Advancing Protein Ensemble Predictions Across the Order–Disorder Continuum

Michele Invernizzi[1*], Sandro Bottaro[1], Julian O. Streit[2], Bruno Trentini[3,4], Niccolò Alberto Elia Venanzi[3], Danny Reidenbach[3], Youhan Lee[3], Christian Dallago[3,6], Hassan Sirelkhatim[3], Bowen Jing[5], Fabio Airoldi[1], Kresten Lindorff-Larsen[2], Carlo Fisicaro[1*], Kamil Tamiola[1*]

[1]Peptone Ltd., 41-43 Chalton Street, London, NW1 1JD, UK.

[2]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen, 2200, DK.

[3]NVIDIA Corp., 2788 San Tomas Expy, Santa Clara, 95051, CA, USA.

[4]University of Oxford, Wellington Sq., Oxford, OX1 2JD, UK.

[5]Computer Science & Artificial Intelligence Laboratory, MIT, 32 Vassar St, Cambridge, 02139, MA, USA.

[6]Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, 27710, NC, USA.

*Corresponding author(s). E-mail(s): michele@peptone.io; carlo@peptone.io; kamil@peptone.io;

Abstract

While deep learning has transformed structure prediction for ordered proteins, intrinsically disordered proteins remain poorly predicted due to systematic underrepresentation in training data, despite constituting approximately 30% of eukaryotic proteomes. We introduce PeptoneBench, the first benchmark enabling systematic assessment of ensemble generators across both ordered and disordered proteins, integrating diverse experimental observables. Our analysis reveals that existing evaluation metrics exhibit systematic bias toward the structured spectrum of the proteome. Assessment of popular predictors (AlphaFold2, ESMFlow, Boltz2) confirms high accuracy on ordered proteins but shows performance degradation with increasing disorder. We further present PepTron, a flow-matching ensemble generator trained on data augmented with synthetic disordered protein ensembles. On our benchmark PepTron matches BioEmu on disordered regions while maintaining competitive accuracy on ordered protein benchmarks. Our data augmentation approach demonstrates that targeted training strategies can approach the performance of computationally expensive simulation-based methods, establishing a generalizable framework applicable to other protein generative models. All datasets, models, and code are openly available.

1 Introduction

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs), which lack stable tertiary structure under native conditions, play central roles in signaling, transcriptional regulation, and a wide range of human diseases [1]. Computational analyses estimate that approximately 30% of residues in eukaryotic proteomes are disordered, yet these regions remain difficult to characterize experimentally and computationally [2].

The flexible nature of IDPs complicates analysis by traditional structure determination methods such as X-ray crystallography and cryogenic sample electron microscopy, necessitating ensemble-averaged approaches like nuclear magnetic resonance (NMR) spectroscopy and small-angle X-ray scattering (SAXS), and other biophysical techniques. These experimental limitations, in turn, influence computational modeling: modern protein structure prediction methods inherit a bias toward ordered structures

from their reliance on the Protein Data Bank (PDB) [3, 4], where disordered regions are systematically underrepresented [5, 6]. Proteome-wide analyses consistently report substantially higher fractions of disordered residues than appear in structural databases [2, 7]. Furthermore, this bias is compounded by poor sequence conservation and sparse multiple sequence alignments (MSAs) of IDRs, limiting the effectiveness of MSA-driven features for disorder modeling [8]. Consequently, models such as AlphaFold2, trained on PDB, struggle to generalize to the substantial fraction of proteins that contain IDRs [9].

Recent methodological advances have begun to address these limitations. While AlphaFold2's predictions for disordered regions do not fully capture biophysical reality, its per-residue local distance difference test (pLDDT) score is remarkably effective for disorder identification [10]. AlphaFold3 employs cross-distillation using AlphaFold-Multimer predictions as training signals to represent disordered regions as extended structures, thereby mitigating hallucinations [11]. Similarly, RosettaFold-3 adopts artificially extended backbone representations for disordered regions during training [12]. Beyond single-structure predictors, ensemble generation methods directly emulate protein dynamics. AlphaFlow and ESMFlow [13] combine flow-matching frameworks with structure prediction models to generate conformational ensembles, while BioEmu [14] integrates over 200 milliseconds of molecular dynamics simulations alongside ordered structures and experimental stability data at training stage to generate protein ensembles at scale.

Complementing deep learning approaches, physics-based methods offer alternative strategies to address IDPs[15, 16], while others utilize generative adversarial or diffusion models trained on molecular simulations [17, 18]. Unified predictors capable of capturing both ordered and disordered regimes have also emerged [19, 20].

Regardless of the predictive method employed, evaluating predictions for disordered proteins presents inherent difficulties: no single native structure exists, and metrics such as root-mean-square deviation (RMSD) and global distance test (GDT-TS) are ill-posed. Meaningful validation requires ensemble-oriented approaches that compare predicted conformational ensembles to experimentally accessible observables: chemical shifts (CS), SAXS profiles, residual dipolar couplings (RDCs), and paramagnetic relaxation enhancements (PREs) [21–23], aligning deep-learning modeling with practices long established in force-field development [24–26].

In this work, we introduce two principal contributions to address these gaps:

- **Evaluation framework:** We introduce PeptoneBench, a systematic protocol with accompanying code for ensemble evaluation that compares experimental observables to ensemble-averaged predictions, including maximum entropy reweighting [27–30], leveraging a curated dataset of SAXS, NMR, RDC, and PRE data.
- **Model development:** We present PepTron, a generative model trained on an augmented dataset enriched with synthetic IDR ensembles. PepTron is benchmarked against leading predictors (AlphaFold2, Boltz2, BioEmu) via our framework, demonstrating state-of-the-art consistency with experimental ensemble properties across both ordered and disordered regimes.

Together, these contributions establish a promising framework for more accurate, ensemble-based protein structure prediction, improving representation of both ordered and disordered configurations. All datasets, models, and code are available under the CC BY or Apache 2.0 license, as detailed in Sec. 3.

2 Results

2.1 PeptoneBench: Disorder-aware Protein Ensembles Evaluation

The Protein Data Bank (PDB) [3] is foundational for training and validating deep learning models in protein structure prediction. However, it is notably depleted in IDRs and IDPs, introducing a bias toward ordered structures (Fig. 1a). Our analysis indicates that only ∼3% of PDB entries are classified as disordered, compared to 20–30% in the human proteome, consistent with previous estimates [31, 32]. This underrepresentation motivates the use of alternative databases: the IDRome database [33], designed to encompass all disordered regions in the human proteome, is highly enriched in disorder (approximately 77%), providing complementary coverage. However, experimentally determined protein structures for the IDRome are not available.

Disorder is quantified here using the protein mean G-score, which ranges from 0 (fully ordered) to 1 (fully disordered) [34]. G-scores are derived from secondary chemical shifts (CS), which directly report on the presence or absence of local secondary structure [35]. For proteins lacking experimental CS data, we predict G-scores using ADOPT2 [36], a machine learning framework trained on the TriZOD dataset [34].

