GeoGraph: Geometric and Graph-based Ensemble Descriptors for Intrinsically Disordered Proteins

Anonymous Author(s)
Affiliation
Address
email

Abstract

While deep learning has revolutionized the prediction of rigid protein structures, modelling the conformational ensembles of Intrinsically Disordered Proteins (IDPs) remains a key frontier. Current AI paradigms present a trade-off: Protein Language Models (PLMs) capture evolutionary statistics but lack explicit physical grounding, while generative models trained to model full ensembles are computationally expensive. In this work we critically assess these limits and propose a path forward. We introduce GeoGraph, a simulation-informed surrogate trained to predict ensemble-averaged statistics of residue–residue contact-map topology directly from sequence. By featurizing coarse-grained molecular dynamics simulations into residue- and sequence-level graph descriptors, we create a robust and information-rich learning target. Our evaluation demonstrates that this approach yields representations that are more predictive of key biophysical properties than existing methods.

1 Introduction

Proteins are the cell's molecular machines: sequence-encoded biopolymers which catalyze reactions, regulate processes, and shape cellular architecture. Recent years have witnessed a paradigm shift in protein modelling, driven by advances in experimental techniques and the maturation of deep learning. In particular, the rapid growth of high-throughput sequencing has been pivotal [41]. On the one hand it has enabled language-modelling approaches, especially Masked Language Modelling (MLM), to learn the statistical patterns of evolution directly from vast, unannotated sequence databases [34, 26]. On the other, Multiple Sequence Alignments (MSAs), coupled with decades of structure determination experiments [5], underpin deep learning models like AlphaFold [24] and RosettaFold [4], which now achieve near-experimental accuracy for a broad class of structured proteins.

With static structures largely tractable, the frontier of computational structural biology is advancing toward a more fundamental problem: modelling the full conformational ensemble – the Boltzmann distribution of conformations under physiological solution conditions. To frame this challenge, we can identify three regimes along the structural order-disorder continuum: (i) proteins that adopt a single, highly stable fold; (ii) dynamic proteins that interconvert among a few metastable states; and (iii) Intrinsically Disordered Proteins (IDPs), which manifest a broad, heterogeneous set of rapidly fluctuating conformations [47, 43]. The first regime is where models trained on protein crystal structures excel. The second is well-captured by Markov State Models (MSM), which characterise the ensemble by the populations of metastable states and the kinetic rates between them, typically inferred from long Molecular Dynamics (MD) simulations [33, 9, 19]. The third regime of IDPs is, however, particularly challenging, and provides the focus for this work. Beyond the inherent complexity of modelling a heterogeneous ensemble, these proteins also face significant experimental and evolutionary hurdles. Experimentally, obtaining data is laborious, and their dynamic nature means measurements are typically averaged across the entire ensemble and/or over time. Evolutionarily,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

they exhibit poor sequence conservation, a characteristic thought to derive from the lack of a stable structure required to maintain function [8].

A recent line of work aims to use deep generative models, especially diffusion models, to map sequence directly to a full conformational distribution [25, 22, 20, 49]. While useful, this strategy faces practical and statistical hurdles: generating, storing, and analyzing thousands of conformers per protein is expensive, and for many downstream tasks such high-dimensional stochastic detail can obscure the underlying biophysical signal. From a statistical-physics perspective, fluctuations faster than the timescale of interest are effectively marginalized as entropy, making the explicit modelling of fine-grained, high-frequency detail counterproductive. Indeed timescale separation underpins MSM coarse-graining, which emphasizes slow, kinetically relevant transitions between states rather than the noisy internal motions within them [9, 19].

Here we take a different approach: rather than modelling entire ensembles explicitly, we model their aggregate properties directly. Specifically, we propose to extract essential biophysical content of an IDP ensemble from the statistics of its transient residue–residue contacts [7, 3]. The power of this approach has recently been demonstrated by WARIO [15], which uses contact-based descriptors to cluster simulation trajectories of IDPs into structurally coherent states. Our work leverages this same insight for a different purpose: instead of post-hoc analysis of a single ensemble, our aim is high-throughput prediction directly from sequence. To achieve this, we convert conformations from simulation into residue-level contact-map graphs, compute a diverse set of graph-theoretic descriptors, and use their ensemble-averaged values as the direct prediction targets for our model. This approach acts as a deliberate information bottleneck, filtering high-frequency fluctuations while preserving the stable signature of the ensemble.

A key design choice is resolution. We operate at the residue level–a natural middle ground between whole-sequence and all-atom representations. Unlike models that predict a few global aggregates and lose positional detail, we learn a rich vector of aggregate properties per residue, capturing biophysical characteristics across the protein sequence.

Finally, our physics-driven residue-level objective allows a direct test of a core premise behind Protein Language Models (PLMs): that evolutionary co-variation provides a statistical proxy for underlying physics [31, 29, 26]. However, this premise relies on stable structural constraints that are largely absent in dynamic IDPs, limiting the effectiveness of purely evolutionary models. By constructing representations grounded in the explicit physical properties of conformational ensembles, our work offers a framework to directly compare what is learnable from the language of evolution versus the language of physics – or at least a coarse-grained MD version of it.

