its dynamical extension, maximum caliber, allow the inference of dynamic behavior consistent with experimental observations^{166,167}. These approaches are still in their early stages but provide a path toward dynamic ensemble models that more realistically capture time-dependent IDP function.

Guidelines for method selection

To support practitioners, we propose a decision framework for method selection, summarized in Supplementary Fig. 3 and Supplementary Tables 2 and 3. The process begins with evaluation of available experimental data. When no structural information exists, molecular dynamics simulations can provide baseline hypotheses. For systems constrained only by local data, knowledge-based or machine learning generators coupled with reweighting are recommended. When both local and global data such as smFRET and SAXS are available, hybrid restrained simulations or Bayesian and maximum-entropy approaches offer the best balance between fidelity and generalizability.

Biological objectives should also guide model resolution. All-atom ensembles are preferred for site-specific predictions such as binding

interfaces and mutational effects, while coarse-grained models are more suitable for large-scale processes such as phase behavior or scaling laws, where computational efficiency and statistical representativeness are critical. Resource availability further shapes method choice. On a desktop workstation, knowledge-based sampling or accelerated Monte Carlo approaches are most practical. Access to modest GPU clusters allows replica-exchange simulations or hybrid reweighting, whereas high-performance computing facilities make possible extensive all-atom replica simulations and advanced machine learning refinement strategies.

Conclusion

In the face of growing interest in disordered proteins and the understating of their structural properties, this Perspective establishes a comprehensive and practical framework to guide IDP ensemble determination, benchmarking and interpretation. By integrating community expertise across experimental and computational domains, we chart a roadmap for IDP ensemble determination, emphasize the need for uncertainty quantification and propose actionable benchmarking strategies. To implement this roadmap, we established the IDP Ensemble Benchmarking Challenge, following the examples of CASP and CAID, which greatly contributed to major advances in computational structural biology. This collaborative effort not only provides clarity on the current landscape of conformational ensemble modeling but also defines a shared vision for its future. As the field moves toward consensus and interoperability, we hope this work will serve as both a reference and a call to action, catalyzing rigorous, reproducible and biologically relevant studies of disordered protein conformational space.

References

- Piovesan, D. et al. MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Res.* **53**, D495–D503 (2025).
- Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024).
- Waterbury, A. L. et al. An autoinhibitory switch of the LSD1 disordered region controls enhancer silencing. *Mol. Cell* **84**, 2238–2254 (2024).
- Tesei, G. et al. Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).
- González-Foutel, N. S. et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022).
- Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequences. *Nat. Methods* <https://doi.org/10.1038/s41592-023-02159-5> (2024).
- Abragam, A. & Hebel, L. C. The principles of nuclear magnetism. *Am. J. Phys.* **29**, 860–861 (1961).
- Alderson, T. R., Lee, J. H., Charlier, C., Ying, J. & Bax, A. Propensity for *cis*-proline formation in unfolded proteins. *Chembiochem* **19**, 37–42 (2018).
- Murrall, M. G., Piai, A., Bermel, W., Felli, I. C. & Pierattelli, R. Proline fingerprint in intrinsically disordered proteins. *Chembiochem* **19**, 1625–1629 (2018).
- Mateos, B. et al. The ambivalent role of proline residues in an intrinsically disordered protein: from disorder promoters to compaction facilitators. *J. Mol. Biol.* **432**, 3093–3111 (2020).
- Redfield, C. Using nuclear magnetic resonance spectroscopy to study molten globule states of proteins. *Methods* **34**, 121–132 (2004).
- Alderson, T. R. & Kay, L. E. Unveiling invisible protein states with NMR spectroscopy. *Curr. Opin. Struct. Biol.* **60**, 39–49 (2020).
- Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci.* **15**, 2795–2804 (2006).
- Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* **51**, 2224–2231 (2012).
- Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR* **70**, 141–165 (2018).
- Neal, S., Nip, A. M., Zhang, H. & Wishart, D. S. Rapid and accurate calculation of protein ^1H , ^{13}C and ^{15}N chemical shifts. *J. Biomol. NMR* **26**, 215–240 (2003).
- Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
- Li, D.-W. & Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **54**, 257–265 (2012).
- Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* **131**, 13894–13895 (2009).
- Karplus, M. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.* **30**, 11–15 (1959).
- Salvador, P. Dependencies of *J*-couplings upon dihedral angles on proteins. *Annu. Rep. NMR Spectrosc.* **81**, 185–227 (2014).
- Reif, B., Hennig, M. & Griesinger, C. Direct measurement of angles between bond vectors in high-resolution NMR. *Science* **276**, 1230–1233 (1997).
- Braun, D. et al. Local structure propensities in disordered proteins from cross-correlated NMR spin relaxation. *J. Biomol. NMR* **79**, 115–127 (2025).
- Marsh, J. A., Baker, J. M. R., Tollinger, M. & Forman-Kay, J. D. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J. Am. Chem. Soc.* **130**, 7804–7805 (2008).
- Delaforge, E., Cordeiro, T. N., Bernadó, P. & Sibille, N. Conformational characterization of intrinsically disordered proteins and its biological significance. in *Modern Magnetic Resonance* (ed. Webb, G. A.) 381–399 (Springer International Publishing, 2018). https://doi.org/10.1007/978-3-319-28388-3_52
- Camilloni, C. & Vendruscolo, M. A tensor-free method for the structural and dynamical refinement of proteins using residual dipolar couplings. *J. Phys. Chem. B* **119**, 653–661 (2015).
- Bernadó, P. et al. A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proc. Natl. Acad. Sci. USA* **102**, 17002–17007 (2005).
- Abyzov, A. et al. Identification of dynamic modes in an intrinsically disordered protein using temperature-dependent NMR relaxation. *J. Am. Chem. Soc.* **138**, 6240–6251 (2016).
- Adamski, W. et al. A unified description of intrinsically disordered protein dynamics under physiological conditions using NMR spectroscopy. *J. Am. Chem. Soc.* **141**, 17817–17829 (2019).
- Salvi, N., Abyzov, A. & Blackledge, M. Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J. Phys. Chem. Lett.* **7**, 2483–2489 (2016).
- Bertoncini, C. W. et al. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc. Natl. Acad. Sci. USA* **102**, 1430–1435 (2005).

32. Mateos, B., Konrat, R., Pierattelli, R. & Felli, I. C. NMR characterization of long-range contacts in intrinsically disordered proteins from paramagnetic relaxation enhancement in ^{13}C direct-detection experiments. *Chembiochem* **20**, 335–339 (2019).
33. Kawasaki, R. & Tate, S. -I. Impact of the hereditary P301L mutation on the correlated conformational dynamics of human tau protein revealed by the paramagnetic relaxation enhancement NMR experiments. *Int. J. Mol. Sci.* **21**, 3920 (2020).
34. Salmon, L. et al. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* **132**, 8407–8418 (2010).
35. Cordeiro, T. N. et al. Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression. *Structure* **27**, 1270–1285 (2019).
36. Du, Z. et al. The sequence-structure-function relationship of intrinsic ER α disorder. *Nature* **638**, 1130–1138 (2025).
37. Kurzbach, D. et al. NMR probing and visualization of correlated structural fluctuations in intrinsically disordered proteins. *Phys. Chem. Chem. Phys.* **19**, 10651–10656 (2017).
38. Serber, Z., Ledwidge, R., Miller, S. M. & Dötsch, V. Evaluation of parameters critical to observing proteins inside living *Escherichia coli* by in-cell NMR spectroscopy. *J. Am. Chem. Soc.* **123**, 8895–8901 (2001).
39. Serber, Z. et al. High-resolution macromolecular NMR spectroscopy inside living cells. *J. Am. Chem. Soc.* **123**, 2446–2447 (2001).
40. Theillet, F. -X. et al. Structural disorder of monomeric α -synuclein persists in mammalian cells. *Nature* **530**, 45–50 (2016).
41. Felli, I. C., Gonnelli, L. & Pierattelli, R. In-cell ^{13}C NMR spectroscopy for the study of intrinsically disordered proteins. *Nat. Protoc.* **9**, 2005–2016 (2014).
42. Lerner, E. et al. FRET-based dynamic structural biology: challenges, perspectives and an appeal for open-science practices. *eLife* **10**, e60416 (2021).
43. Metskas, L. A. & Rhoades, E. Single-molecule FRET of intrinsically disordered proteins. *Annu. Rev. Phys. Chem.* **71**, 391–414 (2020).
44. Chowdhury, A., Nettels, D. & Schuler, B. Interaction dynamics of intrinsically disordered proteins from single-molecule spectroscopy. *Annu. Rev. Biophys.* **52**, 433–462 (2023).
45. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.* **45**, 207–231 (2016).
46. Holmstrom, E. D. et al. Accurate transfer efficiencies, distance distributions, and ensembles of unfolded and intrinsically disordered proteins from single-molecule FRET. *Methods Enzymol.* **611**, 287–325 (2018).
47. Zosel, F., Soranno, A., Buholzer, K. J., Nettels, D. & Schuler, B. Depletion interactions modulate the binding between disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. USA* **117**, 13480–13489 (2020).
48. Galvanetto, N. et al. Extreme dynamics in a biomolecular condensate. *Nature* **619**, 876–883 (2023).
49. Nasir, I., Onuchic, P. L., Labra, S. R. & Deniz, A. A. Single-molecule fluorescence studies of intrinsically disordered proteins and liquid phase separation. *Biochim. Biophys. Acta Proteins Proteom.* **1867**, 980–987 (2019).
50. Sustarsic, M. & Kapanidis, A. N. Taking the ruler to the jungle: single-molecule FRET for understanding biomolecular structure and dynamics in live cells. *Curr. Opin. Struct. Biol.* **34**, 52–59 (2015).
51. Ferreón, A. C. M., Gambin, Y., Lemke, E. A. & Deniz, A. A. Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proc. Natl. Acad. Sci. USA* **106**, 5645–5650 (2009).
52. Melo, A. M. et al. A functional role for intrinsic disorder in the tau-tubulin complex. *Proc. Natl. Acad. Sci. USA* **113**, 14336–14341 (2016).
53. Elbaum-Garfinkle, S. & Rhoades, E. Identification of an aggregation-prone structure of tau. *J. Am. Chem. Soc.* **134**, 16607–16613 (2012).
54. Sisamakris, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate single-molecule FRET studies using multiparameter fluorescence detection. *Methods Enzymol.* **475**, 455–514 (2010).
55. Nikić, I. & Lemke, E. A. Genetic code expansion enabled site-specific dual-color protein labeling: superresolution microscopy and beyond. *Curr. Opin. Chem. Biol.* **28**, 164–173 (2015).
56. Zosel, F., Holla, A. & Schuler, B. Labeling of proteins for single-molecule fluorescence spectroscopy. *Methods Mol. Biol.* **2376**, 207–233 (2022).
57. Hohlbein, J., Craggs, T. D. & Cordes, T. Alternating-laser excitation: single-molecule FRET and beyond. *Chem. Soc. Rev.* **43**, 1156–1171 (2014).
58. Hofmann, H. Understanding disordered and unfolded proteins using single-molecule FRET and polymer theory. *Methods. Appl. Fluoresc.* **4**, 042003 (2016).
59. Miller, J. J. et al. Accounting for fast vs slow exchange in single molecule FRET experiments reveals hidden conformational states. *J. Chem. Theory Comput.* **20**, 10339–10349 (2024).
60. Naudi-Fabra, S., Tengo, M., Jensen, M. R., Blackledge, M. & Milles, S. Quantitative description of intrinsically disordered proteins using single-molecule FRET, NMR, and SAXS. *J. Am. Chem. Soc.* **143**, 20109–20121 (2021).
61. Nettels, D. et al. Single-molecule FRET for probing nanoscale biomolecular dynamics. *Nat. Rev. Phys.* **6**, 587–605 (2024).
62. Koch, M. H., Vachette, P. & Svergun, D. I. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* **36**, 147–227 (2003).
63. Bernadó, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* **8**, 151–167 (2012).
64. Receveur-Brechot, V. & Durand, D. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.* **13**, 55–75 (2012).
65. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577 (2015).
66. Trewthella, J. Recent advances in small-angle scattering and its expanding impact in structural biology. *Structure* **30**, 15–23 (2022).
67. Konuma, T. et al. Highly collapsed conformation of the initial folding intermediates of β -lactoglobulin with non-native α -helix. *J. Mol. Biol.* **427**, 3158–3165 (2015).
68. Shkumatov, A. V., Chinnathambi, S., Mandelkow, E. & Svergun, D. I. Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering. *Proteins* **79**, 2122–2131 (2011).
69. Mahieu, E. & Gabel, F. Biological small-angle neutron scattering: recent results and development. *Acta Crystallogr. D Struct. Biol.* **74**, 715–726 (2018).
70. Krueger, S. Small-angle neutron scattering contrast variation studies of biological complexes: challenges and triumphs. *Curr. Opin. Struct. Biol.* **74**, 102375 (2022).
71. Doniach, S. Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem. Rev.* **101**, 1763–1778 (2001).
72. Flory, P. J. *Principles of Polymer Chemistry* (Cornell University Press, 1953).
73. Bernadó, P. & Blackledge, M. A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophys. J.* **97**, 2839–2845 (2009).

74. Riback, J. A. et al. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
75. Bremer, A. et al. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207 (2022).
76. Heller, W. T. Influence of multiple well defined conformations on small-angle scattering of proteins in solution. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 33–44 (2005).
77. Bernadó, P. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur. Biophys. J.* **39**, 769–780 (2010).
78. Cordeiro, T. N. et al. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr. Opin. Struct. Biol.* **42**, 15–23 (2017).
79. Masson, G. R. et al. Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **16**, 595–602 (2019).
80. Mitra, G. Emerging role of mass spectrometry-based structural proteomics in elucidating intrinsic disorder in proteins. *Proteomics* **21**, e2000011 (2021).
81. Beveridge, R. & Calabrese, A. N. Structural proteomics methods to interrogate the conformations and dynamics of intrinsically disordered proteins. *Front. Chem.* **9**, 603639 (2021).
82. Pan, X., Tran, T., Kirsch, Z. J., Thompson, L. K. & Vachet, R. W. Diethylpyrocarbonate-based covalent labeling mass spectrometry of protein interactions in a membrane complex system. *J. Am. Soc. Mass. Spectrom.* **34**, 82–91 (2023).
83. Biehn, S. E., Limpikirati, P., Vachet, R. W. & Lindert, S. Utilization of hydrophobic microenvironment sensitivity in diethylpyrocarbonate labeling for protein structure prediction. *Anal. Chem.* **93**, 8188–8195 (2021).
84. Barth, M., Bender, J., Kundlacz, T. & Schmidt, C. Evaluation of NHS-acetate and DEPC labelling for determination of solvent accessible amino acid residues in protein complexes. *J. Proteomics* **222**, 103793 (2020).
85. Miles, A. J., Drew, E. D. & Wallace, B. A. DichroIDP: a method for analyses of intrinsically disordered proteins using circular dichroism spectroscopy. *Commun. Biol.* **6**, 823 (2023).
86. Nagy, G., Hoffmann, S. V., Jones, N. C. & Grubmüller, H. Reference data set for circular dichroism spectroscopy comprised of validated intrinsically disordered protein models. *Appl. Spectrosc.* **78**, 897–911 (2024).
87. Drescher, M. EPR in protein science: intrinsically disordered proteins. *Top. Curr. Chem.* **321**, 91–119 (2012).
88. Le Breton, N. et al. Exploring intrinsically disordered proteins using site-directed spin labeling electron paramagnetic resonance spectroscopy. *Front. Mol. Biosci.* **2**, 21 (2015).
89. Weickert, S., Cattani, J., & Drescher, M. Intrinsically disordered proteins (IDPs) studied by EPR and in-cell EPR. In *Electron Paramagnetic Resonance Vol. 26* (eds Chechik, V. & Murphy, D. M.) 1–37 (2018); <https://doi.org/10.1039/9781788013888-00001>
90. Ghafouri, H. et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res.* **52**, D536–D544 (2024).
91. Gomes, G. -N. W. et al. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
92. Borthakur, K. et al. Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution. *Nat. Commun.* **16**, 9036 (2025).
93. Borgia, A. et al. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
94. Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: a perfect match?. *Science* **361**, 355–360 (2018).
95. Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. MetaInference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
96. Pesce, F. & Lindorff-Larsen, K. Refining conformational ensembles of flexible proteins against small-angle X-ray scattering data. *Biophys. J.* **121**, 1576–1579 (2022).
97. Lindorff-Larsen, K., Best, R. B. & Vendruscolo, M. Interpreting dynamically-averaged scalar couplings in proteins. *J. Biomol. NMR* **32**, 273–280 (2005).
98. Lindorff-Larsen, K. & Kragelund, B. B. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167196 (2021).
99. Ozenne, V. et al. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **28**, 1463–1470 (2012).
100. Feldman, H. J. & Hogue, C. W. V. Probabilistic sampling of protein conformations: new hope for brute force?. *Proteins* **46**, 8–23 (2002).
101. Teixeira, J. M. C. et al. IDPConformerGenerator: a flexible software suite for sampling the conformational space of disordered protein states. *J. Phys. Chem. A* **126**, 5985–6003 (2022).
102. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA* **115**, E4758–E4766 (2018).
103. Best, R. B., Zheng, W. & Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
104. Zhou, R. Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol. Biol.* **350**, 205–223 (2007).
105. Schnapka, V. et al. Atomic resolution ensembles of intrinsically disordered proteins with AlphaFold. *Nat. Commun.* <https://doi.org/10.1038/s41467-026-69172-y> (2026).
106. Brotzakis, Z. F., Zhang, S., Murtada, M. H. & Vendruscolo, M. AlphaFold prediction of structural ensembles of disordered proteins. *Nat. Commun.* **16**, 1632 (2025).
107. Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **14**, e1005941 (2018).
108. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. USA* **118**, e2111696118 (2021).
109. Cao, F., von Bülow, S., Tesei, G. & Lindorff-Larsen, K. A coarse-grained model for disordered and multi-domain proteins. *Protein Sci.* **33**, e5172 (2024).
110. Chew, P. Y., Joseph, J. A., Collepardo-Guevara, R. & Reinhardt, A. Thermodynamic origins of two-component multiphase condensates of proteins. *Chem. Sci.* **14**, 1820–1836 (2023).
111. Emenecker, R. J., Guadalupe, K., Shamoon, N. M., Sukenik, S. & Holehouse, A. S. Sequence-ensemble-function relationships for disordered proteins in live cells. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.29.564547> (2023).
112. Pesce, F. et al. Design of intrinsically disordered protein variants with diverse structural properties. *Sci. Adv.* **10**, eadm9926 (2024).
113. Krueger, R. K., Brenner, M. P. & Shrinivas, K. Generalized design of sequence–ensemble–function relationships for intrinsically disordered proteins. *Nat. Comput. Sci.* <https://doi.org/10.1038/s43588-025-00881-y> (2025).

114. Harmon, T. S. et al. GADIS: algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **29**, 339–346 (2016).
115. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
116. Lewis, S. et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* <https://doi.org/10.1126/science.adv9817> (2025).
117. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLoS Comput. Biol.* **20**, e1012144 (2024).
118. Zhang, O. et al. Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *J. Chem. Phys.* **158**, 174113 (2023).
119. Norgaard, A. B., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **94**, 182–192 (2008).
120. An, Y., Webb, M. A. & Jacobs, W. M. Active learning of the thermodynamics-dynamics trade-off in protein condensates. *Sci. Adv.* **10**, eadj2448 (2024).
121. von Bülow, S., Tesei, G., Zaidi, F. K., Mittag, T. & Lindorff-Larsen, K. Prediction of phase-separation propensities of disordered proteins from sequence. *Proc. Natl Acad. Sci. USA* **122**, e2417920122 (2025).
122. Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
123. Stöckelmaier, J. & Oostenbrink, C. Combining simulations and experiments - a perspective on maximum entropy methods. *Phys. Chem. Chem. Phys.* **27**, 14704–14717 (2025).
124. Bottaro, S., Bengtsen, T. & Lindorff-Larsen, K. Integrating molecular simulation and experimental data: a bayesian/maximum entropy reweighting approach. *Methods Mol. Biol.* **2112**, 219–240 (2020).
125. Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
126. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).
127. Nodet, G. et al. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.* **131**, 17908–17918 (2009).
128. Ferrie, J. J. & Petersson, E. J. A unified de novo approach for predicting the structures of ordered and disordered proteins. *J. Phys. Chem. B* **124**, 5538–5548 (2020).
129. Jussupow, A., Bartley, D., Lapidus, L. J. & Feig, M. COCOMO2: a coarse-grained model for interacting folded and disordered proteins. *J. Chem. Theory Comput.* **21**, 2095–2107 (2025).
130. Pesce, F. et al. Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophys. J.* **122**, 310–321 (2023).
131. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
132. Mehdi, S., Smith, Z., Herron, L., Zou, Z. & Tiwary, P. Enhanced sampling with machine learning. *Annu. Rev. Phys. Chem.* **75**, 347–370 (2024).
133. Lindorff-Larsen, K. & Ferkinghoff-Borg, J. Similarity measures for protein ensembles. *PLoS ONE* **4**, e4203 (2009).
134. Lyle, N., Das, R. K. & Pappu, R. V. A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.* **139**, 121907 (2013).
135. Lazar, T. et al. Distance-based metrics for comparing conformational ensembles of intrinsically disordered proteins. *Biophys. J.* **118**, 2952–2965 (2020).
136. Aina, A., Hsueh, S. C. C. & Plotkin, S. S. PROTHON: a local order parameter-based method for efficient comparison of protein ensembles. *J. Chem. Inf. Model.* **63**, 3453–3461 (2023).
137. Ting, D. et al. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput. Biol.* **6**, e1000763 (2010).
138. González-Delgado, J. et al. WASCO: a Wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. *J. Mol. Biol.* **435**, 168053 (2023).
139. McClendon, C. L., Hua, L., Barreiro, A. & Jacobson, M. P. Comparing conformational ensembles using the Kullback-Leibler divergence expansion. *J. Chem. Theory Comput.* **8**, 2115–2126 (2012).
140. Scalvini, B. et al. Circuit topology approach for the comparative analysis of intrinsically disordered proteins. *J. Chem. Inf. Model.* **63**, 2586–2602 (2023).
141. Lotthammer, J. M. & Holehouse, A. S. Disentangling folding from energetic traps in simulations of disordered proteins. *J. Chem. Inf. Model.* **65**, 2897–2910 (2025).
142. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024).
143. Hofmann, H. et al. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. USA* **109**, 16155–16160 (2012).
144. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **112**, 16–21 (2017).
145. Del Conte, A. et al. RING 4.0: faster residue interaction networks with novel interaction types across over 35,000 different chemical structures. *Nucleic Acids Res.* **52**, W306–W312 (2024).
146. González-Delgado, J., Bernadó, P., Neuvial, P. & Cortés, J. Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins. *Bioinformatics* **40**, btac627 (2024).
147. Thomasen, F. E. & Lindorff-Larsen, K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochem. Soc. Trans.* **50**, 541–554 (2022).
148. Tompa, P. & Varadi, M. Predicting the predictive power of IDP ensembles. *Structure* **22**, 177–178 (2014).
149. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026 (2004).
150. Iconaru, L. I. et al. Discovery of small molecules that inhibit the disordered protein, p27^{Kip1}. *Sci. Rep.* **5**, 15686 (2015).
151. Basu, S. et al. Rational optimization of a transcription factor activation domain inhibitor. *Nat. Struct. Mol. Biol.* **30**, 1958–1969 (2023).
152. Chen, Q.-H. & Krishnan, V. V. Identification of ligand binding sites in intrinsically disordered proteins with a differential binding score. *Sci. Rep.* **11**, 22583 (2021).
153. Sciolino, N., Burz, D. S. & Shekhtman, A. In-cell NMR spectroscopy of intrinsically disordered proteins. *Proteomics* **19**, e1800055 (2019).
154. Moses, D. et al. Structural biases in disordered proteins are prevalent in the cell. *Nat. Struct. Mol. Biol.* **31**, 283–292 (2024).
155. Kryshchuk, Schwede, A., Topf, T., Fidelis, M. & Moul, K. J. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins* **91**, 1539–1549 (2023).

156. Mehdiabadi, M. et al. Critical assessment of protein intrinsic disorder round 3 - predicting disorder in the era of protein language models. *Proteins* **94**, 414–424 (2026).
157. Cavender, C. E. et al. Structure-based experimental datasets for benchmarking protein simulation force fields. Preprint at <https://arxiv.org/abs/2303.11056> (2025).
158. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOLOG – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
159. Trehwella, J. et al. A round-robin approach provides a detailed assessment of biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking. *Acta Crystallogr. D Struct. Biol.* **78**, 1315–1336 (2022).
160. Hekstra, D. R. Emerging time-resolved X-ray diffraction approaches for protein dynamics. *Annu. Rev. Biophys.* **52**, 255–274 (2023).
161. Ragavan, M., Iconaru, L. I., Park, C. -G., Kriwacki, R. W. & Hilty, C. Real-time analysis of folding upon binding of a disordered protein by using dissolution DNP NMR spectroscopy. *Angew. Chem. Int. Ed. Engl.* **56**, 7070–7073 (2017).
162. Joshi, A. et al. Single-molecule FRET unmasks structural subpopulations and crucial molecular events during FUS low-complexity domain phase separation. *Nat. Commun.* **14**, 7331 (2023).
163. Seetalo, N., Zacharopoulou, M., Stephens, A. D., Kaminski Schierle, G. S. & Phillips, J. J. Millisecond hydrogen/deuterium-exchange mass spectrometry approach to correlate local structure and aggregation in α -synuclein. *Anal. Chem.* **94**, 16711–16719 (2022).
164. Martin, E. W. et al. A multi-step nucleation process determines the kinetics of prion-like domain phase separation. *Nat. Commun.* **12**, 4513 (2021).
165. Le Ferrand, H., Duchamp, M., Gabryelczyk, B., Cai, H. & Miserez, A. Time-resolved observations of liquid-liquid phase separation at the nanoscale using in situ liquid transmission electron microscopy. *J. Am. Chem. Soc.* **141**, 7202–7210 (2019).
166. Kolloff, C. & Olsson, S. Rescuing off-equilibrium simulation data through dynamic experimental data with dynAMMo. *Mach. Learn. Sci. Technol.* **4**, 045050 (2023).
167. Ji, X., Wang, H. & Liu, W. Experiment-guided refinement of milestone network. *J. Chem. Theory Comput.* **21**, 1078–1088 (2025).

Acknowledgements

This work was supported by: COST Action ML4NGP (CA21160), supported by COST (European Cooperation in Science and Technology); European Union through NextGenerationEU; PNRR project ELIXIRxNextGenIT (IR0000010); National Center for Gene Therapy and Drugs based on RNA Technology (CN00000041); European Union (101160233; HORIZON-Twinning project IDP2Biomed)

and under grant agreement no. 101182949 (HORIZON-MSCA-SE project IDPfun2); Ministry of Education, Youth and Sport of the Czech Republic, within program Inter-Excellence II, INTER-COST (LUC23081); 2022.01454.PTDC (from The Foundation for Science and Technology (FCT), Portugal); Maria de Sousa Award 45/2022 (Bial Foundation and PT Medical Association); French National Research Agency (ANR) grants ANR-22-CE45-0003 and ANR-24-CE11-2382-03; NKFIH RGH_24 (RGH 151464); ItaliaDomani PNRR project 'Potentiating the Italian Capacity for Structural Biology Services in Instruct-ERIC' (ITACA.SB, no. IR0000009); and NKFIH RGH_24 (RGH 151464).

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

All authors contributed to the discussion and the initial draft of this manuscript. H.G. wrote the final document with the help of the co-authors. All authors edited and refined the final manuscript. A.M.M. and S.C.E.T. coordinated the project.

Competing interests

K.L.-L. holds stock options in and is a consultant for Peptone. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-026-03003-2>.

Correspondence should be addressed to Silvio C. E. Tosatto or Alexander Miguel Monzon.

Peer review information *Nature Methods* thanks Anand Srivastava and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2026

¹Department of Biomedical Sciences, University of Padova, Padova, Italy. ²Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic. ³National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic. ⁴iBB—Institute for Bioengineering and Biosciences and Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal. ⁵Associate Laboratory i4HB—Institute for Health and Bioeconomy at Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal. ⁶Centre de Biologie Structurale (CBS), Université de Montpellier, INSERM and CNRS, Montpellier, France. ⁷LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France. ⁸Department of Biochemistry, ELTE Eötvös Loránd University, Budapest, Hungary. ⁹Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. ¹⁰Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark. ¹¹Johannes Kepler University, Institute of Biochemistry, Linz, Austria. ¹²Department of Molecular Immunology and Toxicology and the National Tumor Biology Laboratory, National Institute of Oncology, Budapest, Hungary. ¹³Department of Anatomy and Histology, HUN-REN-UVMB Laboratory of Redox Biology Research Group, University of Veterinary Medicine, Budapest, Hungary. ¹⁴Chemistry Coordinating Institute, University of Debrecen, Debrecen, Hungary. ¹⁵Semmelweis University, Budapest, Hungary. ¹⁶Baruch S. Blumberg Institute, Doylestown, PA, USA. ¹⁷The Wistar Institute, Philadelphia, PA, USA. ¹⁸HUN-REN Office for Supported Research Groups (TKI), Cell Cycle Laboratory,

National Institute of Oncology, Budapest, Hungary. ¹⁹Magnetic Resonance Center and Department of Chemistry “Ugo Schiff”, University of Florence, Sesto Fiorentino, Florence, Italy. ²⁰Department of Biochemistry, University of Zurich, Zurich, Switzerland. ²¹Department of Physics, University of Zurich, Zurich, Switzerland. ²²Centre de Biologie Structurale (CBS), CNRS - Univ. Montpellier - Inserm, Montpellier, France. ²³Institute of Molecular Life Sciences, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary. ²⁴Yusuf Hamied Department of Chemistry, Centre for Misfolding Diseases, University of Cambridge, Cambridge, UK. ²⁵Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czech Republic. ²⁶Structural Biology Brussels, Bio-engineering Department, Brussels, Belgium. ²⁷Interuniversity Institute of Bioinformatics in Brussels, Vrije Universiteit Brussel, Brussels, Belgium. ²⁸AI Lab, Vrije Universiteit Brussel, Brussels, Belgium. ²⁹Chemistry department, Vrije Universiteit Brussel, Brussels, Belgium. ³⁰Biomedical sciences, Vrije Universiteit Brussel, Brussels, Belgium. ✉ e-mail: silvio.tosatto@unipd.it; alexander.monzon@unipd.it



Advances in the determination of disordered protein ensemble

Hamidreza Ghafouri¹, Silvio C. E. Tosatto^{1,2} and Alexander Miguel Monzon¹

Intrinsically disordered proteins (IDPs) play essential roles in regulation, signaling, and phase separation, yet their structural complexity cannot be captured by a single conformation. Instead, they populate dynamic ensembles that encode a context-dependent function. Recent advances in experimental techniques coupled with physics-based simulations, coarse-grained models, and machine learning, have transformed our ability to generate and interpret IDP ensembles. Integrative frameworks now combine complementary data with computational approaches to refine ensembles at both local and global levels. Nevertheless, challenges remain in benchmarking, error estimation, and modeling assemblies involving protein–protein and protein–nucleic acid interactions. We highlight recent progress and outline the emerging directions that will shape the next generation of ensemble determination methods.

Addresses

¹ Department of Biomedical Sciences, University of Padova, Via Ugo Bassi 58/B, 35131, Padova, Italy

² Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, 70126, Italy

Corresponding authors: Tosatto, Silvio C.E. (silvio.tosatto@unipd.it); Monzon, Alexander Miguel (alexander.monzon@unipd.it)

Current Opinion in Structural Biology 2026, 96:103198

This review comes from a themed issue on **Folding, Binding and Protein Design** (2026)

Edited by **John Christodoulou and Michele Vendruscolo**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online xxx

<https://doi.org/10.1016/j.sbi.2025.103198>

0959-440X/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Approximately 70 % of human proteins contain at least one segment of 30 or more amino acids lacking stable secondary or tertiary structure, and ~5 % are entirely disordered [1]. These intrinsically disordered proteins and regions (IDPs/IDRs) play central roles in regulation, signaling, and phase separation, yet their behavior

cannot be captured by a single static structure. Instead, they populate dynamic conformational ensembles that encode functional versatility and context-dependent interactions [1]. While state-of-the-art predictors such as AlphaFold [2] excel for folded proteins, they fail to represent IDP heterogeneity, underscoring the need for dedicated approaches. Ensemble determination requires an integrative workflow that combines experimental techniques, computational generation, and validation (Figure 1) [3]. Advances in biophysical methods, physics-based simulations, knowledge-driven strategies, and machine learning (ML) now allow ensembles constrained at multiple scales. By merging data integration frameworks with artificial intelligence (AI), the field is moving toward quantitative and predictive descriptions of disordered protein behavior. Here, we review recent progress in experimental and computational methodologies for IDP ensemble determination.