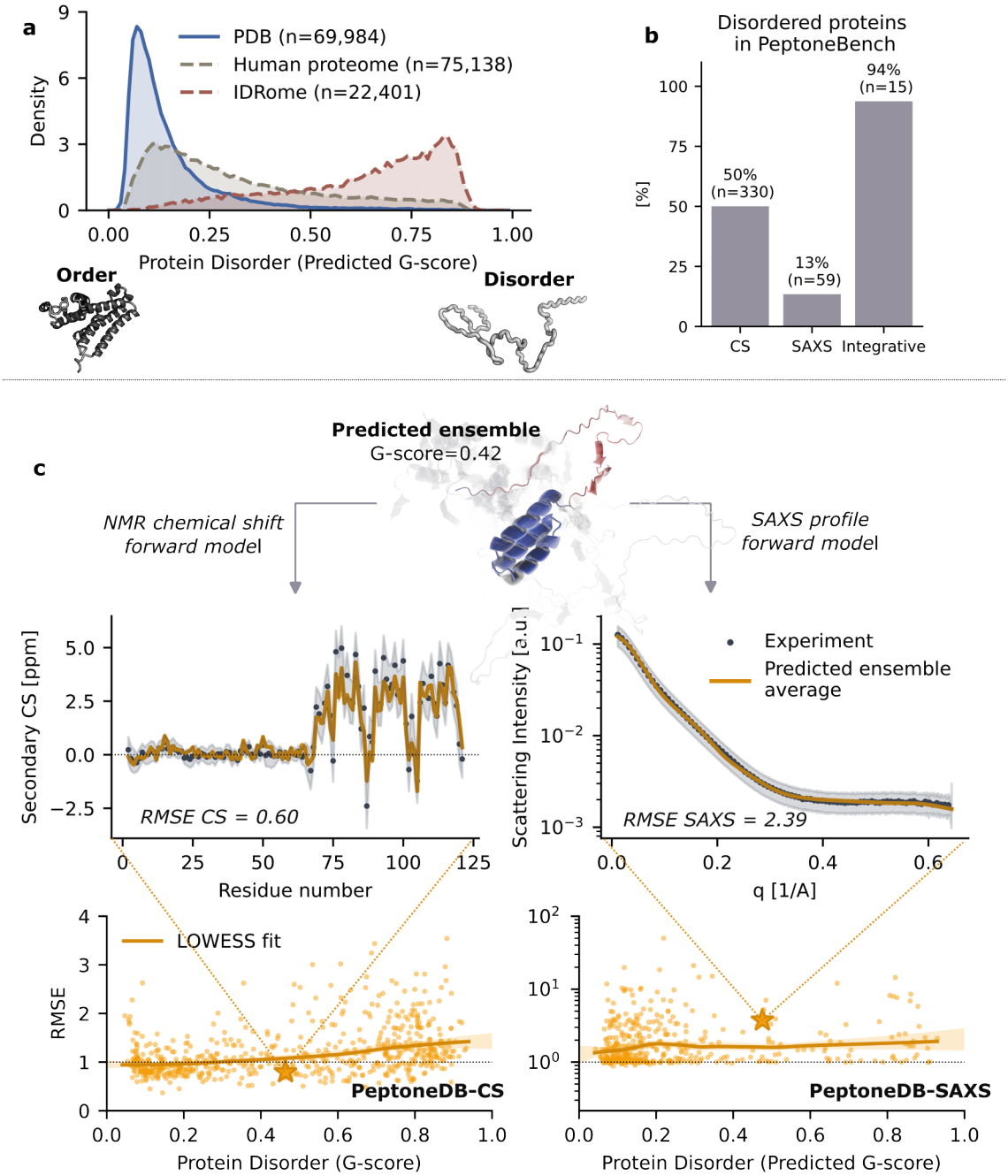

Fig. 1 (**a**) Distribution of predicted protein disorder (mean G-score) in the PDB, human proteome, and IDRome datasets, with examples of ordered and disordered structures. Each reported dataset was clustered with MMseqs2, with sizes indicated in parentheses. (**b**) Fraction of disordered proteins (average G-score > 0.5) in PeptoneBench databases. (**c**) Overview of the benchmarking workflow. For each protein sequence, an ensemble is predicted using the chosen model (in this case PepTron), enabling calculation of relevant observables (e.g., secondary CS, SAXS profiles) depending on the dataset. Ensemble-averaged predictions (with or without maximum entropy reweighting) are compared to experimental measurements, and uncertainty-normalized RMSE is computed for each PeptoneDB-CS and PeptoneDB-SAXS entry. Normalized RMSE values are visualized as a function of disorder content (G-score), with data aggregated via LOWESS smoothing and uncertainty estimated by bootstrapping. Stars denote the same protein present in both PeptoneDB-CS and PeptoneDB-SAXS datasets.

Quantifying disorder makes it clear that, in order to evaluate the quality of generated protein ensembles, one has to go beyond the PDB, and look for different sources of experimental validation. To this end we introduce three curated datasets PeptoneDB-CS (NMR chemical shifts from BMRB [37]), PeptoneDB-SAXS (SAXS profiles from SASBDB [38]), and PeptoneDB-Integrative (entries with multiple complementary experimental measurements [15, 25, 39]; see Table 1) and Fig. 1b. These datasets cover the whole order-disorder continuum and provide complementary information: CS strongly correlate

with local secondary structure content, while SAXS curves are related to the overall shape of the protein ensemble. The PeptoneDB-Integrative focuses on IDPs and provides for each entry an orthogonal measurement (RDCs for CS and PREs for SAXS, see Table A1) that can be used for cross-validation.

We use these datasets to implement *PeptoneBench*, a systematic evaluation framework for quantifying predicted protein ensembles agreement with experimental data (Fig. 1c), across the order-disorder continuum. Prediction models generate ensembles that are pre-processed using standardized criteria (Appendix A.1). Experimental observables are computed from predicted structures using appropriate forward models (Appendix A.2, A.3, A.4) [40]. Agreement between prediction and experiment is assessed using normalized root mean squared error (RMSE), Eq. A6, accounting for uncertainties from both forward models and experimental data [41–43]. RMSE values are visualized as a function of protein disorder (G-score), with data aggregated via locally weighted scatterplot smoothing (LOWESS) and uncertainty estimated by bootstrapping. The area under the LOWESS curves (AUC) provides a condensed accuracy metric across the disorder content. For convenience, we define a PeptoneBench summary score by combining the AUC for CS and SAXS, see Eq. D12.

While RMSE is indicative of accuracy, it is insufficient on its own: broader ensembles with higher RMSE may better capture experimental distributions if improved by maximum entropy reweighting [27, 30]. Therefore, PeptoneBench also computes RMSE with post-refinement results, to determine not only how well ensembles match experiments initially, but also how much their agreement can be improved through maximum entropy reweighting. This approach distinguishes between ensembles that contain the right conformations but need better weighting (*reweightable errors*), and those that lack important conformations entirely (*irrecoverable deficiencies*). A description of the robust reweighting procedure developed for PeptoneBench is provided in the Appendix A.5.

Ensemble refinement is particularly relevant when dealing with IDP ensembles, which are highly sensitive to experimental conditions (e.g., temperature, pH, ionic strength). Generative models that sample diverse conformations across a range of conditions, similarly to multicanonical approaches in molecular dynamics, could be reweighted to match specific experimental parameters, greatly enhancing their predictive utility.