2 Related work

Our work lies at the intersection of modelling protein conformational ensembles, machine learning for intrinsically disordered proteins, and graph-theoretic representations of protein structure.

Following the success of the third-generation AlphaFold3 structure prediction model employing a diffusion architecture [1], a growing line of work is addressing the challenge of modelling full conformational distributions directly from sequence. Prominently, BioEmu emulates equilibrium ensembles at scale from MD data, reporting agreement with thermodynamic readouts across diverse proteins [25]. P2DFlow employs SE(3) flow-matching for sequence-conditioned ensemble sampling [22]. However, these models rely on the quality of MSA and their capacity to capture co-evolutionary signal, which limits their performance for poorly-conserved IDPs.

IDPs, owing to their intrinsic structural and sequential heterogeneity, remain challenging for such general models, and several works specialize directly in sampling IDP ensembles. STARLING employs a sequence-conditioned latent diffusion architecture acting on residue-residue distance maps, enabling rapid prediction of coarse-grained IDP ensembles [32]. idpGAN conditions a transformer-based GAN on sequence to generate coarse-grained IDP conformations [21], and idpSAM adopts a latent-diffusion formulation to improve transfer across sequences [20]. Diffusion-based conditional sampling has also been explored for IDPs at all-atom resolution in IDPFold [49]. Complementing these pure generators, bAIes integrates AlphaFold2 distograms with an atomistic random-coil prior in a Bayesian framework to sample atomic-resolution IDP ensembles with uncertainty modelling of restraints [35].

In line with the present work, several methods learn sequence-to-aggregate mappings for IDP ensemble statistics based on simulated training sets. ALBATROSS is a family of recurrent neural network (RNN) models which each separately predict the average radius of gyration, end-to-end distance, asphericity, and Flory scaling parameters from sequence [28]. Our work is closely aligned with this approach, but targets a richer set of residue-resolved and sequence-level statistics derived from transient contact patterns. From a PLM angle, IDP-BERT is a fine-tuning ProtBERT on the prediction of average radius of gyration, end-to-end decorrelation time, and heat capacity [30], without further language model training on IDP-specific sequences. In contrast, our IDP-ESM models, presented in Sec. 4.3, are first fine-tuned on the MLM objective using a curated dataset of IDP sequences before the prediction head is trained.

Finally, residue interaction/contact networks have a long history as mesoscopic abstractions linking local contacts to global organization. Early studies established contact-graph formalisms and connected network topology to stability [7] and functional residues [3]. Closer to the theme of our work, WARIO employs contact maps for characterizing IDP ensembles [15]. They take a sophisticated approach to defining contacts, introducing a novel continuous function which incorporates residue type, sequence separation, and relative orientation. The focus of their work, clustering conformations of an individual ensemble so as to reveal rare functionally relevant patterns, is complementary to our ensemble-averaged approach.

3 GeoGraph

Our goal is to learn residue-level representations of IDPs directly from MD trajectory data. Our hope is that this will capture essential physical principles – principles that are missed by PLMs, and inaccessible to methods like AlphaFold3, which are trained on structured, folded proteins with deep MSAs.

For a given protein sequence, a conformational ensemble can be sampled from a sufficiently long MD trajectory after an initial equilibration period. A key question, however, concerns the required accuracy of the simulation. While all-atom force fields produce a high-fidelity view of protein dynamics, their immense computational cost makes them unsuitable for generating the large datasets of equilibrated trajectories required for deep learning.

Coarse-grained methods offer a pragmatic and powerful alternative. In particular, CALVADOS-2 is a state-of-the-art[1], one-bead-per-residue coarse-grained force field specifically developed and parameterized to reproduce experimental data on IDPs [40]. This coarse-graining comes at a cost: fine details such as secondary structure are lost. Instead, CALVADOS takes a top-down approach, based on an effective description of non-bonded interactions designed to capture transient residue–residue contact patterns.

Building on this, we hypothesise that these transient residue-residue contacts encode context-rich, robust, and generalizable physicochemical correlations. To formalize this, we analyze properties of these contacts aggregated across the conformational ensemble. For each conformation we construct a contact map graph, where residues are nodes and the edges connect pairs of residues within an 8Å distance cutoff, but omitting edges between adjacent residues. From each graph, we then compute a diverse set of features at both the node level and the graph level (our selection is presented below in Sec. 3.2). Finally, these feature vectors are averaged across the full ensemble to produce a stable, statistical fingerprint of the protein's dynamic structure.

3.1 Architecture

GeoGraph is a sequence-to-sequence model that maps a protein's amino-acid sequence to feature vectors describing aggregate physical properties at both the sequence- and residue-level. The backbone is a transformer encoder [44], chosen for its ability to capture long-range dependencies and produce context-rich embeddings. We build on the Hugging Face implementation of ESM-2 [26, 46], which uses standard modern components such as Pre-Layer Normalization (Pre-LN) [48] and Rotary Position Embeddings (RoPE) [37].

[1]We comment that its successor CALVADOS-3 is an adaptation to incorporate multi-domain proteins, and is equivalent to CALVADOS-2 for IDPs.