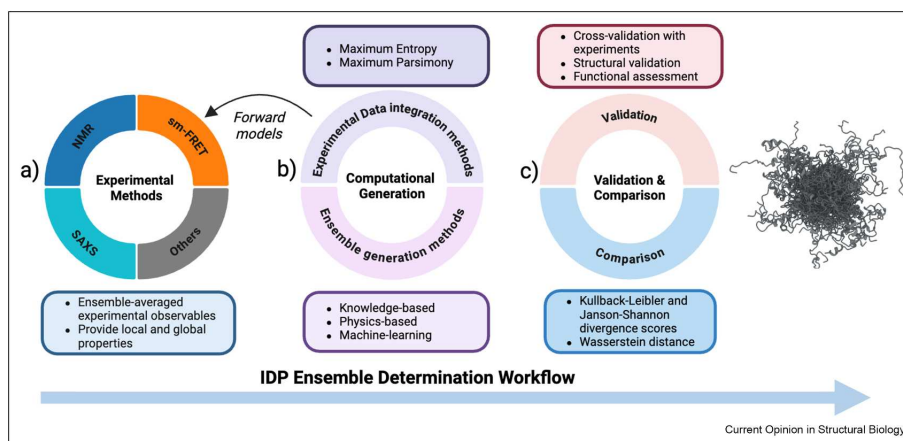
Recent advances in experimental techniques

The structural characterization of IDPs relies on experimental approaches that capture the broad conformational heterogeneity and dynamic nature of these systems. In recent years, methodological advances in nuclear magnetic resonance (NMR) spectroscopy, single-molecule Förster resonance energy transfer (smFRET), and small-angle scattering (SAS), complemented by other biophysical techniques, have significantly expanded our ability to determine conformational ensembles with higher resolution, and physiological relevance [3] (Figure 2).

NMR remains the gold standard for atomic-level insight into IDPs, capturing ensemble-averaged features such as secondary structure propensities, backbone torsion angles, and long-range contacts with minimal perturbation [3]. Recent developments in relaxation dispersion and paramagnetic techniques (e.g. PRE) have enhanced detection of sparsely populated or transiently folded states [4]. Moreover, in-cell NMR continues to mature, extending studies to physiologically relevant environments [5] (Figure 2b).

smFRET offers nanometer-scale distance measurements and uniquely resolves subpopulations and

Figure 1



Roadmap for the determination of IDP conformational ensembles. Experimental data (e.g. NMR, smFRET, SAXS) provide observables that guide computational ensemble generation. Integration methods refine ensembles using forward models, while comparison and validation ensure consistency with independent data. IDP, intrinsically disordered protein; NMR, nuclear magnetic resonance; SAXS, small-angle X-ray scattering; smFRET, single-molecule Förster resonance energy transfer.

dynamic transitions that are obscured in ensemble-averaged methods. Applied to IDPs, smFRET has characterized global chain dimensions, conformational substates, and structural rearrangements in conditions ranging from dilute solutions to crowded cellular milieus and biomolecular condensates [6]. Technical advances in site-specific fluorophore labeling, alternating/pulsed excitation schemes, and multiparameter detection have improved quantitative analysis, while integration with polymer physics models and molecular simulations has refined the extraction of full inter-dye distance distributions and dynamic reconfiguration times [7] (Figure 2a).

SAS, including both X-ray (SAXS) and neutron (SANS) scattering, provides global information on molecular size, shape, and compactness with minimal perturbation to the system in solution. For IDPs, SAS delivers key global descriptors such as the radius of gyration (R_g), scaling exponent (ν), and pairwise distance distribution $P(r)$, complementing the residue-level detail of NMR and smFRET [8]. Recent advances include enhanced beamline sensitivity, rapid data acquisition enabling kinetic studies from microseconds to hours, and improved analysis frameworks capable of extracting polymer scaling behavior directly from scattering profiles [9] (Figure 2c).

Complementary techniques have gained traction in probing additional facets of IDP structure and dynamics. Hydrogen–deuterium exchange, particularly coupled to

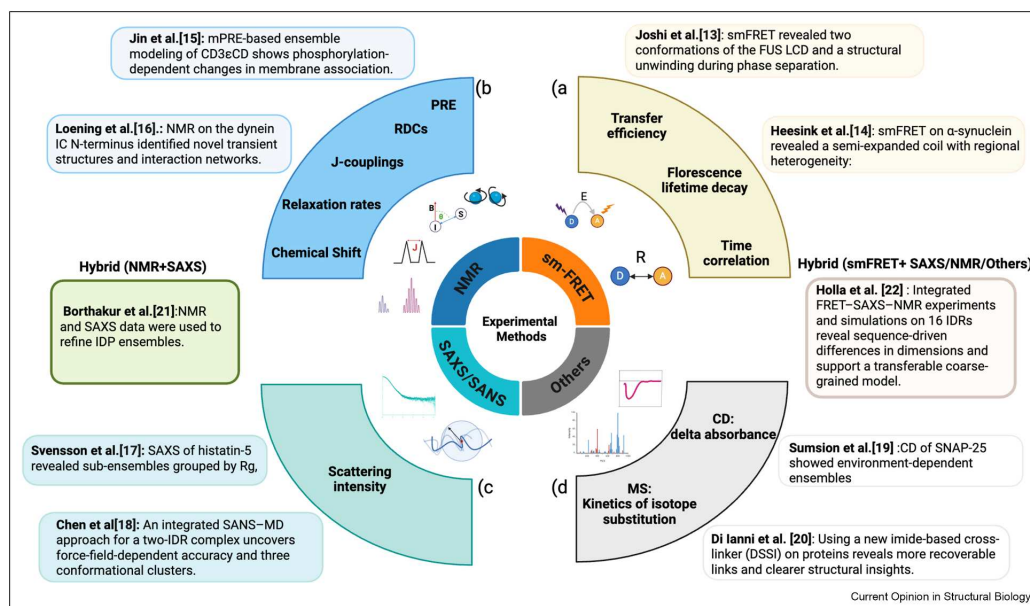
mass spectrometry, maps transient secondary structure and conformational flexibility, while covalent labeling and hydroxyl radical footprinting capture solvent exposure at side chain resolution [10] (Figure 2d).

Collectively, these methodological advances have reshaped the experimental landscape of IDP research, fostering increasingly integrative approaches to decipher sequence-encoded interactions, conformational dynamics, and the resulting structural transitions and phase separation. Recent entries in the Protein Ensemble Database (PED) [11] often combine SAXS and NMR data and/or smFRET measurements [12], frequently complemented by other orthogonal techniques such as electron paramagnetic resonance and circular dichroism (CD), to generate ensembles constrained at both global and local levels.

Recent advances in computational methods

Computational modeling of IDPs as conformational ensembles is a widely used approach for interpreting experimental data and gaining insight into their conformational heterogeneity [23]. There are three main strategies for generating conformational ensembles of IDPs (Table 1): (a) physics-based methods, primarily using molecular dynamics (MD) simulations; (b) knowledge-based techniques, which rely on heuristics or statistical patterns derived from known protein structures; and (c) ML-based approaches, particularly

Figure 2



Experimental methods for the determination of conformational ensembles of IDPs. For each technique we highlight the most important observables and recent studies to showcase the potential of each method for the determination of conformational ensembles. (a) smFRET provides intramolecular distance distributions at the nanometer scale [13,14], (b) NMR yields residue-specific information on local structure and dynamics through chemical shifts, couplings, relaxation parameters [15,16], (c) SAXS/SANS reports on overall chain dimensions and global conformational properties [17,18], (d) other techniques such as CD and MS-based techniques probe secondary structure content and long-range residue-residue distance restraints [19,20]. Moreover, **hybrid techniques** integrate complementary observables from different experimental methods to generate ensembles consistent with multiple structural restraints [21,22]. CD, circular dichroism; IDP, intrinsically disordered protein; MS, mass spectrometry; NMR, nuclear magnetic resonance; SANS, small-angle neutron scattering; SAXS, small-angle X-ray scattering; smFRET, single-molecule Förster resonance energy transfer.

generative models, which infer structural features from trained models without relying on explicit physical principles. Another category of computational techniques for determining IDP ensembles involves integrating experimental data into computational models to obtain ensembles that more accurately reflect experimental observations. These methods are broadly divided into two classes, maximum entropy and maximum parsimony [24], explained later in this section.

Ensemble generation techniques

Knowledge-based methods

Knowledge-based methods offer the fastest way to generate conformational ensembles. They construct conformers by sampling residue fragments from non-redundant, high-resolution experimental structures, using the observed torsion angle distributions in these datasets as empirical descriptors of allowable backbone geometries rather than as explicit statistical potentials or derived force fields. Early examples include TraDES [25], which used probabilistic sampling with secondary

structure biases, and flexible-meccano [26], which added IDP-specific conformational propensities and back-calculation of experimental observables. Recent tools, such as IDPConformerGenerator [27], extend earlier approaches by incorporating fragment-based $\phi/\psi/\omega$ sampling, all-atom modeling, integration of experimental restraints, and pipelines for ensemble analysis and reweighting. The main limitation of such methods is their neglect of time-dependent physical dynamics, which prevents them from capturing kinetically relevant processes and may lead to ensembles that do not fully reflect the underlying energy landscape.

Physics-based methods

In physics-based approaches, MD simulations use force fields to model protein motions over time, offering detailed insights into the dynamic behavior of IDPs.

The choice of method depends strongly on the biological question and the available computational resources. In particular, all-atom MD simulations with force fields

Table 1

Recent ensemble generation techniques, grouped into knowledge-based, physics-based, and machine learning approaches, with applications, strengths, and limitations.

Ensemble Generation Strategy	Technique	Underlying concept	Application context	Advantages	Drawbacks
Knowledge-based	MoMA [74]	Generates conformers through motion-planning algorithms that apply steric constraints and may integrate energy models.	When extensive sampling with realistic geometry is required	Incorporates all-atom representations with physically realistic structures.	Physical dynamics are not accounted for
	IDPConformer Generator [27]	Draws torsion angles from fragment libraries, potentially refined through restraints	When incorporating both sequence data and experimental constraints	Modular framework that allows reweighting and incorporation of SAXS/NMR constraints	Relies on the quality of fragments and the effectiveness of scoring functions.
	DIPEND [75]	Generates ensembles by sampling statistical dihedral angles conditioned on neighboring residues and applying structural filters	For rapid backbone ensemble generation directly from sequence with minimal requirements	Flexible and efficient, incorporating experimental priors with rapid clash resolution and refinement.	Relies on heuristics and rotamer libraries without modeling physical dynamics, offering only backbone-level sampling.
Physics-based	Amber ff99SB ws/ff03ws [76]	All-atom modeling framework refined to better capture protein–solvent interactions	For simulating IDPs where protein–water interactions are accurately balanced.	Enhanced accuracy of global dimensions and protein–protein interactions	Computationally intensive
	DES-amber [29]	All-atom force field optimized to capture protein–protein interactions	Suitable for simulating IDPs, folded proteins, and protein–protein complexes	Enhanced accuracy in global dimensions, transient helix formation, and protein–protein interactions	Limited to amber ecosystem
	MARTINI-3 [77]	Coarse-grained approach for studying complex biomolecular structures	Suitable for studying large biomolecular or membrane-associated systems	Scales efficiently to simulate large assemblies	Simplified interactions
	CALVADOS [40,41]	Implicit-solvent coarse-grained force field for IDPs and multi-domain proteins, parameterized using a top-down approach.	Efficient modeling approach that accounts for sequence effects in IDPs and multi-domain proteins	Efficient; accurate predictions of IDP compaction	Provides residue-level resolution without accounting for short-lived secondary structure or hydrogen bonds
Machine-Learning	idpGAN [42]	GAN trained on CG-MD ensembles to produce new conformers that reproduce the statistical properties of the training data	When rapid ensemble generation is needed	Reproduces realistic ensemble variability and is highly scalable	Constrained by training data patterns from CG simulations; generates C α -only ensemble representations.
	idpSAM [78]	Latent diffusion approach (transformer autoencoder plus diffusion) trained on ABSINTH implicit-solvent ensembles to generate new conformers	Designed to generate ensembles of short IDRs (<60 residues) and generalize to unseen sequences.	Strong ability to generalize to new sequences with robust training stability.	Depends on large amounts of simulation-derived training data.
	BioEmu [79]	Diffusion-based model trained on extensive datasets combining simulations and experimental measurements	Fast production of equilibrium ensembles directly from the sequence.	Rapid and sequence-aware, generating diverse ensembles characteristic of disordered proteins.	Restricted to backbone sampling without side chains; tends to overestimate Rg for longer IDPs (>100 residues).
Hybrid	AlphaFold-MetaInference [57]	Integrates AlphaFold-derived constraints with Bayesian ensemble	Aimed at integrating AI-based insights into the	Incorporates prior knowledge	Depends on AlphaFold-derived priors and confidence metrics.

Table 1 (continued)

Ensemble Generation Strategy	Technique	Underlying concept	Application context	Advantages	Drawbacks
	DynamiCE [50]	modeling, blending ML and physics-based methods. Employs a Generative recurrent neural network to create conformers, refined with experimental data using a Bayesian reward function (X-EISD).	refinement of ensembles. For situations where fixed conformer pools are inadequate and diverse, data-driven IDP ensembles are required.	Allows <i>de novo</i> generation of conformers consistent with data, integrating physical and experimental constraints with iterative improvement.	Local torsion-based sampling limits accuracy in meeting distance restraints such as NOEs and PREs.
	bAles [56]	Incorporates AlphaFold2 distograms as distance restraints into a random coil prior using Bayesian integration.	Designed to efficiently generate atomistic IDP ensembles, particularly under limited experimental data.	Rapid, atomistic-level modeling that is resilient to AF2 errors and captures both local and long-range features.	Relies on AF2 accuracy and is less effective for environment-dependent or transient interactions.
	AFlecto [80]	Constructs IDP ensembles through fragment-based sampling guided by AlphaFold models, incorporating secondary structure elements (SSEs) and steric constraints.	For efficient conformer generation directed by AlphaFold predictions with optional control over SSEs.	Rapid and adjustable, integrates partial structural information without relying on MD or force fields.	Lacks dynamic sampling; side chains are represented by pseudo-atoms.

AF2, AlphaFold2; CD, circular dichroism; IDP, intrinsically disordered protein; IDR, intrinsically disordered proteins and region; MD, molecular dynamics; NMR, nuclear magnetic resonance; Rg, radius of gyration; SAXS, small-angle X-ray scattering; smFRET, single-molecule Förster resonance energy transfer. GAN, generative adversarial network; X-EISD, extended experimental inferential structure determination; NOEs, nuclear overhauser effects; PRE, paramagnetic relaxation enhancement.

adapted for IDPs [28–30] have proven valuable for unraveling binding mechanisms with protein partners [31], clarifying sequence–ensemble–function relationships [32], assessing the impact of small-molecule drugs [33], and elucidating the molecular processes underlying liquid–liquid phase separation [34]. Although computationally demanding, recent advances in GPU-accelerated MD simulations have greatly increased the accessibility of this approach [35]. In addition, enhanced sampling techniques now allow a more efficient exploration of the complex free energy landscapes of IDPs, yielding improved convergence and more reliable ensemble descriptions [36]. Altogether, ongoing research is addressing force field inaccuracies and expanding conformational sampling through AI and ML techniques [37,38]. For further details on atomistic MD simulations applied to IDP systems, we refer readers to this review paper [39].

On the other hand, coarse-grained (CG) MD simulations provide an efficient approach to study global parameters of IDPs across larger timescales and system sizes. Recent CG models, particularly those employing data-driven force fields optimized with SAXS and NMR

data [40,41], have proven highly effective in capturing realistic global ensemble features such as radius of gyration (Rg), end-to-end distance (Ree), and the apparent Flory scaling exponent (ν). However, their reduced resolution limits the accuracy of local structural features and specific residue–residue interactions. Taken together, physics-based methods remain constrained by limited accessible timescales and the inherent approximations of current force fields which make it challenging to achieve convergence in IDP simulations.

Machine learning methods

Machine learning–based methods have almost resolved the challenge of predicting stable folded protein structures, whose conformational ensembles exhibit only limited heterogeneity. Thanks to AlphaFold [2] and other state-of-the-art models, it is now possible to predict the 3D structures of folded proteins with near-experimental accuracy using only the primary sequence. The next major challenge for AI-based approaches, however, lies in predicting protein conformational dynamics, where each sequence maps to a highly heterogeneous distribution of structural ensembles. A

growing number of ML models have recently been developed to overcome the limitations of conventional experimental and computational approaches in determining IDP ensembles. Broadly, these models can be classified into three main categories:

- a. **Sampling techniques** — it includes models that generate structural ensembles either *de novo* or by enhancing molecular simulations. These range from generative models that directly sample conformational space [42–44] to ML approaches that accelerate or guide MD trajectories by learning effective energy landscapes or proposing long timescale transitions [45,46].
- b. **Feature predictors** — it comprises models that learn to link sequence features to ensemble-averaged properties. Rather than generating full ensembles, these models predict global or local descriptors of conformational behavior, such as chain compaction, phase separation propensity, or transient secondary structure content [47–49].
- c. **Experimentally-aware models** — it includes models that integrate experimental restraints into the ML framework, either during training (e.g. incorporating NMR, SAXS, or spectroscopy data) or at the inference time by biasing the sampling process. Such approaches allow generated ensembles to be directly tuned toward agreement with experimental observables [50,51].

Detailed descriptions about these methods are available in these two reviews [52,53].

Despite these advances, ML-based approaches still struggle to fully capture IDP conformational heterogeneity as they depend heavily on training data quality and may miss physically realistic dynamics or rare yet functionally important states.

Experimental data integration methods

After generating conformational ensembles with the aforementioned methods, a common practice is to assess their agreement with experimental data (Figure 3). This process involves computing experimental observables from each conformer using *forward models* (Figure 3a) and comparing the ensemble-averaged predictions with experimental measurements (Figure 3b). In practice, however, the calculated ensembles often show poor agreement with experimental data. This mismatch typically arises from three main factors: insufficient sampling of the protein energy landscape, inaccuracies and approximations in force fields, and uncertainties or sparsity in both forward models and experimental measurements [54]. To overcome these limitations, experimental data integration techniques are employed to reweight ensembles or guide their generation, thereby ensuring that the resulting conformers more accurately

reflect experimental observations (Figure 3c). Notably, two main approaches exist for integrating experimental data with ensemble generation: maximum entropy (ME) and maximum parsimony (MP). ME methods reweight ensembles to remain as close as possible to the original prior distribution while enforcing agreement with experimental observations. This strategy preserves the inherent heterogeneity of IDPs and is particularly powerful in maintaining ensemble diversity in high-entropy systems. In contrast, MP methods aim to identify the smallest subset of conformers that can explain the data, thereby simplifying the representation of conformational space. While ME emphasizes unbiased refinement, MP favors minimal models and is best suited for systems dominated by a few major conformations. Further details on these approaches can be found in a recent review [24]. Suppl. Table 1 summarizes the recent integration methods and their respective use cases.

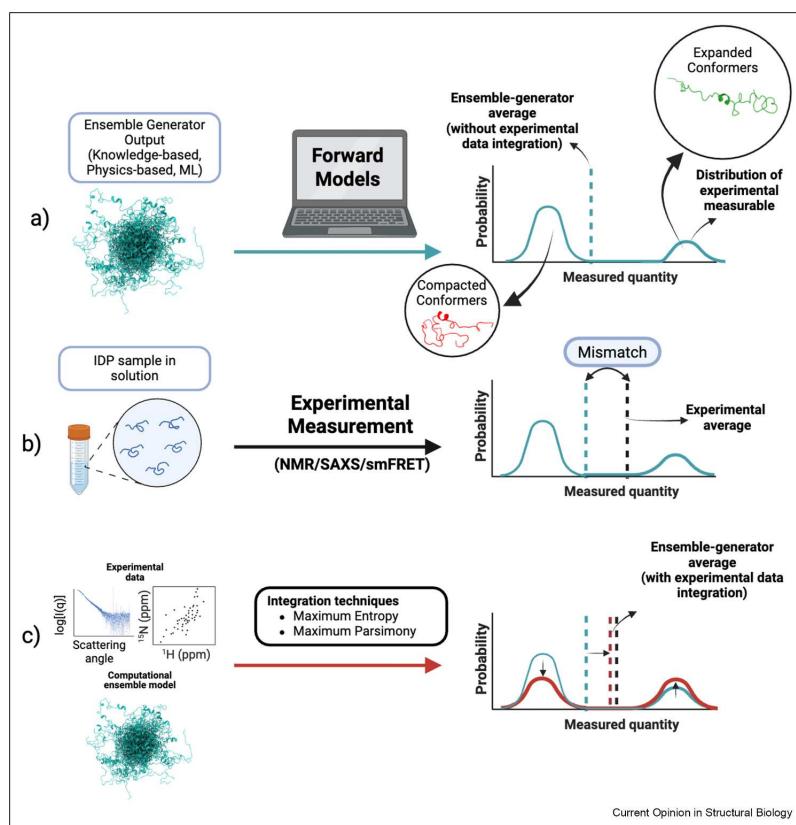
Hybrid techniques

Some computational strategies for modeling IDPs adopt hybrid approaches that combine complementary methods for ensemble generation and data integration. One direction integrates physics-based and knowledge-based sampling to balance computational efficiency with conformational diversity. A common example is fragment-based protocols refined with energy functions, which leverage both empirical knowledge and physical accuracy [55]. Another strategy merges ML with physics-based modeling, where deep learning predictions are refined within Bayesian or simulation frameworks [56,57]. A further class of approaches directly couples ensemble generation with experimental restraints, iteratively adjusting conformers to fit NMR or other measurements [50]. An additional hybrid framework combines hierarchical chain growth with MD simulations, enabling statistically rigorous and computationally efficient sampling of IDP ensembles that can be refined through integrative modeling or experimental reweighting [58]. Such flexible pipelines exploit the complementary strengths of different paradigms, enabling more realistic ensembles in cases where a single approach proves insufficient.

Comparison and validation techniques

As discussed earlier, conformational ensembles of IDPs can be derived from a variety of experimental and computational approaches. However, resolving these ensembles is inherently an ill-posed problem [3] as sparse experimental data often permit multiple solutions. This highlights the need for robust methods to compare ensembles and quantify differences between those obtained from distinct data sources. In the context of IDPs, an effective comparison metric should capture both global ensemble properties and local structural variations by reflecting differences in the underlying

Figure 3



Schematic representation of experimental data integration strategies. **(a)** Forward models calculate observables from the coordinates of each conformer in the ensemble and provide the distribution of experimental measurable in the ensemble. In this figure, we have shown R_g as an example **(b)** Ensemble-averaged value is compared to experimental measurement. In this step, we normally observe mismatches between them. The reasons are mainly insufficient sampling, uncertainties in experimental measurements and forward models, and approximations in force fields **(c)** Integration methods reweight or refine ensembles to fit the computational ensemble model to experimental data. Two main approaches for integration of experimental data are maximum entropy and maximum parsimony.

distributions rather than solely ensemble-averaged information. Such approaches are essential not only for validating ensembles derived from diverse experimental inputs but also for extracting mechanistic insights, for example, identifying binding regions or dynamic linkers, comparing homologous IDPs, and evaluating the effects of solution conditions. Next, we outline recent advances in IDP ensemble comparison techniques and introduce complementary approaches for validating conformational ensembles.

Ensemble comparison methods

A variety of approaches have been developed to compare conformational ensembles of IDPs, each capturing different aspects of their structural variability. Scalar

descriptors provide a simple measure of conformational heterogeneity, allowing ensembles to be compared via single values, though at the cost of detail [59]. Distribution-based methods extend this by analyzing differences in probability distributions of structural features, such as inter-residue distances, using similarity measures like Kullback–Leibler or Jensen–Shannon divergence, Hellinger distance, or Wasserstein distance, which also account for geometric relationships but at higher computational cost [60]. Descriptor-based approaches rely on specific structural parameters: atomic coordinates with root-mean-square deviation (RMSD), torsion angles, or other local properties, sometimes combining local and global descriptors for a more integrated view [61]. Finally, alternative

frameworks employ concepts such as topology or the detection of local energy traps to highlight unique structural features [62]. Together, these strategies enable ensemble comparison across multiple levels of resolution and complexity.

Conformational ensemble validation

Ensemble validation techniques focus on testing the reliability of IDP ensembles by combining structural checks, cross-validation against experimental data, and assessments of biological relevance. Structural validation normally begins with stereochemical checks such as bond lengths, torsion angles, Ramachandran distributions, and clash avoidance, criteria not always satisfied in PED ensembles [11]. It then extends to global parameters like radius of gyration, end-to-end distance, asphericity, and scaling exponents, derived from SAXS, smFRET, and NMR measurements [63]. Local descriptors, including inter-residue contact probabilities [64], secondary structure content (CD and NMR), and residue-specific torsion preferences (NMR: CSs, RDCs, and NOEs), provide finer evaluation. Cross-validation with experimental data can validate IDP ensembles by testing robustness to data removal [65] and by verifying the ability to predict independent observables [66]. Applications to α -synuclein, tau, and viral IDPs using NMR, SAXS, and smFRET highlight this strategy [67]. Finally, functional validation examines biological relevance, such as binding interfaces [68], conserved linker dimensions [69], or disease mechanisms revealed through small-molecule binding and structure-activity relationship (SAR) analysis [70]. Emerging in-cell approaches [71] and sequence conservation analyses [72] provide ultimate physiological benchmarks.

Future perspectives

Despite recent progress, several challenges remain before conformational ensembles of IDPs can be routinely determined and applied across biology. A first priority is the establishment of robust community-wide critical assessment (e.g. IDP-bench [3]), which will allow systematic evaluation of generation methods, forward models, and integration strategies. Notably, a recent study proposed a benchmarking protocol based on a curated set of experimental data, primarily SAXS and NMR, to assess the performance of several ensemble generation approaches including machine learning and physics-based techniques [73]. Equally important is the development of data standards for reporting and exchanging experimental information, including rigorous annotation of uncertainties and error estimates. Without explicit error modeling, ensemble refinement risks overfitting and limited reproducibility. Finally, the field must expand beyond monomeric systems to include the structural heterogeneity of protein–protein, protein–RNA, and protein–DNA assemblies, which

are central to cellular signaling and phase separation. Addressing these challenges will require coordinated efforts across experimentalists, biocurators and computational scientists, ultimately enabling more predictive, transferable, and biologically relevant ensemble descriptions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by: ELIXIR, the research infrastructure for life-science data; COST Action ML4NGP [CA21160], supported by the COST (European Cooperation in Science and Technology); European Union through NextGenerationEU; PNRR project ELIXIRxNextGenIT [IR0000010]; the National Center for Gene Therapy and Drugs based on RNA Technology [CN00000041]; co-funded by the European Union under grant agreement no. 101182949 (HORIZON-MSCA-SE project IDPfun2); views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Funding for open access charge: University of Padua.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.sbi.2025.103198>.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- * of special interest
 - ** of outstanding interest
- Holehouse AS, Kragelund BB: **The molecular basis for cellular function of intrinsically disordered protein regions.** *Nat Rev Mol Cell Biol* 2024, **25**:187–211.
 - Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al.: **Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596**:583–589.
 - Ghafari H, Kadeřávek P, Melo AM, Aspromonte MC, Bernadó P, Cortes J, Dosztányi Z, Erdos G, Feig M, Janson G, et al.: **Towards a unified framework for determining conformational ensembles of disordered proteins.** 2025, <https://doi.org/10.48550/arXiv.2504.03590>.
 - Mateos B, Konrat R, Pierattelli R, Felli IC: **NMR characterization of long-range contacts in intrinsically disordered proteins from paramagnetic relaxation enhancement in 13 C direct-**

This community-driven perspective outlines a unified framework for determining conformational ensembles of disordered proteins by integrating experimental data, computational modeling, and validation, highlighting key challenges such as force field accuracy, sampling efficiency, and benchmarking standards.

- detection experiments.** *Chembiochem Eur J Chem Biol* 2019, **20**:335–339.
5. Adamski W, Salvi N, Maurin D, Magnat J, Milles S, Jensen MR, Abyzov A, Moreau CJ, Blackledge M: **A unified description of intrinsically disordered protein dynamics under physiological conditions using NMR spectroscopy.** *J Am Chem Soc* 2019, **141**:17817–17829.
 6. Chowdhury A, Nettels D, Schuler B: **Interaction dynamics of intrinsically disordered proteins from single-molecule spectroscopy.** *Annu Rev Biophys* 2023, **52**:433–462.
- Chowdhury et al. (2023) reviewed how single-molecule fluorescence spectroscopy provides unique insights into the conformational dynamics and interaction mechanisms of IDPs, revealing disorder-facilitated binding modes, kinetic pathways, and links to processes such as aggregation and phase separation.
7. Nettels D, Galvanetto N, Ivanović MT, Nüesch M, Yang T, Schuler B: **Single-molecule FRET for probing nanoscale bi-molecular dynamics.** *Nat Rev Phys* 2024, **6**:587–605.
 8. Bernadó P, Svergun DI: **Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering.** *Mol Biosyst* 2012, **8**:151–167.
 9. Trehwella J: **Recent advances in small-angle scattering and its expanding impact in structural biology.** *Struct Lond Engl* 1993 2022, **30**:15–23.
 10. Mitra G: **Emerging role of mass spectrometry-based structural proteomics in elucidating intrinsic disorder in proteins.** *Proteomics* 2021, **21**, e2000011.
 11. Ghafouri H, Lazar T, Del Conte A, Tenorio Ku LG, Consortium PED, Tompa P, Tosatto SCE, Monzon AM: **PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins.** *Nucleic Acids Res* 2024, **52**:D536–D544.
 12. Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes G-NW, Gradinaru CC, Forman-Kay JD, Head-Gordon T: **Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states.** *Commun Chem* 2020, **3**:1–12.
 13. Joshi A, Walimbe A, Avni A, Rai SK, Arora L, Sarkar S, Mukhopadhyay S: **Single-molecule FRET unmasks structural subpopulations and crucial molecular events during FUS low-complexity domain phase separation.** *Nat Commun* 2023, **14**: 7331.
 14. Heesink G, Marseille MJ, Fakhree MAA, Driver MD, van Leijenhorst-Groener KA, Onck PR, Blum C, Claessens MMAE: **Exploring Intra- and inter-regional interactions in the IDP α -Synuclein using smFRET and MD simulations.** *Bio-macromolecules* 2023, **24**:3680–3688.
 15. Jin H, Liu D, Ni Y, Wang H, Long D: **Quantitative ensemble interpretation of membrane paramagnetic relaxation enhancement (mPRE) for studying membrane-associated intrinsically disordered proteins.** *J Am Chem Soc* 2024, **146**: 791–800.
 16. Loening NM, Jara KA, Barbar EJ: **NMR approaches to identify transient structure and interactions of intrinsically disordered dynein intermediate chain.** *J Mol Biol* 2025, <https://doi.org/10.1016/j.jmb.2025.169380>.
 17. Svensson O, Bakker MJ, Skepö M: **Deeper insight of the conformational ensemble of intrinsically disordered proteins.** *J Chem Inf Model* 2024, **64**:6105–6114.
- Svensson et al. (2024) combined SAXS with atomistic MD simulations to dissect the conformational ensemble of the IDP histatin-5 by grouping conformers by radius of gyration, providing detailed insight into structural heterogeneity and secondary structure propensities within the ensemble.
18. Chen SH, Weiss KL, Stanley C, Bhowmik D: **Structural characterization of an intrinsically disordered protein complex using integrated small-angle neutron scattering and computing.** *Protein Sci Publ Protein Soc* 2023, **32**, e4772.
 19. Sumsion JS, Shumway SW, Blocker TM, Weed TD, Chambers TM, Poland RJ, Woodbury DJ: **Multiple structural states in an intrinsically disordered protein, SNAP-25, using circular dichroism.** *Biophys J* 2025, **124**:1828–1842.
 20. Ianni AD, Ihling CH, Vranka T, Matoušek V, Sinz A, Iacobucci C: **Evaluating imide-based mass spectrometry-cleavable cross-linkers for structural proteomics studies.** *JACS Au* 2024, **4**: 2936–2943.
 21. Borthakur K, Sisk TR, Panei FP, Bonomi M, Robustelli P: **Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution.** *Nat Commun* 2025, **16**:903.
- Borthakur et al. (2024) introduced a maximum entropy reweighting procedure that integrates extensive experimental data with MD simulations to refine atomic-resolution ensembles of IDPs, achieving force field-independent conformational distributions consistent with NMR and SAXS data.
22. Holla A, Martin EW, Dannenhoffer-Lafage T, Ruff KM, König SLB, Nüesch MF, Chowdhury A, Louis JM, Soranno A, Nettels D, et al.: **Identifying sequence effects on chain dimensions of disordered proteins by integrating experiments and simulations.** *JACS Au* 2024, **4**:4729–4743.
- Holla et al. (2024) systematically investigated how sequence composition shapes the conformational ensembles of IDRs by combining single-molecule FRET, SAXS, CD, NMR, and simulations, revealing sequence-dependent differences in chain dimensions and advancing coarse-grained modeling frameworks.
23. Bottaro S, Lindorff-Larsen K: **Biophysical experiments and biomolecular simulations: a perfect match?** *Science* 2018, **361**:355–360.
 24. Stöckelmaier J, Oostenbrink C: **Combining simulations and experiments - a perspective on maximum entropy methods.** *Phys Chem Chem Phys PCCP* 2025, <https://doi.org/10.1039/d5cp01263e>.
- Stöckelmaier and Oostenbrink (2025) reviewed maximum entropy methods for integrating simulations with experimental data to optimize conformational ensembles of proteins, providing a unified mathematical framework to enhance transparency and understanding of these approaches.
25. Feldman HJ, Hogue CWV: **Probabilistic sampling of protein conformations: new hope for brute force?** *Proteins: Struct, Funct, Bioinf* 2002, **46**:8–23.
 26. Ozenne V, Bauer F, Salmon L, Huang J-R, Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M: **Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables.** *Bioinforma Oxf Engl* 2012, **28**:1463–1470.
 27. Teixeira JMC, Liu ZH, Namini A, Li J, Vernon RM, Krzeminski M, Shamandy AA, Zhang O, Haghighatlari M, Yu L, et al.: **IDPConformerGenerator: a flexible software suite for sampling the conformational space of disordered protein states.** *J Phys Chem A* 2022, **126**:5985–6003.
 28. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD: **CHARMM36m: an improved force field for folded and intrinsically disordered proteins.** *Nat Methods* 2017, **14**:71–73.
 29. Robustelli P, Piana S, Shaw DE: **Developing a molecular dynamics force field for both folded and disordered protein states.** *Proc Natl Acad Sci U S A* 2018, **115**:E4758–E4766.
 30. Piana S, Robustelli P, Tan D, Chen S, Shaw DE: **Development of a force field for the simulation of single-chain proteins and protein-protein complexes.** *J Chem Theor Comput* 2020, **16**: 2494–2507.
 31. Ghosh C, Nagpal S, Muñoz V: **Molecular simulations integrated with experiments for probing the interaction dynamics and binding mechanisms of intrinsically disordered proteins.** *Curr Opin Struct Biol* 2024, **84**, 102756.
- Ghosh, Nagpal, and Muñoz (2024) reviewed how integrative approaches combining molecular simulations with experiments reveal the transient structures, interaction dynamics, and binding mechanisms of IDPs, highlighting implications for drug design and biomolecular engineering.