Dataset	**Data type**	**Source**	**# proteins (ordered/disordered)**	**# residues (ordered/disordered)**
PeptoneDB-CS	NMR chemical shifts	BMRB	329 / 330	32,908 / 31,507
PeptoneDB-SAXS	SAXS profiles	SASBDB	380 / 59	92,577 / 18,372
PeptoneDB-Integrative	NMR chemical shifts RDCs, SAXS, PREs	Literature	1 / 15	151 / 2,042

Table 1 Curated protein datasets used in this study, showing data sources and counts of ordered and disordered proteins and residues. Proteins are classified as ordered (mean G-score < 0.5) or disordered (mean G-score > 0.5). For PeptoneDB-SAXS, G-scores are predicted when chemical shift data are unavailable.*Note:* PaaA2 in PeptoneDB-Integrative (mean G-score $= 0.43$; see Table A1) is classified as ordered by this threshold, but is commonly reported as disordered in the literature [44].

2.2 PepTron: Augmenting Training Data with Disordered Regions for Improved Protein Ensemble Prediction

We introduce *PepTron*, a protein structure generator based on the flow-matching ESMFlow architecture [13], and designed to generate diverse conformational ensembles for proteins spanning from ordered to disordered states.

PepTron is fine-tuned on both, experimentally determined structures from the PDB, and synthetically generated ensembles of disordered proteins. This augmentation exposes the model to the full order-disorder continuum, enabling more accurate predictions across diverse protein types. Remarkably, even computationally inexpensive methods for generating disordered ensembles substantially improve performance. Performance was assessed using PeptoneBench, ensuring evaluation on experimental data fully withheld from training. We benchmarked PepTron together with ESMFold, ESMFlow, AlphaFold2, Boltz2, and BioEmu, and performed dedicated evaluations on the IDP-focused PeptoneDB-Integrative dataset.

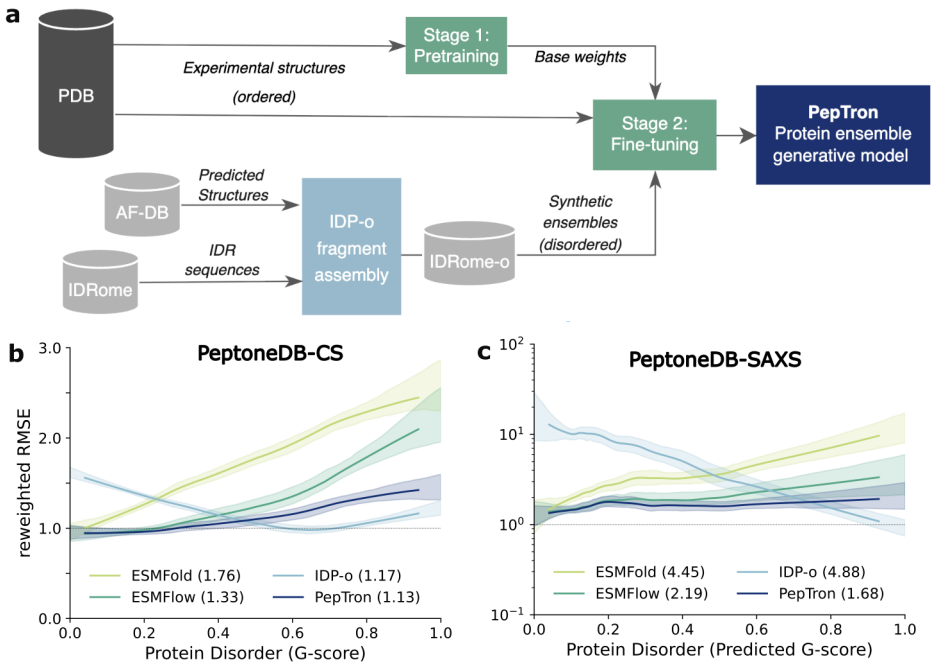

Fig. 2 (**a**) Schematic overview of the PepTron model. Structural data from the PDB are used to pretrain the PepTron model which employs the ESMFlow architecture. Synthetic datasets of disordered proteins are generated using IDP-o, drawing sequences from the IDRome and fragments from the AlphaFold UniProt Database (AlphaFold DB). The resulting synthetic dataset is jointly used for fine-tuning, integrating ordered and disordered states. (**b**) Agreement with experimental CS data and (**c**) SAXS profiles. Reported values correspond to per-protein reweighted RMSE fitted via LOWESS regression with bootstrap-estimated uncertainties for ESMFold, ESMFlow, IDP-o, and PepTron. For each model and dataset, we report in parentheses the area under the LOWESS curve as an overall score (lower is better).

2.2.1 Synthetic IDP Dataset Generation

To address the lack of structural data for disordered regions, we generate a synthetic structural dataset, called IDRome-o, from the sequences in IDRome [45] (Fig. 1a). To this end, we developed a fragment-based protein structure generator, *IDP-o*, which can generate physically plausible IDP ensembles at scale and with reduced computational cost. IDP-o harnesses principles from fragment-based approaches [46] and hierarchical chain-growth techniques [47]. Specifically, it constructs protein configurations by aligning smaller fragments taken from a large structure database. Contrary to previous methods which used the PDB as source for the fragments, IDP-o leverages the full AlphaFold Protein Structure Database (214 million predicted structures), which is orders of magnitude larger. This allows IDP-o to use larger six-residue fragments, and better capture transient helicity content, as shown in the comparison with other IDP-specialised models in Appendix B).

It is important to note that by design IDP-o does not sample a specific equilibrium ensemble distribution, but rather it produces diverse conformational ensembles which represent plausible configurations for a given sequence. Accordingly, these ensembles are most effective when combined with reweighting strategies, or as starting points for further molecular dynamics simulations [47].

2.2.2 Model Architecture and Training Mixture

PepTron is based on ESMFlow architecture, but was implemented within the NVIDIA BioNeMo framework, for improved computational performances [48]. It integrates cuEquivariance triangle attention and uses the Modular Co-Design BioNeMo sub-package for implementing flow matching, as described in Appendix C. Both at training and inference time PepTron uses BioNeMo distributed best practices, leveraging Megatron-style [49] tensor-pipeline-context-expert parallelism, mixed precision, activation checkpointing, and optimized input pipelines to scale efficiently across multi-GPU systems. At inference, PepTron runs from a single checkpoint without MSAs or external ESM weights [50] and directly generates structure ensembles.

We start from the ESMFold weights and retrain the flow matching module using the PDB, as described in the ESMFlow paper [13], but with a twice longer crop length of 512. We then fine-tune PepTron using a mixture of PDB and IDRome-o data, as described in detail in Appendix C.

2.3 Comparative Benchmarking

To assess the performance improvement of PepTron over the baseline models ESMFlow and IDP-o, we predicted ensembles for all sequences in the PeptoneDB-CS and PeptoneDB-SAXS datasets, and computed the normalized RMSE following the PeptoneBench procedure. Figure 2b summarizes the agreement as a function of protein disorder content on the PeptoneDB-CS set for ESMFlow, IDP-o, and PepTron. For comparison, we also include single-structure predictions from ESMFold [51].

A consistent trend emerges: both ESMFold and ESMFlow yield accurate predictions for ordered systems (low G-score, normalized RMSE ≈ 1), but their accuracy declines as disorder increases, with ESMFold degrading more rapidly. In contrast, IDP-o displays the opposite pattern, with improved accuracy at higher disorder levels, in line with expectations. PepTron maintains more stable performance across both ordered and disordered protein ensembles. Similar trends are observed when comparing the models against experimental SAXS profiles and upon reweighting (Fig. 2c). This balanced performance suggests that training on diverse synthetic disordered ensembles enables the model to capture conformational heterogeneity without sacrificing accuracy on well-folded structures.