We use a 4-layer transformer with hidden size 256, 4 attention heads, and a feed-forward expansion factor of 2 (FFN dimension 512), for a total of ≈ 2.2M parameters. The output of the transformer is a sequence of residue-level embeddings. A single sequence-level embedding is then obtained by taking the mean of these residue-level embeddings across the sequence length.

To predict targets, we attach separate heads for sequence-level and residue-level features. Each head is a shallow MLP with a single hidden layer of dimension 128 and a dropout probability of 0.1, so that performance primarily reflects the backbone's context-aware embeddings.

To ensure robust training, the transformer backbone is also regularized with dropout of 0.1 on both the FFN activations and the attention probabilities. We use the AdamW optimizer with a weight decay of 0.01 and a cosine learning rate scheduler with warmup. Full training details are presented in Appendix A.1.

3.2 Ensemble descriptors

In this work we consider two flavours of descriptors, which we refer to as geometric and graph-based. The geometric descriptors are all sequence-level features, while for the graph-based descriptors we consider both sequence- and residue-level features. For training our models we standardise all target features to have zero mean and unit variance.

3.2.1 Geometric features

The geometric features we consider are commonly employed measures of the conformational ensembles of IDPs, and are computable from MD simulation frames as detailed in Appendix B. First, **end-to-end distance** (R_e) is the Euclidean distance between the first and last residues. The **radius of gyration** (R_g) and **asphericity** (Δ) are characteristics of a protein's (mass-weighted) gyration tensor. Lastly, in line with polymer physics, IDPs exhibit Flory scaling. This describes a power-law relationship between a polymer's size (e.g. R_g or R_e) and its length N, which is parametrized by an **exponent** (ν) and a **prefactor** (A_0), as in $R_g \propto A_0 N^{\nu}$ [14]. In practice we compute the scaling parameters for a given IDP by fitting a power-law relationship between the Euclidean distance of residue pairs and their sequence separation.

Both R_g and R_e can be experimentally determined, and in turn used to determine the Flory prefactor and exponent [2]. Small Angle X-ray Scattering (SAXS) yields the ensemble-averaged radius of gyration $\langle R_g \rangle$, whereas Fluorescence resonance energy transfer (FRET) spectroscopy yields $\langle R_e \rangle$, or even R_e distributions in the case of single-molecule FRET [18].

3.2.2 Graph-based features

In selecting graph-based features for a conformations's contact map graph we are non-discriminative, aiming for a diverse collection. Overall we selected 8 sequence-level features and 7 residue-level features as below. We compute these using python's NetworkX package [17] (with default settings).

Sequence-level: As not all graphs were connected we computed **fragmentation index** as the fraction of nodes in the Largest Connected Component (LCC); **average shortest path length** on the LCC and **global efficiency** on the full graph to quantify compactness/communication; **average clustering** and **transitivity** as measures of local triadic closure; and **degree assortativity** as well as **charge assortativity** and **hydrophobicity assortativity** to assess mixing patterns. The latter two were computed by endowing nodes with attribute values determined by corresponding amino acid identity.

Residue-level: Here we included **degree centrality** (local contact density), **betweenness centrality** (bridging propensity), **harmonic centrality** (inverse-distance reachability), **PageRank** [6], **core number**, **local clustering coefficient**, and as well as an **in-largest-connected-component** indicator.

3.3 MD training data

To train and evaluate our models, we use the Human–IDRome dataset [39], containing simulated conformational ensembles for 28,058 intrinsically disordered regions from the human proteome. To our knowledge, this is the largest publicly available dataset of its kind. The ensembles were generated using the CALVADOS-2 coarse-grained force field, with each sequence represented by 1,000 weakly correlated conformational frames sampled from the simulation trajectory [39].

	R_e	R_g	Δ	ν	A_0
STARLING	0.914	0.951	-0.460	0.261	0.386
STARLING (retrained)	0.983	0.992	0.314	0.677	0.539
ALBATROSS	0.899	0.932	0.441	0.275*	-0.471*
ALBATROSS (retrained)	0.970	0.984	0.790	0.698	0.513
GeoGraph	0.993 (0)	0.996 (0)	0.899 (5)	0.893 (6)	0.875 (16)
Geo	0.991 (2)	0.994 (1)	0.875 (13)	0.856 (14)	0.787 (30)
Geo-zero	0.596 (33)	0.603 (33)	0.584 (6)	0.505 (7)	0.389 (13)
Graph + GeoHead	0.992 (1)	0.996 (0)	0.864 (13)	0.854 (15)	0.793 (32)
ESM-8M + GeoHead	0.983 (1)	0.991 (1)	0.754 (8)	0.684 (8)	0.523 (19)
IDP-ESM-8M + GeoHead	0.982 (1)	0.987 (1)	0.783 (2)	0.767 (5)	0.643 (14)
ESM-150M + GeoHead	0.984 (1)	0.991 (1)	0.792 (2)	0.763 (4)	0.637 (5)
IDP-ESM-150M + GeoHead	0.980 (1)	0.986 (1)	0.786 (6)	0.777 (4)	0.660 (7)

Table 1: R^2 scores for the IDP property prediction task on our Human–IDRome test set. Where parentheses are shown, the results are the mean of 5 models with different random seeds, along with the standard error on the final digits. We highlight with (*) that the R^2 scores of the pretrained ALBATROSS models for ν and A_0 are affected by differences in computation of the scaling parameters between our work and theirs (see Appendix C.2.4).