32. Hicks A, Escobar CA, Cross TA: **Zhou H-X: Seqsequence-dependent correlated segments in the intrinsically disordered region of ChIZ.** *Biomolecules* 2020, **10**:946.
33. Zhao J, Blayney A, Liu X, Gandy L, Jin W, Yan L, Ha J-H, Canning AJ, Connelly M, Yang C, et al.: **EGCG binds intrinsically disordered N-terminal domain of p53 and disrupts p53-MDM2 interaction.** *Nat Commun* 2021, **12**:986.
34. Zhang Y, Prasad R, Su S, Lee D, Zhou H-X: **Amino acid-dependent phase equilibrium and material properties of tetrapeptide condensates.** *Cell Rep Phys Sci* 2024, **5**, 102218.
35. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC: **Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald.** *J Chem Theor Comput* 2013, **9**:3878–3888.
36. Gong X, Zhang Y, Chen J: **Advanced sampling methods for multiscale simulation of disordered proteins and dynamic interactions.** *Biomolecules* 2021, **11**:1416.
37. Gupta A, Dey S, Hicks A, Zhou H-X: **Artificial intelligence guided conformational mining of intrinsically disordered proteins.** *Commun Biol* 2022, **5**:610.
38. Zhu J-J, Zhang N-J, Wei T, Chen H-F: **Enhancing conformational sampling for intrinsically disordered and ordered proteins by variational autoencoder.** *Int J Mol Sci* 2023, **24**: 6896.
39. Muhammedkuty FNK, MacAinsh M, Zhou H-X: **Atomistic molecular dynamics simulations of intrinsically disordered proteins.** *Curr Opin Struct Biol* 2025, **92**, 103029.
40. Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K: **Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties.** *Proc Natl Acad Sci U S A* 2021, **118**, e2111696118.
41. Cao F, von Bülow S, Tesei G, Lindorff-Larsen K: **A coarse-grained model for disordered and multi-domain proteins.** *Protein Sci Publ Protein Soc* 2024, **33**, e5172.
42. Janson G, Valdes-Garcia G, Heo L, Feig M: **Direct generation of protein conformational ensembles via machine learning.** *Nat Commun* 2023, **14**:774.
43. Zhu J, Li Z, Zheng Z, Zhang B, Zhong B, Bai J, Hong X, Wang T, Wei T, Yang J, et al.: **Precise generation of conformational ensembles for intrinsically disordered proteins via fine-tuned diffusion models.** 2024, <https://doi.org/10.1101/2024.05.05.592611>.
44. Zhang O, Liu ZH, Forman-Kay JD, Head-Gordon T: **Deep learning of proteins with local and global regions of disorder.** 2025, <https://doi.org/10.48550/arXiv.2502.11326>.
45. Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AYK, et al.: **Scalable emulation of protein equilibrium ensembles with generative deep learning.** *Science* 2025, **389**: eadv9817.
- Lewis et al. (2025) presented BioEmu, a generative deep learning system that emulates protein equilibrium ensembles by integrating MD simulations, static structures, and stability data, enabling scalable prediction of conformational landscapes and thermodynamic properties with near-experimental accuracy.
46. Janson G, Jussupow A, Feig M: **Deep generative modeling of temperature-dependent structural ensembles of proteins.** *BioRxiv Prepr Serv Biol* 2025, <https://doi.org/10.1101/2025.03.09.642148>.
47. Tesei G, Trolle AI, Jonsson N, Betz J, Knudsen FE, Pesce F, Johansson KE, Lindorff-Larsen K: **Conformational ensembles of the human intrinsically disordered proteome.** *Nature* 2024, <https://doi.org/10.1038/s41586-023-07004-5>.
- Tesei et al. (2024) developed a coarse-grained molecular modeling framework to predict conformational ensembles for nearly all human IDRs, uncovering links between sequence features, chain compaction, cellular function, and evolutionary conservation, and releasing a proteome-wide database as a community resource.
48. Lotthammer JM, Ginell GM, Griffith D, Emenecker RJ, Holehouse AS: **Prediction of intrinsically disordered protein conformational properties from sequences.** *Nat Methods* 2024, <https://doi.org/10.1038/s41592-023-02159-5>.
- Lotthammer et al. (2024) introduced ALBATROSS, a deep learning model that predicts conformational properties of IDRs—such as radius of gyration, end-to-end distance, and scaling exponent—directly from sequence, enabling proteome-wide analysis of sequence–ensemble relationships.
49. von Bülow S, Tesei G, Zaidi FK, Mittag T, Lindorff-Larsen K: **Prediction of phase-separation propensities of disordered proteins from sequence.** *Proc Natl Acad Sci U S A* 2025, **122**, e2417920122.
50. Zhang O, Haghighatlari M, Li J, Liu ZH, Namini A, Teixeira JMC, Forman-Kay JD, Head-Gordon T: **Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data.** *J Chem Phys* 2023, **158**, 174113.
51. Fadini A, Li M, McCoy AJ, Terwilliger TC, Read RJ, Hekstra D, AlQuraishi M: **AlphaFold as a prior: experimental structure determination conditioned on a pretrained neural network.** *BioRxiv Prepr Serv Biol* 2025, <https://doi.org/10.1101/2025.02.18.638828>.
52. Janson G, Feig M: **Generation of protein dynamics by machine learning.** *Curr Opin Struct Biol* 2025, **93**, 103115.
53. von Bülow S, Tesei G, Lindorff-Larsen K: **Machine learning methods to study sequence-ensemble-function relationships in disordered proteins.** *Curr Opin Struct Biol* 2025, **92**, 103028.
54. Bonomi M, Vendruscolo M: **Determination of protein structural ensembles using cryo-electron microscopy.** *Curr Opin Struct Biol* 2019, **56**:37–45.
55. Ferrie JJ, Petersson EJ: **A Unified De Novo Approach for Predicting the Structures of Ordered and Disordered Proteins.** *J Phys Chem B* 2020, **124**:5538–5548.
56. Schnapka V, Morozova T, Sen S, Bonomi M: **Atomic resolution ensembles of intrinsically disordered and multi-domain proteins with alphafold.** 2025, <https://doi.org/10.1101/2025.06.18.660298>.
- Schnapka et al. (2025) introduced a Bayesian framework that integrates AlphaFold2 predictions with molecular mechanics force fields to generate atomic-resolution ensembles of disordered and multi-domain proteins, achieving experimental accuracy at substantially reduced computational cost.
57. Brotzakis ZF, Zhang S, Murtada MH, Vendruscolo M: **AlphaFold prediction of structural ensembles of disordered proteins.** *Nat Commun* 2025, **16**:1632.
- Brotzakis et al. (2025) developed AlphaFold-Metainference, which integrates AlphaFold-derived restraints into MD simulations to generate structural ensembles of ordered and disordered proteins, extending deep learning approaches to capture IDP conformational heterogeneity.
58. Pietrek LM, Stelzl LS, Hummer G: **Structural ensembles of disordered proteins from hierarchical chain growth and simulation.** *Curr Opin Struct Biol* 2023, **78**, 102501.
59. Lyle N, Das RK, Pappu RV: **A quantitative measure for protein conformational heterogeneity.** *J Chem Phys* 2013, **139**, 121907.
60. Aina A, Hsueh SCC, Plotkin SS: **PROTHON: a local order parameter-based method for efficient comparison of protein ensembles.** *J Chem Inf Model* 2023, **63**:3453–3461.
61. González-Delgado J, Sagar A, Zanon C, Lindorff-Larsen K, Bernadó P, Neuvial P, Cortés J: **WASCO: a Wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins.** *J Mol Biol* 2023, **435**, 168053.
62. Lotthammer JM, Holehouse AS: **Disentangling folding from energetic traps in simulations of disordered proteins.** *J Chem Inf Model* 2025, **65**:2897–2910.
63. Schuler B, Soranno A, Hofmann H, Nettels D: **Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins.** *Annu Rev Biophys* 2016, **45**:207–231.
64. Del Conte A, Camagni GF, Clementel D, Minervini G, Monzon AM, Ferrari C, Piovesan D, Tosatto SCE: **RING 4.0: faster residue interaction networks with novel interaction types across over 35,000 different chemical structures.** *Nucleic Acids Res* 2024, **52**:W306–W312.

65. Thomasen FE, Lindorff-Larsen K: **Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins.** *Biochem Soc Trans* 2022, **50**:541–554.
66. Tompa P, Varadi M: **Predicting the predictive power of IDP ensembles.** *Struct Lond Engl* 1993 2014, **22**:177–178.
67. Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, Jensen MR, Biernat J, Becker S, Mandelkow E, et al.: **Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α -synuclein in solution from NMR and small angle scattering.** *Struct Lond Engl* 2014, **22**: 238–249. 1993.
68. Piovesan D, Monzon AM, Tosatto SCE: **Intrinsic protein disorder and conditional folding in AlphaFoldDB.** *Protein Sci Publ Protein Soc* 2022, **31**, e4466.
69. González-Foutel NS, Glavina J, Borchers WM, Safranchik M, Barrera-Vilarmau S, Sagar A, Estaña A, Barozet A, Garrone NA, Fernandez-Ballester G, et al.: **Conformational buffering underlies functional selection in intrinsically disordered protein regions.** *Nat Struct Mol Biol* 2022, **29**:781–790.
70. Basu S, Martínez-Cristóbal P, Frigolé-Vivas M, Pesarro-dona M, Lewis M, Szulc E, Bañuelos CA, Sánchez-Zarzalejo C, Bielskutė S, Zhu J, et al.: **Rational optimization of a transcription factor activation domain inhibitor.** *Nat Struct Mol Biol* 2023, **30**:1958–1969.
71. Sciolino N, Burz DS, Shekhtman A: **In-Cell NMR spectroscopy of intrinsically disordered proteins.** *Proteomics* 2019, **19**, e1800055.
72. Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B: **The mutational landscape of a prion-like domain.** *Nat Commun* 2019, **10**:4162.
73. Invernizzi M, Bottaro S, Streit JO, Trentini B, Venanzi NAE, Reidenbach D, Lee Y, Dallago C, Sirelkhatim H, Jing B, et al.: **Advancing protein ensemble predictions across the order-disorder.** *Continuum* 2025, <https://doi.org/10.1101/2025.10.18.680935>.
74. Barozet A, Molloy K, Vaisset M, Zanon C, Fauret P, Siméon T, Cortés J: **MoMA-LoopSampler: a web server to exhaustively sample protein loop conformations.** *Bioinformatics* 2022, **38**: 552–553.
75. Harmat Z, Dudola D, Gáspári Z: **DIPEND: an open-source pipeline to generate ensembles of disordered segments using neighbor-dependent backbone preferences.** *Bio-molecules* 2021, **11**:1505.
76. Best RB, Zheng W, Mittal J: **Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association.** *J Chem Theor Comput* 2014, **10**:5113–5124.
77. Souza PCT, Alessandri R, Barnoud J, Thalmair S, Faustino I, Grünewald F, Patmanidis I, Abdizadeh H, Bruininks BMH, Wassenaar TA, et al.: **Martini 3: a general purpose force field for coarse-grained molecular dynamics.** *Nat Methods* 2021, **18**:382–388.
78. Janson G, Feig M: **Transferable deep generative modeling of intrinsically disordered protein conformations.** *PLoS Comput Biol* 2024, **20**, e1012144.
79. Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AYK, Satorras VG, Abidin O, Veeling BS, Zaporozhets I, et al.: **Scalable emulation of protein equilibrium ensembles with generative deep learning.** *Science* 2025, **389**. eadv9817.
80. Pajkos M, Clerc I, Zanon C, Bernadó P, Cortés J: **AFlecto: a web server to generate conformational ensembles of flexible proteins from AlphaFold models.** *J Mol Biol* 2025, <https://doi.org/10.1016/j.jmb.2025.169003>.

Chapter 3: Computational Tools for the Analysis of Conformational Ensembles

This chapter contains one article published in Protein Science.

1. Ghafouri H, Janson G, Tosatto SC, Monzon AM. IDPEnsembleTools: An open-source library for analysis of conformational ensembles of disordered proteins. Protein Science. 2026 Jan;35(1):e70427.(The appendix associated with this paper is provided in **Appendix D** at the end of the thesis.)

IDPEnsembleTools: An open-source library for analysis of conformational ensembles of disordered proteins

Hamidreza Ghafouri¹ | Giacomo Janson² | Silvio C. E. Tosatto^{1,3}  | Alexander Miguel Monzon¹

¹Department of Biomedical Sciences, University of Padova, Padova, Italy

²Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

³Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy

Correspondence

Silvio C. E. Tosatto and Alexander Miguel Monzon, Department of Biomedical Sciences, University of Padova, Padova 35121, Italy.
Email: silvio.tosatto@unipd.it; alexander.monzon@unipd.it

Funding information

HORIZON EUROPE Excellent Science, Grant/Award Number: 101182949; National Center for Gene Therapy and Drugs based on RNA Technology, Grant/Award Number: CN00000041; Elixir, the research infrastructure for life-science data; European Union through NextGenerationEU, Grant/Award Number: IR0000010; European Cooperation in Science and Technology, Grant/Award Number: COST Action ML4NGP [CA21160]

Review Editor: Nir Ben-Tal

Abstract

Intrinsically disordered proteins (IDPs) lack stable tertiary structure and instead exist as dynamic ensembles of conformations, playing essential roles in cellular regulation, signaling, and disease. As structural ensembles of IDPs become increasingly available through databases such as the Protein Ensemble Database (PED) and various computational generation methods, the need for systematic tools to analyze and compare these ensembles has grown. Here, we present IDPET (Intrinsically Disordered Protein Ensemble Tools), an open-source Python library designed to facilitate comprehensive analysis of IDP conformational ensembles. IDPET enables users to load and process ensembles from various sources and formats in parallel, compute global and local structural features, perform dimensionality reduction and clustering, and compare ensembles quantitatively using metrics based on Jensen–Shannon divergence (JSD). To demonstrate the package's functionalities, we analyze three ensembles of the unfolded drkN SH3 domain deposited in PED. This example illustrates how IDPET can extract structural descriptors, visualize conformational diversity, assess global and local features, and quantify differences between ensembles generated using distinct experimental and computational methods. By providing a reproducible and extensible framework, IDPET supports systematic exploration of ensemble features in IDPs. It is compatible with atomistic and coarse-grained models and can be easily integrated with community resources.

KEYWORDS

conformational ensembles, dimensionality reduction, intrinsically disordered proteins, Protein Ensemble Database (PED), structural ensemble analysis

1 | INTRODUCTION

Intrinsically disordered proteins and regions (collectively referred to as IDPs) play vital roles in diverse biological and pathological processes across both prokaryotic and eukaryotic organisms. Unlike folded proteins, IDPs lack stable tertiary structures under

physiological conditions, allowing them to participate in dynamic cellular functions such as signal transduction, genome maintenance, cell division, circadian rhythm regulation, and homeostasis (Cuylen et al. 2016; Duncker et al. 2001; Pelham et al. 2020; Wright and Dyson 2015).

Due to their dynamic nature, IDPs are best represented by conformational ensembles, which capture their structural variability. These ensembles consist of

Hamidreza Ghafouri and Giacomo Janson contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

rapidly interconverting states, often assigned statistical weights to reflect their relative significance (Bonomi et al. 2017). From these ensembles, various global and local descriptors such as radius of gyration (R_g), transient secondary structure content, Flory scaling exponent, solvent-accessible surface area (SASA), and end-to-end distances (R_{ee}) can be extracted (Holehouse and Kragelund 2023). These properties capture how sequence composition and contextual factors shape the conformational behavior of IDPs (Martin et al. 2020).

Importantly, ensemble properties can reveal evolutionary and functional patterns that are not readily apparent from primary sequences alone. For instance, Gonzalez-Foutel et al. (González-Foutel et al. 2022) demonstrated that the mean end-to-end distance is conserved across orthologs of a specific disordered linker, despite limited sequence similarity. Thus, while traditional multiple sequence alignments (MSAs) often fail to detect conservation in IDPs (Ruff and Pappu 2021) ensemble-level features offer a promising alternative for exploring sequence–ensemble–function relationships. This, in turn, depends on generating accurate conformational ensembles.

This typically requires the integration of experimental data, including nuclear magnetic resonance (NMR) (Gibbs et al. 2017), small-angle X-ray and neutron scattering (SAXS and SANS) (Kikhney and Svergun 2015), single-molecule Förster resonance energy transfer (sm-FRET) (Schuler et al. 2016), and cryogenic electron microscopy (Cryo-EM) (Bonomi and Vendruscolo 2019), with computational techniques such as atomistic simulations, coarse-grained models, Monte Carlo methods, and knowledge-based approaches (Joseph et al. 2021; Piana et al. 2020; Tesi et al. 2024; Vitalis and Pappu 2009). More recently, machine learning (ML) models and data-driven force fields have enabled ensemble generation directly from sequence data (Janson et al. 2023; Tesi and Lindorff-Larsen 2023), even in the absence of explicit experimental inputs.

However, this process is inherently ambiguous: most experimental measurements reflect average behaviors rather than unique conformations. Consequently, multiple distinct ensembles may be consistent with the same data (Ravera et al. 2016). A well-known example is the Sic1 protein, which has over 10 distinct ensembles archived in the Protein Ensemble Database (PED) (Ghafouri et al. 2024), each generated using different experimental or computational protocols.

To systematically analyze and compare such ensembles, we developed IDPET, “Intrinsically Disordered Protein Ensemble Tool,” a Python package designed for structural and statistical analysis of IDP ensembles. IDPET enables users to load multiple ensembles simultaneously, compute ensemble-level

properties, and visualize key features through interactive plots. The tool also includes modules for dimensionality reduction, clustering, and ensemble comparison and can retrieve relevant datasets through integrated database APIs. These capabilities allow users to quantify differences between ensembles generated by different methods or conditions. In the following sections, we describe the structure and functionality of IDPET and demonstrate its application using an example IDP from the PED.

2 | RESULTS

This section describes the general workflow and the various analyses modules in IDEPT. We use the package to analyze three IDP ensembles of the SH3 domain of drkN protein, highlighting its core features. To further demonstrate the versatility of IDPET, additional analyses were performed on systems with diverse structural and experimental characteristics. These include (i) multidomain protein ensembles containing both folded and intrinsically disordered regions and (ii) five α -synuclein ensembles encompassing both atomistic and coarse-grained representations. These examples, available as Jupyter Notebooks at <https://github.com/BioComputingUP/EnsembleTools/tree/main/notebooks>, confirm that IDPET efficiently handles a wide range of protein architectures, sizes, and ensemble types.

2.1 | IDPET workflow

The IDPET package is organized into three main modules: *ensemble*, *ensemble_analysis*, and *visualization*. The *ensemble* module is responsible for loading structural ensembles into the package. IDPET uses MDTraj (McGibbon et al. 2015) for parsing coordinate data, allowing it to support various formats, including .xtc, .dcd, and .pdb. Users can also specify a particular chain or region of interest by providing the chain ID and residue indices. Currently, only one chain per ensemble can be analyzed at a time. Only protein atoms are loaded in IDPET (e.g., solvent is excluded). Both atomistic and coarse-grained (C α -only) protein ensembles are supported. User-provided trajectory files are assumed to be free of discontinuities caused by periodic boundary conditions. If unit-cell information is available, users can let IDPET reconstruct molecules that are split across periodic boundaries via MDTraj's built-in functionalities.

Three dimensional coordinate data can be imported either from local files or directly from external databases. Currently, IDPET supports API requests from PED and ATLAS (Vander Meersche et al. 2024), allowing users to retrieve ensembles using their entry IDs in these databases (Figure 1a).

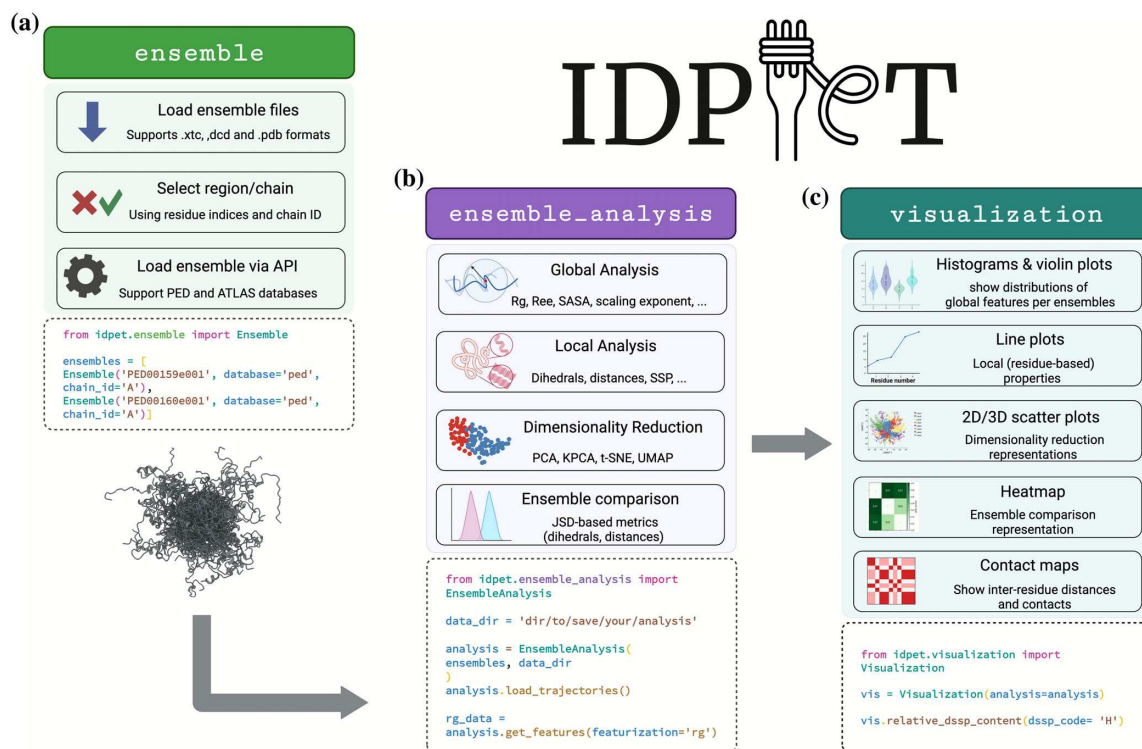


FIGURE 1 Overview of IDPET architecture and analysis workflow. (a) The *ensemble* module handles input of structural ensembles from local files (.xtc, .dcd, .pdb) or directly via API from PED and ATLAS. Users can specify chain and residue selections for targeted analysis. (b) The *ensemble_analysis* module extracts structural features at global (e.g., R_g , scaling exponent) and local levels (e.g., backbone dihedrals, distances, secondary structure propensity). It supports dimensionality reduction (PCA, KPCA, t-SNE, UMAP) and ensemble comparison using Jensen–Shannon divergence metrics. (c) The *visualization* module generates interactive plots including global feature distributions (e.g., histograms, violin plots), residue-level properties (e.g., line plots), dimensionality reduction projections (2D/3D scatter plots), contact maps, and ensemble similarity heatmaps. Example Python code snippets illustrate how each module can be used.

Once the ensemble is loaded, the *ensemble_analysis* module performs the core analysis through four main pipelines (Figure 1b):

- **Global analysis:** Extracts overall structural features, such as R_g , R_{ee} , asphericity, prolateness, global solvent-accessible surface area (SASA), and the Flory scaling exponent (Alston et al. 2023).
- **Local analysis:** Computes residue-level properties, including intramolecular distances, dihedral angles, and site-specific order and flexibility parameters that reflect local disorder (Jeschke 2024).
- **Dimensionality reduction:** Implements three algorithms, t-SNE, UMAP, and PCA, to project high-dimensional structural data (e.g., distances, angles, root mean square deviation) into 2D or 3D space. This helps compare ensembles and visualize differences in internal geometry. Additionally, conformations can be clustered using the k-nearest neighbors (KNN) algorithm (Appadurai et al. 2023).
- **Ensemble comparison:** Introduces three comparison scores based on Jensen–Shannon divergence (JSD),

which quantify differences between distributions of structural features (e.g., dihedral angles, distances) across ensembles. These scores are presented as comparison matrices.

Following this stage, data are organized either as 1D arrays (for global features, yielding one value per conformation) or as 2D arrays (for residue-level descriptors, providing one value per residue per conformation), enabling downstream analysis and visualization.

The third module, *visualization* (Figure 1c), handles graphical representation of the analysis results. All outputs from the *ensemble_analysis* module can be visualized and saved through this interface. For global features, IDPET provides histograms and violin plots. Distance-based data can be visualized using contact probability maps, with customizable thresholds and color schemes. Local features like site-specific order, flexibility, or secondary structure content are shown as line plots, where the x-axis corresponds to residue number and the y-axis to the property value. For dimensionality reduction, IDPET generates scatter plots in 2D

or 3D space. Points can be colored using various labels, such as ensemble identity, clustering results, or global features like R_g , R_{ee} , asphericity, prolateness, and SASA.

2.2 | Studying the global and local features of conformational ensembles of drkN SH3 domain

The drkN SH3 domain, derived from the Drosophila Downstream of receptor kinase (Drk) adaptor protein, is a well-characterized model system in protein folding and disorder studies. In its folded state, it adopts a canonical SH3 β -barrel structure, while its unfolded state serves as a prototypical intrinsically disordered protein (IDP). This duality makes it an ideal benchmark for assessing conformational ensemble modeling techniques. In this work, we analyze three ensembles of the unfolded drkN SH3 domain deposited in the Protein Ensemble Database (PED): PED00156e001, PED00157e001, and PED00158e001. These ensembles were generated and characterized by Lincoff et al. (Lincoff et al. 2020), each using different conformer generation strategies followed by refinement through the Extended Experimental Inferential Structure Determination (X-EISD) approach (Lincoff et al. 2020).

The ensemble PED00156e001 was constructed by selecting 100 structures from a large pool of $\sim 100,000$ conformers generated using TraDES (Feldman and Hogue 2002), a tool that systematically unfolds the native structure without applying experimental restraints. This “random” ensemble represents a broad sampling of disordered conformational space, ranging from compact to highly extended structures. PED00157e001, by contrast, consists of 100 conformers drawn from a smaller pool ($n = 1700$) biased toward fitting SAXS and NMR chemical shift data, using the ENSEMBLE (Krzeminski et al. 2013) approach to guide structure selection. The third ensemble, PED00158e001, includes 88 conformations derived from a 50:50 mixture of the TraDES and ENSEMBLE pools, refined against the same experimental data. This hybrid strategy combines broad conformational diversity with experimentally informed structure selection.

To demonstrate the capabilities of IDPET in capturing ensemble-level diversity, we initially conducted a two-part analysis focusing on both global and local structural features.

2.2.1 | Global analysis

We first examined global descriptors such as R_g , R_{ee} , asphericity, and SASA. These parameters offer a macroscopic view of each ensemble’s structural landscape. As shown in Figure 2a, the three ensembles exhibit roughly similar ensemble-averaged values for most

global features; however, their distributions differ markedly. The ensemble PED00156e001 shows broader and more heterogeneous distributions across all global metrics, reflecting its derivation from a large and unconstrained conformational pool. This suggests a greater diversity in compactness and shape, with populations spanning collapsed, intermediate, and extended states. In contrast, PED00157e001 and PED00158e001, both biased by SAXS and NMR data, exhibit narrower distributions centered around more compact, helical-like conformations. These ensembles display reduced standard deviations and lower asphericity, indicative of more homogeneous structural populations.

2.2.2 | Local analysis

Next, we investigated local conformational features at the residue level. Using IDPET, we computed the per-residue helical content for each ensemble, revealing distinct regions of transient secondary structure. PED00157e001 and PED00158e001 display prominent α -helical propensities in sequence regions 16–20, 30–45, and 50–55 (Figure 2b). These features are absent in PED00156e001, which retains the disordered coil-like characteristics typical of a purely statistical ensemble. To further probe local heterogeneity, we calculated site-specific flexibility indices based on the circular variance of backbone ϕ and ψ dihedral angles. The regions with increased helicity in PED00157e001 and PED00158e001 also show reduced flexibility (Figure 2c), suggesting the presence of partial ordering induced by the NMR chemical shift restraints used during refinement.

We complemented these analyses with Ramachandran plots to visualize the dihedral angle distributions (Figure 2d). While PED00156e001 shows a broad spread of conformations typical of disordered regions, PED00157e001 and PED00158e001 exhibit distinct enrichment in the α -helical regions of ϕ/ψ space. These observations confirm that the application of experimental restraints during ensemble generation has a significant effect on local conformational preferences, even if global averages remain similar.

In summary, our global and local analyses of the drkN SH3 domain ensembles highlight the impact of initial conformer pools and experimental data integration on the structural features of refined ensembles. IDPET enables systematic comparison of such differences and provides both macroscopic and microscopic insights into IDP ensemble characteristics.

2.3 | Mapping the conformational space of drkN SH3 domain ensembles via t-SNE

To explore how different ensemble generation strategies influence the conformational diversity of the

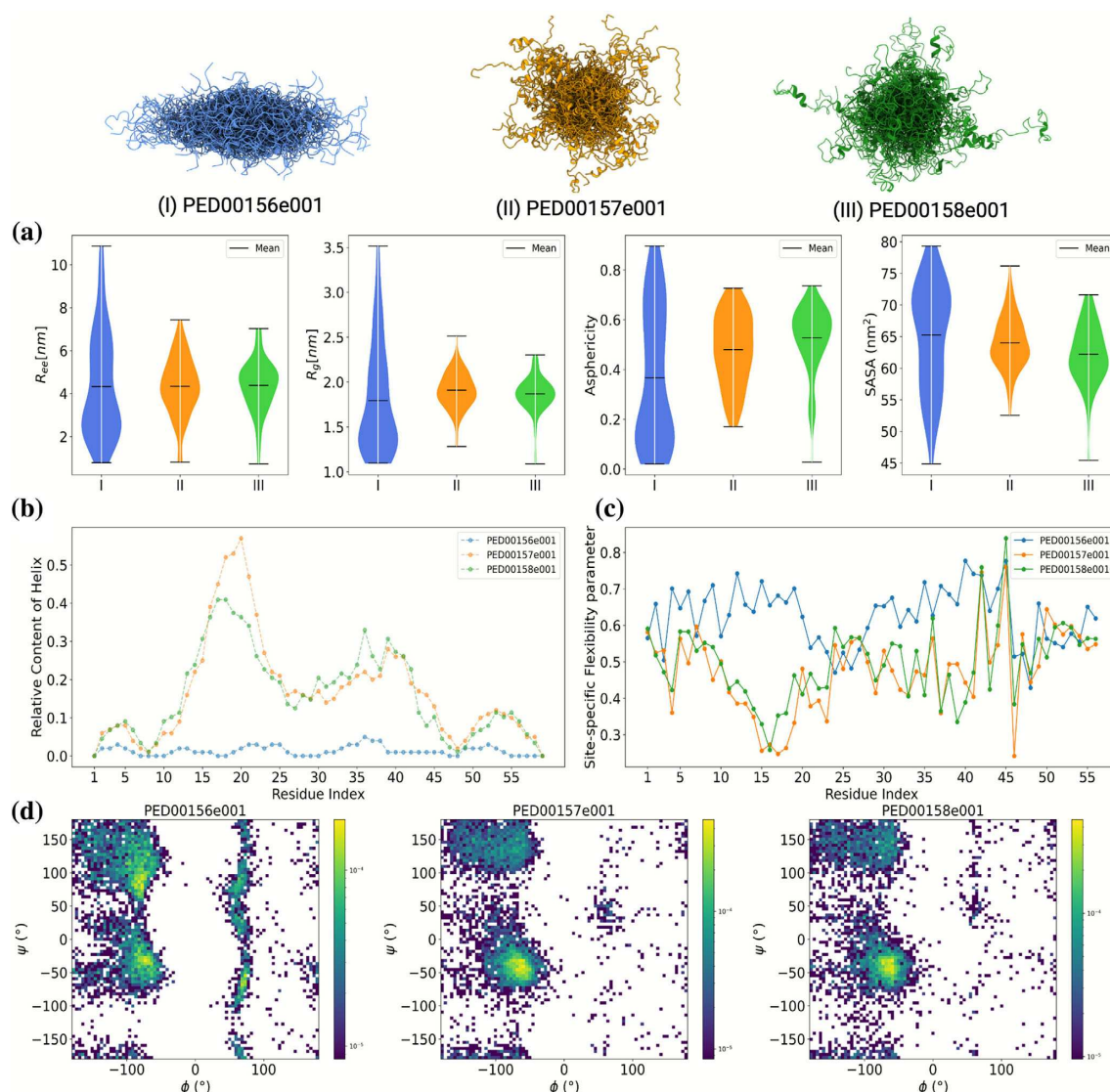


FIGURE 2 Analyzing the global and local features of the ensembles of the drkN SH3 domain in the unfolded state deposited in PED. Cartoon representations (top) illustrate the conformational ensembles PED00156e001 (I, blue), PED00157e001 (II, orange), and PED00158e001 (III, green), with colors corresponding to those used in the plots. (a) Global conformational properties analyzed with IDPET, showing distributions of the (R_g), R_{ee} , asphericity, and SASA. While the three ensembles exhibit roughly similar mean values for each descriptor, their distributions differ, reflecting distinct structural sampling. Notably, PED00156e001 was generated from a randomly generated pool by progressive unfolding of the native structure, while PED00157e001 and PED00158e001 were derived from experimentally biased pools. (b) Relative content of Alpha Helix per residue, highlighting higher local helicity in PED00157e001 and PED00158e001. (c) Site-specific flexibility profiles indicate increased conformational variability for PED00156e001, particularly in regions spanning residues 15–20 and 30–45. (d) Ramachandran plots (2D histograms of backbone dihedral angles ϕ and ψ) reveal enrichment of α -helical conformations in PED00157e001 and PED00158e001, in contrast to the broader conformational spread observed in PED00156e001.

unfolded drkN SH3 domain, we applied dimensionality reduction using *t*-distributed stochastic neighbor embedding (t-SNE) as implemented in IDPET. This technique allows high-dimensional structural data to be projected into a two-dimensional space (Appadurai

et al. 2023), offering an interpretable view of the ensemble landscape and facilitating comparisons across datasets.