To contextualize the performance of PepTron model on PeptoneBench, we compare it against some of the most widely used protein generative models: AlphaFold2 [11], Boltz2 [52], and BioEmu [14]. Figure 3 shows the overall agreement of these models with experimental CS and SAXS data. For convenience, we also provide the PeptoneBench summary scores in Table 2. Consistent with previous findings for ESMFold, both Boltz2 and AlphaFold2 demonstrate strong predictive accuracy for ordered protein systems (low G-score), yet their performance deteriorates as disorder content increases. This systematic decrease highlights a key limitation of current leading models, which, despite their groundbreaking successes in structure prediction, remain challenged by the inherent heterogeneity of IDPs and proteins featuring significant disordered regions.

By contrast, PepTron and BioEmu exhibit robust performance across the entire order-disorder spectrum, maintaining comparable levels of accuracy for both ordered and disordered proteins. The consistently high agreement observed for BioEmu aligns with its design, as the model has been trained explicitly on extensive molecular dynamics simulations of unfolded proteins that provide rich coverage of conformational diversity well suited to the characteristics of IDPs. This observed trend is consistent across both NMR and SAXS benchmark datasets, highlighting its reliability and generality. A more comprehensive comparative analysis including expanded benchmarking across models, both before and after maximum-entropy reweighting, is provided in the Appendix (Figs. D6, D7, and D8). For convenience, we also provide the PeptoneBench summary scores in Table 2. The PeptoneBench results on a larger set of and Fig. D10.

We further benchmarked the predictors on the canonical ordered datasets CAMEO22 and CASP14 (Table 2). AlphaFold2 and Boltz2 consistently achieve the lowest RMSD values and highest LDDT and TM-scores, establishing them as the top-performing methods for modeling ordered proteins. PepTron and ESMFlow display very similar results, with PepTron matching ESMFlow in both RMSD and accuracy metrics across all benchmarks. This parity indicates that the extensive retraining and fine-tuning applied to PepTron for disordered protein ensembles does not compromise its performance on ordered structures, supporting the effectiveness of our IDR augmentation strategy. While BioEmu lags slightly behind PepTron and ESMFlow on the CAMEO22 benchmark, it shows improved performance on CASP14, with TM-scores and RMSD values closer to the best-performing models, AlphaFold2 and Boltz2.

Model	**PeptoneBench**		**CAMEO22**			**CASP14**		
	Plain↓	Reweight↓	RMSD ↓	LDDT ↑	TM-Score ↑	RMSD ↓	LDDT ↑	TM-Score ↑
AlphaFold2	3.07	3.07	3.58/1.86	0.82/0.86	0.86/0.94	5.03/3.02	0.78/0.82	0.85/0.91
Boltz2	3.10	3.80	3.53/1.56	0.86/0.89	0.87/0.95	5.77/3.18	0.81/0.84	0.83/0.90
ESMFlow	2.12	2.74	4.53/2.69	0.70/0.75	0.82/0.89	10.50/6.97	0.53/0.54	0.63/0.68
BioEmu	1.49	2.22	4.92/2.84	0.52/0.54	0.80/0.88	7.25/4.17	0.49/0.52	0.74/0.80
PepTron	1.65	2.54	4.72/2.75	0.71/0.75	0.81/0.89	10.71/7.15	0.54/0.57	0.63/0.70

Table 2 Performance of protein structure prediction on PeptoneBench and the CAMEO22 and CASP14 benchmarks. For the PeptoneBench we report summary scores (see Eq. D12) both with and without ensemble refinement via maximum entropy reweighting. For CAMEO22 and CASP14 we report the average and median over samples for each metric (values for ESMFlow and AlphaFold2 are taken from Ref. [53]).

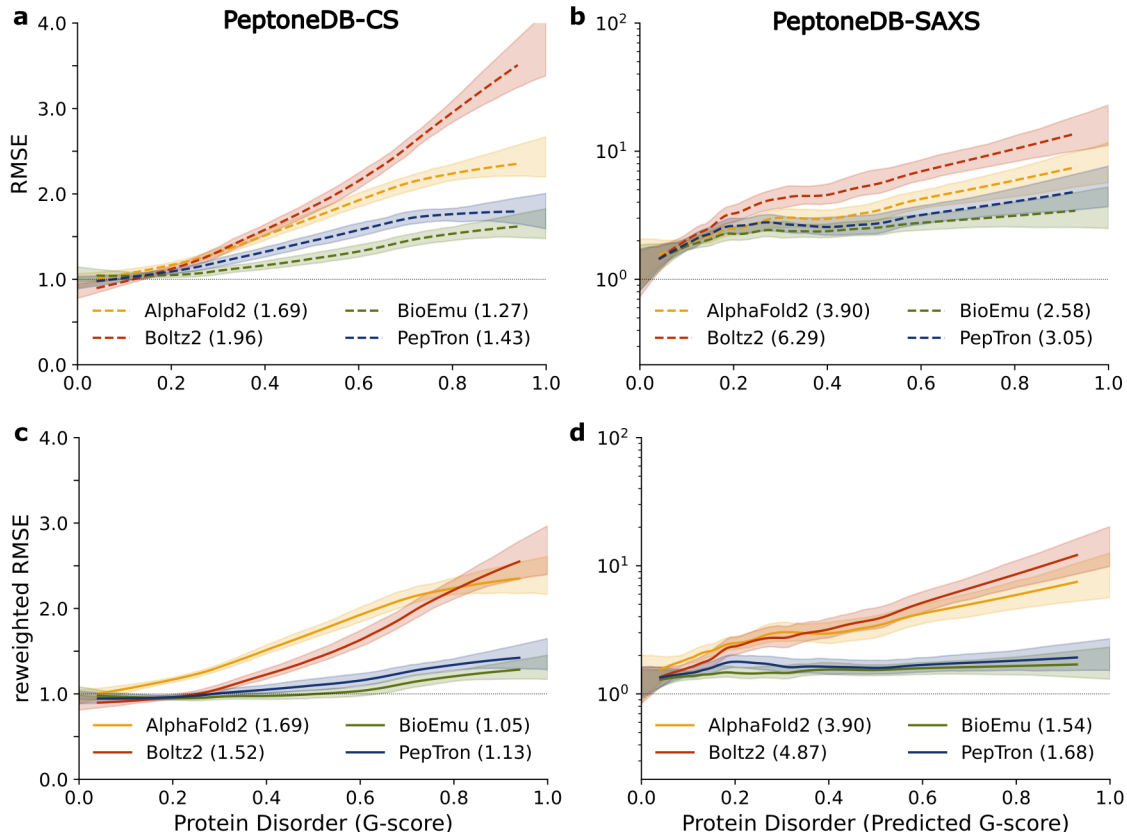

Fig. 3 Agreement with experimental CS data (**a**) and SAXS profiles (**b**). Reported values correspond to per-protein normalized RMSE fitted via LOWESS regression with bootstrap-estimated uncertainties for PepTron, AlphaFold2, Boltz2, and BioEmu. For each model and dataset, we report in parentheses the area under the LOWESS curve as a summary score (lower is better). Panels **c** and **d** report the same quantities after maximum entropy reweighting.