We partitioned the dataset based on sequence similarity into 80/10/10 splits for training, validation, and testing. To ensure fair comparison with prior work, this split was performed using MMseqs2 [36] with parameters (min_seq_id=0.7, coverage=0.8, cov_mode=1), identical to the parameters used by STARLING [32]. Finally, we filtered the dataset to sequences with a maximum length of 256 residues.

4 Evaluation

We evaluate models on their ability to predict the five geometric features $(R_e, R_g, \Delta, \nu, A_0)$ described in Sec. 3.2.1, which are well-studied, experimentally relevant measures of IDP conformational ensembles.

4.1 Baseline IDP models

First, we evaluate two prominent methods for IDP property prediction: ALBATROSS [28] and STARLING [32]. ALBATROSS is a family of 5 recurrent neural network models, each trained to independently predict one of the ensemble-averaged geometric features $\{R_e, R_g, \Delta, \nu, A_0\}$ directly from sequence. STARLING is a generative diffusion model which generates a conformational ensemble of IDPs by denoising a latent representation of residue-residue distance maps for each conformation. We follow the method used by [32] for property prediction with STARLING: we sample 1000 conformations using 25 DDIM steps, then using the generated ensemble to calculate the ensemble-averaged geometric feature values for each sequence.

We evaluate the publicly released models for both methods on our test set, however we also note that the IDP datasets used to train ALBATROSS and STARLING notably differ from our training dataset Human–IDRome. In particular, their datasets contain synthetic as well as biological IDP sequences, and the conformational ensembles were generated via coarse-grained MD using an adapted version of the Mpipi force field [23] rather than CALVADOS-2. We therefore additionally retrained these models from scratch on the Human–IDRome dataset, and report results using both the pretrained and retrained versions of these models in Table 1. We include additional details related to retraining and comparison in Appendix C.

In addition to the results presented in Table 1, we evaluated the published IDP-BERT model [30] for predicting R_g on our test set, which achieved an R^2 score of 0.949. For a more detailed evaluation of PLM embeddings we use the ESM-2 model and an IDP fine-tuning of it below in Sec. 4.3.

We also attempted to evaluate BioEmu, a large-scale general-purpose ensemble emulator [25] which uses a diffusion model to generate conformational ensembles conditioned on the MSA of a sequence. Due to computational constraints, we were not able to generate sufficiently large ensembles with

	R_e	R_g	Δ	ν	A_0
GeoGraph (4 layers)	0.993 (0)	0.996 (0)	0.899 (5)	0.893 (6)	0.875 (16)
– 6 layers	0.993 (1)	0.996 (1)	0.897 (3)	0.891 (4)	0.872 (15)
– 2 layers	0.992 (1)	0.996 (0)	0.890 (6)	0.883 (3)	0.848 (10)
– 1 layer	0.991 (2)	0.994 (2)	0.864 (26)	0.859 (14)	0.794 (31)
– w/o sequence graph features	0.993 (1)	0.996 (1)	0.896 (9)	0.886 (9)	0.858 (15)
– w/o residue graph features	0.988 (2)	0.992 (2)	0.858 (10)	0.856 (15)	0.806 (34)
– w/o residue centralities	0.993 (1)	0.996 (0)	0.886 (8)	0.880 (12)	0.848 (26)
– w/o residue pagerank	0.993 (1)	0.996 (1)	0.896 (10)	0.889 (7)	0.868 (18)
– w/o residue clustering	0.993 (1)	0.996 (0)	0.897 (4)	0.886 (6)	0.861 (16)

Table 2: R^2 scores on our Human–IDRome test set for several ablations on the GeoGraph model. The results are the mean of 5 models with different random seeds, along with the standard error on the final digits in parentheses.

BioEmu on our test set to make a fair comparison. In a small experiment where we generated 1000 conformers/sequence for 100 randomly-sampled test sequences, we observed very poor performance ($R^2 < 0$ for all features), which is consistent with recent work evaluating BioEmu for IDPs [35].

4.2 GeoGraph models

We trained multiple variants of our GeoGraph model so as to clearly dissect its behaviour. These can be grouped as follows:

Main model

- **GeoGraph:** trained with both sequence-level and residue-level prediction heads on the sets of geometric and graph features.

Baselines

- **Geo:** trained to predict only the sequence-level geometric features, i.e. those used as benchmarks. This serves as an analogue to ALBATROSS, up to the change in architecture and the use of a single model to predict all features.
- **Geo-zero:** a greatly simplified variant of Geo where the transformer backbone has zero layers. This tests the performance of contextless token embeddings, and provides a minimal performance floor.

Transferability Test

- **Graph + GeoHead:** a variant designed to assess the transferability of the learned embeddings. It is trained in two stages: first, the full backbone is trained to predict only the graph-based features; second, the backbone weights are frozen, and a new prediction head is trained to predict the geometric features from the resulting embeddings.

In addition to these variants, we performed a set of feature ablations on the main GeoGraph model, presented in Table 2.