For each ensemble (PED00156e001, PED00157e001, PED00158e001), we constructed a feature matrix from

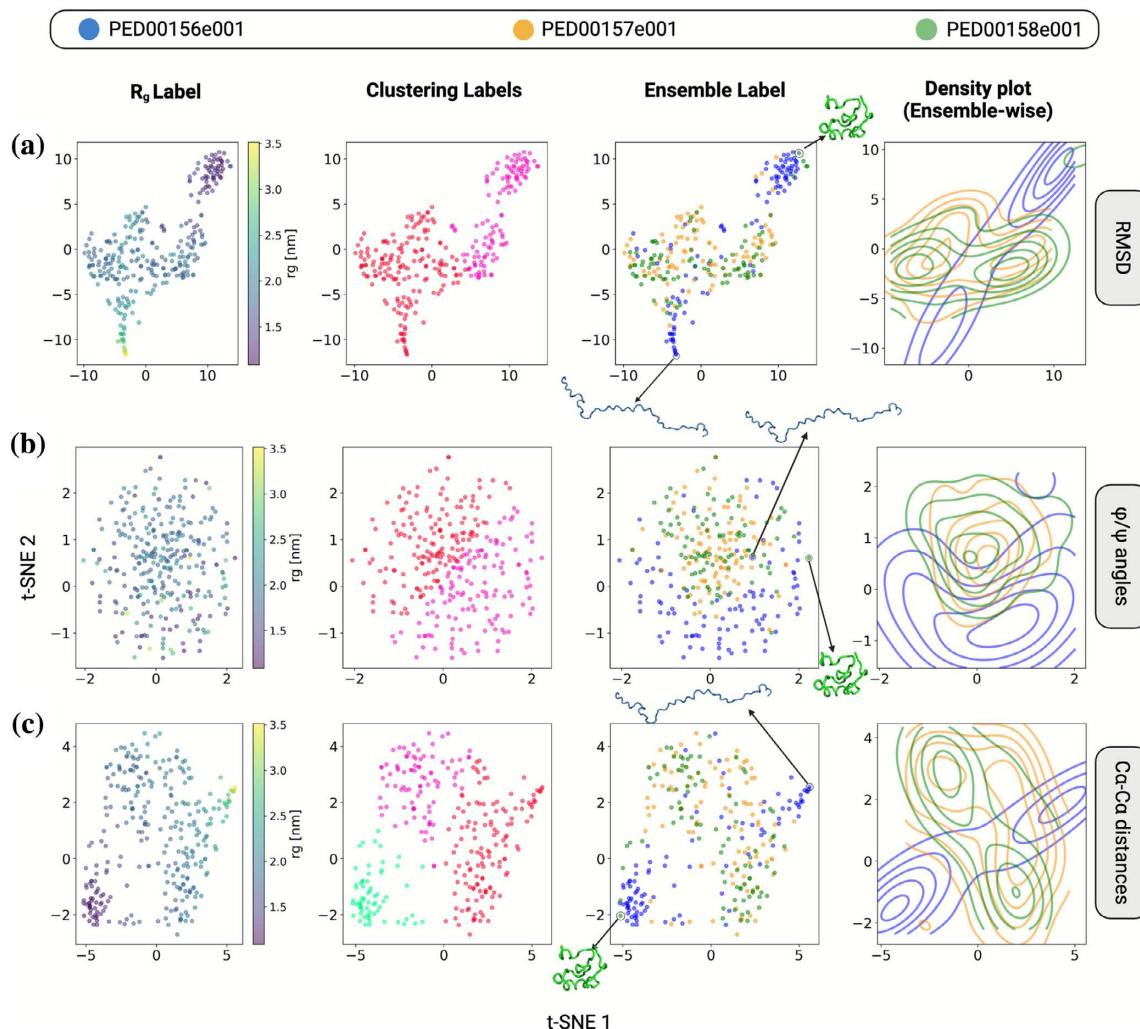


FIGURE 3 Visualization of drkN SH3 conformational ensembles in reduced-dimensional space using t-SNE projections based on different structural descriptors. Each row corresponds to a different input feature used to compute pairwise distances between conformers: (a) Ca–Ca distances, (b) RMSD, and (c) backbone dihedral angles (ϕ , ψ). These features were used to generate 2D t-SNE embeddings for three PED ensembles (PED00156e001, PED00157e001, PED00158e001), shown as columns from left to right: (1) conformers colored by radius of gyration (R_g), (2) K-nearest neighbor clustering labels, (3) original ensemble labels, and (4) kernel density estimates (KDEs) for each ensemble in t-SNE space. Panels in the third column include the same two representative structures, corresponding to a compact and an extended conformation from PED00158e001 and PED00156e001, respectively, overlaid across all rows to illustrate how these structural extremes are positioned within the t-SNE embeddings.

three types of structural descriptors: (i) pairwise root mean square deviation (RMSD) between conformers, (ii) backbone ϕ and ψ dihedral angles, and (iii) inter-residue Ca–Ca distances. These features provide complementary perspectives on structural variability, capturing both global geometric differences and local conformational states.

When using RMSD-based features, each structure is represented as a vector of its distances to all others in the ensemble. The resulting t-SNE projection (Figure 3a)

reveals that PED00157e001 and PED00158e001 occupy compact and overlapping regions in the 2D embedding, consistent with their narrower global feature distributions. In contrast, PED00156e001 spans a wider region, reflecting its broader structural heterogeneity. This observation aligns with earlier analyses, in which this ensemble showed greater variance in radius of gyration and asphericity.

To determine whether these structural embeddings could be used to discriminate between ensembles, we

applied KNN clustering to the t-SNE output. Although the RMSD-based embedding broadly separated structures by their compactness, it did not fully resolve the three ensembles, suggesting that RMSD alone lacks the resolution to capture finer structural differences.

We therefore repeated the dimensionality reduction using torsional features, specifically, the backbone ϕ and ψ dihedral angles, as input (Figure 3b). This representation more effectively distinguished the ensembles. PED00156e001 adopts a distinct distribution, while PED00157e001 and PED00158e001, both refined using similar experimental restraints, show substantial overlap. The improved resolution highlights the discriminatory power of local angular features, which capture subtle conformational biases introduced by ensemble refinement protocols.

Finally, we assessed a third feature set based on C α –C α inter-residue distances. The resulting t-SNE projection resembled the RMSD-based embedding in its global structure (Figure 3c), capturing differences in ensemble compactness but offering limited discrimination at finer levels. Clustering on this embedding produced three partially overlapping groups, which correlated well with global structural descriptors such as radius of gyration and end-to-end distance.

Taken together, these dimensionality reduction results underscore the importance of feature selection in comparative ensemble analysis. While global distance-based descriptors are useful for capturing broad structural trends, local features such as torsional angles offer enhanced sensitivity to protocol-specific biases. IDPET provides flexible access to both types of representations and facilitates systematic evaluation of conformational space across ensembles.

2.4 | Quantitative ensemble comparison using Jensen–Shannon divergence

To enable systematic comparison between conformational ensembles of disordered proteins, IDPET implements three complementary metrics based on Jensen–Shannon divergence (JSD) (Lin 2002) (Ensemble comparison section in Data S1). These scores assess the dissimilarity between feature distributions extracted from different ensembles, capturing both global and local structural differences. The three implemented metrics are:

- **adaJSD** (alpha distance average JSD): compares the distributions of all pairwise C α –C α distances, reflecting global structural differences.
- **ramaJSD** (Ramachandran JSD): evaluates the joint ϕ/ψ backbone dihedral angle distributions, sensitive to local backbone geometry. This metric applies only to all-atom ensembles.
- **ataJSD** (alpha torsion average JSD): quantifies differences in α -torsion angles formed by four consecutive

C α atoms, offering a coarse-grained measure of local structure that is compatible with both all-atom and coarse-grained ensembles.

We applied these metrics to compare the three ensembles of the drkN SH3 domain: PED00156e001 (I), PED00157e001(II), and PED00158e001(III). As shown in Figure 4, each metric provides a distinct perspective on ensemble similarity.

The adaJSD values (Figure 4a) reveal that ensembles II and III are the most similar in terms of global C α –C α distance distributions, consistent with the result of dimensionality reduction analysis. In contrast, ensemble I shows higher divergence when compared to both II and III, reflecting its broader conformational variability. This suggests that while the overall shape and compactness remain in a comparable range, ensemble I explores a more diverse set of global conformations.

By contrast, the ramaJSD scores (Figure 4b) show greater divergence, particularly in comparisons involving ensemble I. This indicates that local backbone dihedral angle distributions differ substantially between the randomly sampled ensemble and those refined using experimental restraints. These results align with our earlier observations that ensemble II and III exhibit enriched helical propensities, while ensemble I explores a broader range of disordered conformations.

The ataJSD metric (Figure 4c) reveals a similar trend to ramaJSD, though with lower absolute values. This behavior reflects the coarser resolution of α -torsion angles compared to ϕ/ψ angles, yet still confirms meaningful local structural divergence, especially in distinguishing PED00156e001 from the other two ensembles. Importantly, ataJSD offers robustness against noise in atomistic detail, making it a useful alternative when working with lower-resolution or coarse-grained models.

To assess the statistical significance of these differences, we implemented a bootstrapping strategy within IDPET. For each pair of ensembles, the metric is recalculated over multiple iterations using subsampled conformers, and the distributions are compared via a one-sided Mann–Whitney *U* test. This allows for estimation of intra- versus inter-ensemble divergence, even for ensembles with small sample sizes (Figures S1 and S2 and Bootstrap section in Data S1).

Finally, while panels a–c in Figure 4 show average JSD scores that summarize global differences between ensembles, panels d–f provide a more detailed, residue-level view. These plots show how C α –C α distances (adaJSD), Ramachandran angles (ramaJSD), and α -torsions (ataJSD) vary across the sequence when comparing ensemble I with ensemble III, as highlighted by the red boxes in Figure 4a–c. The largest differences appear in the N-terminal (residues 10–20) and C-terminal (residues 45–58) regions, which correspond to segments with transient helices in ensemble

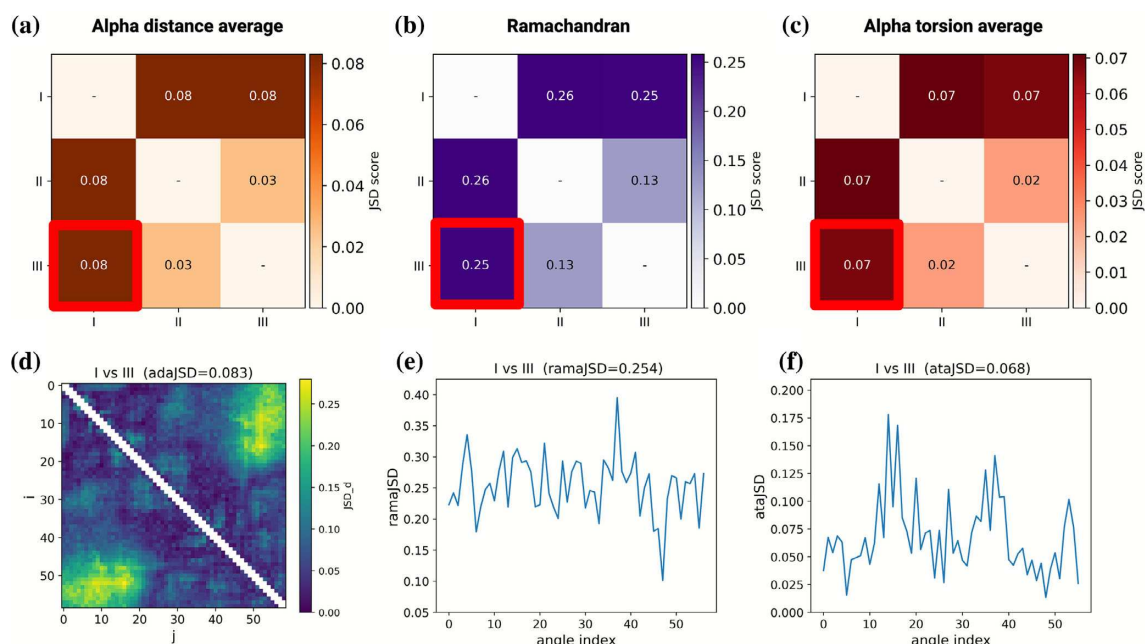


FIGURE 4 Quantitative comparison of drkN SH3 ensembles using Jensen–Shannon divergence (JSD) across structural features. Pairwise comparison between three PED ensembles (I: PED00156e001, II: PED00157e001, III: PED00158e001) using Jensen–Shannon divergence (JSD) to assess differences in structural descriptors. Top row: heatmaps showing average pairwise JSD scores between ensembles based on (a) C α –C α distances, (b) backbone dihedral (ϕ/ψ) angles, and (c) α -torsion angles. Bottom row: residue-level or feature-wise JSD profiles for the most divergent pair (I vs. III, highlighted in red). Panels show (d) a heatmap of per-residue JSD scores for C α –C α distances, (e) residue-wise JSD for dihedral angles, and (f) residue-wise JSD for α -torsion angles.

III. This suggests that using experimental restraints can lead to more structured regions in specific parts of the sequence. In contrast, the middle region is relatively consistent between the two ensembles. Together, these local profiles help explain where and how ensembles differ, beyond what global metrics can capture.

Overall, this ensemble comparison framework demonstrates that while global properties may converge across ensembles, local structural features, particularly backbone dihedral distributions, are more sensitive to the underlying modeling strategy. IDPET thus provides a multiresolution toolkit for ensemble comparison, capable of capturing subtle yet functionally relevant differences in IDP conformational landscapes.

3 | DISCUSSION

The highly dynamic nature of IDPs is best captured by structural ensembles that represent the diversity of their conformational states. Traditionally, such ensembles have been derived from molecular dynamics (MD) simulations, often combined with ensemble refinement techniques to integrate experimental data (e.g., NMR, SAXS, smFRET) (Bottaro and Lindorff-Larsen 2018).

Alternatively, knowledge-based methods such as TraDES (Feldman and Hogue 2002) and Flexible-Meccano (Ozenne et al. 2012) generate large pools of random conformers, which can then be reweighted to match experimental observations (Bernadó et al. 2007). More recently, machine learning and data-driven approaches have made it possible to generate conformational ensembles even in the absence of direct experimental input. For example, idpGAN (Janson et al. 2023) uses a generative framework to produce realistic ensembles directly from sequence data. Furthermore, new methods have emerged that generate ensembles based on AlphaFold2 (Jumper et al. 2021) predictions (Brotzakis et al. 2025; Schnapka et al. 2025; von Bülow et al. 2025a). For instance, Afflecto (Pajkos et al. 2025) utilizes AlphaFold predictions for the globular regions of proteins while sampling conformations for the disordered segments, such as loops, linkers, and terminal tails. Alternatively, data-driven force fields like CALVADOS (Tesei et al. 2021; von Bülow et al. 2025b) which are optimized using SAXS and NMR measurements of IDPs, enable the generation of realistic coarse-grained ensembles.

These recent advances have significantly expanded both experimentally derived and computationally

predicted structural ensembles. For example, the number of entries in the PED has grown from 87 in 2021 to over 500 today. In parallel, large-scale efforts have also led to the publication of predicted conformational ensembles for the entire human intrinsically disordered proteome (Tesei et al. 2024).

This increasing diversity and volume of ensembles now demand analytical frameworks that are not only flexible but also tailored to the unique challenges of IDPs, including low-resolution representations and high conformational heterogeneity.

While general-purpose libraries like MDTraj (McGibbon et al. 2015) and MDAnalysis (Michaud-Agrawal et al. 2011) offer extensive functionality for MD trajectory analysis, they are not tailored to the particularities of IDP ensemble evaluation. This has led to the development of more specialized tools. SOURSOP (Lalmansingh et al. 2023), for instance, provides IDP-specific analysis pipelines built on top of MDTraj, supporting tasks such as secondary structure assignment, contact analysis, and the calculation of various global and local properties from MD trajectories of IDP simulations. MMMx (Jeschke 2024) focuses on integrative ensemble modeling guided by DEER spectroscopy, particularly well-suited for spin-labeling studies. EnGens (Conev et al. 2023) facilitates generation, featurization, and dimensionality reduction of ensembles, streamlining the analysis of large-scale conformational datasets. A detailed comparison between IDPET and SOURSOP is provided in Table S2.

In this context, IDPET introduces a complementary and modular toolkit designed for the structural ensemble analysis of disordered proteins. Its key features include support for simultaneous multi-ensemble analysis, a comprehensive set of global and local structural descriptors, and flexible pipelines for dimensionality reduction and visualization. These capabilities make IDPET particularly well-suited for comparing ensembles generated using different computational or experimental methods. In addition to single-domain IDP ensembles, IDPET also supports multidomain proteins and disordered regions within folded proteins, which can be loaded by specifying residue ranges or chain selections. This enables targeted analyses of flexible linkers and domain–domain interactions using the same set of structural descriptors. Notably, it integrates JSD-based metrics and bootstrapping procedures to enable statistically robust comparisons of ensemble similarity, even in the presence of noise or limited sample sizes. While the drkN SH3 example serves as a validation case, the same analysis pipelines can be applied to functionally or evolutionarily related IDPs to assess how ensemble properties (e.g., compaction, flexibility, transient structure) correlate with conserved biological roles, regulatory mechanisms, or environmental conditions.

Despite these advances, important challenges remain. The accuracy and interpretability of ensemble-derived features are deeply influenced by the underlying

model quality and the choice of descriptors. Ensemble properties can vary depending on the simulation method, refinement strategy, and biophysical context, complicating direct comparisons (Ghafouri et al. 2025). Feature selection for dimensionality reduction and similarity analysis is particularly nontrivial. Linear methods like PCA may miss important nonlinear relationships, while nonlinear techniques like t-SNE and UMAP are sensitive to initialization and hyperparameter tuning. Moreover, one- or two-dimensional similarity metrics such as JSD may overlook higher-order correlations between structural features, an issue partly addressed by more computationally intensive approaches like WASCO (González-Delgado et al. 2023). IDPET addresses some of these limitations by supporting multiple structural representations—such as RMSD matrices, backbone torsions, and inter-residue distances—and by offering modular analysis pipelines adaptable to both atomistic and coarse-grained data. Its bootstrapping-based statistical evaluation framework further enhances the robustness of ensemble comparisons. Future extensions of IDPET will include integration of sequence-based descriptors to connect sequence features (e.g., charge patterning, hydropathy, and aromatic content) with ensemble-level structural properties. This will further expand IDPET into a unified framework for correlating sequence composition with conformational behavior in disordered and hybrid proteins.

4 | CONCLUSION

In summary, IDPET provides an easy-to-use and comprehensive toolbox for studying structural ensembles of IDPs. The modular characteristics of the package also give this possibility to continuously update the package based on the coming state-of-the-art methods corresponding to featurization, comparison, multidimensional projection, and new visualization settings. Moreover, the package can be deployed as an analysis pipeline for developing standalone web application interfaces. Future developments will include the integration of IDPET analysis pipelines within the PED, enabling automatic computation and visualization of ensemble-level descriptors for all deposited entries. This integration will strengthen the connection between data deposition and analysis, facilitating large-scale comparative studies across disordered protein ensembles.

5 | METHODS

IDPET was developed as a modular Python library for analyzing conformational ensembles of intrinsically disordered proteins. Its architecture comprises three main modules: *ensemble* for loading data, *ensemble_analysis* for extracting structural features, and *visualization*

for plotting outputs. Ensemble properties are computed using a combination of MDTraj-based and custom algorithms. Global features include radius of gyration, end-to-end distance, and solvent-accessible surface area; local features include dihedral angles, flexibility indices, and residue-level order parameters.

Dimensionality reduction is performed using PCA, Kernel PCA, UMAP, and t-SNE on diverse structural representations (e.g., ϕ/ψ angles, α -torsions, pairwise RMSD). To assess ensemble similarity, we implemented three Jensen–Shannon divergence (JSD)-based scores: adaJSD (global distances), ataJSD (α -torsions), and ramaJSD (backbone dihedrals). Statistical significance of comparisons can be evaluated through a bootstrapping procedure with Mann–Whitney U testing. All figures were generated using matplotlib and Plotly, with interactive options available. Additional implementation details, feature definitions, dimensionality reduction strategies, and ensemble comparison formulas are provided in detail in Methods section in Data S1.

IDPET's computational efficiency scales with its NumPy, MDTraj, and scikit-learn dependencies. On standard workstations, typical analyses such as ensemble loading, featurization, computation of global descriptors, dimensionality reduction via PCA, and JSD-based ensemble comparisons complete within seconds to minutes for ensembles comprising up to several thousand conformers. More computationally intensive tasks, including SASA calculations and t-SNE projections, may require longer runtimes that increase nonlinearly with ensemble size. Benchmarks performed on large PED ensembles (~3000 conformers) are provided in Table S1 and detailed in the online documentation.

Comprehensive documentation for the IDPET package (<https://biocomputingup.github.io/EnsembleTools/>) was developed using Sphinx, ensuring accessibility and clarity. A dedicated demo section was included to illustrate the package's capabilities in analyzing conformational ensembles of disordered proteins. Several Jupyter notebooks are also provided in the GitHub repository (<https://github.com/BioComputingUP/EnsembleTools/tree/main/notebooks>), offering examples of how to extract ensemble properties for diverse systems, including all-atom, coarse-grained, and multidomain proteins. In addition, the notebooks provide step-by-step guidance on loading ensembles into IDPET, comparing ensembles, and applying dimensionality-reduction techniques to highlight structural differences.

AUTHOR CONTRIBUTIONS

Hamidreza Ghafouri: Conceptualization; investigation; methodology; validation; visualization; writing – original draft; writing – review and editing; software; formal analysis; data curation. **Giacomo Janson:** Conceptualization; investigation; writing – original draft; methodology;

validation; visualization; writing – review and editing; software; formal analysis; data curation. **Silvio C. E. Tosatto:** Supervision; resources; project administration; writing – original draft; writing – review and editing; conceptualization; funding acquisition. **Alexander Miguel Monzon:** Writing – original draft; writing – review and editing; project administration; formal analysis; supervision; conceptualization; methodology.

ACKNOWLEDGMENTS

We gratefully acknowledge Prof. Michael Feig for his insightful discussions and guidance throughout this work. Open access publishing facilitated by Università degli Studi di Padova, as part of the Wiley – CRUI-CARE agreement.

FUNDING INFORMATION

This work has been supported by ELIXIR, the research infrastructure for life-science data; COST Action ML4NGP [CA21160], supported by COST (European Cooperation in Science and Technology); European Union through NextGenerationEU; PNRR project ELIX-IRxNextGenIT [IR0000010]; National Center for Gene Therapy and Drugs based on RNA Technology [CN00000041]; co-funded by the European Union under grant agreement no. 101182949 (HORIZON-MSCA-SE project IDPfun2); views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Funding for open access charge: University of Padua.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GitHub at <https://github.com/BioComputingUP/EnsembleTools>.

ORCID

Silvio C. E. Tosatto  <https://orcid.org/0000-0003-4525-7793>

REFERENCES

- Alston JJ, Ginell GM, Soranno A, Holehouse AS. The analytical Flory random coil is a simple-to-use reference model for unfolded and disordered proteins. *J Phys Chem B*. 2023;127:4746–60.
- Appadurai R, Koneru JK, Bonomi M, Robustelli P, Srivastava A. Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. *J Chem Theory Comput*. 2023;19:4711–27.
- Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc*. 2007;129:5656–64.

- Bonomi M, Heller GT, Camilloni C, Vendruscolo M. Principles of protein structural ensemble determination. *Curr Opin Struct Biol*. 2017;42:106–16.
- Bonomi M, Vendruscolo M. Determination of protein structural ensembles using cryo-electron microscopy. *Curr Opin Struct Biol*. 2019;56:37–45.
- Bottaro S, Lindorff-Larsen K. Biophysical experiments and biomolecular simulations: a perfect match? *Science*. 2018;361:355–60.
- Brotzakis ZF, Zhang S, Vendruscolo M. AlphaFold prediction of structural ensembles of disordered proteins. *Nat Commun*. 2025;16:1632.
- Conev A, Rigo MM, Devaurs D, Fonseca AF, Kalavadwala H, de Freitas MV, et al. EnGens: a computational framework for generation and analysis of representative protein conformational ensembles. *Brief Bioinform*. 2023;24:bbad242.
- Cuylen S, Blaukopf C, Politi AZ, Müller-Reichert T, Neumann B, Poser I, et al. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature*. 2016;535:308–12.
- Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19:26–59.
- Feldman HJ, Hogue CWV. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins Struct Funct Bioinform*. 2002;46:8–23.
- Ghafouri H, Kadeřávek P, Melo AM, Aspromonte MC, Bernadó P, Cortes J, et al. Towards a unified framework for determining conformational ensembles of disordered proteins. 2025. [bioRxiv. <https://doi.org/10.48550/arXiv.2504.03590>](https://doi.org/10.48550/arXiv.2504.03590)
- Ghafouri H, Lazar T, del Conte A, Tenorio Ku LG, PED Consortium, Aspromonte MC, et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res*. 2024;52:D536–44.
- Gibbs EB, Cook EC, Showalter SA. Application of NMR to studies of intrinsically disordered proteins. *Arch Biochem Biophys*. 2017;628:57–70.
- González-Delgado J, Sagar A, Zanon C, Lindorff-Larsen K, Bernadó P, Neuvial P, et al. WASCO: a Wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. *J Mol Biol*. 2023;435:168053.
- González-Foutel NS, Glavina J, Borchers WM, Safranchik M, Barrera-Vilarmau S, Sagar A, et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat Struct Mol Biol*. 2022;29:781–90.
- Holehouse AS, Kragelund BB. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol*. 2023;25:187–211. <https://doi.org/10.1038/s41580-023-00673-0>
- Janson G, Valdes-Garcia G, Heo L, Feig M. Direct generation of protein conformational ensembles via machine learning. *Nat Commun*. 2023;14:774.
- Jeschke G. Protein ensemble modeling and analysis with MMMx. *Protein Sci*. 2024;33:e4906.
- Joseph JA, Reinhardt A, Aguirre A, Chew PY, Russell KO, Espinosa JR, et al. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci*. 2021;1:732–43.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.
- Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett*. 2015;589:2570–7.
- Krzeminski M, Marsh JA, Neale C, Choy W-Y, Forman-Kay JD. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*. 2013;29:398–9.
- Lalmansingh JM, Keeley AT, Ruff KM, Pappu RV, Holehouse AS. SOURSOP: a Python package for the analysis of simulations of intrinsically disordered proteins. *J Chem Theory Comput*. 2023;19:5609–20.
- Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory*. 2002;37:145–51.
- Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes G-NW, Gradinaru CC, et al. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun Chem*. 2020;3:1–12.
- Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. 2020;367:694–9.
- McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys J*. 2015;109:1528–32.
- Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem*. 2011;32:2319–27.
- Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, et al. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*. 2012;28:1463–70.
- Pajkos M, Clerc I, Zanon C, Bernadó P, Cortés J. Afflecto: a web server to generate conformational ensembles of flexible proteins from AlphaFold models. *J Mol Biol*. 2025;437:169003. <https://doi.org/10.1016/j.jmb.2025.169003>
- Pelham JF, Dunlap JC, Hurley JM. Intrinsic disorder is an essential characteristic of components in the conserved circadian circuit. *Cell Commun Signal*. 2020;18:181. <https://doi.org/10.1186/s12964-020-00658-y>
- Piana S, Robustelli P, Tan D, Chen S, Shaw DE. Development of a force field for the simulation of single-chain proteins and protein–protein complexes. *J Chem Theory Comput*. 2020;16:2494–507.
- Ravera E, Sgheri L, Parigi G, Luchinat C. A critical assessment of methods to recover information from averaged data. *Phys Chem Chem Phys*. 2016;18:5686–701.
- Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433:167208.
- Schnapka V, Morozova T, Sen S, Bonomi M. Atomic resolution ensembles of intrinsically disordered and multi-domain proteins with AlphaFold. 2025. [bioRxiv. <https://doi.org/10.1101/2025.06.18.660298>](https://doi.org/10.1101/2025.06.18.660298)
- Schuler B, Soranno A, Hofmann H, Nettels D. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu Rev Biophys*. 2016;45:207–31.
- Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res Eur*. 2023;2:94.
- Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc Natl Acad Sci U S A*. 2021;118:e2111696118.
- Tesei G, Trolle AI, Jonsson N, Betz J, Knudsen FE, Pesce F, et al. Conformational ensembles of the human intrinsically disordered proteome. *Nature*. 2024;626:897–904.
- Vander Meersche Y, Cretin G, Gheeraert A, Gelly J-C, Galochkina T. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res*. 2024;52:D384–92.
- Vitalis A, Pappu RV. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J Comput Chem*. 2009;30:673–99.
- von Bülow S, Johansson KE, Lindorff-Larsen K. AF-CALVADOS: AlphaFold-guided simulations of multi-domain proteins at the proteome level. 2025a. [bioRxiv. <https://doi.org/10.1101/2025.10.19.683306>](https://doi.org/10.1101/2025.10.19.683306)

von Bülow S, Yasuda I, Cao F, Schulze TK, Trolle AI, Rauh AS, et al. Software package for simulations using the coarse-grained CALVADOS model. 2025b. bioRxiv. <https://doi.org/10.48550/arXiv.2504.10408>

Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2015;16:18–29.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ghafouri H, Janson G, Tosatto SCE, Monzon AM. IDPEnsembleTools: An open-source library for analysis of conformational ensembles of disordered proteins. *Protein Science*. 2026;35(1):e70427. <https://doi.org/10.1002/pro.70427>

3– Discussion

The Protein Ensemble Database: Development, Expansion, and Methodological Trends

PED was introduced approximately ten years ago with the goal of collecting conformational ensembles of intrinsically disordered and unfolded proteins[88]. The first release of PED contained 60 conformational ensembles determined using experimentally constrained computational approaches. In this initial release, the experimental techniques most frequently used to derive conformational ensembles were NMR spectroscopy and SAXS. From a computational perspective, conformational ensembles have been derived either from MD simulations or from statistical conformer generators, such as Flexible-meccano[89], followed by ensemble refinement using reweighting or selection approaches implemented in methods such as ENSEMBLE[90] or ASTEROIDS[91, 92]. Approximately seven years after the first release, the second major update of PED was published in 2021, coinciding with the transfer of the database to the University of Padova[93]. In this version, a new submission service was introduced for the first time to ensure data quality through a combination of automated validation pipelines and manual curation steps. Moreover, new fields were added to each entry, including cross-references to other databases such as BMRB[94], SASBDB[95] and DisProt[74]. Most notably, this release expanded the number of experimentally defined ensembles from 60 to 215. Moreover, ensembles derived from multiple experimental restraints were more prevalent in this release. For instance, Gomes *et al.* deposited conformational ensembles for the Sic1 protein defined through the integration of three different experimental techniques, including NMR, SAXS, and smFRET[96].