2.4 PeptoneDB-Integrative Insights

While PeptoneDB-CS and PeptoneDB-SAXS provide broad coverage of proteins across the order-disorder continuum, validation on IDPs with multiple orthogonal experimental measurements (PeptoneDB-Integrative) offers a more stringent test of ensemble quality. For this dataset, we performed two complementary analyses. First, we performed maximum entropy fitting of CS data, tuning the regularization parameter to avoid overfitting, ensuring that improving agreement with CS data did not degrade agreement with the orthogonal RDC measurements (Fig. D12). Note that for RDCs, the quality factor (Q-factor) is used to quantify performance (see Appendix A.4).

While the modest number of proteins in this dataset limits definitive ranking, some clear trends emerged (Fig. 4). The disorder-specialized predictor, IDP-o, consistently outperformed all other models after reweighting, yielding lower RMSE and Q-factor values and thus reflecting better agreement with experiment. Closer inspection indicates that models with relatively high RMSE and Q-factors (2–3 and >1, respectively) often fail to capture key helical propensities, as confirmed by per-protein, per-model analyses (Figs. D14 and D15). Notably, both PepTron and BioEmu exhibited similar performance on RDCs (mean Q-factor ∼0.8), although BioEmu performed better on chemical shifts. In contrast, AlphaFold2, ESMFold, and Boltz2 demonstrated consistently poor agreement across both local experimental metrics.

The complementary analysis using experimental data reporting on more global structural features, i.e., fitting to SAXS profiles with cross-validation restraints from PRE data (Figs. 4, D11 and D13) confirmed these trends. IDP-o again yielded the best performance, followed by BioEmu. PepTron and ESMFlow also performed well, with slightly worse agreement than that achieved by BioEmu. IdpGAN [18] delivered strong agreement for global observables (SAXS and PREs), but does not capture secondary structure preferences (Figs. D14 and D15).

Overall, similar results were obtained when comparing the models without reweighting (Fig. D11). In this case, IDP-o performed best on chemical shifts, RDCs and PREs, and second best behind IdpGAN on SAXS data. IDP-o thus appears to be particularly well suited to both capture reasonable local and global properties of disordered proteins, including in the absence of experimental reweighting. Conversely, AlphaFold2, ESMFold and Boltz2 had the worst agreement with all types of experimental data, suggesting

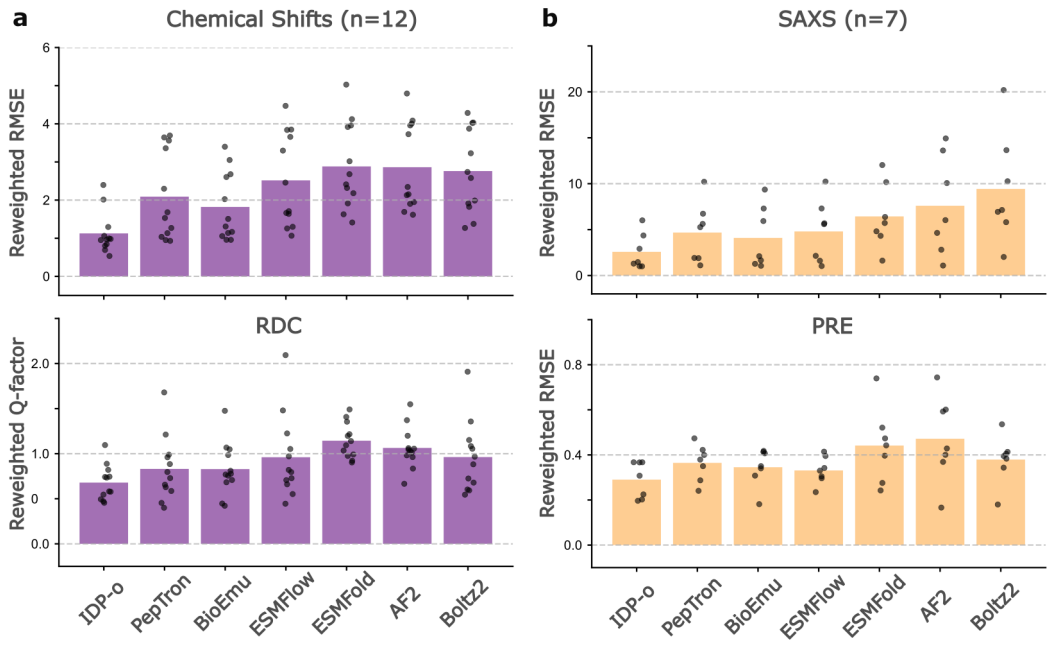

Fig. 4 (a) RMSE between predicted and experimental chemical shifts (top panel) for ensembles generated by seven different models following maximum entropy reweighting using CS data. In the bottom panel, Q-factors measuring agreement between predicted and observed residual dipolar couplings (RDCs) across the same models and CS-reweighted ensembles are shown. Each bar represents mean performance for n=12 intrinsically disordered systems for which both CS and RDC are available; individual points correspond to results for each individual system. Lower RMSE and Q-factor values indicate improved agreement with experimental data. (b) RMSE between predicted and experimental SAXS profiles (top panel) following maximum entropy reweighting using SAXS data. In the bottom panel, RMSE of PRE data for the same models are shown. Each bar represents mean performance for n=7 intrinsically disordered systems for which both SAXS and PREs are available. Note that the RMSE values for PREs refer to the root mean squared error of predicted PRE intensity ratios (ranging from 0 to 1) relative to experimental values.

these models do not accurately capture local or global properties of IDPs. Other ensemble-based predictors applicable to both ordered and disordered sequences typically achieve performances intermediate between those of IDP-o and AlphaFold2/ESMFold/Boltz2. The assessment of this benchmark dataset thus suggests that for typical unfolded proteins IDP-o is most suitable. Notably, RDC Q-factors achieved by IDP-o are comparable to those achieved by state-of-the-art molecular dynamics simulations of disordered proteins[25], with Q-factors ranging from ∼0.4 to 1.0.

Despite the overall improvements in modern ensemble predictors, this integrative benchmark also highlights persistent challenges for all models. For example, PRE experiments reveal that subtle long-range contact preferences remain difficult to predict and capture with reweighting. Residual secondary structure biases retained by some models, including for example BioEmu and PepTron, can also lead to poor agreement with experimental data (e.g., reflected in chemical shift and RDC predictions, Figs. D14 and D15). Another previously noted limitation of models such as AlphaFold2, BioEmu and PepTron is their inability to capture unfolded states of destabilized sequences that may fold into stable structures under some conditions[20]. While IDP-o captures these disordered states well, a folding bias of the other models results in large secondary chemical shifts and poor agreement with experiments (Fig. D16).

3 Discussion

A central contribution of this work is the explicit incorporation of disorder as a dimension in protein structure prediction. This is a critical factor widely acknowledged in the biophysics community but not yet fully integrated into current machine learning frameworks. Building on this concept, our work makes three key advances: (1) we introduce PeptoneBench, an experimentally anchored benchmark for evaluating ensemble predictions using standardized pre- and post-reweighting metrics (Fig. 1); (2) we demonstrate that fine-tuning on a disorder-focused dataset markedly improves ensemble modeling accuracy in our novel PepTron model, which utilizes a deliberately standard neural architecture (Fig. 2); and (3) we systematically benchmark leading structure predictors, including PepTron, AlphaFold2, Boltz2, and BioEmu, with particular emphasis on performance in disordered regions (Figs. 3 and 4). Our evaluation demonstrates that the predictive accuracy of AlphaFold2 and Boltz2 systematically declines as

protein disorder increases, as commonly observed. In contrast, PepTron and BioEmu sustain high performance across both ordered and intrinsically disordered protein regions. Benchmarking on ordered single-structure datasets (CAMEO22 and CASP14) highlights both Boltz2 and AlphaFold2 as leading predictors (Table 2). Collectively, our benchmark establishes robust, disorder-dependent performance baselines for the field.

