TRUSTWORTHY BIO-AI

Bio-Foundation Models Are Not Yet Robust

to Biologically Plausible Perturbations and ML Transformations

A SYSTEMATIC ROBUSTNESS STUDY

11 Bio-FMs · 7 datasets · 2,128 experiments

The question we ask:

Are Bio-FMs robust enough for real-world use?

Which perturbations compromise their reliability?

When do these failures pose deployment risk?

Jinhao Duan*, Ruichen Zhang*, and collaborators

Across 11 Bio-FMs · 7 datasets · sequence / structure / image

Research talk · 10-minute external presentation

Bio-Foundation Models are reshaping biomedicine

From sequence to structure to images — Bio-FMs sit at the heart of high-stakes pipelines.

Protein structure & design

AlphaFold-2/3, ESM-3, ProteinMPNN, ESM-IF1 power drug discovery and antibody design.

Function & fitness prediction

GearNet, ProNet, SaProt, S3F predict enzymatic function, GO terms, mutational fitness.

Cryo-EM reconstruction

CryoDRGN, CryoNeRF turn 2D micrographs into 3D molecular structures.

These models are deployed in real lab pipelines — and increasingly chained together.

- Antibody design: **ProteinMPNN + AlphaFold3 + Rosetta** in one workflow.
- Function annotation, enzyme screening, vaccine design, *therapeutic candidate ranking*.
- But: tiny shifts in inputs — noise, missing residues, format glitches — happen every day in real labs.
- If a Bio-FM is brittle, errors *propagate downstream* and become silent deployment risk.

Are Bio-FMs robust to small-but-real shifts?

Prior work studies content safety; robustness to realistic perturbations is largely underexplored.

WHAT WE ASK

Are Bio-FMs robust enough for real-world use? What perturbations compromise their reliability? When do failures pose risk in deployment?

Prior work focuses on…

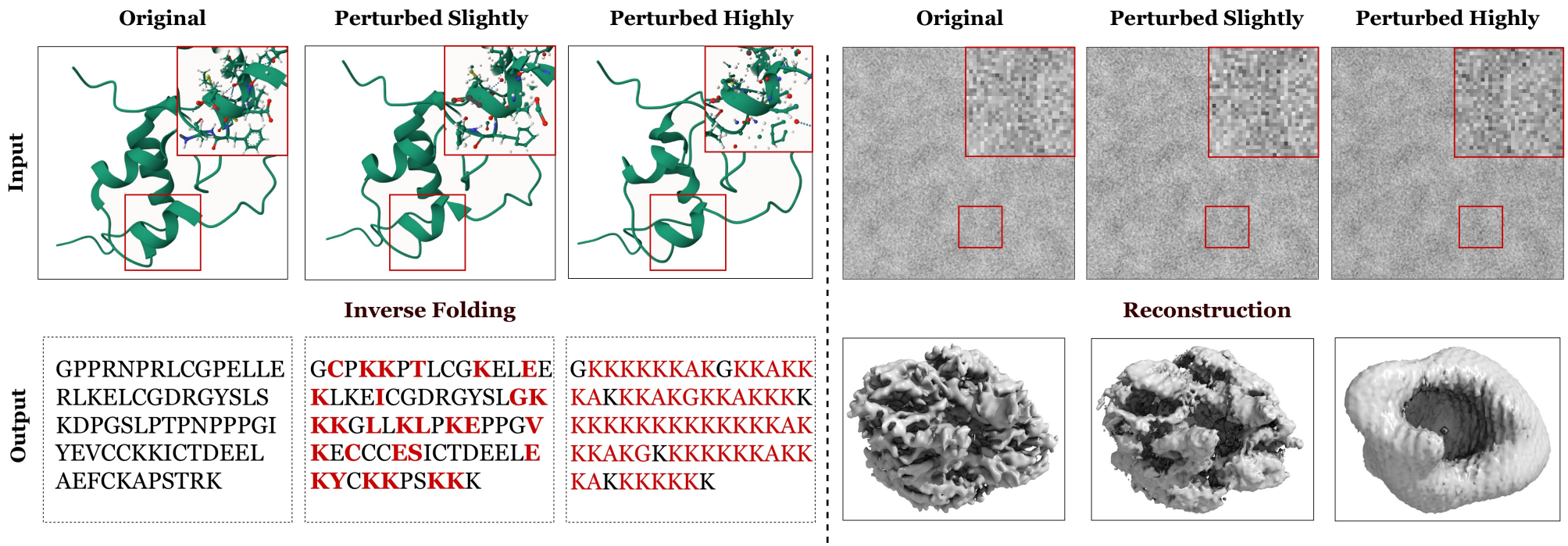
- **Content safety** and benign-usage of FMs in biology
- Adversarial robustness against **synthesized** perturbations
- Task-specific reliability (e.g., AlphaFold folding errors)