Finally, the latest update of PED was released in 2024[80], to which I contributed as part of my PhD work. In this update, the PED deposition service was substantially improved, not only at the UI/UX front-end level but also at the back-end. Specifically, the deposition pipeline was integrated with DRMAAtic[97], which acts as a mediator between the deposition service and high-performance computing (HPC) clusters to coordinate and automate multiple tasks, such as the execution of validation pipelines required during the deposition process. In addition, the deposition service was connected to the ORCID system to improve contributor identification and attribution. Another major update in this release was the integration of NMR-derived ensembles containing substantial disordered regions from the PDB into PED. To achieve this, we performed a systematic search in MobiDB[75], using the mobility feature[98] as a key filtering criterion, and subsequently deposited the selected NMR ensembles into PED following manual inspection.

Another novel aspect of this update was the integration of conformational ensembles generated using machine-learning and data-driven approaches, including idpGAN[99] and the IDPConformerGenerator (IDPCG)[100]. In total, 14 ensembles generated with IDPCG and 47 ensembles generated with idpGAN were deposited, demonstrating the capability of such methods to produce conformational ensembles of IDPs without the direct use of experimental data. The accuracy of these ensembles with respect to experimental data could be further evaluated whenever the experimental

data is available. For example, the accuracy of different ensembles for α -synuclein showing that the ensembles generated with IDPCG (PED00426) and a99SB-disp force field (PED00545) are in good agreement with the experimental SAXS data. The details of this evaluation are provided in figure 11.

Currently, the number of experimentally derived ensembles in PED has increased to 628. While NMR and SAXS remain the most commonly used experimental techniques for the determination of conformational ensembles, the growing use of smFRET represents a notable emerging trend in recent PED entries. Moreover, the recurrent combination of SAXS, NMR, and smFRET highlights the importance of integrating multiple sources of information to obtain more reliable structural ensembles of IDPs[80].

Lastly, in addition to integrating multiple sources of experimental data, advances on the computational side, particularly the development of force fields specifically designed and optimized for IDPs, have played a significant role in determining more accurate conformational ensembles. A representative example is the recent deposition of 15 ensembles by Borthakur *et al.*[101], in which three different force fields optimized for IDPs were combined with experimental data from NMR and SAXS to determine conformational ensembles for five different IDPs. These ensembles are available in PED under accession numbers PED00531 to PED00545.

Scalable Deposition and Integration of Predicted Conformational Ensembles:

While the primary focus of PED has traditionally been the storage and annotation of experimentally derived conformational ensembles, recent advances in artificial intelligence-based methods[78, 102, 99, 103, 104, 105] have introduced computational techniques capable of learning the conformational behavior of IDPs from data derived from molecular dynamics simulations and experimental measurements. These approaches can subsequently be used to generate conformational ensembles for large sets of proteins at scale. To accommodate this development within PED, I implemented a batch deposition pipeline that enables the automated submission of large numbers of ensembles. From a back-end perspective, this pipeline leverages DRMAA^{tic} to maximize the use of available HPC resources by distributing multiple jobs, including validation pipelines, while using the PED deposition APIs to insert and manage the associated information for each ensemble.

To demonstrate the capability of this pipeline, we collaborated with Prof. Julie Forman-Kay’s laboratory to deposit a large-scale dataset comprising conformational ensembles for approximately 7,000 human proteins, generated using an AI-based method[78], into PED. All ensembles within this category are accessible by searching for the “AlphaFlex” tag in PED. It should be noted that, since these ensembles are not explicitly compared with experimental data, any mechanistic interpretation or analysis of their ensemble properties should be made with caution.

In another collaboration with the UniProt[106] consortium, PED entries were linked to the corresponding UniProt entries for proteins for which conformational ensembles are available in PED. This integration allows users, in addition to accessing predicted structures from AlphaFold[107] and experimentally derived structures from the PDB,

to directly identify and explore available conformational ensemble data through UniProt when such data are present in PED.

Next steps and future perspective of PED:

As the PED continues to grow in size, scope, and methodological diversity, a key challenge for its future development is to ensure clarity, interpretability, and usability of the deposited ensemble data. In particular, the increasing coexistence of experimentally derived and purely predicted conformational ensembles highlights the need for clearer separation, annotation, and downstream support for different ensemble types. Addressing these challenges is essential for maintaining PED as a reliable and scientifically meaningful resource for the community.

A first important future direction is the explicit distinction between experimentally derived ensembles and predicted ensembles within the PED framework. While both ensemble types are valuable, they differ fundamentally in terms of data provenance, validation, and appropriate use cases. Experimentally derived ensembles are constrained directly by experimental measurements and therefore support mechanistic interpretation within the limits of the underlying data. In contrast, predicted ensembles, particularly those generated using machine-learning-based approaches, are not necessarily validated against experimental observables and should primarily be viewed as hypotheses or exploratory representations of conformational behavior. Providing clearly separated sections or standardized labels for these two ensemble categories would help users immediately assess the level of experimental support and avoid inappropriate interpretation. Such a distinction would also facilitate benchmarking studies and method comparisons by enabling systematic filtering based on ensemble origin.

A second major perspective concerns the analysis and visualization of ensemble properties. At present, PED primarily serves as a repository for storing ensemble coordinates and associated metadata. While this is a necessary foundation, the increasing volume and complexity of ensemble data call for richer, standardized ways to summarize and compare ensemble properties directly within the database. Integrating ensemble-level descriptors, such as global properties, flexibility profiles, and inter-residue distance statistics, would substantially enhance the interpretability of PED entries, particularly for users who are not specialists in ensemble modeling. In this context, the ensemble analysis framework developed in this thesis (IDPET) (66) provides a basis for future extensions of PED. IDPET enables systematic extraction of global and local ensemble properties, quantitative comparison between ensembles, and statistically grounded assessment of ensemble similarity. Incorporating selected IDPET-derived descriptors into PED would allow users to explore ensemble behavior without downloading and post-processing large coordinate sets, thereby lowering the barrier to ensemble-based analysis. Such integration would also promote consistency across studies by encouraging the use of standardized metrics and representations. Beyond static descriptors, interactive visualization tools represent another promising avenue for development. Visual summaries of ensemble dimensions, residue-resolved flexibility, or distance distributions could provide intuitive entry points into complex ensemble data, complementing traditional coordinate-based visualization. Impro-

tantly, these tools should be designed to emphasize ensemble distributions rather than representative conformers, reinforcing the ensemble-centric perspective that underlies PED. Another important direction for the future development of PED is the inclusion of quantitative fitting scores reporting the agreement between deposited ensembles and available experimental data. Whenever possible, PED could provide standardized metrics indicating how well an ensemble reproduces experimental observables. Such scores would allow users to quickly assess the experimental consistency of an ensemble and facilitate comparison between alternative models of the same protein.

An illustrative example is α -synuclein, for which multiple ensembles are available in PED (PED00426, PED00477, PED00543, PED00544, and PED00545). Although these entries correspond to the same protein, they were generated using different approaches: PED00426 using IDPConformerGenerator [100] biased by chemical shift data, PED00477 using the idpGAN [99] machine-learning model, and PED00543–PED00545 using molecular dynamics simulations with different force fields (C36m, C22, and a99SB-disp) [101]. When evaluated against experimental SAXS data, the agreement of these ensembles differs substantially. In a simple comparison using unweighted ensembles, PED00426 shows the best agreement with the experimental SAXS profile, whereas the ensemble generated with the C22 force field shows the poorest agreement.

Providing such fitting metrics directly within PED entries would help users assess how well different ensembles reflect experimental data and would strengthen PED’s role as a platform for comparative evaluation of ensemble quality.

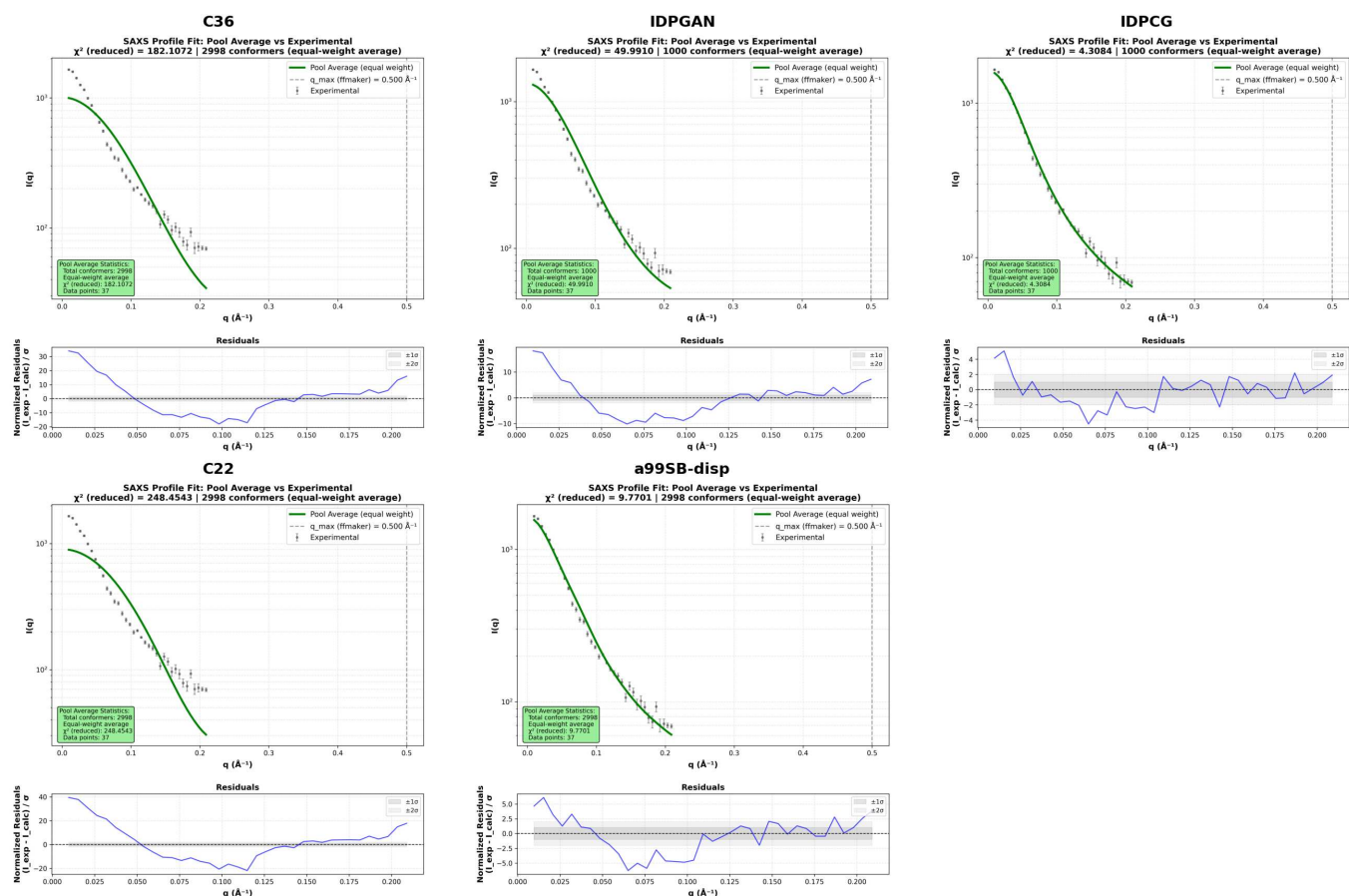


Figure 11: Experimental SAXS profiles are compared with ensemble-averaged scattering curves calculated from five ensembles generated using different approaches: C36m molecular dynamics (PED00543), idpGAN (PED00477), IDPConformerGenerator with chemical-shift bias (PED00426), C22 molecular dynamics (PED00544), and a99SB-disp molecular dynamics (PED00545). Reduced χ^2 values and residuals show substantial differences in agreement with the experimental data, with the IDPConformerGenerator ensemble exhibiting the best fit and the C22 force field ensemble the poorest agreement.

Toward a Unified Framework for Determining Conformational Ensembles:

Four years ago, when I began my master’s thesis on studying IDPs, I sought to understand how conformational ensembles of IDPs are determined in practice. Through an extensive review of the literature, it became evident that different studies relied on substantially different experimental, computational, and integrative approaches. It was at this stage that I identified a clear gap in the field: despite the widespread use of conformational ensembles, there was no unified framework describing how such ensembles should be generated, constrained, validated, and interpreted. This is particularly important for researchers who are new to the field, as it provides a global overview of the different ways in which conformational ensembles are determined.

Moreover, the availability of a standardized framework can promote the generation of more reliable ensembles by explicitly accounting for experimental and computational uncertainties, reducing the risk of overfitting to experimental data, providing a basis for benchmarking practices, and ultimately facilitating the development of more robust machine-learning-based ensemble generators.

After starting my PhD, I had the opportunity to participate in the ML4NGP (Machine Learning for Non-Globular Proteins) workshop on structural ensembles of intrinsically disordered proteins, held in Prague (<https://ml4ngp.eu/workshop-prague/>). This workshop brought together researchers from diverse experimental and computational backgrounds to discuss the challenges associated with determining conformational ensembles of IDPs and to advocate for the development of a unified framework in this area. These discussions, together with insights from the literature, motivated the formulation of a unified framework that can be decomposed into a set of core conceptual and methodological components.

Summary of the major components of the ensemble determination framework:

As described comprehensively in chapter 2 [62], the major components of the framework can be summarized as follows: 1- experimental methods, 2- computational generation and 3- validation and comparison. In the following, I provide a brief discussion of each component of the framework.

1. **Experimental methods:** The experimental characterization of IDPs differs fundamentally from that of globular proteins. Owing to the highly dynamic and heterogeneous nature of IDPs, techniques such as X-ray crystallography are generally not suitable, as they rely on the formation of stable crystals and typically capture a single, ordered conformation. Instead, the study of IDPs requires experimental methods that operate in solution and are capable of reporting on conformational heterogeneity. Moreover, most experimental methods provide an average and sparse information on the dynamics of IDPs conformational ensemble.

Within this module, NMR, SAXS, and smFRET are identified as the primary experimental techniques for the characterization of IDP ensembles, while other methods such as circular dichroism (CD) and hydrogen-deuterium exchange mass spectrometry (HDX-MS) are considered complementary. It is also emphasized that each technique provides information on distinct aspects of IDP ensembles. For example, NMR offers residue-level information, including transient secondary structure propensities and local dynamics; smFRET reports on intramolecular distance distributions; and SAXS captures global ensemble properties such as overall dimensions and the radius of gyration.

Ultimately, the experimental observables obtained from these techniques, together with their associated uncertainties, define the constraints that guide ensemble determination. Rather than directly specifying unique structures, these measurements provide ensemble-averaged

or distribution-based information that must be interpreted within an appropriate modeling framework. Recognizing the intrinsic limitations and uncertainty ranges of each experimental method is therefore essential to avoid overfitting and overinterpretation. Within the unified framework, experimental data are treated as probabilistic constraints that inform ensemble generation and refinement, highlighting the need for integrative approaches that combine complementary techniques to achieve a more complete and reliable description of IDP ensembles.

2. **Computational generation:** As highlighted in the previous section, computational techniques are essential for deconvoluting experimentally measured, ensemble-averaged observables into explicit conformational ensembles. Within the unified framework, this module can be divided into two components: (a) ensemble generation methods and (b) experimental data integration methods. Ensemble generation methods aim to sample the conformational space accessible to an IDP in the absence of, or prior to, experimental constraints. These approaches define the initial conformational pool, or prior, from which ensembles are subsequently refined. Common strategies include physics-based molecular dynamics simulations using all-atom or coarse-grained force fields, as well as statistical conformer generators that produce large pools of structures based on backbone torsion angle libraries and sequence-dependent propensities. More recently, data-driven and machine-learning-based methods have been introduced to generate ensembles by learning statistical features of IDP conformational behavior from simulation data or curated structural datasets. While these approaches differ substantially in resolution, computational cost, and underlying assumptions, they all serve the same conceptual role within the framework: defining a plausible and sufficiently diverse representation of the conformational landscape.

The second component, experimental data integration, addresses the incorporation of experimental observables into the generated conformational pool to obtain ensembles that are consistent with measured data. Because experimental techniques report on averages or distributions rather than on individual structures, this step typically involves reweighting, selection, or refinement procedures. Two broad classes of approaches are commonly distinguished. Maximum-entropy-based methods adjust ensemble weights to minimally perturb the prior distribution while enforcing agreement with experimental observables, thereby preserving ensemble diversity. In contrast, maximum-parsimony or selection-based approaches identify a reduced subset of conformers that best reproduces the data, often at the cost of reduced heterogeneity. Each strategy entails different trade-offs between interpretability, diversity, and sensitivity to noise, and their suitability depends on the nature and amount of available experimental data.

Importantly, the unified framework emphasizes that ensemble genera-

tion and data integration are conceptually distinct but tightly coupled steps. The properties of the final ensemble depend not only on the experimental data but also on the characteristics of the initial conformational pool and on the assumptions embedded in the integration method. Explicitly separating these two stages clarifies methodological choices, facilitates reproducibility, and enables meaningful comparison between different ensemble-determination strategies. This modular view also provides a foundation for the validation and comparison procedures discussed in the next section.

3. **Comparison and Validation:** Validation and comparison constitute a critical component of the unified framework for determining conformational ensembles of IDPs. Given the intrinsic heterogeneity of ensemble representations, multiple distinct ensembles can often reproduce the same set of experimental observables within uncertainty. Validation therefore cannot be understood as the identification of a single “correct” ensemble, but rather as the assessment of whether an ensemble is consistent with available data, physically plausible, and robust with respect to methodological choices. A key principle of the framework is the distinction between *fitting* and *validation*. Agreement with the experimental data used to generate or refine an ensemble is necessary but not sufficient to establish its reliability. Whenever possible, ensembles should be evaluated against independent experimental observables that were not employed during the refinement process. Such cross-validation provides a more robust test of predictive power and reduces the risk of overfitting to a limited or noisy dataset[96]. Explicit consideration of experimental uncertainties and forward-model errors is essential in this context, as overinterpretations can arise when uncertainties are underestimated or ignored[108].

Beyond validation against experimental data, consistency with physical principles represent additional evaluation criteria. Ensembles should be assessed for artifacts arising from force-field biases, sterical clashes, insufficient sampling, or overly restrictive refinement procedures. Analysis of ensemble diversity, population distributions, and sensitivity to perturbations, such as changes in input data, weighting schemes, or integration parameters[109], can reveal whether an ensemble is robust or dominated by a small number of conformations. These analyses are particularly important for IDPs, where excessive collapse or artificial structuring may reflect methodological bias rather than genuine biophysical behavior.

Comparison between ensembles is another essential element of the framework, especially as the number of available ensemble models continues to grow. Quantitative comparison enables assessment of how different experimental inputs, computational priors, or integration strategies influence ensemble properties. Rather than relying on single

structural descriptors, ensemble comparison should be performed using distribution-based metrics that capture both global and local features, such as distance distributions, secondary structure propensities, or measures of conformational heterogeneity. Divergence-based measures and statistical distances provide a robust way to compare ensembles while accounting for their intrinsic variability[70].

Importantly, validation and comparison are not end points but iterative components of ensemble determination. Insights gained from validation may motivate refinement of experimental strategies, adjustments to computational models, or reconsideration of integration assumptions. Within the unified framework, this iterative process promotes transparency and reproducibility by making explicit which aspects of an ensemble are well supported by data and which remain uncertain. By treating validation and comparison as main components rather than post hoc checks, the framework provides a structured and reliable approach for interpreting conformational ensembles of IDPs.

The importance of benchmarking of conformational ensembles of IDPs as the next step:

As part of the work presented in Chapter 2, we also emphasized the importance of benchmarking ensemble generation methods. In particular, following the emergence of machine-learning-based techniques, it has become increasingly important to establish comprehensive and unbiased benchmarking strategies to critically assess conformational ensembles against diverse types of experimental data.

In this context, we proposed the IDP Ensemble Benchmarking Challenge (IDP-Bench) as a conceptual and practical response to the current lack of standardized evaluation procedures. The central message of this initiative is that ensemble quality cannot be judged using a single observable or method-specific criterion. Instead, meaningful benchmarking must be community-driven, method-agnostic, and explicitly designed to assess agreement across multiple experimental modalities and length scales[110]. By drawing inspiration from successful community assessments in related fields, like CASP (critical assessment of protein structure prediction)[111], IDP-Bench highlights the value of blind evaluation, transparent scoring protocols, and publicly accessible results as essential elements for fair comparison.

More broadly, the IDP-Bench concept underscores a shift in how ensemble prediction methods should be evaluated. Rather than ranking approaches based on their ability to reproduce a limited subset of experimental data, benchmarking should focus on robustness, generalization, and consistency across independent observables. This is particularly critical for machine-learning-based ensemble generators, which may achieve excellent agreement with training data while failing to capture physically meaningful conformational behavior outside their training set. A structured benchmarking framework therefore serves not only as a means of comparison but also as a diagnostic tool to identify methodological biases and limitations.

Ultimately, the discussion around IDP-Bench reflects a broader need in the field: to

move from ad hoc validation toward systematic, reproducible assessment of ensemble models. Establishing such benchmarks will be essential for building confidence in ensemble-based representations, guiding methodological development, and ensuring that advances in computational modeling translate into biologically meaningful insight.

The key important steps for future of modeling of conformational ensembles of IDPs:

The future progress of conformational ensemble modeling for IDPs will depend on addressing several interconnected challenges that extend beyond the development of individual methods. Progress in ensemble modeling critically depends on recognizing uncertainty, dynamical behavior, and methodological choice as fundamental components of the modeling process, rather than as secondary concerns.

A critical step forward is the explicit incorporation of uncertainty quantification into all stages of ensemble determination. As highlighted in this work, ensemble models are influenced not only by experimental noise but also by forward-model inaccuracies and modeling assumptions. Without a rigorous treatment of these uncertainties, apparent agreement with experimental data may reflect overfitting rather than genuine physical insight. Embedding uncertainty-aware evaluation, through variance propagation, cross-validation, and probabilistic scoring, provides a more realistic assessment of ensemble quality and enables fair comparison between fundamentally different modeling strategies. In this context, uncertainty-aware benchmarking frameworks, such as those envisioned within IDP-Bench, represent an essential foundation for future method development.

Beyond equilibrium descriptions, an important frontier for the field lies in capturing the dynamic and non-equilibrium behavior of IDPs. Many IDPs operate through transient conformational switching, kinetic regulation, and time-dependent interactions that are not adequately described by static equilibrium ensembles. While experimental access to such dynamics has historically been limited, emerging time-resolved techniques are beginning to provide constraints on the temporal evolution of IDP conformations. Integrating these data with modeling approaches rooted in statistical mechanics, such as maximum entropy and maximum caliber frameworks, offers a promising route toward dynamic ensemble representations that more closely reflect biological function. Advancing this direction will require conceptual shifts in how ensembles are defined, validated, and interpreted, moving from purely structural descriptions toward time-resolved, process-oriented models.

A representative example is the coupled folding and binding of the intrinsically disordered ACTR domain to NCBP, where time-resolved experiments have shown that binding does not proceed as a simple direct transition from an unbound ensemble to a fully folded complex. Instead, the interaction involves a transient encounter complex and microsecond transition-path dynamics, revealing that intermediate states along the binding pathway can play an important mechanistic role [112, 113]. In such systems, modeling must account not only for the structural properties of the initial and final states, but also for the rates of interconversion, the lifetimes of intermediates, and the possibility of parallel pathways or kinetic heterogeneity [114].

These features are difficult to validate because models that agree with the same equilibrium-averaged observables may still imply very different dynamic mechanisms or timescales[115] .

Another key step for the field is the development of clear and pragmatic guidelines for method selection. The rapid expansion of available computational tools, ranging from physics-based simulations to machine-learning-driven generators, has created both opportunities and confusion, particularly for non-expert users. Progress will therefore depend not only on improving individual methods but also on providing structured guidance that links experimental data types, biological questions, and computational constraints to appropriate modeling strategies. Such guidance helps prevent misuse of methods, reduces redundant effort, and promotes more consistent interpretation of ensemble results across studies.

Importantly, these future directions are tightly interdependent. Robust uncertainty handling enables meaningful benchmarking; benchmarking clarifies method strengths and limitations; dynamic modeling challenges current validation practices; and method-selection guidelines depend on transparent performance assessments. Addressing these aspects in isolation is unlikely to yield lasting progress. Instead, advancing the field will require coordinated efforts that integrate uncertainty-aware modeling, dynamic ensemble representations, standardized benchmarking, and user-oriented decision frameworks within a coherent methodological ecosystem.

In summary, the next phase of IDP ensemble modeling will be defined less by the introduction of individual new methods and more by the maturation of the field toward rigor, transparency, and integration. By explicitly accounting for uncertainty, embracing dynamics, and guiding method selection through standardized frameworks, the community can move toward ensemble models that are not only computationally sophisticated but also biologically interpretable and reproducible.

Review of recent state-of-the-art methods for determining conformational ensembles:

In parallel with the development of the unified framework, and motivated by the breadth and fragmentation of the existing literature, we conducted a comprehensive review of recent advances in conformational ensemble determination. This review was informed by an extensive survey of experimental, computational, and integrative approaches, and it was explicitly structured around the core components of the framework described above. By organizing the literature according to experimental methods, ensemble generation strategies, and validation and comparison procedures, the review provides a coherent overview of the current state of the art and highlights emerging trends, methodological strengths, and remaining challenges in the field.

Computational Methods for the Analysis of Conformational Ensembles:

The rapid expansion of experimentally derived and computationally predicted conformational ensembles has shifted a major bottleneck in IDP research from ensemble generation to ensemble interpretation. While the availability of ensembles has in-

creased substantially through advances in integrative modeling, machine learning, and large-scale prediction efforts, the lack of standardized and accessible analytical frameworks has limited the ability to extract comparable, reproducible, and biologically meaningful insights from these datasets. This challenge is particularly critical for IDPs, whose ensembles are inherently heterogeneous, often low-resolution, and sensitive to both modeling protocols and environmental context.

Traditional MD analysis libraries, such as MDTraj[116] and MDAnalysis[117], provide powerful and general tools for trajectory handling and structural analysis, but they were primarily developed for folded proteins and long MD simulations. As a result, they do not directly address several central challenges of ensemble-based IDP analysis: comparison between independently generated ensembles, analysis across heterogeneous resolutions (all-atom vs coarse-grained), and quantification of ensemble similarity in a statistically meaningful way. Moreover, many existing analysis workflows implicitly assume a single reference structure or a dominant basin, assumptions that are not valid for most disordered systems.

In response to these limitations, the work presented in this thesis introduces IDPET (Intrinsically Disordered Protein Ensemble Tools) as a dedicated computational framework for the systematic analysis of conformational ensembles of disordered proteins[66]. Rather than focusing on trajectory-centric analysis, IDPET adopts an ensemble-centric perspective, treating each ensemble as a statistical object whose properties emerge from distributions of conformations rather than from individual structures. This design choice aligns directly with the sequence–ensemble–function paradigm discussed in earlier chapters and provides a consistent computational counterpart to ensemble-focused databases such as the PED.

A central contribution of IDPET is the unification of global, local, and comparative analyses within a single, reproducible framework. At the global level, IDPET enables extraction of ensemble-averaged and distributional properties such as radius of gyration, end-to-end distance, asphericity, solvent-accessible surface area, and apparent scaling exponent. These descriptors provide a coarse-grained view of ensemble compactness and shape, facilitating comparisons across conditions, sequences, or modeling strategies. At the local level, IDPET computes residue-resolved features, including backbone dihedral distributions, transient secondary structure propensities, and flexibility indices, allowing direct mapping of ensemble heterogeneity onto sequence position.

Beyond feature extraction, a key methodological advance lies in the quantitative comparison of ensembles. IDPET implements multiple ensemble similarity metrics based on Jensen–Shannon divergence, applied to different structural representations, including inter-residue distance distributions, backbone dihedral angles, and coarse-grained torsional descriptors. This multi-resolution approach enables discrimination between ensembles that may appear similar at the level of global averages but differ substantially in local conformational preferences. Importantly, the inclusion of bootstrapping procedures allows statistical assessment of inter-ensemble differences even for ensembles with limited sample sizes, addressing a common practical limitation in IDP studies.

Dimensionality reduction and clustering constitute another critical component of

ensemble analysis, particularly for visualizing high-dimensional conformational spaces. IDPET integrates several commonly used linear and nonlinear techniques, including PCA, t-SNE[118], and UMAP[119], while explicitly exposing the dependence of results on feature choice and parameterization. By supporting multiple structural representations as input, the framework enables users to evaluate how different descriptors emphasize global versus local variability, an aspect that is especially important for disordered systems where no single coordinate system is universally optimal.

From a broader perspective, IDPET is designed to function not as an isolated analysis tool, but as a computational layer within an ecosystem of ensemble-based resources. Its direct integration with PED through programmatic access allows users to retrieve deposited ensembles and immediately apply standardized analyses, closing the loop between data deposition and data interpretation. This tight coupling between infrastructure and analysis reflects a central theme of this thesis: ensemble-based protein science requires coordinated development across data representation, computational workflows, and analytical methodologies.

Despite these advances, several challenges remain. Ensemble-derived descriptors are unavoidably influenced by the underlying generation protocol, force field, or experimental restraints, complicating direct biological interpretation. Feature selection for dimensionality reduction and similarity analysis remains non-trivial, and different metrics may emphasize distinct aspects of conformational behavior. Furthermore, current ensemble comparison approaches largely rely on pairwise distributional metrics, which may not fully capture higher-order correlations or collective motions. Addressing these limitations will require continued methodological development, including integration of sequence-derived descriptors, improved statistical models, and closer coupling between ensemble analysis and functional annotation.

In summary, the computational methods developed and applied in this thesis provide a structured and extensible framework for analyzing conformational ensembles of IDPs. By shifting the focus from individual structures to ensemble-level properties and by enabling statistically grounded comparisons across datasets, these tools support the transition toward a mature sequence–ensemble–function paradigm. Together with the infrastructure and ensemble-determination frameworks discussed in previous sections, they contribute to a coherent methodological foundation for future large-scale and integrative studies of protein conformational heterogeneity.

4– Conclusion

This thesis has addressed an important challenge in modern protein science: how to represent, determine, analyse, and share the structural behaviour of proteins whose biological function does not arise from a single stable structure, but from conformational heterogeneity. With a particular focus on IDPs and conformationally dynamic systems, this work contributes to the ongoing shift from static, structure-centred descriptions toward ensemble-based representations of protein behaviour.

At a conceptual level, the thesis supports and further develops the sequence–ensemble–function paradigm. Instead of treating conformational heterogeneity as a limitation,

this framework recognises conformational ensembles as the most appropriate physical description for many protein systems. Throughout the thesis, this perspective has been used to describe both folded and disordered proteins within a unified framework, emphasising that biological function often depends on population distributions, dynamic fluctuations, and context-dependent changes in conformational states, rather than on a single dominant structure.

At the infrastructural level, this thesis has contributed to improving the resources required to support ensemble-based protein science. In particular, the work on the PED has strengthened its role as a dedicated repository for conformational ensembles, through improvements in deposition workflows, validation pipelines, metadata organisation, and large-scale integration strategies. The integration of DRMAA^{tic} as a backend workflow manager further enables scalable and reproducible processing of ensemble data on heterogeneous computing infrastructures. Together, these developments address key limitations of traditional structural databases when applied to intrinsically disordered and flexible proteins.

At the methodological level, this thesis has contributed to clarifying how conformational ensembles are determined in practice. By proposing a unified framework that separates experimental characterisation, computational generation, and validation and comparison, this work provides a clear and structured view of ensemble determination strategies. The framework highlights common principles across different approaches, makes methodological assumptions explicit, and treats uncertainty and heterogeneity as intrinsic aspects of ensemble modelling. This perspective supports reproducibility, facilitates benchmarking, and provides guidance for method selection, especially in the context of emerging machine-learning-based ensemble generation methods.

At the level of data analysis, the development of IDPET addresses a major limitation in current ensemble-based studies, namely the lack of standardised tools for ensemble-level analysis and comparison. By treating conformational ensembles as statistical objects rather than as collections of representative structures, IDPET enables robust analysis of global and local ensemble properties, quantitative comparison between independently generated ensembles, and statistically meaningful assessment of ensemble similarity. The integration of these tools with PED further connects ensemble deposition with ensemble interpretation, reinforcing the importance of analysis as a core component of ensemble infrastructure.

More broadly, this thesis highlights the need for uncertainty-aware modelling, independent validation, and community-driven benchmarking as essential steps for the future development of ensemble-based protein modelling. As ensemble prediction methods continue to increase in scale and complexity, systematic evaluation of robustness, generalisation, and physical plausibility will become increasingly important. Initiatives such as the proposed IDP Ensemble Benchmarking Challenge reflect this need and provide a possible route toward transparent and reproducible assessment practices.

In conclusion, this thesis advances ensemble-based protein science through coordinated contributions to infrastructure, methodology, and computational analysis. By promoting standardisation, transparency, and reproducibility, the work supports a

more rigorous and interpretable use of conformational ensembles in both fundamental research and applied contexts, including drug discovery. Overall, the results presented here aim to facilitate broader adoption of the sequence–ensemble–function paradigm as a practical framework for understanding protein behaviour in its dynamic and heterogeneous nature.