Although we present a summary PeptoneBench score with and without reweighting (Table 2), we are well aware that summarizing model performance with a single score can be misleading. For this reason, we provide a range of evaluation metrics and plots as output of our benchmark, and we encourage researchers to engage with different metrics to capture distinct aspects of accuracy across datasets and levels of disorder. Moving forward, we intend to further explore the richness and nuances of the presented datasets, for example analysing how well models handle different types of IDRs, such as flexible linkers or disordered terminal domains.

While this work represents a step toward more accurate and general protein ensemble prediction, several limitations remain. First, the realism of IDR training data could be further enhanced by using either curated databases of structural ensembles [21] or advanced computational methods, such as long-timescale molecular dynamics (MD) simulations with improved force fields and sampling techniques [54]. Recently developed MD datasets such as ATLAS [55], mdCATH [56], DynaREPO [57], and MISATO [58] have greatly expanded the availability of simulation-based protein ensembles. However, these predominantly capture local fluctuations near the folded state and often fall short in sampling the broad conformational landscape characteristic of IDRs. The BioEmu MD datasets stands out by providing extensive simulations of unfolded proteins and by training explicitly on equilibrium properties [14], as reflected in its strong performance on PeptoneBench (Figs. 3 and 4).

A further challenge in ensemble validation lies in the use of forward models to translate predicted structures into experimental observables with accurate uncertainty quantification. Many such models have been optimized and validated primarily on structured systems, raising concerns about their reliability when extrapolated to disordered proteins [59, 60]. For instance, CS predictors may exhibit systematic biases in highly flexible regions, while SAXS forward models may not fully capture the effects of excluded volume in extended conformations. Comprehensive cross-validation, leveraging multimodal datasets with diverse observables, is essential to mitigate these issues; yet assembling such datasets remains difficult and resource-intensive [21, 26, 61]. Future work should prioritize the development and validation of forward models specifically calibrated for disordered protein ensembles.

Finally, it is important to acknowledge that the conformational preferences of IDPs are highly sensitive to external conditions, including temperature, pH, ionic strength, and the presence of binding partners or small molecules. Numerous studies have shown that these environmental factors can drive disorder-to-order transitions, modulate structural heterogeneity, alter phase separation behavior, and impact biological function. Achieving predictive accuracy across varying conditions is therefore critical, not just for understanding protein biophysics but also for applications in therapeutic development and drug discovery [62]. Although modeling environmental effects lies beyond the scope of this work, we highlight this challenge as a key priority for the next generation of protein structure prediction methods. Approaches that integrate physics-based modeling with learned representations, or that explicitly condition predictions on experimental variables, may prove particularly valuable in this regard.

Beyond these technical challenges, our results carry broader implications for the field. The success of relatively simple data augmentation strategies in improving disorder prediction suggests that architectural complexity and the sheer amount of training data may not be the only limiting factor. Rather, the diversity and representativeness of training data appear critical. This finding has practical implications for resource allocation in model development and highlights opportunities for community-driven efforts to curate and share high-quality ensemble datasets [63].

Data and Code Availability

The PeptoneDB datasets, the IDRome-o dataset, the PepTron weights, and all the generated protein ensembles evaluated in the paper are available under CCBY on Zenodo at https://zenodo.org/records/17306061.

The code to reproduce the results and figures of the paper and to perform the PeptoneBench benchmark is available under Apache 2 license at https://github.com/peptoneltd/peptonebench. The code for inference and training of the PepTron model is available under Apache 2 license at https://github.com/peptoneltd/peptron. The IDP-o code will be made available in the near future.

Author Contributions

MI and CF led ideation, model development, and data analysis; JOS curated the PeptoneDB-integrative dataset, and performed analyses; FA generated the data. SB analysed data. BJ and BT contributed to research; BT and NAEV to model development, benchmarking. Additional model-development contributions from NAEV, DR, YL, and HS; All authors contributed to writing the paper.

Acknowledgments

We thank Oliver Dutton for contributing to the ideation and implementation of IDP-o. We acknowledge Francesco Farina, Simon Chu, Farhad Ramezanghorbani, Neha Tadimeti, Emine Kucukbenli, Cedric Steenbeke, and Saee Paliwal from NVIDIA for supporting this project. We also thank the PDB, BMRB, SASBDB, UniProt and all authors of experimental studies who have publicly shared their data enabling the benchmarks presented in this manuscript.

References

[1] Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* **25**, 187–211 (2024).

[2] Pritišanac, I. *et al.* A functional map of the human intrinsically disordered proteome. biorxiv [preprint](2024) (2025).

[3] Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

[4] Burley, S. K. *et al.* Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic acids research* **53**, D564–D574 (2025).

[5] Gall, T. L., Romero, P. R., Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in the protein data bank. *Journal of Biomolecular Structure and Dynamics* **24**, 325–341 (2007).

[6] Erdős, G. & Dosztanyi, Z. Deep learning for intrinsically disordered proteins: From improved predictions to deciphering conformational ensembles. *Current Opinion in Structural Biology* **89**, 102950 (2024).

[7] Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Biophysical Journal* **123**, 43a (2024).

[8] Riley, A. C., Ashlock, D. A. & Graether, S. P. The difficulty of aligning intrinsically disordered protein sequences as assessed by conservation and phylogeny. *Plos one* **18**, e0288388 (2023).

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

[10] Piovesan, D., Monzon, A. M. & Tosatto, S. C. Intrinsic protein disorder and conditional folding in alphafolddb. *Protein Science* **31**, e4466 (2022).

[11] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

[12] Corley, N. *et al.* Accelerating biomolecular modeling with atomworks and rf3. *bioRxiv* 2025–08 (2025).

[13] Jing, B., Berger, B. & Jaakkola, T. AlphaFold meets flow matching for generating protein ensembles. *NeurIPS* (2023). URL https://openreview.net/forum?id=yQcebEgQfH. Three versions, only the first one mentions IDPs.

[14] Lewis, S. *et al.* Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* eadv9817 (2025).

[15] Cao, F., von Bülow, S., Tesei, G. & Lindorff-Larsen, K. A coarse-grained model for disordered and multi-domain proteins. *Protein Science* **33**, e5172 (2024).

[16] Joseph, J. A. *et al.* Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nature computational science* **1**, 732–743 (2021).

[17] Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Computational Biology* **20**, e1012144 (2024).

[18] Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nature Communications* **14**, 774 (2023).

[19] Zhang, O., Liu, Z. H., Forman-Kay, J. D. & Head-Gordon, T. Deep learning of proteins with local and global regions of disorder. *arXiv preprint arXiv:2502.11326* (2025).

[20] Schnapka, V., Morozova, T. I., Sen, S. & Bonomi, M. Atomic resolution ensembles of intrinsically disordered proteins with alphafold. *bioRxiv* 2025.06.18.660298 (2025).

