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

Eoin Quinn
InstaDeep
London, UK
e.quinn@instadeep.com

Marco Carobene
InstaDeep
Berlin, Germany
m.carobene@instadeep.com

Jean Quentin
InstaDeep
Paris, France
j.quentin@instadeep.com

Sebastien Boyer
InstaDeep
Paris, France
s.boyer@instadeep.com

Miguel Arbesú
InstaDeep
Berlin, Germany
m.arbesu@instadeep.com

Oliver Bent
InstaDeep
Paris, France
o.bent@instadeep.com

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 [33]. 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 [27, 21]. On the other, Multiple Sequence Alignments (MSAs), coupled with decades of structure determination experiments [4], underpin deep learning models like AlphaFold [19] and RosettaFold [3], 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 [38, 34]. 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

AI for Science Workshop, NeurIPS 2025.

inferred from long Molecular Dynamics (MD) simulations [26, 8, 14]. 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, they exhibit poor sequence conservation, a characteristic thought to derive from the lack of a stable structure required to maintain function [7].

A recent line of work aims to use deep generative models, especially diffusion models, to map sequence directly to a full conformational distribution [20, 17, 15, 40]. 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 metastable states, rather than the noisy internal motions within them [8, 14].

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 [6, 2]. The power of this approach has recently been demonstrated by WARIO [11], 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.

2 Related work

Our work builds on several research threads at the intersection of machine learning and protein science. One major line of work uses deep generative models to sample full conformational ensembles from sequence, for both general proteins [20, 17**?**] and specifically for IDPs [25, 16, 15, 40]. While powerful, these methods can be computationally expensive, and general-purpose models often rely on co-evolutionary signals from MSAs that are absent in IDPs. An alternative approach, more aligned with our own, is to predict ensemble-averaged aggregate properties directly from sequence. For example, ALBATROSS [23] predicts five global geometric properties of IDPs, while IDP-BERT [24] fine-tunes a protein language model for similar tasks. We extend this strategy by introducing GeoGraph, a model that learns a richer representation by predicting a diverse set of residue-level geometric and graph-theoretic descriptors. This method is inspired by the long history of using residue contact networks to analyze protein structure and stability [6, 2] and is complementary to the recent method WARIO [11] which uses contact maps for post-hoc characterization of individual IDP ensembles.

3 GeoGraph

Our goal is to learn residue-level representations of IDPs from molecular dynamics (MD) simulations to capture essential physical principles missed by protein language models (PLMs) and methods trained on static, folded proteins.

Generating the vast simulation data required for deep learning is computationally prohibitive with high-fidelity all-atom force fields. We therefore use CALVADOS-2 [32], a state-of-the-art one-bead-per-residue coarse-grained force field designed for IDPs, and experimentally validated on SAXS

and FRET measurements. While this approach sacrifices fine-grained detail, its design is based on an effective description of non-bonded interactions, which enables it to excel at capturing transient residue-residue contact patterns.

We hypothesise that these transient contacts encode rich physicochemical information, which we formalize by analyzing their aggregate properties. For each conformation, we construct a residue-contact graph (8Å cutoff), compute a diverse set of node- and graph-level features, and average these across the full ensemble. This yields a stable statistical fingerprint of the protein's dynamic structure which serves as a direct prediction target.

Our model, GeoGraph, employs a sequence-to-sequence architecture with a 4-layer transformer encoder backbone ($\sim$ 2M parameters) that maps an amino acid sequence to residue-level embeddings. These embeddings are fed as input to separate shallow MLP heads to predict properties at both the sequence- and residue-level. For full details see Appendix A.

We consider two flavours of descriptors, which we refer to as *geometric* and *graph*-based. The geometric descriptors are commonly-employed sequence-level measures of IDP conformational ensembles: end-to-end distance (R_e), radius of gyration (R_g), asphericity (Δ), and the Flory scaling exponent (ν) and prefactor (A_0). Both R_g and R_e can be experimentally determined, and in turn used to determine the Flory prefactor and exponent [1]. 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 [13].

For the graph-based descriptors we consider both sequence- and residue-level features, which capture diverse properties such as network compactness (global efficiency), mixing patterns (assortativity), and residue importance (degree and betweenness centrality), providing a rich, physics-informed learning signal. For full details see Appendix A.2.