Appendix A: DRMAAAtic supplementary materials

Supplementary Material

DRMAAAtic: dramatically improve your cluster potential

Alessio Del Conte¹, Hamidreza Ghafouri¹, Damiano Clementel¹, Ivan Mičetić¹,
Damiano Piovesan¹, Silvio C.E. Tosatto^{1,2,*}, Alexander Miguel Monzon^{1,*}

¹Department of Biomedical Sciences, University of Padova, Padova 35121, Italy

²Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari 70126, Italy

*Corresponding authors. Department of Biomedical Sciences, University of Padova, Padova 35121, Italy. E-mails: alexander.monzon@unipd.it (A.M.M.) and silvio.tosatto@unipd.it (S.C.E.T.)

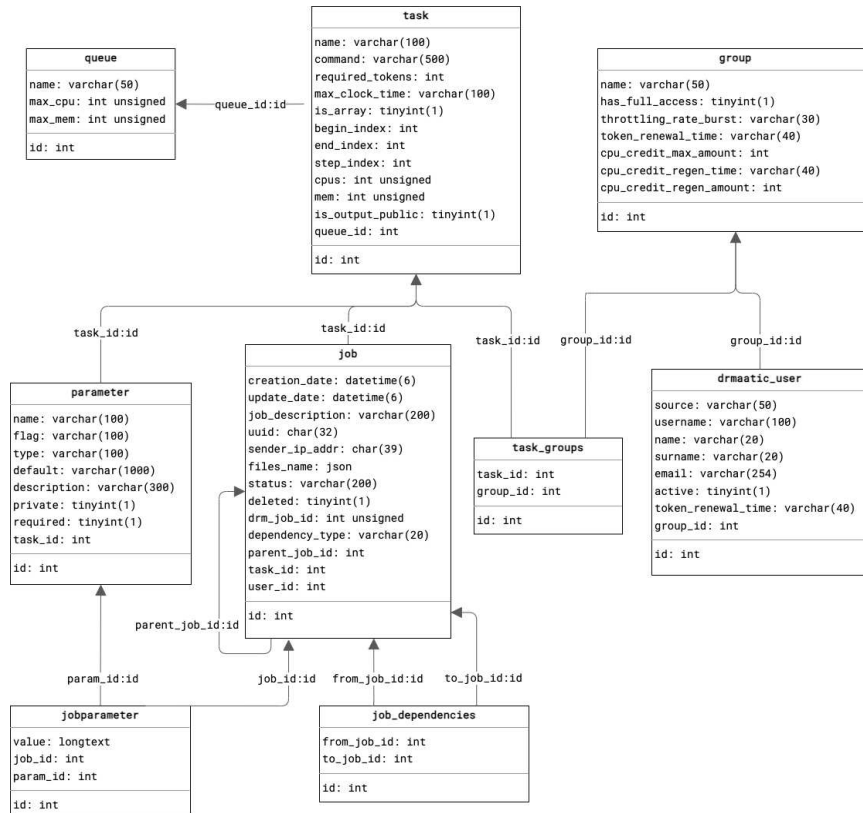


Figure 1. Database schema of DRMAAAtic with all the models represented as rectangles, with a list of their fields. Arrows indicate the relation between two models, the label on the arrow is the foreign key that is used for the relation.

A

Request

```
POST /job HTTP/1.1
Host: drmaatic-example.com
Content-Type: application/json
Content-Length: [length]
{
  "task_name": "wait_task",
  "job_description": "wait task example"
  "parent_job": "c796003d-7150-461d-b2ebdae7ac76",
  "dependencies": [
    "c796003d-7150-461d-b2ebdae7ac76"
  ],
  "dependency_type": "afterok",
  "wait_time": "10"
}
```

B

Response

```
HTTP/1.1 200 OK
Date: Wed, 03 Apr 2024 14:40:37 GMT
Content-Type: application/json
Content-Length: [length]
{
  "uuid": "d8f3e7ea-91b5-4848-af91-fd4378bfe6f2",
  "task": "wait_task",
  "depends_on": [
    "c796003d-7150-461d-b2ebdae7ac76"
  ],
  "dependency_type": "afterok",
  "status": "job is queued and active",
  "creation_date": "2024-04-03T14:39:09.785085+02:00",
  "params": [
    {
      "name": "wait_time",
      "value": "10"
    }
  ]
}
```

Figure 2. Example of job submission with a POST method, where the user wants to execute the *wait_task* that has a *parent_job* which also depends on, meaning that it will be executed in the same execution directory of the parent and only if the parent job will complete without any errors. The *wait_time* is a parameter for this specific task, that will be passed and used by the task script. In panel (A) is the request made from the client, in panel (B) the response with the information about the job.

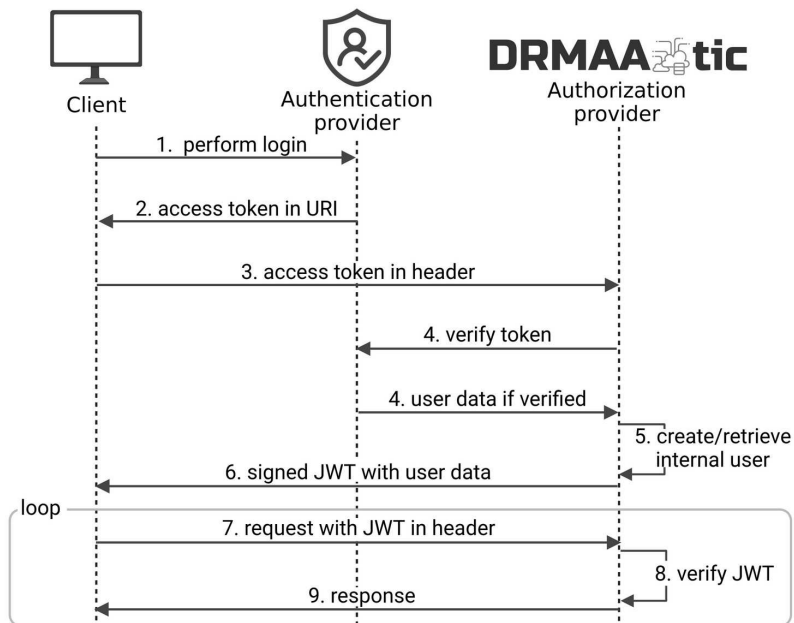


Figure 3. Scheme of the OAuth 2.0 implicit flow integrated in DRMAAtic with a JWT as authorization token. The authentication provider in this first implementation is the ORCID platform. URI: Uniform resource identifier.

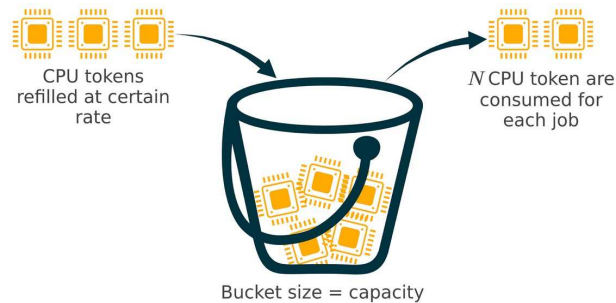


Figure 4. Scheme of how the CPU token bucket throttling algorithm works. To each user, identified by their IP address or authentication, a fixed number of CPU tokens is assigned. To each task there is a number of required tokens that need to be consumed upon job submission. If there are not enough tokens, the request is throttled. The bucket has a maximum size that gets replenished at a fixed rate.

Add task

Name: Command:

Required tokens:

Queue: CPUs per task: Memory per task (MB):

☐ Is output public

max clock time:

Groups:

Hold down "Control", or "Command" on a Mac, to select more than one.

☐ Is array Begin index: End index: Step index:

PARAMETERS							
NAME	FLAG	TYPE	DEFAULT	DESCRIPTION	PRIVATE	REQUIRED	DELETE?
		String			☐	☐	x

Add another Parameter

Figure 5. DRMAAatic Administration Interface: Adding a New Task. The administrator is prompted to complete the necessary fields (highlighted in bold). Subsequently, a parameter list can be configured to be passed the command, requiring each parameter to have a designated name and type. Additionally, if a flag is specified, it will be placed before the parameter value.

Action:

Update DRM status

Go

0 of 100 selected

☐	UUID	TASK	STATUS	OUTPUTS	DELETED	CREATION DATE	USER
☐	da67dd28-1af...	task-3	Created	out / err / files		March 12, 2024, 2:40 p.m.	-
☐	e248e159-89c...	task-8	Done	out / err / files		March 12, 2024, 2:15 p.m.	0000-0012-3400-0000
☐	4cbe630d-b82...	task-6	Done	out / err / files		March 12, 2024, 12:57 p.m.	0000-0012-3412-0034
☐	5d719e01-ecf...	task-3	Created	out / err / files		March 12, 2024, 12:15 p.m.	-
☐	e8e3cc5a-c92...	task-5	Failed	out / err / files		March 12, 2024, 12:15 p.m.	1234-0012-3400-0000

FILTER

→ By task

→ By status

→ By deleted

creation date

From date

To date

Search

Reset

→ By user

Change job

d44f079e-882d-42bb-b3be-d2ae0106acdc - task-3

job description:

Uuid:

d44f079e-882d-42bb-b3be-d2ae0106acdc

Task:

task-3

Creation date:

March 12, 2024, 12:08 p.m.

Dependencies:

Dependency type:

-

JOB PARAMETERS

PARAM	VALUE
input_file -i : input_file.pdb	
input_file -i	input_file.pdb
all_chains --all_chains : true	
all_chains --all_chains	true
all_models --all_models : true	
all_models --all_models	true
seq_sep -g : 3	
seq_sep -g	3

Figure 6. Interface for Job Administration: (A) presents a comprehensive list of submitted jobs, displaying task execution status, current state, outputs, and relevant details, with available filter options. (B) offers detailed job information, while (C) highlights input parameter values.

Appendix B: Towards a unified framework for determination of conformational ensembles

Supplementary Table 1: Additional experimental techniques used to study IDP structural features.

Summary of experimental methods that complement classical biophysical approaches for IDP analysis, including their forward models, information content, level of structural resolution, and system applicability (e.g., in vitro vs. in cell).

Technique	Experimental measurement	Forward Models / Computational tools / Theoretical models	Information content	Structural interpretation (e.g. residue level vs. global)	System perturbation	Environment (In-vivo / in-vitro / in-cell)	Reference database
Neutron Spin Echo (NSE) spectroscopy	Intermediate scattering function	Diffusion models	Provides dynamic information on nanosecond to microsecond timescales	Internal dynamics and diffusion	None required	In-vitro	N/A
Hydrogen deuterium exchange mass spectrometry (HDX-MS)	Kinetics of isotope substitution	Under development; HDXsimulator	Rate of proton/deuterium exchange	Exposure to the solvent, local stability, hydrogen bond formation	Gas phase, heavy water, back exchange	In-vitro / in-cell	Under development; ProteomeXchange
Circular dichroism (CD) spectroscopy	Differential absorption of left and right circularly polarized light	Secondary structure prediction algorithms	Provides secondary structure content	Secondary structure content	None required	In-vitro	PCDDb
Fluorescence anisotropy	Anisotropy	Perrin equation or molecular simulation	Local flexibility, rotational motion	Rotational correlation time	Dye labeling	In-vitro/in-cell	N/A
Fluorescence Correlation Spectroscopy (FCS)	Fluorescence intensity correlation	Stokes-Einstein or Stokes-Einstein-Debye	Rotational and translational motion	Hydrodynamic radius	Dye labeling	in -vitro/in-cell	N/A
Fourier Transform Infrared Spectroscopy (FTIR)	Infrared absorption spectrum	Vibrational modes analysis	Provides information on secondary structure and chemical bonds	Secondary structure, functional groups	None required	In-vitro	N/A
Limited Proteolysis coupled to Mass Spectrometry (LiP-MS)	Proteolytic cleavage patterns analyzed by mass spectrometry	Protease cleavage patterns and MS-detectable peptides	Provides insights into protein structure and conformational changes	Protein conformation and dynamics	Protease treatment	In-vitro	N/A
Dynamic light scattering (DLS)	Intensity fluctuations of scattered light	Stokes-Einstein equation	Provides size distribution of particles in solution	Hydrodynamic radius	None required	In-vitro	N/A
Cryogenic Electron Microscopy (CryoEM)	Electron scattering images	Electron density map modelling	Provides near-atomic resolution structures	High-resolution structures	Non required	In-vitro (frozen-hydrated samples)	EMDB
Laser-THz Emission Microscope (LTEM)	Electron scattering images	THz emission modelling	Similar to CryoEM	High-resolution structures	Non required	In-vitro (cryogenic temperatures)	N/A
Electron Paramagnetic Resonance (EPR)	Absorption of microwave radiation by unpaired electrons	Spin Hamiltonian analysis	Distance distributions, local dynamics	Local environment of unpaired electrons	Spin labeling	In-vitro/in-cell	N/A
Atomic Force Microscopy (AFM)	Force-distance curves	Force spectroscopy models	Provides high-resolution surface images and mechanical properties	Surface topography at the nanoscale	None required	In-vitro	N/A
Mass Spectrometry (MS)	Thermal Proteome Profiling (TPP)	Thermal stability profiles	Provides information on protein stability and interactions	Protein stability and folding states	Temperature changes	In vivo	N/A

Supplementary Table 2. Summary of conformational ensemble generation techniques.

Classification of computational strategies for conformational ensemble generation based on the underlying modeling paradigm (knowledge-based, physics-based, machine learning, or hybrid), with evaluation of their strengths, limitations, and suitable applications.

Ensemble Generation Strategy	Method/Force field	Core Principle	When to use	Strength	Limitation
Knowledge-based	Flexible Meccano ¹	Random coil-based sampling of backbone dihedrals using statistical distributions	When fast, generic IDP ensembles are needed without added data	Fast; simple; minimal input	No statistical weights; lacks side chains and atomistic detail
	TraDES ²	Generates ensembles via statistical coil models and steric filtering	For baseline comparisons or when no structural data is available	Very efficient; customizable chain properties	No statistical weights; lacks experimental integration
	MoMA ³	Builds conformers using motion planning algorithms with steric constraints and possibly incorporating energy models	When large-scale exploration with realistic geometry is desired	Includes full-atom models; physically plausible structures	Dynamics are not considered
	IDPConformerGenerator ⁴	Samples torsion angles from fragment libraries, optionally guided by restraints	When integrating sequence and experimental restraints	Modular; supports reweighting and SAXS/NMR constraints	Depends on fragment quality and scoring functions
	RANCH ⁵	Stochastic generation of flexible linkers and IDRs between rigid bodies, based on geometric constraints	When modeling large-scale flexibility or preparing SAXS ensembles	Very efficient; compatible with SAXS; minimal input required	No statistical weights; lacks atomistic detail; not suitable for local structure prediction
	DIPEND ⁶	Generates ensembles using neighbor-dependent statistical dihedral angle distributions and structural filters	When fast backbone ensemble generation is needed from sequence with minimal input	Efficient and customizable; uses experimental priors; fast clash resolution and refinement steps	Lacks physical dynamics; relies on heuristics and rotamer libraries; limited to backbone sampling
Physics-based	Amber ff99SBws/ff03ws ⁷	All-atom force field with improved water-protein interactions	For simulations of IDPs with balanced protein-water interactions	Improved global dimensions and protein-protein association	Computationally intensive
	Amber ff99SB-disp ⁸	Dispersion-corrected all-atom force field for IDPs and folded proteins	For simulations of IDPs and folded proteins	Improved global dimensions and transient helicity	Computationally intensive
	DES-Amber ⁹	All-atom force field with improved description of protein-protein complexes	For simulations of IDPs, folded proteins, and protein-protein complexes	Improved global dimensions, transient helicity, and protein-protein complexes	Limited to Amber ecosystem
	REMD ¹⁰	Parallel simulations at different temperatures for enhanced sampling	To improve conformational sampling in rugged energy landscapes	Better sampling of rare conformers	Requires many replicas; costly
	Metadynamics ¹¹	Biases applied to accelerate rare event sampling	When slow transitions hinder sampling of	Efficient barrier crossing	Choice of the collective variables may affect efficiency

			conformers		
	ABSINTH ¹²	Implicit-solvent all-atom force field for IDPs	Fast all-atom modeling of long IDPs	Efficient; all-atom resolution	Lacks explicit solvent detail
	MARTINI ¹³	Widely used coarse-grained model for large biomolecular systems	For large-scale or membrane-associated systems	Scalable to large complexes	Simplified interactions
	Mpipi ¹⁴	Implicit-solvent coarse-grained force field for IDPs; bottom-up and top-down parameterization	Fast and sequence-dependent modeling of IDPs	Efficient; accurate predictions of IDP compaction	Residue-level resolution; does not capture transient secondary structure or hydrogen bonding
	CALVADOS ^{15,16}	Implicit-solvent coarse-grained force field for IDPs and multi-domain proteins; top-down parameterization.	Fast and sequence-dependent modeling of IDPs and multi-domain proteins	Efficient; accurate predictions of IDP compaction	Residue-level resolution; does not capture transient secondary structure or hydrogen bonding
	HPS ¹⁷	Implicit-solvent coarse-grained force field for IDPs and multi-domain proteins; bottom-up parameterization	Fast and sequence-dependent modeling of IDPs	Efficient; captures trends in IDP compaction	Residue-level resolution; does not capture transient secondary structure or hydrogen bonding; may overestimate IDP compaction
Machine-Learning	idpGAN ¹⁸	GAN trained on CG-MD generated ensembles to generate new conformers with statistical characteristics of training data	When need rapid ensemble generation	Captures realistic ensemble diversity; scalable	Limited to patterns learned; learns from CG training data; C α -only ensembles
	idpSAM ¹⁹	Latent diffusion model (transformer autoencoder + diffusion) trained on implicit-solvent ensembles (ABSINTH) to sample new structures	For generating ensembles of short IDRs (< 60 residues), transfers to sequences absent in training set	High generalization to unseen sequences; stable training	Requires substantial training data from simulations
	STARLING ²⁰	Diffusion/generative approach leveraging physics-based insights to produce hundreds of conformers in seconds	When needing rapid, large-scale ensemble predictions from sequence	Extremely fast (< seconds); works on both GPU & CPU	Newer tool; may lack extensive validation across diverse sequences
	IDPFold ²¹	Fine-tuned diffusion model generating backbone-level conformers directly from sequence, without MSAs or experimental input	When backbone ensemble prediction is needed rapidly from sequence alone	Bypasses need for MSA or experiments; provides diverse structural features	Focuses on backbone only; may miss side-chain dynamics and solvent effects
	Phanto-IDP ²²	Diffusion-based generative model trained on sequence-conditioned structural features of IDPs	When rapid ensemble generation is needed from sequence alone	Physics-informed generative model; captures IDP-specific features; fast sampling	Relies on training data quality; may lack fine-resolution structural detail
	BioEmu ²³	Diffusion model trained on a large dataset of simulation and experimental data	Rapid generation of equilibrium ensembles from sequence	Fast and sequence-sensitive; produces diverse ensembles consistent with disordered behavior	Backbone-only, lacks sidechain sampling;; overestimates radius of gyration for longer IDPs (>100)

					residues)
Hybrid	AlphaFold-MetaInference ²⁴	Combines AlphaFold constraints with Bayesian ensemble modeling (ML + Physics-based approach)	To integrate AI predictions into ensemble refinement	Incorporates prior knowledge	Requires AlphaFold confidence and priors
	FastFloppyTail ²⁵	Uses Rosetta fragment insertion with modified energy terms and sampling strategies optimized for IDPs (Physics-based + knowledge-based)	For IDPs requiring backbone ensembles with local structure propensity	Combines knowledge-based fragment sampling with physics-informed scoring; accurate for local motifs	Computationally intensive; coarse side-chain modeling; limited by fragment library quality
	DynamICE ²⁶	Uses a Generative Recurrent Neural Network (GRNN) to generate conformers, fine-tune with experimental data via a Bayesian reward function (X-EISD). (ML + ME)	When fixed pools are insufficient and data-driven generation of diverse IDP ensembles is needed.	Enables de novo, data-consistent conformer generation, integrates physical and experimental constraints, and improves iteratively.	Torsion-based (local) sampling makes it suboptimal to accurately satisfy distance restraints like NOEs and PREs.
	AF2-RAVE ²⁷	Combines AlphaFold2 (with reduced MSA) + short MD + machine learning (SPIB)	When diverse functional states are needed from sequence with minimal prior knowledge	Efficiently maps conformational landscape; captures metastable states; integrates physical and statistical modeling	Requires short MD; not suitable for ultra-fast motions; moderate compute setup needed
	bAles ²⁸	Bayesian integration of AlphaFold2 distograms as distance restraints into a random coil model as a prior	For generating atomistic IDP ensembles efficiently, especially when experimental data is limited	Fast; atomistic resolution; robust to AF2 errors; captures local & long-range features	Dependent on AF2 accuracy; limited for environment-sensitive or transient contacts
	AFlecto ²⁹	Builds IDP ensembles using fragment-based sampling informed by AlphaFold models, with secondary structure element (SSE) and steric constraints	When needing efficient conformer generation guided by AlphaFold predictions and optional SSE control	Fast and tunable; incorporates partial structure; no MD or force fields required	No dynamic sampling; side chains approximated with pseudo atoms

Supplementary Table 3. Experimental data integration techniques for conformational ensemble refinement.

Comparison of methods used to integrate experimental data (e.g., SAXS, NMR, smFRET) into structural ensemble modeling. Integration strategies include maximum entropy, maximum parsimony, and hybrid approaches, along with their assumptions and computational features.

Integration Strategy	Method	Core Principle	When to use	Strength	Limitation
Maximum Entropy	X-EISD ³⁰	Bayesian inference using error models to compare ensembles to various experimental data	When integrating NMR/SAXS data with structural ensembles	Handles multiple data types; includes uncertainty; enables reweighting	Needs accurate error models; relies on ensemble diversity
	EROS ³¹	Maximum-entropy reweighting of simulation ensemble to fit SAXS data	Flexible or disordered proteins with SAXS data and simulated ensembles	Avoids overfitting	Computationally demanding
	BioEn ³²	Combines EROS with simulations and reweighting to refine ensembles	When you want accurate ensemble refinement with control over data vs. model balance	Faster convergence; explores multiple confidence levels (θ) without extra simulations	More complex setup; depends on good sampling and reweighting
	COPER ³³	Uses convex optimization to reweight large ensembles to fit experimental averages	When working with very large ensembles and average experimental observables (e.g., SAXS, NMR)	Highly efficient and scalable; guarantees convergence and stability	Does <i>not</i> offer uncertainty estimates for the reweighted ensemble weights, unlike Bayesian approaches.
	BME ³⁴	Bayesian/maximum-entropy reweighting of ensembles using experimental averages with uncertainty	When combining simulations with experimental data (e.g., NMR, SAXS)	Accounts for data uncertainty; balances fit and prior; reusable weights	Requires tuning; depends on ensemble quality
	BEGR ³⁵	Reweights ensembles to match residue-level secondary structure content from chemical shifts	When experimental NMR chemical shifts are available and you aim to refine an existing IDP ensemble to achieve accurate secondary structure distributions	Achieves converged ensemble distributions; directly uses experimental constraints	Depends on quality of initial ensemble and secondary structure predictions from chemical shifts
	Metainference ³⁶	Simultaneously infers structural ensembles and experimental errors by integrating prior distributions and data uncertainty into a Bayesian framework.	Best suited for heterogeneous systems (e.g., IDPs, multi-state folded proteins) and when data noise and structural variability must be jointly modeled.	Metainference rigorously accounts for experimental noise, ensemble heterogeneity, and uncertainty in observables, uses a replica-averaged likelihood, and is integrated with molecular dynamics (e.g., PLUMED) to sample ensembles	Computationally demanding, requires integration with MD simulations, and may be complex for users unfamiliar with Bayesian inference.

				on-the-fly.	
	ENSEMBLE ³⁷	Adjusts weights on a pool of conformers to match experimental data while preserving ensemble diversity	When analyzing IDPs with diverse experimental restraints requiring a flexible, data-consistent ensemble	ENSEMBLE supports a wide range of experimental data types, allows flexible weighting of both conformers and data modules, preserves ensemble diversity, and integrates with external tools for accurate back-calculation.	It requires a large, pre-generated conformer pool, can be computationally intensive for large datasets, and offers lower interpretability due to the complexity of weighted ensembles.
Maximum Parsimony	EOM ⁶	Selects a small subset of conformers from a large random pool to best fit SAXS data, using a genetic algorithm.	When interpreting SAXS data for disordered or flexible proteins, especially when interested in overall shape and size distributions.	EOM is tailored for SAXS analysis, offers intuitive control over ensemble size and fit quality, produces interpretable distributions (e.g., Rg, Dmax), and is fast and easy to use for flexibility profiling.	It supports only SAXS data, relies on pre-generated conformers, and may oversimplify underlying structural heterogeneity due to its reliance on a small subset.
	ASTEROIDS ³⁸	Uses a genetic algorithm to select a fixed-size ensemble of conformers from a large pool that minimizes deviation from experimental data using a multi-component fitness function.	When you need a sparse, interpretable ensemble that fits RDCs, PREs, or NOEs, and want robust, global optimization of ensemble composition.	ASTEROIDS builds minimal ensembles with strong data agreement, uses explicit fitness weighting by data type, and employs genetic diversity mechanisms (mutation, crossover, tournaments) to avoid local minima and improve robustness across 2000+ generations.	It requires a fixed conformational collection size, lacks conformer weighting (binary inclusion only), may underrepresent structural diversity, and is tailored mainly for NMR restraints, limiting applicability to multimodal datasets.
	SES ³⁹	Selects the sparsest possible weighted ensemble from an oversampled conformer pool that fits experimental data within a defined error, using deterministic optimization.	Ideal for systems with underdetermined experimental data (e.g., RDCs, SAXS) where sparse conformational representation is desired and data scaling is uncertain.	SES provides a deterministic and scalable algorithm (Multi-OMP), avoids overfitting via rigorous -curve analysis, recovers robust ensemble weights, and requires no prior assumptions about weight distributions.	Requires an accurate prediction model for experimental observables, depends on quality and coverage of the input conformer pool, and currently supports only linearly averaged data (e.g., RDCs, SAXS).
Other	MMMx ⁴⁰	Reweights precomputed conformer ensembles to match experimental data (e.g., DEER, SAXS) via non-negative least squares or Monte Carlo sampling	When refining large conformer pools with multi-modal experimental restraints	Handles diverse data types; supports block-wise fitting and pruning; efficient for large ensembles	Reweighting can bias ensemble toward overrepresented states if prior is poor

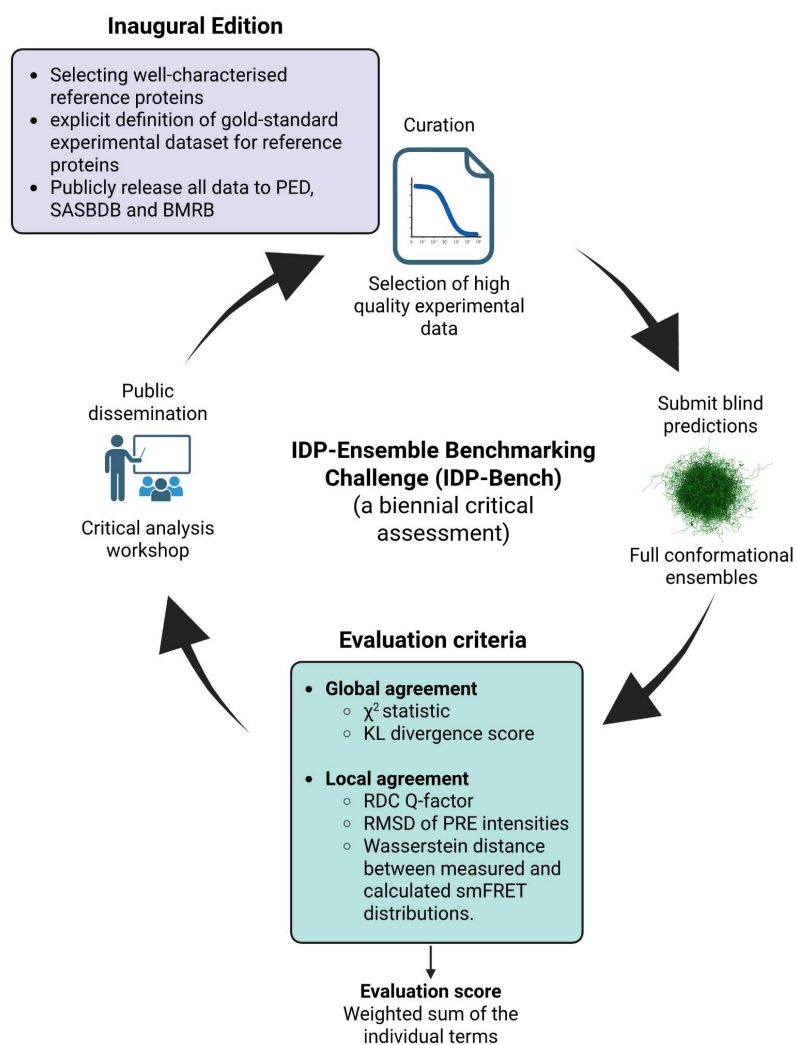
Supplementary Table 4. Summary of conformational ensemble comparison techniques. Summary of methods and metrics for quantifying similarities and differences between conformational ensembles. Includes descriptor-based, probabilistic, topological, and energy-based approaches, with notes on their computational complexity and interpretability.

Method/Metric	Description	Advantages	Disadvantages	Computational cost
Conformational Heterogeneity ⁴¹	Computes inter-residue distances across ensemble conformers to yield a single scalar quantifying conformational diversity.	Simple to implement; provides a quick overview of ensemble heterogeneity.	Overly simplistic; fails to capture local or multimodal structural differences.	Low
Residue Pair Distance Matrices ⁴²	Compares conformational ensembles via statistical analysis of residue-residue distance matrices.	Easy to compute and to interpret.	Comparisons based on median and standard deviation of distances, but not on the distribution of values..	Moderate
Kullback-Leibler (KL) or Jensen-Shannon (JS) Divergence based on Probability Density Estimates ^{43,44}	Estimates the probability density of a conformational space descriptor and quantifies the difference between ensembles based on KL or JS divergence.	Considers relevant descriptors of conformational variability.	Precise estimation of probability density is a challenge for highly dynamic systems.	Moderate to high
Wasserstein Distance Considering Local and Global Descriptors ⁴⁵	It computes the "cost" of transforming one ensemble distribution into another, using local descriptors (backbone angle distributions for each residue) and global descriptors (inter-residue distances).	Provides intuitive interpretation as "minimum effort" to morph one ensemble into another; robust to outliers; identifies the residues/regions that contribute most significantly; alignment-free.	Computationally intensive; scaling to large ensembles or high-dimensional descriptors can be demanding.	High
Topology-Based Metrics ⁴⁶	Compares ensembles by analyzing differences in contact-based graph topology.	Captures global rearrangements; alignment-free.	Lacks geometric detail; may oversimplify structure.	Moderate
Local Energy Trap Analysis ⁴⁷	Identifies residues with persistent low-energy conformers across ensemble	Highlights energetic hotspots; links sequence to ensemble behavior	Force field-dependent; may require good sampling	Moderate

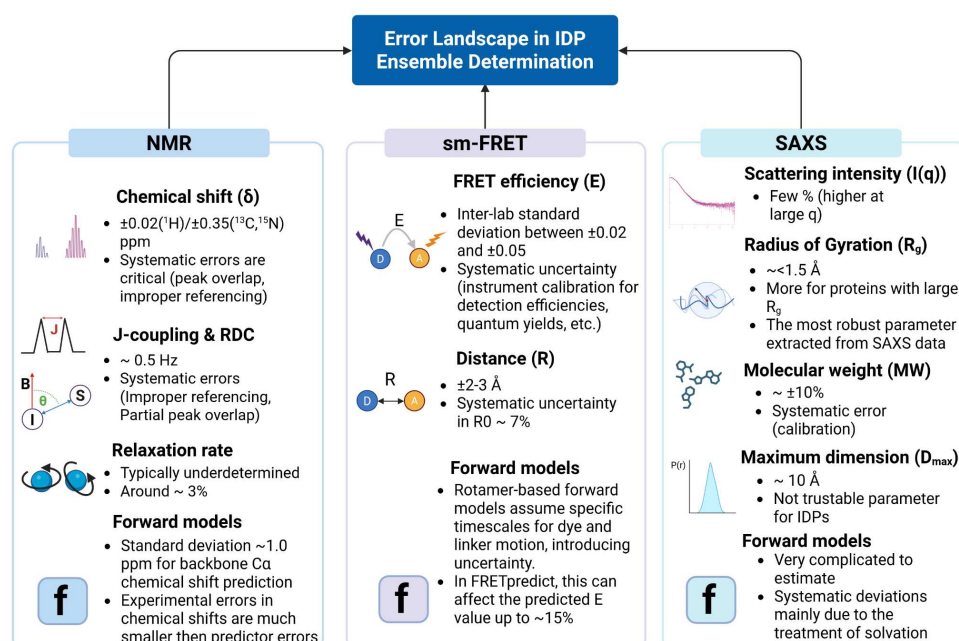
Supplementary Table 5. Proposed Scoring Metrics and Evaluation Criteria for IDP-Bench Submissions

Experimental Technique	Scoring metric	Structural Resolution	Advantages	Limitations
SAXS	χ^2 score between full SAXS profiles ($I(q)$ vs q) of experimental and computed ensemble; residual analysis	Global	Sensitive to overall dimensions and shape; fast and widely available	Low resolution; ensemble averaging; hydration layer and buffer condition effects
smFRET	Wasserstein distance between predicted and measured distance distributions	Local & Global	Reports on full distance distributions; resolves subpopulations	Requires careful control of dye labeling, orientation, and mobility
RDCs	Q-factor	Local	Sensitive to bond vector orientations and secondary structure propensities	May requires alignment tensors; sensitive to artifacts from alignment media
PRE	RMSE of PRE intensities	Local	Captures long-range transient contacts; informative for compactness	Sensitive to spin-label flexibility; forward model uncertainties
Chemical Shifts	RMSE or correlation coefficients	Local	Standardized and broadly available; reflects secondary structure propensity	Sensitive to sequence context and environmental effects
HDX-MS	Protection factor deviation	Local(per-segment)	Reports on transient structural protection; complementary to NMR	Resolution depends on peptide coverage; requires careful experimental setup
Hydrodynamic radius(e.g. from DOSY)	ΔR_h from predicted vs. measured	Global	Fast readout of chain compaction/expansion	Influenced by buffer viscosity, temperature, and shape assumptions

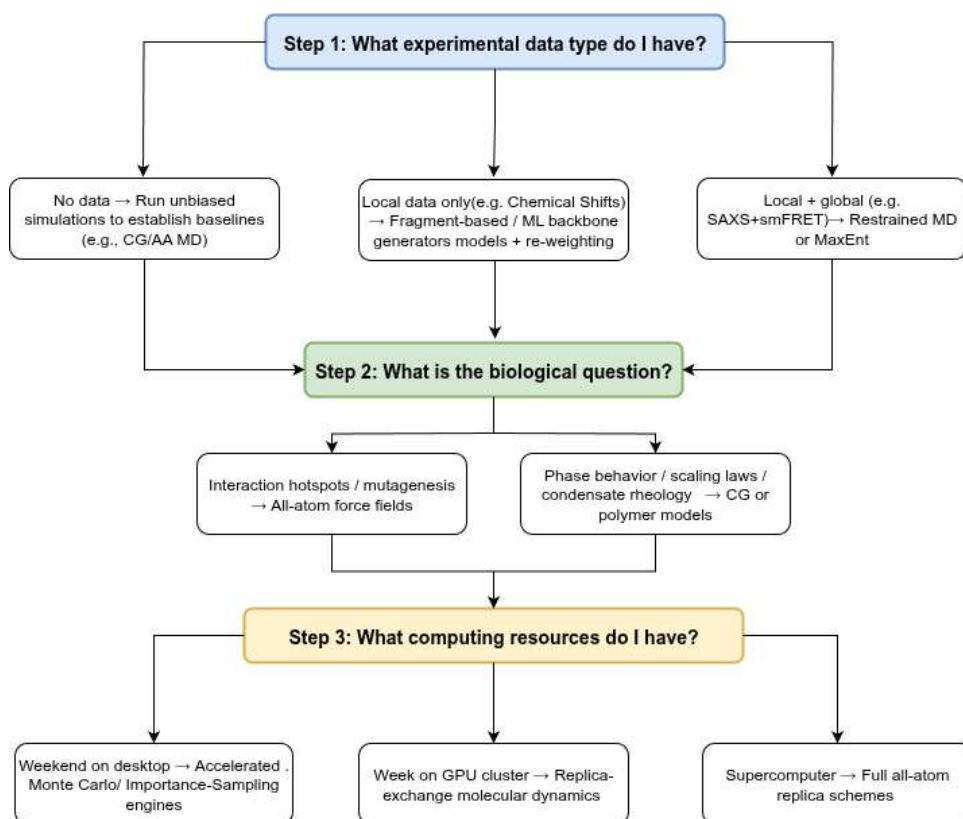
Supplementary Figure 1: Overview of the IDP-Bench benchmarking framework. The four-phase cycle includes data curation, blind ensemble prediction, multi-modal scoring, and public dissemination. Evaluation incorporates diverse experimental inputs and standardized metrics to assess global and local ensemble accuracy. Created in BioRender. Ghafouri, H. (2025) <https://BioRender.com/i36we79>.