[21] Ghafouri, H. *et al.* Ped in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic acids research* **52**, D536–D544 (2024).

[22] Sil, S., Datta, I. & Basu, S. Use of ai-methods over md simulations in the sampling of conformational ensembles in idps. *Frontiers in Molecular Biosciences* **12**, 1542267 (2025).

[23] Ghafouri, H. *et al.* Towards a unified framework for determining conformational ensembles of disordered proteins. *arXiv* (2025).

[24] Norgaard, A. B., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophysical journal* **94**, 182–192 (2008).

[25] Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* **115**, E4758–E4766 (2018).

[26] Borthakur, K., Sisk, T. R., Panei, F. P., Bonomi, M. & Robustelli, P. Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution. *Nature Communications* **16**, 9036 (2025). URL https://doi.org/10.1038/s41467-025-64098-3.

[27] Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of chemical physics* **143** (2015).

[28] Cesari, A., Gil-Ley, A. & Bussi, G. Combining simulations and solution experiments as a paradigm for rna force field refinement. *Journal of chemical theory and computation* **12**, 6192–6200 (2016).

[29] Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. Metainference: A bayesian inference method for heterogeneous systems. *Science advances* **2**, e1501177 (2016).

[30] Bottaro, S., Bengtsen, T. & Lindorff-Larsen, K. *Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach*, 219–240 (Springer US, New York, NY, 2020).

[31] Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics* **30**, 137–149 (2012).

[32] Aspromonte, M. C. *et al.* Disprot in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research* **52**, D434–D441 (2024).

[33] Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).

[34] Haak, M., Senoner, T., Schlensok, J., Wagemann, D. & Rost, B. TriZOD: continuous, per-residue disorder scores from protein NMR experiments. https://github.com/MarkusHaak/trizod (2024).

[35] Nielsen, J. T. & Mulder, F. A. A. Quality and bias of protein disorder predictors. *Scientific Reports* **9**, 5137 (2019).

[36] Redl, I. *et al.* Optimizing protein language models with sentence transformers (2023). URL https://www.mlsb.io/papers_2023/Optimizing_protein_language_models_with_Sentence_Transformers.pdf. Workshop Paper.

[37] Hoch, J. C. *et al.* Biological magnetic resonance data bank. *Nucleic Acids Research* **51**, D368–D376 (2022).

[38] Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Science* **29**, 66–75 (2020).

[39] Thomasen, F. E. *et al.* Rescaling protein-protein interactions improves martini 3 for flexible proteins in solution. *Nature Communications* **15**, 6645 (2024).

[40] Cavender, C. E. *et al.* Structure-based experimental datasets for benchmarking of protein simulation force fields. *arXiv preprint arXiv:2303.11056* (2023).

[41] Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR* **48**, 13–22 (2010).

[42] Li, J., Bennett, K. C., Liu, Y., Martin, M. V. & Head-Gordon, T. Accurate prediction of chemical shifts for aqueous protein structure on "real world" data. *Chemical Science* **11**, 3180–3191 (2020).

[43] Grudinin, S., Garkavenko, M. & Kazennov, A. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle x-ray scattering profiles. *Acta Crystallographica Section D: Structural Biology* **73**, 449–464 (2017).

[44] Sterckx, Y. G. *et al.* Small-angle x-ray scattering-and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin paaa2. *Structure* **22**, 854–865 (2014).

[45] Tesei, G. *et al.* Deer-predict: Software for efficient calculation of spin-labeling epr and nmr data from conformational ensembles. *PLOS Computational Biology* **17**, e1008551 (2021).

[46] Teixeira, J. M. *et al.* Idpconformergenerator: a flexible software suite for sampling the conformational space of disordered protein states. *The Journal of Physical Chemistry A* **126**, 5985–6003 (2022).

[47] Pietrek, L. M., Stelzl, L. S. & Hummer, G. Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations. *Journal of chemical theory and computation* **16**, 725–737 (2019).

[48] John, P. S. *et al.* BioNeMo framework: a modular, high-performance library for AI model development in drug discovery. *arXiv* (2024).

[49] Shoeybi, M. *et al.* Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).

[50] Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022).

[51] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[52] Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv* 2025–06 (2025).

[53] Wang, Y., Lu, J., Jaitly, N., Susskind, J. & Bautista, M. A. Simplefold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480* (2025).

[54] Invernizzi, M., Piaggi, P. M. & Parrinello, M. Unified approach to enhanced sampling. *Physical Review X* **10**, 041034 (2020).

[55] Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C. & Galochkina, T. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research* **52**, D384–D392 (2024).

[56] Mirarchi, A., Giorgino, T. & De Fabritiis, G. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data* **11**, 1299 (2024).

[57] Mokhtari, O., Bignon, E., Khakzad, H. & Karami, Y. Dynarepo: The repository of macromolecular conformational dynamics. *bioRxiv* 2025–08 (2025).

[58] Siebenmorgen, T. *et al.* Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature computational science* **4**, 367–378 (2024).

[59] 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. *Journal of Molecular Biology* **433**, 167196 (2021).

[60] Schneidman-Duhovny, D., Pellarin, R. & Sali, A. Uncertainty in integrative structural modeling. *Current opinion in structural biology* **28**, 96–104 (2014).

[61] Nagy, G., Hoffmann, S. V., Jones, N. C. & Grubmüller, H. Reference data set for circular dichroism spectroscopy comprised of validated intrinsically disordered protein models. *Applied Spectroscopy* **78**, 897–911 (2024).

[62] Wu, K. *et al.* Design of intrinsically disordered region binding proteins. *Science* **389**, eadr8063 (2025).

[63] Amaro, R. E. *et al.* The need to implement fair principles in biomolecular simulations. *Nature methods* 1–5 (2025).

[64] Heo, L. & Feig, M. One particle per residue is sufficient to describe all-atom protein structures. *bioRxiv* (2023).

[65] Visani, G. M., Galvin, W., Pun, M. & Nourmohammad, A. Knowles, D. A. & Mostafavi, S. (eds) *H-packer: Holographic rotationally equivariant convolutional neural network for protein side-chain packing.* (eds Knowles, D. A. & Mostafavi, S.) *Proceedings of the 18th Machine Learning in Computational Biology meeting*, Vol. 240 of *Proceedings of Machine Learning Research*, 230–249 (PMLR, 2024).

[66] Ptaszek, A. L., Li, J., Konrat, R., Platzer, G. & Head-Gordon, T. UCBShift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures. *Journal of the American Chemical Society* **146**, 31733–31745 (2024).

[67] Cavender, C. E. *et al.* Structure-based experimental datasets for benchmarking of protein simulation force fields. *arXiv* (2023).

[68] Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *Journal of Biomolecular NMR* **70**, 141–165 (2018).

[69] Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography* **25**, 495–503 (1992).

[70] Hansen, S. Bayesian estimation of hyperparameters for indirect fourier transformation in small-angle scattering. *Applied Crystallography* **33**, 1415–1421 (2000).

[71] Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).

[72] Pesce, F. & Lindorff-Larsen, K. Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. *Biophysical Journal* **120**, 5124–5135 (2021).

[73] Hou, L. *et al.* Solution nmr studies of the aβ (1- 40) and aβ (1- 42) peptides establish that the met35 oxidation state affects the mechanism of amyloid formation. *Journal of the American Chemical Society* **126**, 1992–2005 (2004).