Our angle: small-but-real shifts

- Realistic **biological curation noise** (PDB / cryo-EM artifacts)
- Inevitable **ML processing choices** (graph, tokenization)
- How do these perturbations *translate into deployment risk*?

Tiny perturbations → dramatic failures

Subtle, often imperceptible input shifts can collapse Bio-FM outputs across modalities.



Left: protein backbone perturbations + inverse-folding outputs. Right: cryo-EM image corruptions + 3D reconstructions.

Inverse folding output

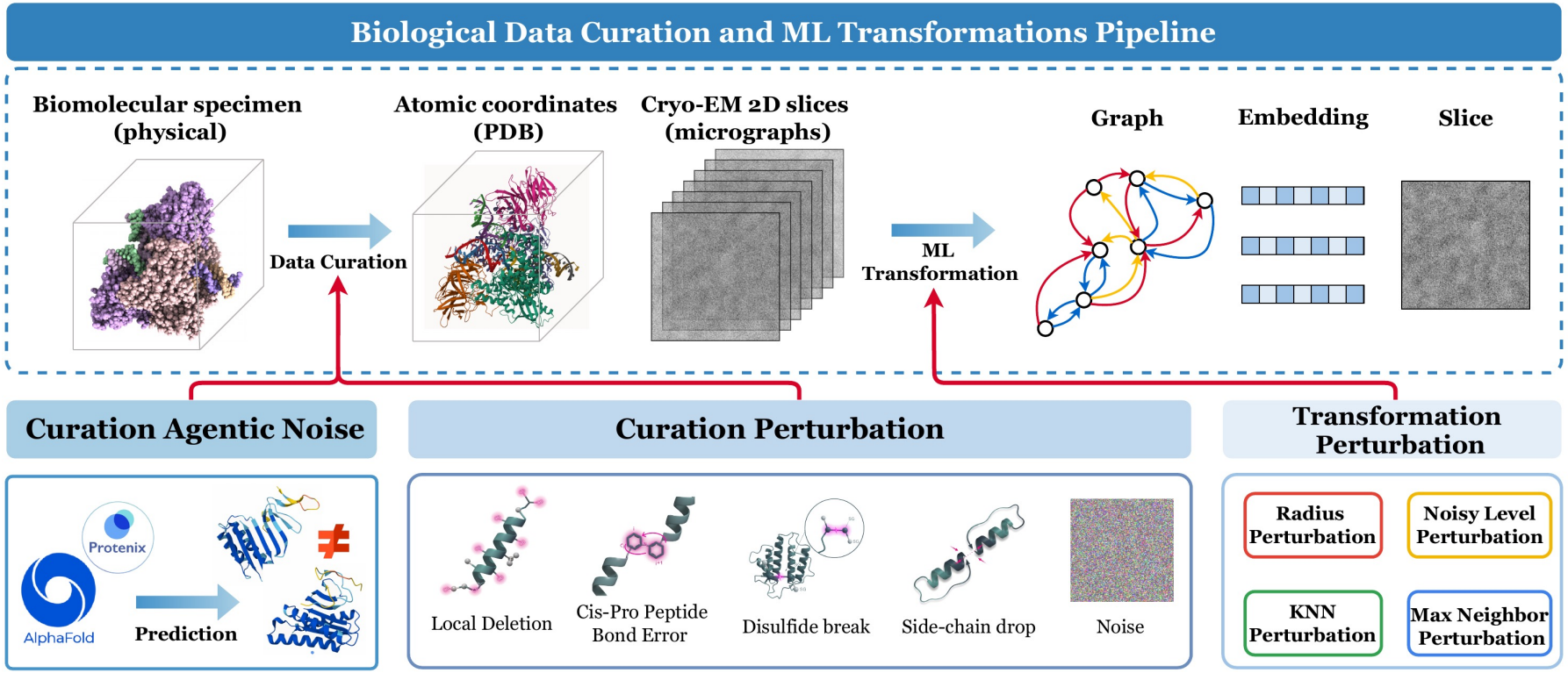
Sequence degenerates into repetitive 'K' residues — fully broken designs.