4.3 Protein Language Model evaluation

Finally, we evaluate the performance of the MLM model ESM-2 on our benchmark tasks. We test the model in two settings: first, using its general-purpose pre-trained embeddings, and second, after fine-tuning it on a large, curated dataset of biological IDP sequences. As we did for the *Graph + GeoHead* model above, we train a prediction head on the mean embeddings across the sequence to predict the geometric features. We report results for 8M and 150M models for each in Table 1.

4.3.1 IDP fine-tuning

For training our fine-tuned versions of ESM-2, **IDP-ESM2-8M** and **IDP-ESM2-150M**, we curated a large dataset of biological IDP sequences. Adopting the approach used by ALBATROSS [28] to

extract biological IDPs from multiple proteomes, we used Metapredict V3 [13] to annotate disordered regions in all 2,764 eukaryotic proteomes available in the UniProt [10] database, as detailed in Appendix A.2. We refined the resulting dataset by filtering out sequences shorter than 30 amino acids, a standard preprocessing step for IDPs [11], and then applied clustering at 90% sequence identity, following the UniRef90 protocol [38], resulting in a dataset of approximately 30 million IDP sequences, which we refer to as **IDP-Euka-90**.

The fine-tuning leads to a significant performance improvement for the 8M model, while having little effect on the 150M model. We hypothesize that ESM2-150M already captures key properties of IDP sequences from its UniRef50 pretraining, and that additional fine-tuning does not significantly enhance its ability to model geometric features.

5 Discussion

From Table 1 we see that GeoGraph achieves highly competitive performance against leading methods for IDP ensemble property prediction. While the end-to-end distance (R_e) and radius of gyration (R_g) appear to be relatively tractable tasks for most models, GeoGraph demonstrates a marked improvement in predicting the more challenging shape descriptors of asphericity (Δ), the Flory scaling exponent (ν), and its prefactor (A_0).

We gain further insight into this success by comparing our model variants: the Geo model, which trains on geometric features alone, and the Graph + GeoHead model, which learns representations solely from graph topology. Both variants perform on par with each other, demonstrating that the rich biophysical information in the contact-map topology is sufficient to create representations as powerful as those learned by direct optimization. The value of these learned representations is confirmed by the far superior performance of Graph + GeoHead relative to our Geo-zero baseline, which lacks this contextual learning. Crucially, the main GeoGraph model outperforms both specialized variants, demonstrating a clear synergistic effect. This supports our central hypothesis: the auxiliary task of predicting contact map characteristics is a highly beneficial component for extracting transferable representations from MD simulation data. Furthermore, ablations in Table 2 reveal that the context-aware, residue-level graph features are the primary drivers of this learning.

In comparing GeoGraph's performance to reference IDP models, we must first highlight a key caveat: our model was trained on the Human–IDRome dataset generated with the CALVADOS-2 force field, whereas the original ALBATROSS and STARLING models were trained on data containing both biological and synthetic sequences, and generated using the Mpipi force field. This difference in training distributions means a direct comparison between the models is imperfect. To mitigate this, we retrained these models on the Human–IDRome dataset, which consistently improved the performance of both methods on our test set across all features.

While the retrained STARLING model performs well on simpler metrics like R_g, its accuracy degrades significantly on the more challenging shape descriptors. Specifically on asphericity (Δ), the generative model performs worse than our Geo-zero baseline, which lacks any contextual transformer layers. This suggests that while the generative model can capture the average size of an ensemble, it has struggled to capture the distribution of more complex descriptors. This suggests that the aggregate graph descriptors distill the essential information of an ensemble's shape more effectively than the raw residue-residue distance maps used by STARLING.

We find that the retrained ALBATROSS models underperform relative to our Geo model, which acts as an analogue of ALBATROSS. There are multiple factors which may contribute to this performance gap, the most prominent of which being that the ALBATROSS models were developed with a focus on enabling high-throughput inference by using very small model sizes (between 34K-107K parameters), whereas the Geo model has $\approx$2.2M parameters.

Our final comparison is against the evolution-informed embeddings of ESM-2, and our MLM fine-tuned variants IDP-ESM-8M and IDP-ESM-150M. Here, the most direct comparison is with our Graph + GeoHead model, which isolates the quality of the embeddings pretrained on graph topology. We find that these simulation-informed embeddings provide a significantly stronger predictive signal for geometric properties than those learned by MLM. While a superior performance is expected when the training objective aligns with the evaluation task, the magnitude of the difference underscores a

key limitation of protein language models: the statistical patterns of evolution are an incomplete and often noisy proxy for the underlying physical properties when applied to IDPs [8].

Overall, we find the results of our evaluation highly encouraging. Nevertheless, we emphasize that this is an exploratory study, and there are several limitations. Firstly, like all comparable methods, GeoGraph is fundamentally an emulator of the underlying coarse-grained simulation. It therefore inherits the limitations of the CALVADOS-2 force field, most notably the absence of all-atom detail. Importantly, we would not expect our current model to learn a meaningful signal for key fine-detail properties such as secondary structure propensity. Future work could address this by training on data from all-atom simulations, which, while computationally expensive, may provide a richer and more physically accurate training signal.