[74] Newby, F. N. *et al.* Structure-free validation of residual dipolar coupling and paramagnetic relaxation enhancement measurements of disordered proteins. *Biochemistry* **54**, 6876–6886 (2015).

[75] Iešmantavičius, V. *et al.* Modulation of the intrinsic helix propensity of an intrinsically disordered protein reveals long-range helix–helix interactions. *Journal of the American Chemical Society* **135**, 10155–10163 (2013).

[76] Kjaergaard, M. *et al.* Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline ii? *Protein Science* **19**, 1555–1564 (2010).

[77] Newcombe, E. A. *et al.* Insight into calcium-binding motifs of intrinsically disordered proteins. *Biomolecules* **11**, 1173 (2021).

[78] Pesce, F. *et al.* Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophysical Journal* **122**, 310–321 (2023).

[79] Due, A. D. *et al.* Hierarchy in regulator interactions with distant transcriptional activation domains empowers rheostatic regulation. *Protein Science* **34**, e70142 (2025).

[80] Rao, J. N., Kim, Y. E., Park, L. S. & Ulmer, T. S. Effect of pseudorepeat rearrangement on α-synuclein misfolding, vesicle binding, and micelle binding. *Journal of molecular biology* **390**, 516–529 (2009).

[81] Bertoncini, C. W. *et al.* Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proceedings of the National Academy of Sciences* **102**, 1430–1435 (2005).

[82] Ahmed, M. C. *et al.* Refinement of α-synuclein ensembles against saxs data: Comparison of force fields and methods. *Frontiers in molecular biosciences* **8**, 654333 (2021).

[83] Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M. & Dobson, C. M. Mapping long-range interactions in α-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *Journal of the American Chemical Society* **127**, 476–477 (2005).

[84] Qu, X. *et al.* Structural basis for the dynamic chaperoning of disordered clients by hsp90. *Nature Structural & Molecular Biology* **31**, 1482–1491 (2024).

[85] Shortle, D. & Ackerman, M. S. Persistence of native-like topology in a denatured protein in 8 m urea. *Science* **293**, 487–489 (2001).

[86] Zhang, O. & Forman-Kay, J. D. Structural characterization of folded and unfolded states of an sh3 domain in equilibrium in aqueous buffer. *Biochemistry* **34**, 6784–6794 (1995).

[87] Marsh, J. A., Baker, J. M., Tollinger, M. & Forman-Kay, J. D. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *Journal of the American Chemical Society* **130**, 7804–7805 (2008).

[88] Choy, W.-Y. *et al.* Distribution of molecular size within an unfolded state ensemble using small-angle x-ray scattering and pulse field gradient nmr techniques. *Journal of molecular biology* **316**, 101–112 (2002).
[89] Marsh, J. A. & Forman-Kay, J. D. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *Journal of molecular biology* **391**, 359–374 (2009).
[90] Du, Z. *et al.* The sequence–structure–function relationship of intrinsic erα disorder. *Nature* **638**, 1130–1138 (2025).
[91] Streit, J. O. *et al.* The ribosome lowers the entropic penalty of protein folding. *Nature* **633**, 232–239 (2024).
[92] Gely, S. *et al.* Solution structure of the c-terminal x domain of the measles virus phosphoprotein and interaction with the intrinsically disordered c-terminal domain of the nucleoprotein. *Journal of molecular recognition* **23**, 435–447 (2010).
[93] Jensen, M. R. *et al.* Intrinsic disorder in measles virus nucleocapsids. *Proceedings of the National Academy of Sciences* **108**, 9839–9844 (2011).
[94] Schedlbauer, A. *et al.* Backbone assignment of osteopontin, a cytokine and cell attachment protein implicated in tumorigenesis. *Biomolecular NMR Assignments* **2**, 29–31 (2008).
[95] Platzer, G. *et al.* The metastasis-associated extracellular matrix protein osteopontin forms transient structure in ligand interaction sites. *Biochemistry* **50**, 6113–6124 (2011).
[96] De Biasio, A. *et al.* p15paf is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophysical journal* **106**, 865–874 (2014).
[97] Sebák, F., Ecsédi, P., Nyitray, L. & Bodor, A. Assignment of the disordered, proline-rich n-terminal domain of the tumour suppressor p53 protein using 1hn and 1hα-detected nmr measurements. *Biomolecular NMR Assignments* **17**, 309–314 (2023).
[98] Wells, M. *et al.* Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proceedings of the National academy of Sciences* **105**, 5762–5767 (2008).
[99] Baker, J. M. *et al.* Cftr regulatory region interacts with nbd1 predominantly via multiple transient helices. *Nature structural & molecular biology* **14**, 738–745 (2007).
[100] Mittag, T. *et al.* Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an scf ubiquitin ligase. *Structure* **18**, 494–506 (2010).
[101] Gomes, G.-N. W. *et al.* Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. *Journal of the American Chemical Society* **142**, 15697–15710 (2020).
[102] Crha, R. *et al.* Hiding in plain sight: complex interaction patterns between tau and 14-3-3ζ protein variants. *International Journal of Biological Macromolecules* **266**, 130802 (2024).
[103] Mukrasch, M. D. *et al.* Structural polymorphism of 441-residue tau at single residue resolution. *PLoS biology* **7**, e1000034 (2009).
[104] Mylonas, E. *et al.* Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochemistry* **47**, 10345–10353 (2008).
[105] Zweckstetter, M. Nmr: prediction of molecular alignment from structure using the pales software. *Nature protocols* **3**, 679–690 (2008).
[106] Kim, H., Mirdita, M. & Steinegger, M. Foldcomp: a library and format for compressing and indexing large protein structure sets. *Bioinformatics* **39**, btad153 (2023).
[107] Dutton, O., Hoffmann, F. & Tamiola, K. Nerfax: An efficient and scalable conversion from the internal representation to cartesian space. *bioRxiv* 2022–05 (2022).
[108] Verkuil, R. *et al.* Language models generalize beyond natural proteins. *BioRxiv* 2022–12 (2022).
[109] Consortium, U. Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research* **49**, D480–D489 (2021).
[110] Ahdritz, G. *et al.* Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature methods* **21**, 1514–1524 (2024).
[111] Webby, M. N. *et al.* Structural Analysis of the Menangle Virus P Protein Reveals a Soft Boundary between Ordered and Disordered Regions. *Viruses* **13**, 1737 (2021).
[112] Robin, X. *et al.* Continuous automated model evaluation (cameo)—perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* **89**, 1977–1986 (2021).
[113] Jing, B. *et al.* Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198* (2023).
[114] Mirdita, M. *et al.* Colabfold: making protein folding accessible to all. *Nature methods* **19**, 679–682 (2022).
[115] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
[116] Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nature Communications* **14**, 774 (2023). idpGAN.
[117] Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *bioRxiv* 2024.02.08.579522 (2024).
[118] Zhu, J. *et al.* Accurate Generation of Conformational Ensembles for Intrinsically Disordered Proteins with IDPFold. *Advanced Science* (2025).
[119] Wohlwend, J. *et al.* Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv* 2024.11.19.624167 (2024).
[120] Redl, I. *et al.* ADOPT: intrinsic protein disorder prediction through deep bidirectional transformers. *NAR Genomics and Bioinformatics* **5** (2023). URL https://doi.org/10.1093/nargab/lqad041. Lqad041.