Cryo-EM reconstruction

3D density loses recognizable secondary structure under heavy noise.

Taxonomy: two complementary perturbation sources

Real-world failures arise both from biology and from ML — separating them gives a complete diagnosis.



Biologically plausible perturbations

Real corruptions during data curation: coordinate noise, residue deletions, sidechain drop, B-factor scrambling, cryo-EM noise.

ML transformations

Inference-time choices inside Bio-FMs: graph radius, k-NN, tokenization, preprocessing, noise level.

Biologically plausible perturbations

11 widely-occurring, unavoidable corruptions grounded in literature and domain expertise.

Protein structure perturbations

Geometric / coordinate-level

- Gaussian coordinate noise (thermal motion)
- Local residue deletions (unresolved loops)
- Sidechain atom drop, disulfide breakage
- Cis-peptide bond, local geometric distortion

Annotation / format-level

- B-factor / occupancy scrambling
- Residue rename / renumber anomalies
- CONECT-record loss, header corruption
- Atom-name / element misalignment

Cryo-EM image perturbations

Noise models

- Gaussian / shot / impulse / speckle noise
- Low signal-to-noise, defocus artifacts

Image-quality & geometry

- Gaussian blur, low contrast
- Rotation, translation, elastic transform

Worst-case

- PGD adversarial perturbations on micrographs

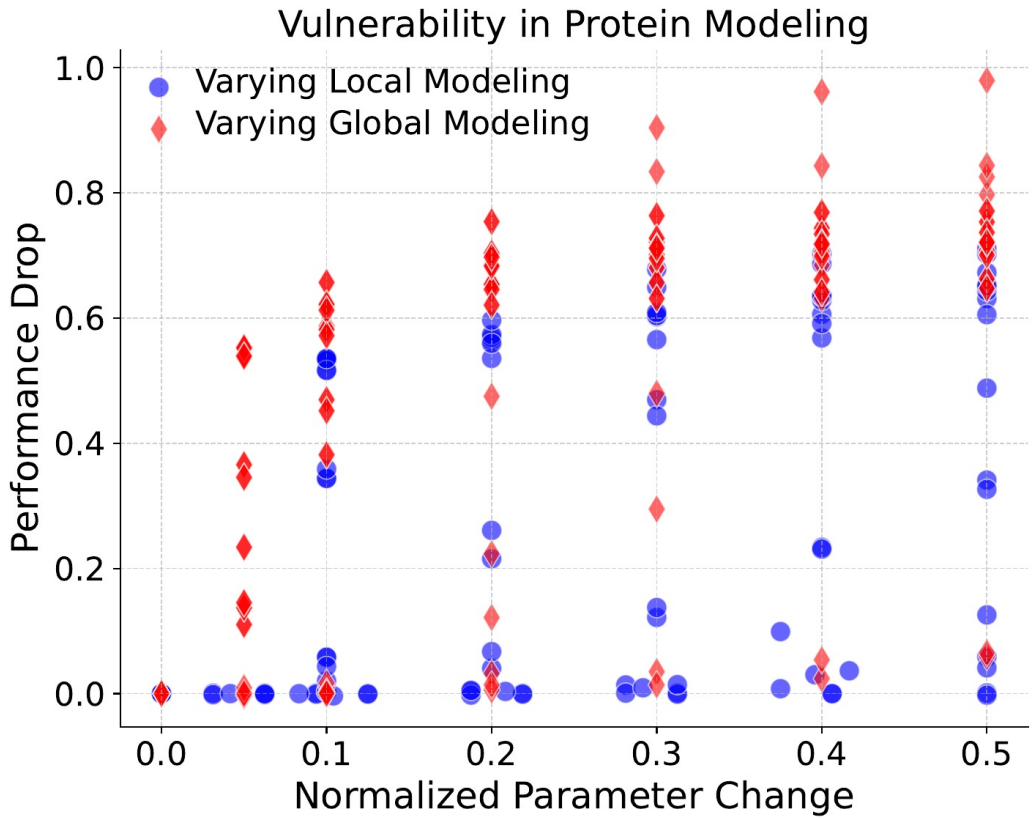
ML transformations: choices inside the model pipeline

Bio-FMs build graphs from coordinates — small hyperparameter choices reshape what the model sees.