Secondly, our methodological choices in featurizing the contact map graphs can be refined. We sought diverse sets of both residue-level (node) and sequence-level (graph) descriptors, but did not attempt to optimise these. In addition there are descriptors we did not consider, a notable example are community structures, which may help to capture modular organization within the ensembles [16]. Our definition of the contact map itself, a hard 8Å cutoff, is also simplistic. It would be interesting to assess the impact of the cutoff distance, or more generally to adopt distribution-based contact definitions, such as that introduced recently by WARIO [15], which may provide a more nuanced learning target.

Finally, our aggregation of features across the ensemble deserves mention. In this work, we focused on predicting the mean of each feature, a choice that imposes a powerful information bottleneck to filter the inherent stochasticity and extract a robust signal. A drawback is that this averaging necessarily loses information about the ensemble's heterogeneity, such as rare functional conformations – often key in IDP biological functionality. A natural next step is to enrich our prediction targets by including higher-order statistics, such as the variance of each feature. This would allow the model to learn not just the average state but also the extent of conformational fluctuations, capturing a deeper layer of the ensemble's physical character without incurring the full cost of a generative model.

6 Conclusion

In this work we introduce GeoGraph, a sequence-to-sequence model trained to predict aggregate properties of IDP conformational ensembles. It achieves this by first featurizing individual conformations from MD simulations into contact-graph topologies, and then learning to predict the ensemble average of these features, at both a residue- and sequence-level. Our evaluation demonstrates that this approach not only achieves highly competitive performance on benchmark tasks but also yields embeddings that are more effective for predicting key experimentally relevant properties than existing methods. Our trained GeoGraph and IDP-ESM2 models, along with the IDP-Euka-90 training dataset, will be publicly released.

Let us conclude by highlighting some directions for future work. Firstly, it is of interest to expand and refine the set of residue-level features that can be extracted from MD simulation data. If one extends the scope to all-atoms simulations one can expect that significantly richer features can be employed. Secondly, while we have focused here on contrasting our approach to MLM, establishing a solid baseline with our IDP fine-tuned IDP-ESM2-8M and IDP-ESM2-150M models, ultimately we anticipate a merging of the two approaches, so as to create powerful sequence representations that combine evolutionary covariance with underlying physics. Finally, we hope that these methods will lead to improved understanding of IDPs, for example serving as useful representations for characterising functional motifs [42].

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, others, Andrew W. Senior, Demis Hassabis, and John Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[2] Mustapha Carab Ahmed, Ramon Crehuet, and Kresten Lindorff-Larsen. Computing, analyzing and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins, August 2019.

[3] Gal Amitai, Ariel Shemesh, Eitan Sitbon, Michal Shklar, Dror Netanely, Inbar Venger, and Shmuel Pietrokovski. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344(4):1135–1146, 2004.

[4] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021.

[5] Helen M. Berman, John Westbrook, Zukang Feng, and et al. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117, 1998.

[7] K. V. Brinda and Saraswathi Vishveshwara. A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, 89(6):4159–4170, 2005.

[8] Celeste J. Brown, Audra K. Johnson, A. Keith Dunker, and Gary W. Daughdrill. Evolution and disorder. *Curr. Opin. Struct. Biol.*, 21(3):441–446, 2011.

[9] John D. Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, 2014.

[10] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024.

[11] Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. New classification of intrinsic disorder in the human proteome. *bioRxiv*, 2018.

[12] R. I. Dima and D. Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B*, 108(21):6564–6570, 2004.

[13] Ryan J. Emenecker, Danielm Griffith, and Alex S. Holehouse. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical Journal*, 120(20):4312–4319, 2021.

[14] Paul. J. Flory and M. Volkenstein. Statistical mechanics of chain molecules. *Biopolymers*, 8(5):699–700, November 1969.

[15] Javier González-Delgado, Pau Bernadó, Pierre Neuvial, and Juan Cortés. Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins. *Bioinformatics*, 40(11):btae627, 2024.

[16] William P Grant and Sebastian E Ahnert. Modular decomposition of protein structure using community detection. *Journal of Complex Networks*, 7(1):101–113, 2019.

[17] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Aug 2008.

[18] Hagen Hofmann, Andrea Soranno, Alessandro Borgia, Klaus Gast, Daniel Nettels, and Benjamin Schuler. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proceedings of the National Academy of Sciences*, 109(40):16155–16160, October 2012.

[19] Brooke E. Husic and Vijay S. Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140(7):2386–2396, 2018.

[20] Giacomo Janson, Alexey Jussupow, and Michael Feig. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Computational Biology*, 20(5):e1012144, 2024.

[21] Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14:774, 2023.

[22] Yaowei Jin, Qi Huang, Ziyang Song, Mingyue Zheng, Dan Teng, and Qian Shi. P2dflow: A protein ensemble generative model with se(3) flow matching. *Journal of Chemical Theory and Computation*, 21(6):3288–3296, 2025.

[23] J.A. Joseph, A. Reinhardt, A. Aguirre, P.Y. Chew, K.O. Russell, J.R. Espinosa, A. Garaizar, and R. Collepardo-Guevara. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci.*, 1(11):732–743, 2021.

