Energy-Guided Conformational Ensemble Generation for Intrinsically Disordered Proteins

Md Wasi Ul Kabir and Md Tamjidul Hoque

Department of Computer Science, University of New Orleans

mkabir3@uno.edu thoque@uno.edu

Intrinsically Disordered Proteins

- Intrinsically Disordered Proteins (IDPs) represent **30-40% of the human proteome** and are involved in critical biological processes.
- Unlike traditional proteins with fixed 3D structures, IDPs exist as dynamic conformational ensembles—constantly shifting between multiple shapes.
- **60-80% of cancer-related proteins** contain disordered regions
- IDPs are key players in neurodegenerative diseases (Alzheimer's, Parkinson's)
- Traditional structure prediction methods (AlphaFold, RosettaFold) **fail for IDPs**
- Experimental characterization is costly and time-consuming

Problem Statement & Goal

- **Problem:** Existing predictors return a single structure; IDPs require **ensembles**.
- **Goal:** Generate **diverse, native-like** conformations consistent with experiments.
- **Approach:** Sample broadly (BioEmu) + select physically realistic states (IDPEnergy).

Proposed Solution

- Combine the generative modeling with physics-based energy scoring to intelligently filter conformational ensembles.
- Previous methods generate thousands of structures without quality control
- Our approach uses physics to guide selection toward native-like conformations
- Result: Higher quality ensembles with fewer computational resources

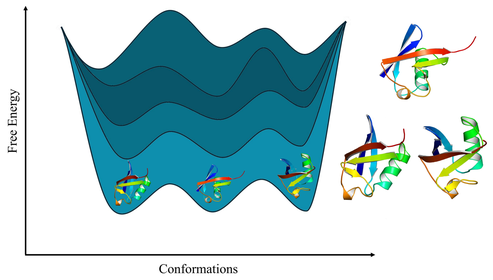

Figure 1. IDP free energy landscape showing multiple minima and diverse conformational states. IDPs lack stable 3D structure and exist as dynamic ensembles.

Dataset

Training Data (IDPEnergy):

- **Source**: Protein Data Bank (PDB)
- **Initial pool**: 8,705 NMR-only structures
- After filtering (<30% seq. identity): 2,806 proteins
- **Independent test sets**: 2 × 86 targets

Validation Data (CEG-IDP):

- **Source**: Protein Ensemble Database (PED)
- **50 IDP proteins** with experimental ensembles
- Independent from training data

Statistical Energy Function (IDPEnergy)

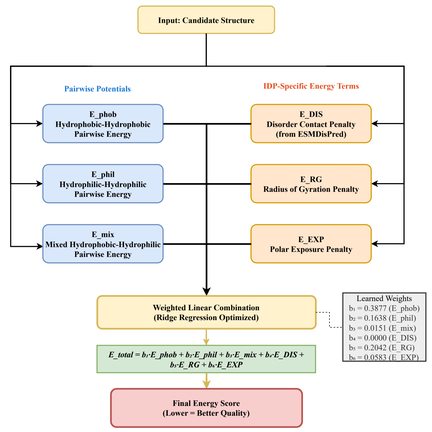

Figure 2. IDPEnergy scoring function architecture combining base pairwise potentials and IDP-specific energy terms.

Total Energy Function:

$$E_{\text{total}} = b_1 \cdot E_{\text{phob}} + b_2 \cdot E_{\text{phil}} + b_3 \cdot E_{\text{mix}} + b_4 \cdot E_{\text{DIS}} + b_5 \cdot E_{\text{RG}} + b_6 \cdot E_{\text{EXP}}$$

Lower energy indicates more physically realistic conformations; coefficients b_i are learned to maximize native–decoy separation on NMR ensembles.

Ridge Regression Optimization:

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \left(\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right)$$

Weights optimized on 2,806 NMR structures to minimize native vs. decoy energy differences.

Conformational Ensemble Generator (CEG-IDP)

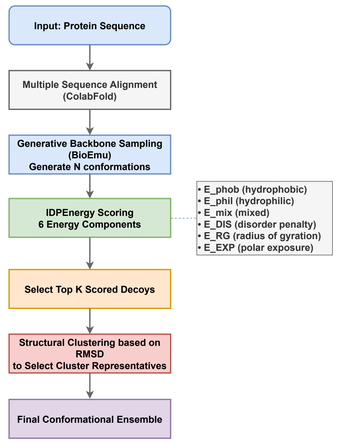

Figure 3. CEG-IDP pipeline: MSA generation, generative sampling, energy scoring, and clustering.