Supplementary Figure 2. Error landscape in integrative ensemble determination of intrinsically disordered proteins (IDPs) using NMR, smFRET, and SAXS. The schematic summarizes typical experimental and modeling uncertainties associated with each method. **NMR:** Chemical shift uncertainties range from ± 0.02 ppm (^1H) to ± 0.35 ppm for $^{13}\text{C}/^{15}\text{N}$, though systematic referencing errors are often dominant. J-coupling and RDC errors are ~ 0.5 Hz; relaxation rates are typically underdetermined and often assigned an arbitrary error ($\sim 3\%$). For chemical shifts, experimental errors are much smaller than predictor errors. Forward models for chemical shift mainly have standard deviations of 1.0 ppm for backbone $\text{C}\alpha$ predictions on their training datasets⁴⁸. **smFRET:** FRET efficiency measurements have inter-lab standard deviations of ± 0.02 to ± 0.05 , systematic error dominated by instrument calibration (corresponding distance uncertainties are $\pm 2\text{--}3$ Å) and the uncertainty in the Förster radius (R_0), which has been estimated to be $\sim 7\%$ ⁴⁹. In sm-FRET forward models based on rotamer libraries, a source of uncertainty is the assumption about the timescales of linker and dye motions⁵⁰. For FRETpredict⁵¹, this uncertainty can affect the predicted FRET efficiency up to $\sim 15\%$. **SAXS:** Radius of gyration (R_g) is the most reliable parameter, with < 1.5 Å uncertainty for folded proteins (slightly higher for large R_g proteins)⁵². Molecular weight (MW) determination from scattering intensity has $\sim 10\%$ error, while maximum dimension (D_{max}) is poorly defined for IDPs and typically underestimates the true value by 10–15%. Systematic errors in SAXS forward models stem largely from solvent treatment and vary with protein features. Created in BioRender. Ghafouri, H. (2025) <https://BioRender.com/c2irt8f>



Supplementary Figure 3. Decision diagram for IDP conformational ensemble determination. A conceptual workflow diagram highlighting key decision points in the selection of computational and experimental strategies for generating and refining structural ensembles of IDPs.



Supplementary References

1. Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinforma. Oxf. Engl.* **28**, 1463–1470 (2012).
2. Feldman, H. J. & Hogue, C. W. V. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins Struct. Funct. Bioinforma.* **46**, 8–23 (2002).
3. Barozet, A. *et al.* MoMA-LoopSampler: a web server to exhaustively sample protein loop conformations. *Bioinformatics* **38**, 552–553 (2022).
4. Teixeira, J. M. C. *et al.* IDPConformerGenerator: A Flexible Software Suite for Sampling the Conformational Space of Disordered Protein States. *J. Phys. Chem. A* **126**, 5985–6003 (2022).
5. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 (2007).
6. Harmat, Z., Dudola, D. & Gáspári, Z. DIPEND: An Open-Source Pipeline to Generate Ensembles of Disordered Segments Using Neighbor-Dependent Backbone Preferences. *Biomolecules* **11**, 1505 (2021).
7. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
8. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
9. Piana, S., Robustelli, P., Tan, D., Chen, S. & Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes. *J. Chem. Theory Comput.* **16**, 2494–2507 (2020).
10. Zhou, R. Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol. Biol. Clifton NJ* **350**, 205–223 (2007).
11. Bonomi, M., Barducci, A. & Parrinello, M. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J. Comput. Chem.* **30**, 1615–1621 (2009).
12. Vitalis, A. & Pappu, R. V. ABSINTH: A new continuum solvation model for simulations

- of polypeptides in aqueous solutions. *J. Comput. Chem.* **30**, 673–699 (2009).
13. Souza, P. C. T. *et al.* Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **18**, 382–388 (2021).
 14. Joseph, J. A. *et al.* Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **1**, 732–743 (2021).
 15. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2111696118 (2021).
 16. Cao, F., von Bülow, S., Tesei, G. & Lindorff-Larsen, K. A coarse-grained model for disordered and multi-domain proteins. *Protein Sci. Publ. Protein Soc.* **33**, e5172 (2024).
 17. Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **14**, e1005941 (2018).
 18. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
 19. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Comput. Biol.* **20**, e1012144 (2024).
 20. Novak, B., Lotthammer, J. M., Emenecker, R. J. & Holehouse, A. S. Accurate predictions of conformational ensembles of disordered proteins with STARLING. 2025.02.14.638373 Preprint at <https://doi.org/10.1101/2025.02.14.638373> (2025).
 21. Zhu, J. *et al.* Precise Generation of Conformational Ensembles for Intrinsically Disordered Proteins via Fine-tuned Diffusion Models. 2024.05.05.592611 Preprint at <https://doi.org/10.1101/2024.05.05.592611> (2024).
 22. Zhu, J. *et al.* Phanto-IDP: compact model for precise intrinsically disordered protein backbone generation and enhanced sampling. *Brief. Bioinform.* **25**, bbad429 (2024).
 23. Lewis, S. *et al.* Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* eadv9817 (2025).
 24. Brotzakis, Z. F., Zhang, S., Murtada, M. H. & Vendruscolo, M. AlphaFold prediction of structural ensembles of disordered proteins. *Nat. Commun.* **16**, 1632 (2025).
 25. Ferrie, J. J. & Petersson, E. J. A Unified De Novo Approach for Predicting the Structures of Ordered and Disordered Proteins. *J. Phys. Chem. B* **124**, 5538–5548 (2020).

26. Zhang, O. *et al.* Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *J. Chem. Phys.* **158**, 174113 (2023).
27. Teng, D., Meraz, V. J., Aranganathan, A., Gu, X. & Tiwary, P. af2rave: protein ensemble generation with physics-based sampling. *Digit. Discov.* (2025).
28. Schnapka, V., Morozova, T., Sen, S. & Bonomi, M. Atomic resolution ensembles of intrinsically disordered and multi-domain proteins with AlphaFold. 2025.06.18.660298 Preprint at <https://doi.org/10.1101/2025.06.18.660298> (2025).
29. Pajkos, M., Clerc, I., Zanon, C., Bernadó, P. & Cortés, J. AFflecto: A web server to generate conformational ensembles of flexible proteins from AlphaFold models. *J. Mol. Biol.* 169003 (2025).
30. Lincoff, J. *et al.* Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun. Chem.* **3**, 1–12 (2020).
31. Różycki, B., Kim, Y. C. & Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Struct. Lond. Engl.* **1993** **19**, 109–116 (2011).
32. Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
33. Leung, H. T. A. *et al.* A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **12**, 383–394 (2016).
34. Bottaro, S., Bengtsen, T. & Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *Methods Mol. Biol. Clifton NJ* **2112**, 219–240 (2020).
35. Ytreberg, F. M., Borchers, W., Wu, H. & Daughdrill, G. W. Using chemical shifts to generate structural ensembles for intrinsically disordered proteins with converged distributions of secondary structure. *Intrinsically Disord. Proteins* **3**, e984565 (2015).
36. Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
37. Choy, W. Y. & Forman-Kay, J. D. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308**, 1011–1032 (2001).
38. Nodet, G. *et al.* Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. *J.*

- Am. Chem. Soc.* **131**, 17908–17918 (2009).
39. Berlin, K. *et al.* Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J. Am. Chem. Soc.* **135**, 16595–16609 (2013).
 40. Jeschke, G. Protein ensemble modeling and analysis with MMMx. *Protein Sci.* **33**, e4906 (2024).
 41. Lyle, N., Das, R. K. & Pappu, R. V. A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.* **139**, 121907 (2013).
 42. Lazar, T. *et al.* Distance-Based Metrics for Comparing Conformational Ensembles of Intrinsically Disordered Proteins. *Biophys. J.* **118**, 2952–2965 (2020).
 43. Aina, A., Hsueh, S. C. C. & Plotkin, S. S. PROTHON: A Local Order Parameter-Based Method for Efficient Comparison of Protein Ensembles. *J. Chem. Inf. Model.* **63**, 3453–3461 (2023).
 44. Lindorff-Larsen, K. & Ferkinghoff-Borg, J. Similarity measures for protein ensembles. *PloS One* **4**, e4203 (2009).
 45. González-Delgado, J. *et al.* WASCO: A Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins. *J. Mol. Biol.* **435**, 168053 (2023).
 46. Scalvini, B. *et al.* Circuit Topology Approach for the Comparative Analysis of Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **63**, 2586–2602 (2023).
 47. Lotthammer, J. M. & Holehouse, A. S. Disentangling Folding from Energetic Traps in Simulations of Disordered Proteins. *J. Chem. Inf. Model.* **65**, 2897–2910 (2025).
 48. Cavender, C. E. *et al.* Structure-Based Experimental Datasets for Benchmarking Protein Simulation Force Fields. Preprint at <https://doi.org/10.48550/arXiv.2303.11056> (2025).
 49. Hellenkamp, B. *et al.* Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nat. Methods* **15**, 669–676 (2018).
 50. Nüesch, M., Ivanović, M. T., Nettels, D., Best, R. B. & Schuler, B. Accuracy of distance distributions and dynamics from single-molecule FRET. *Biophys. J.* (2025).
 51. Montepietra, D. *et al.* FRETpredict: a Python package for FRET efficiency predictions using rotamer libraries. *Commun. Biol.* **7**, 298 (2024).
 52. Trewhella, J. *et al.* A round-robin approach provides a detailed assessment of

biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking. *Acta Crystallogr. Sect. Struct. Biol.* **78**, 1315–1336 (2022).

Appendix C: Advances in the determination of conformational ensembles

Advances in the determination of disordered protein ensemble

Hamidreza Ghafouri¹, Silvio C.E. Tossatto^{1,*}, Alexander Miguel Monzon^{1,*}

¹Department of Biomedical Sciences, University of Padova, Italy. Via Ugo Bassi 58/B, 35131 Padova, Italy

*Corresponding authors

Email addresses: alexander.monzon@unipd.it and silvio.tosatto@unipd.it.

Supplementary Material

Table 1. Recent experimental data integration techniques

Integration Strategy	Tool	Underlying concept	Application Context	Advantages	Limitations
Maximum Entropy	X-EISD[1]	Bayesian inference with explicit error models to refine ensembles with diverse experimental data.	Integrating NMR or SAXS data into structural ensembles.	Can incorporate multiple data types, accounts for uncertainty, enables ensemble reweighting.	Relies on accurate error models and sufficient ensemble diversity.
	EROS[2]	Maximum-entropy reweighting of simulations against SAXS measurements.	Flexible/disordered proteins with SAXS and simulated ensembles.	Reduces risk of overfitting.	Computationally demanding.
	BioEn[3]	Combines EROS with simulation and reweighting for refinement.	When precise refinement and balancing data-model agreement is needed.	Fast convergence; allows tuning confidence levels (θ) without extra simulations.	More complex workflow; requires well-sampled ensembles.
	COPER[4]	Convex optimization for reweighting large ensembles to match averaged experimental data.	Very large ensembles constrained by averages (e.g., SAXS, NMR).	Highly efficient, scalable, guarantees stable convergence.	Provides no uncertainty estimates for weights.
	BME[5]	Bayesian/maximum-entropy integration of simulations with experimental averages.	Combining ensembles with experimental restraints (NMR, SAXS).	Balances fit with priors, accounts for uncertainty, transferable weights.	Requires parameter tuning, sensitive to ensemble quality.
	EnsRefEx[6]	Post-processing MD snapshots by optimizing weights to minimize Kullback–Leibler divergence, while ensuring deviations from experimental observables remain below a set error tolerance.	Post-processing MD snapshots by optimizing weights to minimize Kullback–Leibler divergence, while ensuring deviations from experimental observables remain below a set error tolerance.	Conserves original ensemble distribution; prevents overfitting via error tolerance; computationally flexible.	Relies on accuracy and coverage of the starting ensemble; cannot recover missing conformational states; performance depends on force field quality.

	Metainference[7]	Bayesian framework integrating data uncertainty, noise, and structural heterogeneity via replica-averaged likelihood.	Heterogeneous systems where data noise and ensemble variability must be modeled jointly.	Rigorous treatment of uncertainty; integrates with MD for on-the-fly sampling.	Computationally heavy, requires Bayesian expertise and MD coupling.
	BICePsv2.0[8]	Bayesian reweighting of pre-generated conformational ensembles using posterior sampling to fit experimental observables (e.g., NMR).	When refining ensembles with diverse Experimental data (NOEs, J-couplings, chemical shifts, HDX) without running new simulations.	Rigorous Bayesian framework; supports multiple data types; provides objective BICePs score for model comparison; includes convergence diagnostics.	Dependent on accurate forward models and quality of prior ensemble; requires careful MCMC convergence analysis.
	EOM[9]	Selects a subset of conformers via genetic algorithm to match SAXS profiles.	SAXS-based analysis of disordered/flexible proteins.	Tailored for SAXS, intuitive ensemble control, produces interpretable Rg and Dmax distributions.	Limited to SAXS; may oversimplify heterogeneity.
Maximum Parsimony	ASTEROIDS[10]	Genetic algorithm selects fixed-size ensemble minimizing deviations from experimental data..	Sparse interpretable ensembles constrained by NMR data (RDCs, PREs, NOEs)..	Builds minimal ensembles with strong fit, uses genetic diversity to avoid local minima.	Fixed ensemble size, binary conformer inclusion, limited applicability beyond NMR.

SES[11]	Deterministic optimization selects the sparsest weighted ensemble consistent with data.	Underdetermined systems (e.g., RDCs, SAXS) where sparse representation is needed.	Deterministic, scalable, avoids overfitting, robust weight recovery..	Requires accurate observable prediction; depends on conformer pool quality; limited to linearly averaged data.
MESMER[12]	Uses genetic algorithms to select minimal ensembles from conformer pools that best fit multiple experimental datasets (e.g., SAXS, RDCs, PCS, DEER).	For heterogeneous systems requiring integration of diverse experimental restraints into compact, interpretable ensembles.	Handles multiple data types; modular and extensible; avoids overfitting by identifying several near-optimal solutions.	Dependent on conformer pool quality; can underestimate heterogeneity; computationally intensive for large systems.

Supplementary References

1. Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes G-NW, Gradinaru CC, Forman-Kay JD, Head-Gordon T: **Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states**. *Commun Chem* 2020, **3**:1–12.
2. Różycki B, Kim YC, Hummer G: **SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions**. *Struct Lond Engl* 1993 2011, **19**:109–116.
3. Hummer G, Köfinger J: **Bayesian ensemble refinement by replica simulations and reweighting**. *J Chem Phys* 2015, **143**:243150.
4. Leung HTA, Bignucolo O, Aregger R, Dames SA, Mazur A, Bernèche S, Grzesiek S: **A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content**. *J Chem Theory Comput* 2016, **12**:383–394.
5. Bottaro S, Bengtsen T, Lindorff-Larsen K: **Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach**. *Methods Mol Biol Clifton NJ* 2020, **2112**:219–240.
6. Yamamori Y, Tomii K: **An ensemble reweighting method for combining the information of experiments and simulations**. *Chem Phys Lett* 2021, **779**:138821.
7. Bonomi M, Camilloni C, Cavalli A, Vendruscolo M: **Metainference: A Bayesian inference method for heterogeneous systems**. *Sci Adv* 2016, **2**:e1501177.
8. Raddi RM, Ge Y, Voelz VA: **BICePs v2.0: Software for Ensemble Reweighting Using Bayesian Inference of Conformational Populations**. *J Chem Inf Model* 2023, **63**:2370–2381.
9. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural characterization of flexible proteins using small-angle X-ray scattering**. *J Am Chem Soc* 2007, **129**:5656–5664.
10. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M: **Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings**. *J Am Chem Soc* 2009, **131**:17908–17918.
11. Berlin K, Castañeda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D: **Recovering a representative conformational ensemble from underdetermined macromolecular structural data**. *J Am Chem Soc* 2013, **135**:16595–16609.
12. Ihms EC, Foster MP: **MESMER: minimal ensemble solutions to multiple experimental restraints**. *Bioinforma Oxf Engl* 2015, **31**:1951–1958.

Appendix D: IDPET supplementary materials

IDPEnsembleTools: an open-source library for analysis of conformational ensembles of disordered proteins

Hamidreza Ghafouri^{1,^}, Giacomo Janson^{2,^}, Silvio C.E. Tosatto^{1,3,*}, Alexander Miguel Monzon^{1,*}

¹Department of Biomedical Sciences, University of Padova, Padova 35121, Italy

²Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States

³Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari 70126, Italy

[^] These authors contributed equally to this work.

^{*}Corresponding authors. Department of Biomedical Sciences, University of Padova, Padova 35121, Italy.

E-mails: alexander.monzon@unipd.it (A.M.M.) and silvio.tosatto@unipd.it (S.C.E.T.)

Supplementary Material

Methods

Dimensionality reduction

IDPET offers PCA (Principal Component Analysis) and Kernel PCA (KPCA), t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) as three dimensionality reduction methods for studying the IDPs/IDRs ensembles' properties. These methods are applicable on extracted feature matrices from ensembles. Currently, IDPET can transform ensembles into six different feature matrices based on phi-psi angles, intra atomic Ca distances, pairwise RMSD, trRosetta-style phi angles, trRosetta-style omega angles and alpha angles (**Supplementary Table.1**). Then, based on the selected feature, each conformation in the ensemble will be transformed to a 1D array containing extracted feature data and consequently the ensembles of conformations will be introduced as a 2D matrix to the selected dimensionality reduction method.

Supplementary Table 1: Description of the features that IDPET could extract from the ensembles and proper dimensionality reduction techniques

Feature	Description	Dimensionality reduction methods
phi-psi angles	Phi and Psi are the dihedral angles around the N-C α and C α -C atoms respectively	KPCA, circular UMAP and t-SNE(using circular metric)
intra atomic Ca distances	Interatomic distances between Ca atoms	PCA, UMAP and t-SNE (using euclidean metric)
pairwise RMSD	Generate all-vs-all matrix contains all pairwise RMSD values between the conformations in the ensembles	UMAP and t-SNE (using precomputed metric for symmetric matrices)
trRosetta-style phi angles	The planar angle between three atoms of two residues i and j (Ca[i], Cb[i] and Cb[j])	KPCA, UMAP and t-SNE (using circular metric)
trRosetta-style omega angles	The dihedral angle between the planes formed by four atoms of 2 residues i and j (Ca[i], Cb[i], Cb[j], and Ca[j])	KPCA, UMAP and t-SNE (using circular metric)
alpha angles	The dihedral angle between the planes formed by four Ca atoms of 4 consecutive residues (Ca[i], Ca[i+1], Ca[i+2], and Ca[i+3])	KPCA, UMAP and t-SNE (using circular metric)

a. PCA & KPCA

IDPET implements user-friendly interfaces for PCA and KPCA using the scikit-learn library to explore conformational ensemble properties. KPCA is particularly suited for analyzing non-linear and periodic data such as angular variables. To account for the circular nature of angular features, each angle is transformed into its sine and cosine components before computing the kernel matrix using a custom distance-based kernel function. This transformation preserves angular periodicity and enables meaningful dimensionality reduction in the latent space, offering insight into conformational diversity.

b. t-SNE

As part of its dimensionality reduction toolkit, IDPET includes an adapted implementation of t-SNE based on the pipeline suggested by Robustelli et al. ¹, which combines t-SNE with k-means clustering to effectively group heterogeneous conformational ensembles. Two hyperparameters are tuned: the number of k-means clusters and the t-SNE perplexity value, which balances local and global relationships within the data by approximating the number of nearest neighbors. IDPET performs a grid search over these parameters, selecting the optimal combination using the Silhouette score, a metric that evaluates both intra-cluster cohesion and inter-cluster separation to assess clustering quality.

c. UMAP

IDPET also includes UMAP, a non-linear dimensionality reduction technique based on fuzzy topological representations of data. UMAP generally outperforms linear methods in preserving complex structural relationships and is computationally more efficient than t-SNE, with comparable embedding quality. The two main hyperparameters are: the number of neighbors (analogous to t-SNE's perplexity), which governs the balance between local and global data structure, and the minimum distance between points, which controls cluster compactness versus preservation of global structure. IDPET adopts the same optimization strategy as for t-SNE, tuning these parameters by maximizing the Silhouette score across combinations of neighbor counts and cluster numbers.

Ensemble comparison

IDPET implements a set of scores based on Jensen–Shannon divergence (JSD) to quantitatively compare conformational ensembles of proteins with the same length LLL, even when their sequences differ. These scores evaluate the similarity between probability distributions of structural features, such as interatomic distances or torsion angles, derived from the ensembles. To compute the JSD for continuous variables, the values are first discretized into histograms with N_{bins} , following established practices ². Given two

distributions P and Q corresponding to a specific molecular feature from each ensemble, their JSD is defined as:

$$JSD(P\|Q) = \frac{1}{2}(KLD(P\|M) + KLD(Q\|M))$$

where $M = \frac{1}{2}(P + Q)$ is a mixture, and KLD is the Kullback-Leibler divergence, expressed as:

$$KLD(X\|M) = \sum_k^{N_{bins}} X_k \log\left(\frac{X_k}{M_k}\right)$$

where k is the bin index and X_k and M_k are the frequencies for bin k estimated from histogram data, with $M_k = \frac{1}{2}(P_k + Q_k)$. No pseudo-counts are used and bins with $X_k = 0$ have zero contribution. JSD scores range from 0 (P and Q have identical counts) to a maximum of $\log \log(2) \approx 0.6931$ (no bin has at least one count from both P and Q).

IDPET implements the following JSD based scores:

a. Carbon Alpha Distance Average JSD (adaJSD)

The adaJSD score compares C α -C α distances in two ensembles and is defined as:

$$adaJSD = \frac{1}{N_{pairs}} \sum_{j-i>1} JSD(D_{ij}^{[A]} \| D_{ij}^{[B]})$$

where $D_{ij}^{[A]}$ and $D_{ij}^{[B]}$ are the distance distributions between residue i and j in ensembles A and B , and $N_{pairs} = (L - 1)(L - 2)/2$ is the total number of distances evaluated. For each distance, its histogram range is defined by its minimum and maximum values in the ensembles.

b. Alpha Torsion Average JSD (ataJSD)

The ataJSD score compares α angle distributions:

$$ataJSD = \frac{1}{N_{alpha}} \sum_{i=2} JSD(T_{i-1,i,i+1,i+2}^{[A]} \| T_{i-1,i,i+1,i+2}^{[B]})$$

where $N_{alpha} = L - 3$ is the number α angles in a protein and $T_{i-1,i,i+2,i+3}^{[A]}$ and $T_{i-1,i,i+2,i+3}^{[B]}$ are the distributions of α angles formed by residue i and its neighbors. The histogram range is always $-\pi$ to π .

c. RAMAchandran plot average JSD (ramaJSD)

The ramaJSD (RAMAchandran plot average JSD) score compares joint φ and ψ angle distributions:

$$ramaJSD = \frac{1}{N_{rama}} \sum_{i=2} JSD(R_i^{[A]} \parallel R_i^{[B]})$$

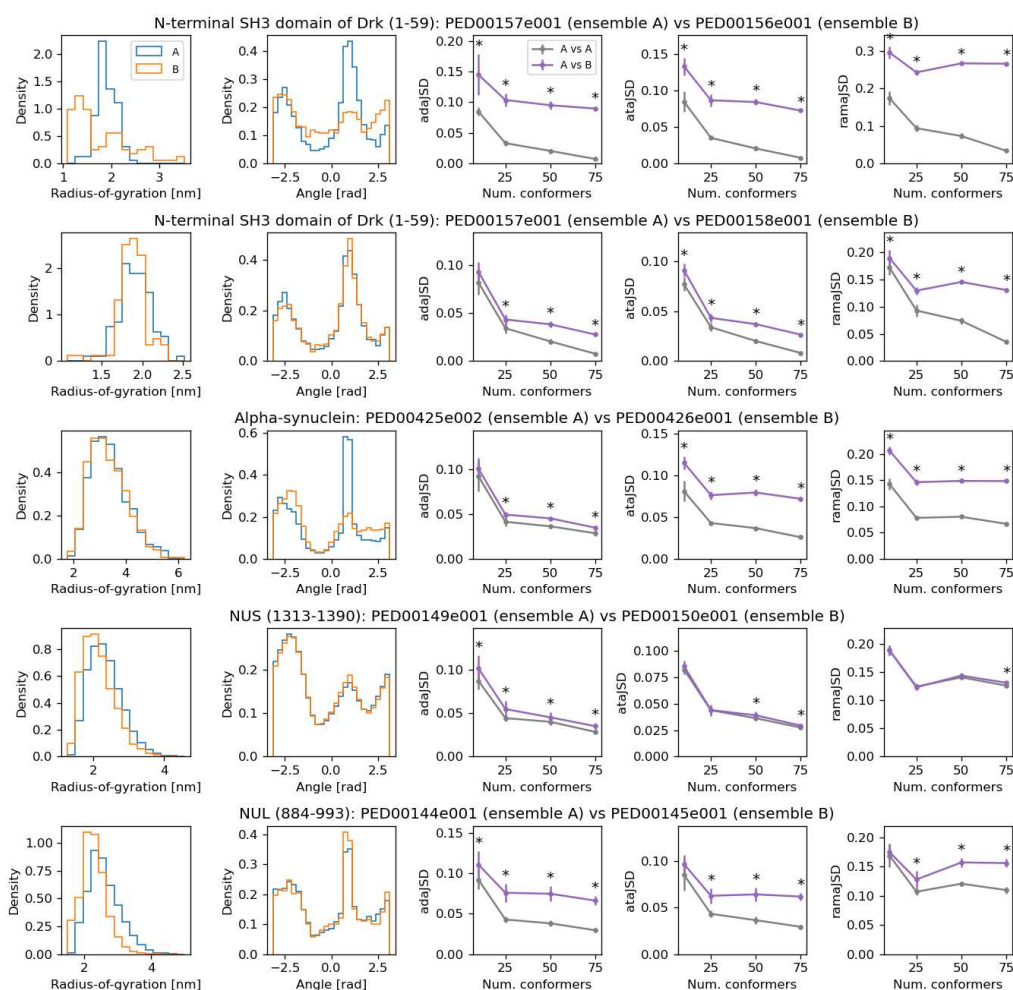
where $N_{rama} = L - 2$ is the number of residues with both φ and ψ values and $R_i^{[A]}$ and $R_i^{[B]}$ are the joint distributions of φ and ψ for residue i . For both angles, the histogram range is always $-\pi$ to π split into N_{bins} , resulting in a 2d histogram with an effective bin number of N_{bins}^2 .

Effect of ensemble size

A crucial parameter for computing above mentioned scores is the number of bins N_{bins} . Large N_{bins} values will lead to inaccurate results if ensembles contain few conformers. While users can manually set N_{bins} , by default IDPET automatically assigns it via the square root rule:

$$N_{bins} = \min(\lceil \sqrt{n} \rceil, 50)$$

where n is the number of conformers in the smaller ensembles. Our benchmarking (see **Supplementary Figure 1**) shows that IDPET scores reliably capture ensemble differences with as few as 25 conformers. However, when comparing smaller ensembles, the discriminative power of the scores significantly decreases.



Supplementary Figure 1. Relationship between IDPET comparison scores and ensemble size.

Each row corresponds to two PED ensembles (A and B) for the same protein. The first two columns show histograms of radius of gyration and α angles, summarizing global and local structural properties, respectively. The remaining columns plot adaJSD, ataJSD, and ramaJSD scores as a function of the number of bootstrapped conformers. Gray curves show A vs A auto-comparisons; purple curves show A vs B comparisons. Each point represents the average of $n = 10$ bootstrap replicates (error bars: standard deviation). Asterisks indicate statistically significant differences (Mann–Whitney U test, $p < 0.05$) between A vs A and A vs B scores for ensemble sizes of 10, 25, 50, and 75 conformers. When A and B share similar structural properties (first two columns), differences in JSD scores become indistinguishable from background variation at low ensemble sizes. These results highlight the need for caution when interpreting scores from small ensembles (fewer than 50 conformers).

Statistical analysis

Due to the finite size of protein ensembles analyzed in IDPET, a non-zero JSD score does not necessarily imply a statistically significant difference between two ensembles. To address this, IDPET includes an optional bootstrap-based statistical analysis, particularly recommended for small ensembles (e.g., fewer than 50 conformers).

This analysis can be applied to all comparison scores implemented in IDPET (adaJSD, ataJSD, and ramaJSD), and proceeds as follows:

1. Inter-ensemble bootstrapping:

Conformers are randomly sampled (with replacement) from ensembles A and B and compared over $N_{bootstrap}$ iterations to generate a distribution of inter-ensemble scores.

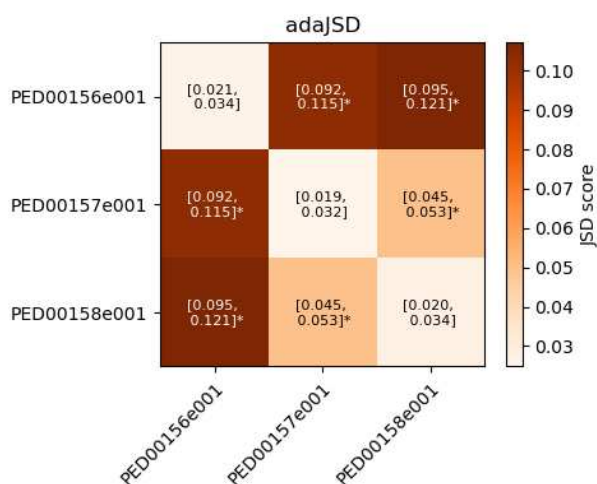
2. Intra-ensemble bootstrapping:

Each ensemble (A and B) is also bootstrapped and compared to itself $N_{bootstrap}$ times. The highest average intra-ensemble score from these self-comparisons is used as the reference.