[24] John Jumper, Richard Evans, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

[25] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761), 2025. eadv9817.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, William Lu, Nikita Smetanin, Robert Verkuil, Ousen Kabeli, Yaniv Shmueli, Allan Santos Costa, Mahyar Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[27] Jeffrey M. Lotthammer, Jorge Hernández-García, Daniel Griffith, Dolf Weijers, Alex S. Holehouse, and Ryan J. Emenecker. Metapredict enables accurate disorder prediction across the tree of life. *bioRxiv*, 2024.

[28] J.M. Lotthammer, G.M. Ginell, D. Griffith, R.J. Emenecker, and A.S. Holehouse. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat Methods*, 21(3):465–476, 2024.

[29] Debora S. Marks, Lucy J. Colwell, Ruth Sheridan, and et al. Protein 3d structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12):e28766, 2011.

[30] Parisa Mollaei, Danush Sadasivam, Chakradhar Guntuboina, and Amir Barati Farimani. Idpbert: Predicting properties of intrinsically disordered proteins using large language models. *The Journal of Physical Chemistry B*, 128(49):12030–12037, 2024.

[31] Faruck Morcos, Andrea Pagnani, Bryan Lunt, and et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, 108(49):E1293–E1301, 2011.

[32] B. Novak, J.M. Lotthammer, R.J. Emenecker, and A.S. Holehouse. Accurate predictions of conformational ensembles of disordered proteins with starling. *bioRxiv 2025.02.14.638373*, 2025.

[33] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, and et al. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.

[34] Alexander Rives, Joshua Meier, Tom Sercu, and et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15):e2016239118, 2021.

[35] Vladislav Schnapka, Tatiana Morozova, Subhadip Sen, and Massimiliano Bonomi. Atomic resolution ensembles of intrinsically disordered and multi-domain proteins with alphafold. *bioRxiv*, 2025. Version 2.

[36] M. Steinegger and J. Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35:1026–1028, 2017.

[37] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. arXiv:2104.09864, 2021.

[38] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 03 2007.

[39] G. Tesei, A.I. Trolle, N. Jonsson, J. Betz, F.E. Knudsen, F. Pesce, K.E. Johansson, and K. Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, 2024.

[40] Giulio Tesei and Kresten Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Research Europe*, 2:94, 2023.

[41] The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025.

[42] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in biochemical sciences*, 33(1):2–8, 2008.

[43] Ruben van der Lee, Marija Buljan, Benjamin Lang, and et al. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, 2014.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[45] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, and et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[47] Peter E. Wright and H. Jane Dyson. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6:197–208, 2005.

[48] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, and et al. On layer normalization in the transformer architecture. arXiv:2002.04745, 2020.

[49] Junjie Zhu, Zhengxin Li, Bo Zhang, Zhuoqi Zheng, Bozitao Zhong, Jie Bai, Xiaokun Hong, Taifeng Wang, Ting Wei, Jianyi Yang, and Hai-Feng Chen. Precise generation of conformational ensembles for intrinsically disordered proteins via fine-tuned diffusion models. *bioRxiv*, 2024.

A Additional details

A.1 Training

A.1.1 GeoGraph and GeoHead models

All models presented in this paper are built with a backbone and a GeoHead: the backbone can either be a GeoGraph transformer encoder or a pretrained ESM2 model. The GeoHead is a regressor that takes as input the hidden representation from the backbone, and processes it with two fully connected layers with a hidden dimension equal to half the dimension of the backbone output (128 for GeoGraph, 160 for ESM2-8M and 320 for ESM2-150M). The GeoHead is trained either with the backbone (GeoGraph, Geo models), or while the backbone is frozen (ESM-2 derived models, Graph + GeoHead model). Joint backbone-GeoHead training is performed with a batch size of 512 and a learning rate of 5e-4, while GeoHead-only training uses the same batch size with a learning rate of 3e-3.

A.1.2 Finetuning ESM Models

We fine-tuned ESM-2 models on the IDP-Euka-90 dataset, using a 1% randomly sampled subset for validation. Fine-tuning was performed on the masked language modeling (MLM) task using four H100 GPUs. We employed a learning rate of 4e-4, consistent with the original ESM pretraining setup. For ESM2-8M, we used a batch size of 700, and for ESM2-150M, a batch size of 96 with 10 gradient accumulation steps. Models were trained for a single epoch to preserve the representations learned during pretraining and avoid overfitting to the downstream dataset.

A.2 IDP-Euka-90 dataset curation

As suggested in the Metapredict V3 paper [27], eukaryotes have significantly more disordered regions than bacteria and euryarchaeota: we hence decided to focus on eukaryotes to extract IDRs. We downloaded all 2764 eukaryota proteomes from UniProt and ran Metapredict V3 command metapredict-predict-idrs [13] with default disorder threshold of 0.5 on each one of them. We removed sequences shorter than 30 amino acids and clustered the dataset with mmseqs2 linclust command, with minimum sequence identity threshold of 0.9, 0.8 coverage in coverage mode 1. This pipeline produced an IDP dataset consisting of 30,337,340 sequences.