We consider multiple variants of GeoGraph so as to clearly dissect its behaviour. Our main model is:

- **GeoGraph:** the full architecture described above, containing the transformer backbone with both sequence-level and residue-level prediction heads, and trained end-to-end to predict the full suite of geometric and graph-based features.

We complement this with baseline variants as follows:

- **Geo:** a baseline model trained to predict only the sequence-level geometric features, i.e. those used as benchmarks. I.e. the prediction of the graph-based features is omitted from the training. 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 the Geo model where the transformer backbone has zero layers. This tests the performance of contextless token embeddings, and provides a naive minimal performance floor.
- **Graph:** a variant of GeoGraph designed to assess the transferability of the learned embeddings. Trained in two stages: first, the full model is trained end-to-end to predict only the graph-based features; second, the backbone weights are frozen, and a new sequence-level prediction head (GeoHead) is trained to predict the geometric features from the learned embeddings.

We train and evaluate our models on the Human–IDRome dataset [31], containing simulated conformational ensembles for 28,058 intrinsically disordered regions from the human proteome. This is the largest publicly available dataset of its kind, which makes it ideal for benchmarking differing approaches. 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 [31]. We partition the dataset using a 80/10/10 split based on sequence similarity, see Appendix A.3.

4 Evaluation

To benchmark performance we evaluate models on their ability to predict the five geometric features $(R_e, R_g, \Delta, \nu, A_0)$, which are well-studied, experimentally relevant measures of IDP conformational ensembles. Results are presented in Table 1, and ablation on the GeoGraph model is provided in

	R_e	R_g	Δ	ν	A_0
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 $\rightarrow$ GeoHead	0.992 (1)	**0.996** (0)	0.864 (13)	0.854 (15)	0.793 (32)
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
ESM2-8M	0.983 (1)	0.991 (1)	0.754 (8)	0.684 (8)	0.523 (19)
IDP-ESM2-8M	0.982 (1)	0.987 (1)	0.783 (2)	0.767 (5)	0.643 (14)
ESM2-150M	0.984 (1)	0.991 (1)	0.792 (2)	0.763 (4)	0.637 (5)
IDP-ESM2-150M	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 may be affected by differences in computation of the scaling parameters between our work and theirs (see Appendix D.2.4).

Table 2. We highlight that while scores for R_e and R_g are high across the board, the true test is the performance on the more complex shape descriptors (Δ, ν, A_0), on which GeoGraph excels.

We compare against two leading IDP methods: STARLING [25], a generative diffusion model, and ALBATROSS [23], an RNN-based direct predictor. Since the original models were trained on data generated with a different force field, we retrained them on our dataset for a fair comparison. See Appendix D for further details.

We also compare the sequence-to-sequence backbone against Protein Language Model (PLM) embeddings. We use ESM-2 [21] and test the model in two settings. Firstly, we used the general-purpose pre-trained embeddings of the 8M and 150M models. Secondly, we curated a dataset of 30 million IDP sequences, which we refer to as *IDP-Euka-90*, and used this to fine-tune two corresponding versions of ESM-2, *IDP-ESM2-8M* and *IDP-ESM2-150M*, see Appendix B for further details. In both cases, we freeze the backbone model and train a sequence-level prediction head for predicting the geometric features – as we did for the Graph model above.

Finally, we attempted to evaluate BioEmu, a large-scale general-purpose ensemble emulator [20] 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 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 [28], and may be explained by the poor sequence conservation of IDPs.

	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. For each task the best performing results (within error) are in bold.

5 Discussion

As shown in Table 1, GeoGraph achieves highly competitive performance against leading methods for IDP ensemble property prediction, in particular on the more complex shape descriptors (Δ, ν, A_0). Critically, our model predicts these descriptors several orders of magnitude faster than it takes to run the CALVADOS-2 simulator that it emulates: GeoGraph can process the entire test set of 2,388 sequences in approximately 1 second on a single GPU (H100), whereas simulation of these ensembles on Google Colab takes on the order of 10 days [31].

The source of GeoGraph's strong performance can be seen by comparing our model variants: the Geo model, which trains on geometric features alone, and the Graph model, which learns representations solely from graph topology (and is evaluated on the GeoHead trained on these). 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 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, the ablation in Table 2 reveals that the context-aware, residue-level graph features are the primary drivers of this learning.