Graph-construction hyperparameters we vary

- GearNet / ESM-GearNet: **radius** cutoff, **k-NN**
- ProNet: **cutoff** distance, **max neighbors**
- S3F: **min-distance**, **radius** for edges
- ProteinMPNN: **num-neighbors**, **noise level**
- Quantify perturbation strength with *graph Jaccard / spectral / Frobenius* similarity

Key idea: ML perturbations act as a controllable proxy for biological noise — same graphs, different processing.



Experimental scope: a comprehensive robustness benchmark

Four task families · three modalities · 2,128 evaluations across leading Bio-FMs.

11
state-of-the-art Bio-FMs

7
benchmark datasets

2,128
evaluation experiments

4
task families × 3 modalities

Function / Structure Prediction

Models

GearNet, ESM-GearNet, ESM-1, ProNet

Dataset · Metric

EC · GO · ProtFunc · TAPE
AUPRC / F1 / Accuracy

Sequence Generation

Models

ESM-3, ProteinMPNN, ESM-IF1

Dataset · Metric

PInvBench
Recovery rate

Protein 3D Reconstruction

Models

CryoDRGN, CryoNeRF

Dataset · Metric

EMPIAR-10049 (RAG1–RAG2)
FSC resolution

Protein Fitness Prediction

Models

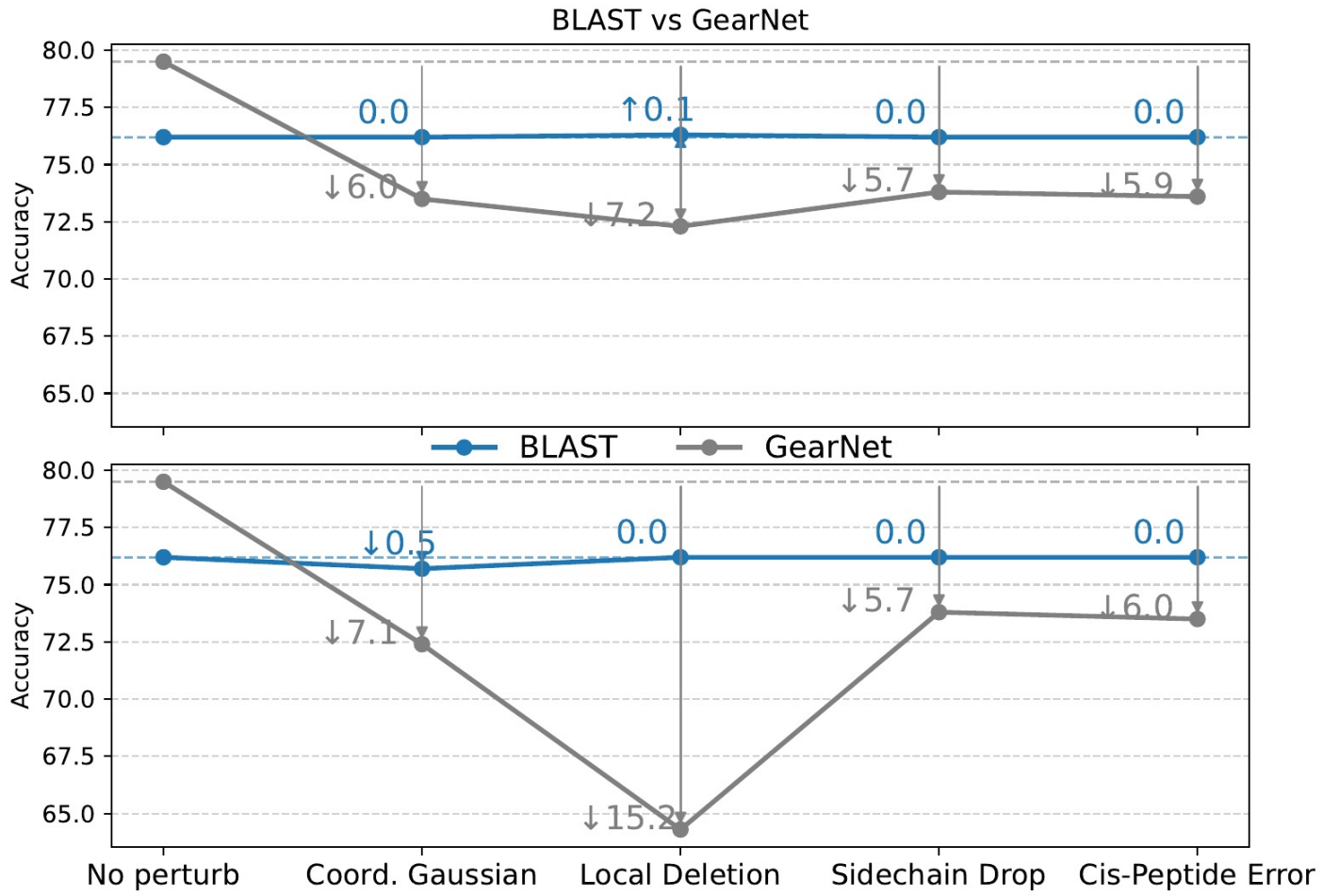
SaProt, ESM-3, ESM-IF1, S3F, ProteinMPNN