Validation Metrics (vs. PED):

- **EMD**$_{Rg}$: Earth Mover's Distance between R_g distributions (captures *global compaction/expansion*).
- **JS**$_{P(r)}$: Jensen–Shannon divergence between pairwise distance histograms $P(r)$ (captures *overall shape / distance profile*).
- **L1**$_{\text{contact}}$: Mean absolute difference between contact probability maps (captures *medium/long-range contacts*; contact if Cα distance < 8 Å).
- **Objective** J: Composite score combining the above:

$$J = w_{Rg} \cdot \frac{\text{EMD}_{Rg}}{\sigma_{Rg}} + w_{\text{contact}} \cdot \text{L1}_{\text{contact}} + w_{P(r)} \cdot \text{JS}_{P(r)}$$

Here σ_{Rg} is the interquartile range (IQR) of the PED R_g distribution for each target. Weights are set automatically as $w_{Rg} = 1$ and $w_{\text{contact}} = w_{P(r)} = \sigma_{Rg}$ to balance metric scales relative to experimental ensemble variability.

Results

IDPEnergy Performance

- Independent test sets: 2 × 86 targets
- Top-1 accuracy: **89.5%** and **90.7%**
- Native Z-scores: **-8.18** and **-9.5**

Large negative Z-scores indicate strong native–decoy discrimination.

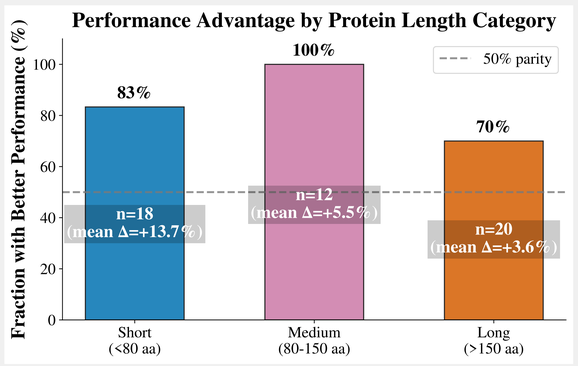

Figure 4. Fraction of proteins with better performance stratified by protein length.

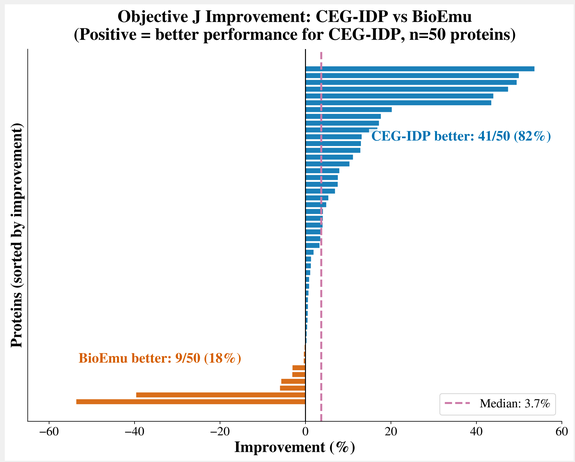

Figure 5. Objective J improvement across 50 PED proteins.

Impact & Applications

- **Cancer Drug Discovery:** Target disordered oncoproteins (p53, c-Myc)
- **Neurodegenerative Diseases:** Model α-synuclein (Parkinson's), tau (Alzheimer's)
- **Protein Engineering:** Design IDPs with specific properties
- **IDP Research:** Understand IDP biology and phase separation

Computational Resources

- Training completed on 1× NVIDIA H100 GPU.
- BioEmu sampling: 3,000 conformations per target.
- Final ensembles reduced to 200 representatives.
- Experiments executed on institutional high-performance computing clusters.

Conclusions

- **Developed CEG-IDP:** an energy-guided ensemble generator that filters generative samples using the IDPEnergy scoring function.
- **Consistent gains on PED:** improved Objective J on **82%** of 50 experimental IDP targets, with strongest gains in global compaction (EMD$_{Rg}$).
- **Efficient ensembles:** achieves higher-quality ensembles with fewer retained conformations after energy-based selection and clustering.
- **Future work:** benchmark against additional IDP ensemble methods (e.g., **IDP-Fold** and **IDPConformer**), and integrate experimental restraints (SAXS/NMR/FRET) for further refinement. Energy-guided ensemble modeling enables structure-aware drug discovery for intrinsically disordered targets.
  Code & Data: Pipeline scripts, scoring functions, and evaluation code will be released upon publication.

Limitations & Future Work

- Extend to **longer IDPs** and multi-domain proteins.
- Add **experimental restraints** (SAXS, NMR, FRET) during selection.
- Benchmark against **IDP-Fold** and **IDPConformer** on shared test sets.
- Speed up scoring via **parallel/GPU** evaluation or a learned energy surrogate.

References

1. Lewis et al. (2025). Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6757).
2. Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583-589.
3. Baek et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871-876.
4. Lin et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123-1130.

Acknowledgments

Funding: Research supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103424-21.

Computing Resources: Louisiana Optical Network Infrastructure (LONI) for high-performance computing access.