When compared to the generative baseline of STARLING, our direct-prediction approach appears more robust for capturing complex shape descriptors. This suggests that inferring these aggregate properties from a generated ensemble can be a less effective approach. The improved performance over ALBATROSS, an analogous direct predictor, can be attributed to our model's larger capacity and richer, residue-level feature set.

For the comparison against PLM embeddings, the most direct reference point is with our Graph $\rightarrow$ GeoHead model. We see that our simulation-informed embeddings provide a significantly stronger predictive signal for geometric properties, even after fine-tuning ESM-2 on a large corpus of IDP sequences. While a superior performance may be expected when the training objective aligns with the evaluation task, the magnitude of the difference underscores a key limitation of protein language models when applied to IDPs: their reliance on evolutionary patterns, which serve as a noisy and incomplete proxy for the physical properties that govern dynamic ensembles [7].

We observe that the IDP fine-tuning of ESM-2 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.

While the results of our evaluation are highly encouraging, this exploratory study has several key limitations. GeoGraph is fundamentally an emulator of the CALVADOS-2 coarse-grained simulation, inheriting its lack of all-atom detail. Furthermore, our contact map featurization is simplistic, and by predicting only the mean of each descriptor, we lose valuable information about the ensemble's heterogeneity. Future work could directly address these points by training on all-atom data and enriching the prediction targets to include higher-order statistics like variance. A next iteration of the GeoGraph approach could learn an optimal graph construction and its most relevant features directly from the data.

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.

References

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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 GeoGraph

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 [35], chosen for its ability to capture long-range dependencies and produce context-rich embeddings. We build on the Hugging Face implementation of ESM-2 [21, 37], which uses Pre-Layer Normalization (Pre-LN) [39] and Rotary Position Embeddings (RoPE) [30].

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. We obtain a single sequence-level embedding by taking the mean of these residue-level embeddings, a simple yet robust method for creating a global representation.

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 Adam optimizer and a cosine learning rate scheduler with warmup, with peak a learning rate of 5e-4, and a batch size of 512.

For the Graph model, where we train the prediction head for the geometric features in a second stage on a frozen backbone, we used the same batch size with a peak learning rate of 3e-3.

A.2 Features

We consider two flavours of descriptors, which we refer to as geometric and graph-based. The geometric descriptors all sequence-level features, while for the graph-based descriptors we consider both sequence- and residue-level features. Graph features are computed using python's NetworkX package (default settings) [12], and for training our models we standardise all target features to have zero mean and unit variance.

Geometric, sequence-level: We consider commonly employed measures of IDP conformational ensembles of computable from MD simulation frames: **end-to-end distance** (R_e), **radius of gyration** (R_g), **asphericity** (Δ), and the Flory scaling **exponent** (ν) and **prefactor** (A_0).

Graph, 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.

Graph, residue-level: Here we included **degree centrality** (local contact density), **betweenness centrality** (bridging propensity), **harmonic centrality** (inverse-distance reachability), **PageRank** [5], **core number**, **local clustering coefficient**, and as well as an **in-largest-connected-component** indicator.

A.3 Human–IDRome dataset

We partitioned the Human–IDRome dataset [31] 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 [29] with parameters (min_seq_id=0.7, coverage=0.8, cov_mode=1), identical to the parameters used by STARLING [25]. Additionally, we filtered the dataset to sequences with a maximum length of 256 residues.

B IDP-ESM

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, which we call *IDP-Euka-90*. As suggested in the Metapredict V3 paper [22], 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 [10] 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.

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.

C 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 [9], we calculate the radius of gyration and asphericity features using the mass-weighted Gyration tensor, $\mathbf{T} \in \mathbb{R}^{3\times 3}$, 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 Euclidean 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. It can be 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 Euclidean distance between residues scales as a function of their spacing in sequence. Following the implementation used by [31], 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 [36] to fit the (ν, A_0) parameters.

D Comparisons with existing IDP models

We evaluate two prominent methods for IDP property prediction: ALBATROSS [23] and STARLING [25]. 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 [25] 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 [18] 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.

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.

D.1 STARLING

Following the preprocessing in STARLING [25], 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 [25] to sequentially retrain the STARLING VAE and DDPM models from scratch using our train and validation splits.

D.2 ALBATROSS

D.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.

D.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 [23] 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.

D.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 [23]. This discrepancy can be partly explained by a difference in the definition of R^2 used between our work and theirs. In [23], 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.

D.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 [31], 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, which may differ from that used by ALBATROSS [23].