Dataset · Metric

ProteinGym (DMS subs/indels)
Spearman / AUC / Recall

Finding 1: Bio-FMs are more brittle than non-FM tools

On enzyme function prediction, BLAST stays flat where GearNet drops up to 15.2 % accuracy.

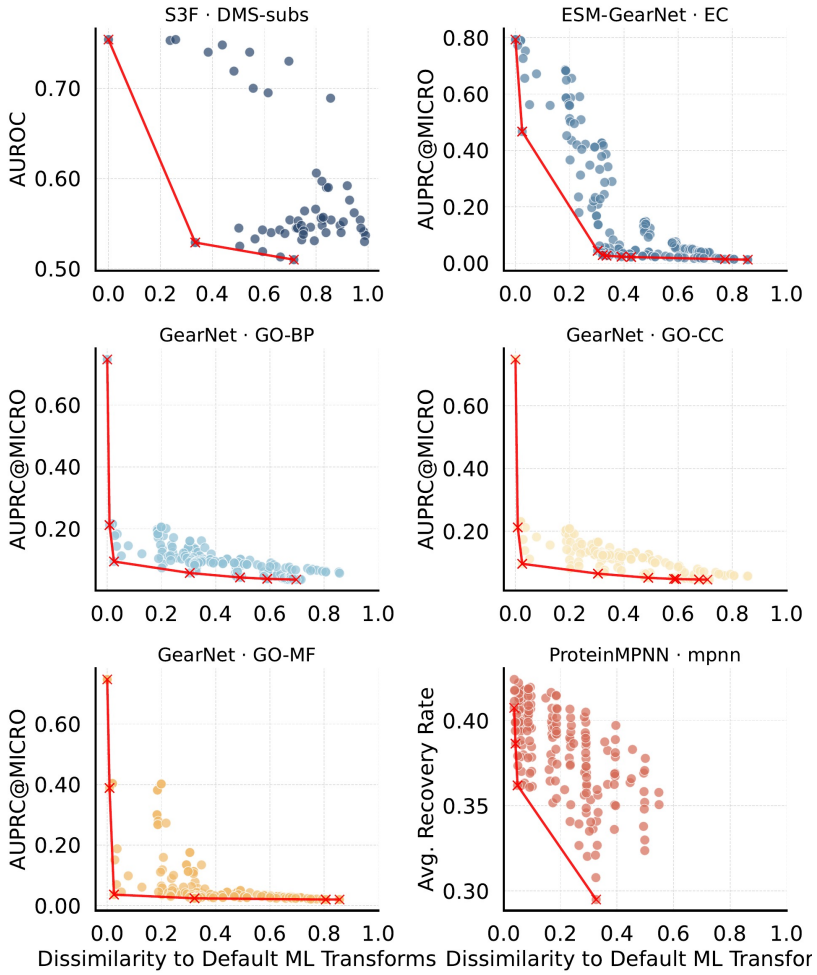


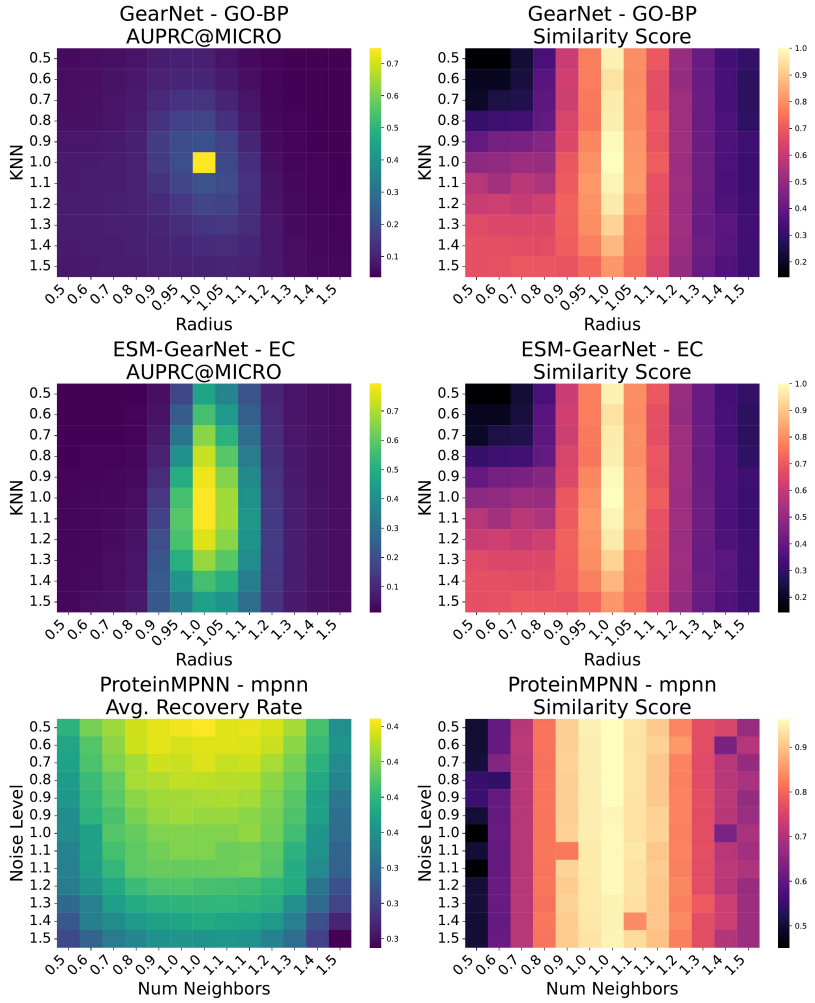
Why this matters

- Bio-FMs are high-capacity & broadly pretrained — *not* explicitly invariant to small biological shifts.
- Subtle shifts push inputs **off the training manifold** and amplify through deep layers.
- BLAST (sequence matching) is mostly **flat** under coordinate noise, deletions, sidechain drop.
- GearNet drops **6–15 %** accuracy on the same perturbations.
- Robustness is **more deployment-critical** for Bio-FMs than for classical tools.

Finding 2: Tiny ML perturbations → catastrophic drops

Worst-case curves (red) collapse within a few percent of graph dissimilarity.