B Geometric feature calculation

We explain here how all geometric features are calculated for a 3D protein structure containing N residues with Cartesian coordinates $\{r_i\}_{i=1}^N$, indexed according to the residue's position in the protein sequence. The features (R_e, R_g, Δ) are computed separately for each conformation then averaged over the ensemble, whereas the Flory scaling parameters (ν, A_0) are fit using the full ensemble (details given below).

As in [12], we calculate the radius of gyration and asphericity features using the mass-weighted Gyration tensor, $\mathbf{T} \in \mathbb{R}^{3\times3}$, computed as

$$T_{\alpha\beta} = \frac{1}{M}\sum_{i=1}^{N} m_i \tilde{r}_{i\alpha}\tilde{r}_{i\beta} \tag{1}$$

where $m_i \in \mathbb{R}$ is the mass of residue i, and $\tilde{\mathbf{r}}_i \in \mathbb{R}^3$ are its coordinates after subtracting the center of mass. We denote with $\{\lambda_j\}_{j=1}^3$ the eigenvalues of the gyration tensor $\mathbf{T}$,

End-to-end distance (R_e) The distance between the first and last residue in the sequence:

$$R_e = |r_1 - r_N| \tag{2}$$

Radius of gyration (R_g) A geometric property that describes how the protein's mass is distributed about its center of mass. Equivalent to the root mean square distance of all atoms from the protein's

center of mass, and calculated using $\mathbf{T}$ as

$$R_g = \sqrt{\text{tr}(\mathbf{T})} \tag{3}$$

Asphericity (Δ) Characterises the degree to which a protein's three-dimensional shape deviates from a perfect sphere. Calculated using $\mathbf{T}$ as

$$\Delta = \frac{3}{2} \frac{\sum_{j=1}^{3} (\lambda_j - \bar{\lambda})^2}{(\text{tr}(\mathbf{T}))^2} \tag{4}$$

Flory scaling exponent and prefactor (ν, A_0) Parametrise the power-scaling-law relationship describing how the Cartesian distance between residues scales as a function of their spacing in sequence. Following the implementation used by [39], we fit this relationship to residues spaced at least 5 amino acids apart:

$$|r_i - r_j| = A_0 |i - j|^{\nu} \ ; \qquad |i - j| > 5 \tag{5}$$

Unlike the other geometric features which are calculated for each conformation separately and then averaged, the Flory scaling parameters are calculated by first averaging the inter-residue distances observed for each spacing across the whole ensemble, then using the `optimize.curve_fit` function provided by SciPy [45] to fit the (ν, A_0) parameters.

C Comparisons with IDP models

C.1 STARLING

C.1.1 Retraining

Following the preprocessing in STARLING [32], we first downsampled the frames for each sequence in our Human–IDRome dataset to reduce the correlation between conformers. We found that keeping every 20th frame was sufficient to stabilise model training, resulting in 50 conformers for each sequence. After downsampling, we used the same hyperparameters and methodology as in [32] to sequentially retrain the STARLING VAE and DDPM models from scratch using our train and validation splits.

C.2 ALBATROSS

C.2.1 Model versions

In our evaluation of the pretrained ALBATROSS models, we use the default (V2) models available via the SPARROW GitHub repository (https://github.com/idptools/sparrow). For predicting R_g and R_e with ALBATROSS, we used the "scaled" versions of these models as recommended.

C.2.2 Retraining

We use the same model architecture hyperparameters (number of hidden layers, hidden size) for each feature as used in the published V2 models. We found that we could improve training stability and performance by replacing the loss function used by [28] with the mean of the L1 loss over a batch rather than the sum, and performing a grid search over batch sizes {64, 128, 256} and learning rates {1e-3, 5e-3, 1e-2} for each model. We report the best test R^2 score achieved over the grid search for each feature in Table 1, and the hyperparameters used for each model in Table 3.

C.2.3 R^2 score calculation

The R^2 scores attained in our evaluation of ALBATROSS are notably lower than those reported in the original ALBATROSS work [28]. This discrepancy can be partly explained by a difference in the definition of R^2 used between our work and theirs. In [28], the authors define the R^2 score as the square of the Pearson correlation coefficient between the true and predicted values, whereas here we define R^2 as the coefficient of determination.

For targets and predictions $\{(y_i, f_i)\}_{i=1}^{N}$ with target mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, we calculate the coefficient of determination (R^2) as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{6}$$

which, in general, is lower than the square of the Pearson correlation coefficient - and can even be negative.

C.2.4 Flory scaling parameters

We compute the Flory scaling parameters by fitting a power-law relationship between the Euclidean distance of residue pairs and their sequence separation. Following the methodology of the Human-IDRome paper [39], we exclude residue pairs with a sequence separation of less than five, as these short-range interactions are governed by local chain stiffness and deviate from the global scaling law. This approach differs from that used by ALBATROSS [28], which does not place a minimum distance constraint on residues included in its calculation.

Predictor	Number of layers	Hidden size	Learning rate	Batch size	# Parameters
R_e	1	55	1e-2	128	34K
R_g	1	55	5e-3	128	34K
Δ	2	55	1e-2	128	107K
ν	2	35	5e-3	64	46K
A_0	1	70	1e-3	128	52K

Table 3: Hyperparameters used for the ALBATROSS (retrained) models.