3. Significance testing:

A one-sided Mann–Whitney U test is applied to compare the distributions of inter-ensemble and intra-ensemble scores. The null hypothesis is that both distributions are equal; the alternative hypothesis is that the inter-ensemble scores are stochastically greater. The resulting p-value is reported to assess statistical significance.

Additionally, IDPET reports bootstrap confidence intervals for the score differences using the percentile method (**Supplementary Figure 2**). This approach allows users to gauge whether differences between ensembles are meaningful or arise from statistical noise. While this analysis is critical for small ensembles, it is generally unnecessary for large ensembles (e.g., >1,000 conformers), where comparison scores are already reliable estimators of structural divergence.



Supplementary Figure 2. It shows the result of bootstrapping by providing the confidence intervals and the statistical significance for both intra-ensemble and inter-ensemble comparisons based on adaJSD score. In this example, while PED00157 and PED00158 have low adaJSD scores, these scores appear to be significantly higher than those obtained by auto-comparisons, hinting at a systematic difference.

Visualization

In IDPET, a dedicated module has been developed specifically for data visualization, offering multiple modes to analyze conformational ensembles effectively. Each visualization method includes customizable options, allowing users to easily adjust and refine the plots according to their needs. The table below summarizes the different visualization modes available in IDPET:

Type of analysis	Type(s) of visualization	Main options supported
<i>Distributions of ensemble properties</i> - <i>R_g</i> - <i>Asphericity</i> - <i>Prolateness</i> - <i>End-to-end distance</i> - <i>Global SASA</i>	- Histogram - Violin plot	- number of bins - mean/median indicator - multiple plots - color
<i>Ensemble comparison</i>	- comparison matrix heatmap	
<i>Dimensionality reduction</i>	- Density plot - Scatter plots	- color by different labels
<i>DSSP</i>	- relative DSSP content plot based on selected DSSP code	- DSSP code among "H" for helix, "E" for strand and "C" for coil
<i>Distance/contact maps</i>	- average distance map through ensembles - contact probability map based on a selected contact threshold	- contact threshold - turn to log-scale
<i>Angles</i>	- Ramachandran plot - Alpha angle	
<i>Site-specific order/flexibility parameters</i>	- Site-specific order/flexibility parameter for each residue in the ensemble	- residue pointer
<i>Per residue average SASA</i>	- plot the average SASA for each residue among all conformations in the ensemble	- residue pointer

Operation	A β 40, n=300	A β 40, n=2998	α -syn, n=300	α -syn, n=2998
Initial download+process from PED. Download PDB files from PED and process them with MDTraj. Performed only the first time a user retrieves the ensembles from PED.	-	105.9	-	388.3
Ensemble loading. Loading trajectory files, performed every time users analyze the ensembles.	-	1.0	-	3.4
Quick global descriptors. Run sequentially these methods of the Visualization class: radius_of_gyration, end_to_end_distances, asphericity, prolateness	0.5	1.3	0.7	2.9
Global SASA. Run the visualization.global_sasa method	15.4	153.3	56.4	562.6
adaJSD. Run visualization.comparison_matrix(score="adaJSD", bootstrap_iters=None)	0.1	0.3	1.3	4.6
adaJSD bootstrap x 25. Run visualization.comparison_matrix(score="adaJSD", bootstrap_iters=25)	4.9	13.7	64.7	223.9
PCA with Ca-dist. Run analysis.extract_features(featurization='ca_dist') and analysis.reduce_features(method='pca')	0.02	0.1	0.2	1.2
kPCA with phi/psi. Run analysis.extract_features(featurization='phi_psi') and analysis.reduce_features(method='kpca', circular=True)	0.07	53.4	0.1	54.2
t-SNE selection with Ca-dist. Run analysis.extract_features(featurization='ca_dist') and analysis.reduce_features(method='kpca', **tsne_params)	25.0	387.6	30.7	445.9
t-SNE selection with phi/psi. Run analysis.extract_features(featurization='phi_psi') and analysis.reduce_features(method='kpca', circular=True, **tsne_params)	38.4	1427.8	46.6	1931.3

Supplementary Table 1. Wall-clock times for common IDPET operations. The analyses were performed using 3 PED ensembles for A β 40 (PED00531-PED00533, with 40 residues) or α -synuclein (PED00543-PED00545, with 141 residues) with all the conformers available (2998 per ensemble) or with smaller subsamples (300 per ensemble). For all t-SNE analyses: perplexity_vals=[10, 20, 50, 100], range_n_clusters=range(2, 5, 1). All tests were run with Python 3.10 on a MacBook Air (2020) platform with a M1 processor and 16 GB RAM. Wall-clock times are reported in seconds.

Feature	IDPET	SOURSOP
Input	Supports multi-model PDB files and trajectory files (.dcd, .xtc).	Supports trajectory files (.dcd, .xtc).
Analysis pipeline	Can handle multiple ensembles, each containing a single chain.	Can handle a single ensemble containing multiple chains.
Ensemble features	Calculates common global and local structural descriptors such as radius of gyration (Rg), end-to-end distance (Ree), scaling exponent, and backbone dihedral angles.	Computes a broader range of structural features,
Visualization	Includes a built-in visualization module for plotting and interactive analysis.	Does not include a dedicated visualization module, but example Jupyter notebooks are provided
Experimental observables	Does not directly compute experimental observables.	Can provide experimental observables such as chemical shifts and PRE values.
Dimensionality reduction	Supports dimensionality reduction methods such as t-SNE and PCA..	Currently does not include dimensionality reduction features.
Ensemble comparison and validation	Provides quantitative ensemble comparison using Jensen–Shannon divergence (JSD) metrics.	Includes functions for assessing sampling quality.

Supplementary Table 2. Comparison between the IDPET and SOURSOP packages for analyzing conformational ensembles of IDPs. Various aspects of the two packages are compared. For a quantitative comparison, a Jupyter notebook demonstrating their functionalities is provided in the GitHub repository of the package.

Supplementary References:

1. Appadurai, R., Koneru, J. K., Bonomi, M., Robustelli, P. & Srivastava, A. Clustering Heterogeneous Conformational Ensembles of Intrinsically Disordered Proteins with t-Distributed Stochastic Neighbor Embedding. *J. Chem. Theory Comput.* **19**, 4711–4727 (2023).
2. Lindorff-Larsen, K. & Ferkinghoff-Borg, J. Similarity measures for protein ensembles. *PloS One* **4**, e4203 (2009).

References

- [1] Louis Rosenfeld. *Origins of Clinical Chemistry: The Evolution of Protein Analysis*. en. Elsevier, Dec. 2012.
- [2] Toyokazu Ishida. “Effects of Point Mutation on Enzymatic Activity: Correlation between Protein Electronic Structure and Motion in Chorismate Mutase Reaction”. In: *Journal of the American Chemical Society* 132.20 (May 2010), pp. 7104–7118. DOI: 10.1021/ja100744h.
- [3] “Fibrous Proteins: New Structural and Functional Aspects Revealed”. en-US. In: *Advances in Protein Chemistry*. Vol. 70. Academic Press, Jan. 2005, pp. 1–10. DOI: 10.1016/S0065-3233(05)70001-2.
- [4] Ferrannini et al. “Insulin: new roles for an ancient hormone”. en. In: *European Journal of Clinical Investigation* 29.10 (1999), pp. 842–852. DOI: 10.1046/j.1365-2362.1999.00536.x.
- [5] Allan Fletcher. “The cell membrane and receptors”. In: *Anaesthesia & Intensive Care Medicine*. Neurosurgical Anaesthesia 18.6 (June 2017), pp. 316–320. DOI: 10.1016/j.mpaic.2017.03.005.
- [6] Mark L. Chiu et al. “Antibody Structure and Function: The Basis for Engineering Therapeutics”. en. In: *Antibodies* 8.4 (Dec. 2019), p. 55. DOI: 10.3390/antib8040055.
- [7] Milton H. Saier Jr et al. “The Transporter Classification Database (TCDB): recent advances”. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D372–D379. DOI: 10.1093/nar/gkv1103.
- [8] Kutti R. Vinothkumar and Richard Henderson. “Structures of membrane proteins”. en. In: *Quarterly Reviews of Biophysics* 43.1 (Feb. 2010), pp. 65–158. DOI: 10.1017/S0033583510000041.
- [9] Rhiannon Morris, Katrina A. Black, and Elliott J. Stollar. “Uncovering protein function: from classification to complexes”. In: *Essays in Biochemistry* 66.3 (Aug. 2022), pp. 255–285. DOI: 10.1042/EBC20200108.
- [10] Prakash Kulkarni et al. “Intrinsically disordered proteins: Ensembles at the limits of Anfinsen’s dogma”. In: *Biophysics Reviews* 3.1 (Mar. 2022), p. 011306. DOI: 10.1063/5.0080512.
- [11] C. B. Anfinsen et al. “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain”. In: *Proceedings of the National Academy of Sciences* 47.9 (Sept. 1961), pp. 1309–1314. DOI: 10.1073/pnas.47.9.1309.
- [12] Christian B. Anfinsen. “Principles that Govern the Folding of Protein Chains”. In: *Science* 181.4096 (July 1973), pp. 223–230. DOI: 10.1126/science.181.4096.223.
- [13] S. Walter Englander and Leland Mayne. “The nature of protein folding pathways”. In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 15873–15880. DOI: 10.1073/pnas.1411798111.
- [14] C. D. Sfatos and E. I. Shakhnovich. “Statistical mechanics of random heteropolymers”. In: *Physics Reports. I.M. Lifshitz and Condensed Matter Theory* 288.1 (Sept. 1997), pp. 77–108. DOI: 10.1016/S0370-1573(97)00022-7.
- [15] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. “Theory of protein folding: the energy landscape perspective”. eng. In: *Annual Review of Physical Chemistry* 48 (1997), pp. 545–600. DOI: 10.1146/annurev.physchem.48.1.545.
- [16] Kiyoto Kamagata, Munehito Arai, and Kunihiro Kuwajima. “Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins”. In: *Journal of Molecular Biology* 339.4 (June 2004), pp. 951–965. DOI: 10.1016/j.jmb.2004.04.015.
- [17] Christopher M. Dobson. “Protein folding and misfolding”. en. In: *Nature* 426.6968 (Dec. 2003), pp. 884–890. DOI: 10.1038/nature02261.
- [18] Peter G. Wolynes. “Evolution, energy landscapes and the paradoxes of protein folding”. In: *Biochimie* 119 (Dec. 2015),

- pp. 218–230. DOI: 10.1016/j.biochi.2014.12.007.
- [19] Jacquelyn S Fetrow and Jeffrey Skolnick. “Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to Glutaredoxins/Thioredoxins and T1Ribonucleases1”. In: *Journal of Molecular Biology* 281.5 (Sept. 1998), pp. 949–968. DOI: 10.1006/jmbi.1998.1993.
- [20] Christian B. Anfinsen. “Principles that Govern the Folding of Protein Chains”. In: *Science* 181.4096 (July 1973), pp. 223–230. DOI: 10.1126/science.181.4096.223.
- [21] Annabel E Todd, Christine A Orengo, and Janet M Thornton. “Evolution of function in protein superfamilies, from a structural perspective1”. In: *Journal of Molecular Biology* 307.4 (Apr. 2001), pp. 1113–1143. DOI: 10.1006/jmbi.2001.4513.
- [22] C. Chothia and A.M. Lesk. “The relation between the divergence of sequence and structure in proteins.” en. In: *The EMBO Journal* 5.4 (Apr. 1986), pp. 823–826. DOI: 10.1002/j.1460-2075.1986.tb04288.x.
- [23] Valerie Vaissier Welborn and Teresa Head-Gordon. “Computational Design of Synthetic Enzymes”. In: *Chemical Reviews* 119.11 (June 2019), pp. 6613–6630. DOI: 10.1021/acs.chemrev.8b00399.
- [24] Christine E. Tinberg et al. “Computational design of ligand-binding proteins with high affinity and selectivity”. en. In: *Nature* 501.7466 (Sept. 2013), pp. 212–216. DOI: 10.1038/nature12443.
- [25] Qi Wen and Paul A. Janmey. “Polymer physics of the cytoskeleton”. In: *Current Opinion in Solid State and Materials Science* 15.5 (Oct. 2011), pp. 177–182. DOI: 10.1016/j.cossms.2011.05.002.
- [26] Tony Hunter. “Signaling—2000 and Beyond”. English. In: *Cell* 100.1 (Jan. 2000), pp. 113–127. DOI: 10.1016/S0092-8674(00)81688-8.
- [27] Mikko Taipale. “Disruption of protein function by pathogenic mutations: common and uncommon mechanisms”. In: *Biochemistry and Cell Biology* 97.1 (Feb. 2019), pp. 46–57. DOI: 10.1139/bcb-2018-0007.
- [28] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. “Theory of protein folding: the energy landscape perspective”. eng. In: *Annual Review of Physical Chemistry* 48 (1997), pp. 545–600. DOI: 10.1146/annurev.physchem.48.1.545.
- [29] J. Monod, J. Wyman, and J. P. Changeux. “ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL”. eng. In: *Journal of Molecular Biology* 12 (May 1965), pp. 88–118. DOI: 10.1016/s0022-2836(65)80285-6.
- [30] Acacia F. Dishman et al. “Evolution of Fold-Switching in a Metamorphic Protein”. In: *Science (New York, N.Y.)* 371.6524 (Jan. 2021), pp. 86–90. DOI: 10.1126/science.abd8700.
- [31] Alex S. Holehouse and Birthe B. Kragelund. “The molecular basis for cellular function of intrinsically disordered protein regions”. en. In: *Nature Reviews Molecular Cell Biology* 25.3 (Mar. 2024), pp. 187–211. DOI: 10.1038/s41580-023-00673-0.
- [32] Peter E Wright and H. Jane Dyson. “Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm”. In: *Journal of Molecular Biology* 293.2 (Oct. 1999), pp. 321–331. DOI: 10.1006/jmbi.1999.3110.
- [33] V. N. Uversky, J. R. Gillespie, and A. L. Fink. “Why are "natively unfolded" proteins unstructured under physiologic conditions?” eng. In: *Proteins* 41.3 (Nov. 2000), pp. 415–427. DOI: 10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7.
- [34] A. Keith Dunker et al. “Intrinsically disordered protein”. In: *Journal of Molecular Graphics and Modelling* 19.1 (Feb. 2001), pp. 26–59. DOI: 10.1016/S1093-3263(00)00138-8.

- [35] H. Jane Dyson and Peter E. Wright. “Coupling of folding and binding for unstructured proteins”. eng. In: *Current Opinion in Structural Biology* 12.1 (Feb. 2002), pp. 54–60. DOI: 10.1016/s0959-440x(02)00289-0.
- [36] Tanja Mittag and Julie D Forman-Kay. “Atomic-level characterization of disordered protein ensembles”. In: *Current Opinion in Structural Biology*. Folding and binding / Protein-nucleic interactions 17.1 (Feb. 2007), pp. 3–14. DOI: 10.1016/j.sbi.2007.01.009.
- [37] P. Romero et al. “Sequence complexity of disordered protein”. eng. In: *Proteins* 42.1 (Jan. 2001), pp. 38–48. DOI: 10.1002/1097-0134(20010101)42:1<38::aid-prot50>3.0.co;2-3.
- [38] Christopher J. Oldfield and A. Keith Dunker. “Intrinsically disordered proteins and intrinsically disordered protein regions”. eng. In: *Annual Review of Biochemistry* 83 (2014), pp. 553–584. DOI: 10.1146/annurev-biochem-072711-164947.
- [39] Norman E. Davey, Martha S. Cyert, and Alan M. Moses. “Short linear motifs - ex nihilo evolution of protein regulation”. eng. In: *Cell communication and signaling: CCS* 13 (Nov. 2015), p. 43. DOI: 10.1186/s12964-015-0120-z.
- [40] Albert H. Mao et al. “Net charge per residue modulates conformational ensembles of intrinsically disordered proteins”. In: *Proceedings of the National Academy of Sciences* 107.18 (May 2010), pp. 8183–8188. DOI: 10.1073/pnas.0911107107.
- [41] Rahul K. Das and Rohit V. Pappu. “Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.33 (Aug. 2013), pp. 13392–13397. DOI: 10.1073/pnas.1304749110.
- [42] J. J. Ward et al. “Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life”. In: *Journal of Molecular Biology* 337.3 (Mar. 2004), pp. 635–645. DOI: 10.1016/j.jmb.2004.02.002.
- [43] Luca Mollica et al. “Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment”. English. In: *Frontiers in Molecular Biosciences* 3 (Sept. 2016). DOI: 10.3389/fmolb.2016.00052.
- [44] Paul Robustelli, Stefano Piana, and David E. Shaw. “Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein”. In: *Journal of the American Chemical Society* 142.25 (June 2020), pp. 11092–11101. DOI: 10.1021/jacs.0c03217.
- [45] San Hadži et al. “The Thermodynamic Basis of the Fuzzy Interaction of an Intrinsically Disordered Protein”. en. In: *Angewandte Chemie International Edition* 56.46 (2017), pp. 14494–14497. DOI: 10.1002/anie.201707853.
- [46] Monika Fuxreiter. “Fuzzy protein theory for disordered proteins”. In: *Biochemical Society Transactions* 48.6 (Nov. 2020), pp. 2557–2564. DOI: 10.1042/BST20200239.
- [47] *Relevance of Electrostatic Charges in Compactness, Aggregation, and Phase Separation of Intrinsically Disordered Proteins*.
- [48] Orkid Coskuner-Weber and Vladimir N. Uversky. “Current Stage and Future Perspectives for Homology Modeling, Molecular Dynamics Simulations, Machine Learning with Molecular Dynamics, and Quantum Computing for Intrinsically Disordered Proteins and Proteins with Intrinsically Disordered Regions”. In: *Current Protein and Peptide Science* 25.2 (Feb. 2024), pp. 163–171. DOI: 10.2174/0113892037281184231123111223.
- [49] Kagistia Hana Utami et al. “The roles of intrinsically disordered proteins in neurodegeneration”. In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1869.4 (Apr. 2025), p. 130772. DOI: 10.1016/j.bbagen.2025.130772.
- [50] M Madan Babu et al. “Intrinsically disordered proteins: regulation and disease”. In: *Current Opinion in*

- Structural Biology* 21.3 (June 2011), pp. 432–440. DOI: 10.1016/j.sbi.2011.03.011.
- [51] Birgit Strodel. “Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins”. In: *Journal of Molecular Biology*. From Protein Sequence to Structure at Warp Speed: How AlphaFold Impacts Biology 433.20 (Oct. 2021), p. 167182. DOI: 10.1016/j.jmb.2021.167182.
- [52] Rafael G. Viegas, Ingrid B. S. Martins, and Vitor B. P. Leite. “Understanding the Energy Landscape of Intrinsically Disordered Protein Ensembles”. In: *Journal of Chemical Information and Modeling* 64.10 (May 2024), pp. 4149–4157. DOI: 10.1021/acs.jcim.4c00080.
- [53] Eduardo Flores et al. “Single-point mutations in disordered proteins: Linking sequence, ensemble, and function”. In: *Current Opinion in Structural Biology* 91 (Apr. 2025), p. 102987. DOI: 10.1016/j.sbi.2025.102987.
- [54] Albert H. Mao, Nicholas Lyle, and Rohit V. Pappu. “Describing sequence–ensemble relationships for intrinsically disordered proteins”. In: *Biochemical Journal* 449.2 (Dec. 2012), pp. 307–318. DOI: 10.1042/BJ20121346.
- [55] Aldo R. Camacho-Zarco et al. “NMR Provides Unique Insight into the Functional Dynamics and Interactions of Intrinsically Disordered Proteins”. In: *Chemical Reviews* 122.10 (May 2022), pp. 9331–9356. DOI: 10.1021/acs.chemrev.1c01023.
- [56] Snigdha Maiti et al. “Experimental methods to study the structure and dynamics of intrinsically disordered regions in proteins”. In: *Current Research in Structural Biology* 7 (Jan. 2024), p. 100138. DOI: 10.1016/j.crstbi.2024.100138.
- [57] Daniel Nettels et al. “Single-molecule FRET for probing nanoscale biomolecular dynamics”. en. In: *Nature Reviews Physics* 6.10 (Oct. 2024), pp. 587–605. DOI: 10.1038/s42254-024-00748-7.
- [58] Pau Bernadó and Dmitri I. Svergun. “Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering”. eng. In: *Molecular bioSystems* 8.1 (Jan. 2012), pp. 151–167. DOI: 10.1039/c1mb05275f.
- [59] Gopa Mitra. “Emerging Role of Mass Spectrometry-Based Structural Proteomics in Elucidating Intrinsic Disorder in Proteins”. eng. In: *Proteomics* 21.3-4 (Feb. 2021), e2000011. DOI: 10.1002/pmic.202000011.
- [60] Fidha Nazreen Kunnath Muhammedkutty, Matthew MacAinsh, and Huan-Xiang Zhou. “Atomistic molecular dynamics simulations of intrinsically disordered proteins”. eng. In: *Current Opinion in Structural Biology* 92 (June 2025), p. 103029. DOI: 10.1016/j.sbi.2025.103029.
- [61] Sandro Bottaro and Kresten Lindorff-Larsen. “Biophysical experiments and biomolecular simulations: A perfect match?” eng. In: *Science (New York, N.Y.)* 361.6400 (July 2018), pp. 355–360. DOI: 10.1126/science.aat4010.
- [62] Hamidreza Ghafouri et al. *Toward a Unified Framework for Determining Conformational Ensembles of Disordered Proteins*. 2026. DOI: 10.1038/s41592-026-03003-2.
- [63] Jeffrey M. Lotthammer et al. “Direct prediction of intrinsically disordered protein conformational properties from sequences”. en. In: *Nature Methods* (Jan. 2024), pp. 1–12. DOI: 10.1038/s41592-023-02159-5.
- [64] Giulio Tesei et al. *Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation*. en. May 2023. DOI: 10.1101/2023.05.08.539815.
- [65] Jared M. Lalmansingh et al. “SOURSOP: A Python Package for the Analysis of Simulations of Intrinsically Disordered Proteins”. In: *Journal of Chemical Theory and Computation* 19.16 (Aug.

- 2023), pp. 5609–5620. DOI: 10.1021/acs.jctc.3c00190.
- [66] Hamidreza Ghafouri et al. “IDPEnsembleTools: An open-source library for analysis of conformational ensembles of disordered proteins”. en. In: *Protein Science* 35.1 (2026), e70427. DOI: 10.1002/pro.70427.
- [67] *Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins / The Journal of Physical Chemistry B*.
- [68] Do-Hyoung Kim and Kyou-Hoon Han. “Transient Secondary Structures as General Target-Binding Motifs in Intrinsically Disordered Proteins”. In: *International Journal of Molecular Sciences* 19.11 (2018). DOI: 10.3390/ijms19113614.
- [69] Wade Borchers et al. “Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells”. en. In: *Nature Chemical Biology* 10.12 (Dec. 2014), pp. 1000–1002. DOI: 10.1038/nchembio.1668.
- [70] Javier González-Delgado et al. “WASCO: A Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins”. eng. In: *Journal of Molecular Biology* 435.14 (July 2023), p. 168053. DOI: 10.1016/j.jmb.2023.168053.
- [71] Rahul K Das, Kiersten M Ruff, and Rohit V Pappu. “Relating sequence encoded information to form and function of intrinsically disordered proteins”. In: *Current Opinion in Structural Biology*. New constructs and expression of proteins / Sequences and topology 32 (June 2015), pp. 102–112. DOI: 10.1016/j.sbi.2015.03.008.
- [72] Alex S. Holehouse et al. “CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins”. eng. In: *Biophysical Journal* 112.1 (Jan. 2017), pp. 16–21. DOI: 10.1016/j.bpj.2016.11.3200.
- [73] Megan C. Cohan et al. “Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins”. In: *Journal of Molecular Biology* 434.2 (Jan. 2022), p. 167373. DOI: 10.1016/j.jmb.2021.167373.
- [74] Maria Victoria Nugnes et al. “DisProt in 2026: enhancing intrinsically disordered proteins accessibility, deposition, and annotation”. In: *Nucleic Acids Research* (Nov. 2025), gkaf1175. DOI: 10.1093/nar/gkaf1175.
- [75] Damiano Piovesan et al. “MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins”. eng. In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D495–D503. DOI: 10.1093/nar/gkae969.
- [76] Satoshi Fukuchi et al. “IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners”. In: *Nucleic Acids Research* 42.D1 (Oct. 2013), pp. D320–D325. DOI: 10.1093/nar/gkt1010.
- [77] Damiano Piovesan et al. “Databases for intrinsically disordered proteins”. In: *Acta Crystallographica. Section D, Structural Biology* 78.Pt 2 (Jan. 2022), pp. 144–151. DOI: 10.1107/S2059798321012109.
- [78] Zi Hao Liu et al. *AlphaFlex: Ensembles of the human proteome representing disordered regions*. en. Nov. 2025. DOI: 10.1101/2025.11.24.690279.
- [79] *Protein Data Bank / Nucleic Acids Research / Oxford Academic*.
- [80] Hamidreza Ghafouri et al. “PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins”. In: *Nucleic Acids Research* 52.D1 (Jan. 2024), pp. D536–D544. DOI: 10.1093/nar/gkad947.
- [81] Holger Dinkel et al. “ELM—the database of eukaryotic linear motifs”. eng. In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D242–251. DOI: 10.1093/nar/gkr1064.
- [82] Iva Pritišanac et al. *A Functional Map of the Human Intrinsically Disordered Proteome*. en. Mar. 2024. DOI: 10.1101/2024.03.15.585291.

- [83] Priyanka Madhu, Norman E. Davey, and Ylva Ivarsson. “How viral proteins bind short linear motifs and intrinsically disordered domains”. In: *Essays in Biochemistry* 66.7 (Dec. 2022). Ed. by Samrat Mukhopadhyay, pp. 935–944. DOI: 10.1042/EBC20220047.
- [84] Andras Hatos et al. “FuzDB: a new phase in understanding fuzzy interactions”. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D509–D517. DOI: 10.1093/nar/gkab1060.
- [85] Kaiqiang You et al. “PhaSepDB: a database of liquid–liquid phase separation related proteins”. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D354–D359. DOI: 10.1093/nar/gkz847.
- [86] Qian Li et al. “LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro”. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D320–D327. DOI: 10.1093/nar/gkz778.
- [87] Xi Wang et al. “LLPSDB v2.0: an updated database of proteins undergoing liquid–liquid phase separation in vitro”. In: *Bioinformatics* 38.7 (Jan. 2022), pp. 2010–2014. DOI: 10.1093/bioinformatics/btac026.
- [88] Mihaly Varadi et al. “pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins”. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D326–D335. DOI: 10.1093/nar/gkt960.
- [89] Valéry Ozenne et al. “Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables”. eng. In: *Bioinformatics (Oxford, England)* 28.11 (June 2012), pp. 1463–1470. DOI: 10.1093/bioinformatics/bts172.
- [90] Mickaël Krzeminski et al. “Characterization of disordered proteins with ENSEMBLE”. In: *Bioinformatics* 29.3 (Feb. 2013), pp. 398–399. DOI: 10.1093/bioinformatics/bts701.
- [91] Loïc Salmon et al. “NMR characterization of long-range order in intrinsically disordered proteins”. eng. In: *Journal of the American Chemical Society* 132.24 (June 2010), pp. 8407–8418. DOI: 10.1021/ja101645g.
- [92] Gabrielle Nodet et al. “Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings”. eng. In: *Journal of the American Chemical Society* 131.49 (Dec. 2009), pp. 17908–17918. DOI: 10.1021/ja9069024.
- [93] Tamas Lazar et al. “PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins”. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D404–D411. DOI: 10.1093/nar/gkaa1021.
- [94] Jeffrey C Hoch et al. “Biological Magnetic Resonance Data Bank”. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D368–D376. DOI: 10.1093/nar/gkac1050.
- [95] Alexey G. Kikhney et al. “SASBDB: Towards an automatically curated and validated repository for biological scattering data”. en. In: *Protein Science* 29.1 (2020), pp. 66–75. DOI: 10.1002/pro.3731.
- [96] Gregory-Neal W. Gomes et al. “Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET”. eng. In: *Journal of the American Chemical Society* 142.37 (Sept. 2020), pp. 15697–15710. DOI: 10.1021/jacs.0c02088.
- [97] Alessio Del Conte et al. “DRMAAtic: dramatically improve your cluster potential”. In: *Bioinformatics Advances* 5.1 (Jan. 2025), vbaf112. DOI: 10.1093/bioadv/vbaf112.
- [98] Alberto J. M. Martin, Ian Walsh, and Silvio C. E. Tosatto. “MOBI: a web server to define and visualize structural mobility in NMR protein ensembles”. In: *Bioinformatics* 26.22 (Nov. 2010), pp. 2916–2917. DOI: 10.1093/bioinformatics/btq537.
- [99] Giacomo Janson et al. “Direct generation of protein conformational ensembles

- via machine learning”. en. In: *Nature Communications* 14.1 (Feb. 2023), p. 774. DOI: 10.1038/s41467-023-36443-x.
- [100] João M. C. Teixeira et al. “IDPConformerGenerator: A Flexible Software Suite for Sampling the Conformational Space of Disordered Protein States”. eng. In: *The Journal of Physical Chemistry. A* 126.35 (Sept. 2022), pp. 5985–6003. DOI: 10.1021/acs.jpca.2c03726.
- [101] Kaushik Borthakur et al. “Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution”. en. In: *Nature Communications* 16.1 (Oct. 2025), p. 9036. DOI: 10.1038/s41467-025-64098-3.
- [102] Sarah Lewis et al. *Scalable emulation of protein equilibrium ensembles with generative deep learning*. en. Feb. 2025. DOI: 10.1101/2024.12.05.626885.
- [103] Giacomo Janson and Michael Feig. “Transferable deep generative modeling of intrinsically disordered protein conformations”. eng. In: *bioRxiv: The Preprint Server for Biology* (Feb. 2024), p. 2024.02.08.579522. DOI: 10.1101/2024.02.08.579522.
- [104] Borna Novak et al. *Accurate predictions of conformational ensembles of disordered proteins with STARLING*. en. Feb. 2025. DOI: 10.1101/2025.02.14.638373.
- [105] Junjie Zhu et al. “Phanto-IDP: compact model for precise intrinsically disordered protein backbone generation and enhanced sampling”. In: *Briefings in Bioinformatics* 25.1 (Jan. 2024), bbad429. DOI: 10.1093/bib/bbad429.
- [106] The UniProt Consortium. “UniProt: the Universal Protein Knowledgebase in 2025”. In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D609–D617. DOI: 10.1093/nar/gkae1010.
- [107] Mihaly Varadi et al. “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences”. In: *Nucleic Acids Research* 52.D1 (Jan. 2024), pp. D368–D375. DOI: 10.1093/nar/gkad1011.
- [108] Massimiliano Bonomi et al. “Metainference: A Bayesian inference method for heterogeneous systems”. In: *Science Advances* 2.1 (Jan. 2016), e1501177. DOI: 10.1126/sciadv.1501177.
- [109] Francesco Pesce and Kresten Lindorff-Larsen. “Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data”. eng. In: *Biophysical Journal* 120.22 (Nov. 2021), pp. 5124–5135. DOI: 10.1016/j.bpj.2021.10.003.
- [110] Michele Invernizzi et al. *Advancing Protein Ensemble Predictions Across the Order–Disorder Continuum*. en. Nov. 2025. DOI: 10.1101/2025.10.18.680935.
- [111] John Moult. “A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction”. In: *Current Opinion in Structural Biology*. Sequences and topology/Nucleic acids 15.3 (June 2005), pp. 285–289. DOI: 10.1016/j.sbi.2005.05.011.
- [112] Kenji Sugase, H Jane Dyson, and Peter E Wright. “Mechanism of coupled folding and binding of an intrinsically disordered protein”. In: *Nature* 447.7147 (2007), pp. 1021–1025.
- [113] Flurin Sturzenegger et al. “Transition path times of coupled folding and binding reveal the formation of an encounter complex”. In: *Nature communications* 9.1 (2018), p. 4708.
- [114] Aritra Chowdhury, Daniel Nettels, and Benjamin Schuler. “Interaction dynamics of intrinsically disordered proteins from single-molecule spectroscopy”. In: *Annual review of biophysics* 52.1 (2023), pp. 433–462.
- [115] Robert Schneider et al. “Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR”. In: *Journal of the American Chemical Society* 137.3 (2015), pp. 1220–1229.
- [116] Robert T. McGibbon et al. “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”. English. In: *Biophysical Journal* 109.8

- (Oct. 2015), pp. 1528–1532. DOI: 10.1016/j.bpj.2015.08.015.
- [117] Richard Gowers et al. “MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations”. en. In: Austin, Texas, 2016, pp. 98–105. DOI: 10.25080/Majora-629e541a-00e.
- [118] Hongyu Zhou, Feng Wang, and Peng Tao. “t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations”. In: *Journal of Chemical Theory and Computation* 14.11 (Nov. 2018), pp. 5499–5510. DOI: 10.1021/acs.jctc.8b00652.
- [119] Francesco Trozzi, Xinlei Wang, and Peng Tao. “UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study”. In: *The Journal of Physical Chemistry B* 125.19 (2021), pp. 5022–5034. DOI: 10.1021/acs.jpcc.1c02081.