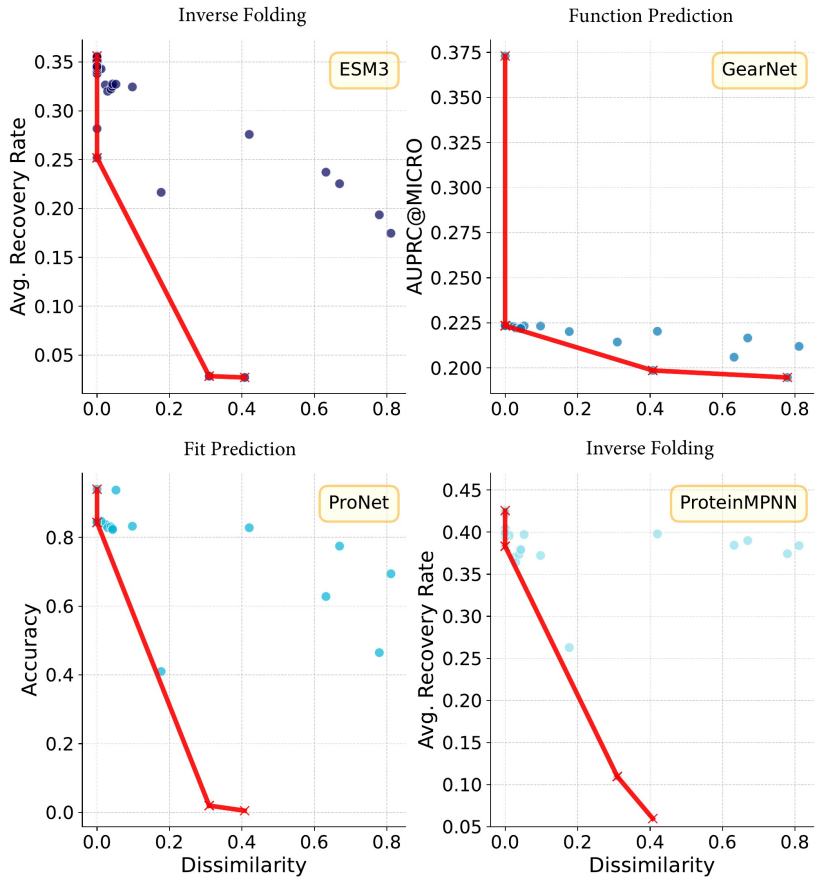


Takeaways

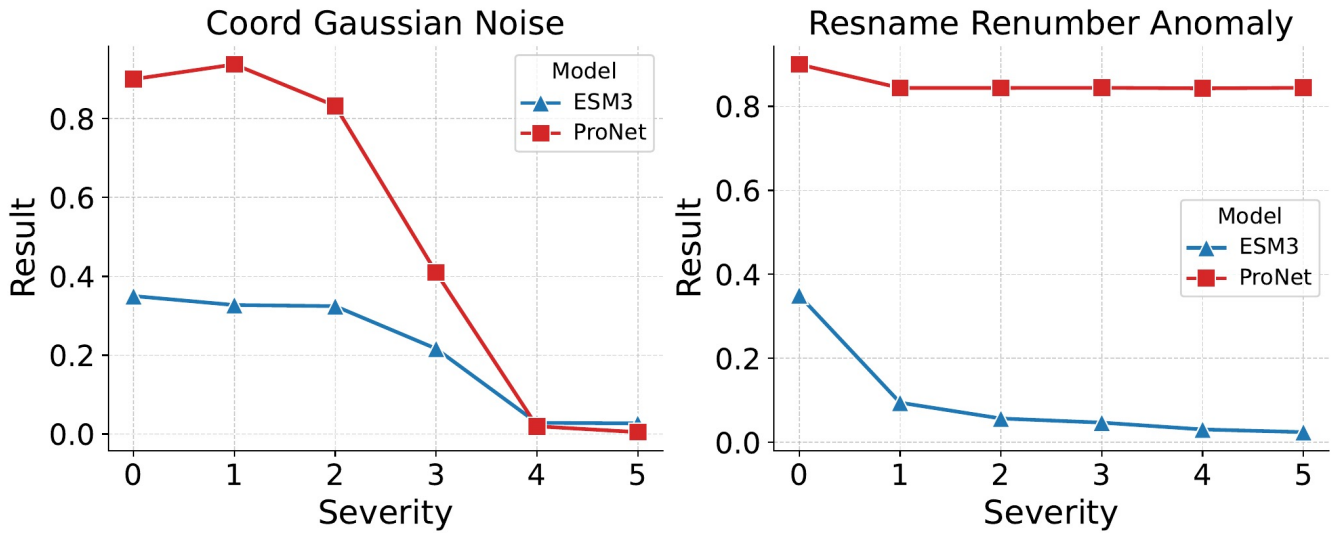
- All Bio-FMs studied have a sharp **robust boundary**.
- GearNet collapses from **0.7 → 0.1** AUPRC@MICRO at *1 %* dissimilarity.
- Graphs remain almost identical — yet model output flips.
- Spatial-radius edges are **more brittle** than k-NN density edges.

Finding 3: Biological perturbations also break Bio-FMs

Different Bio-FMs collapse along different perturbation axes — no single model is uniformly safe.



Right: severity-by-severity behavior of ESM3 vs ProNet.



Coordinate noise

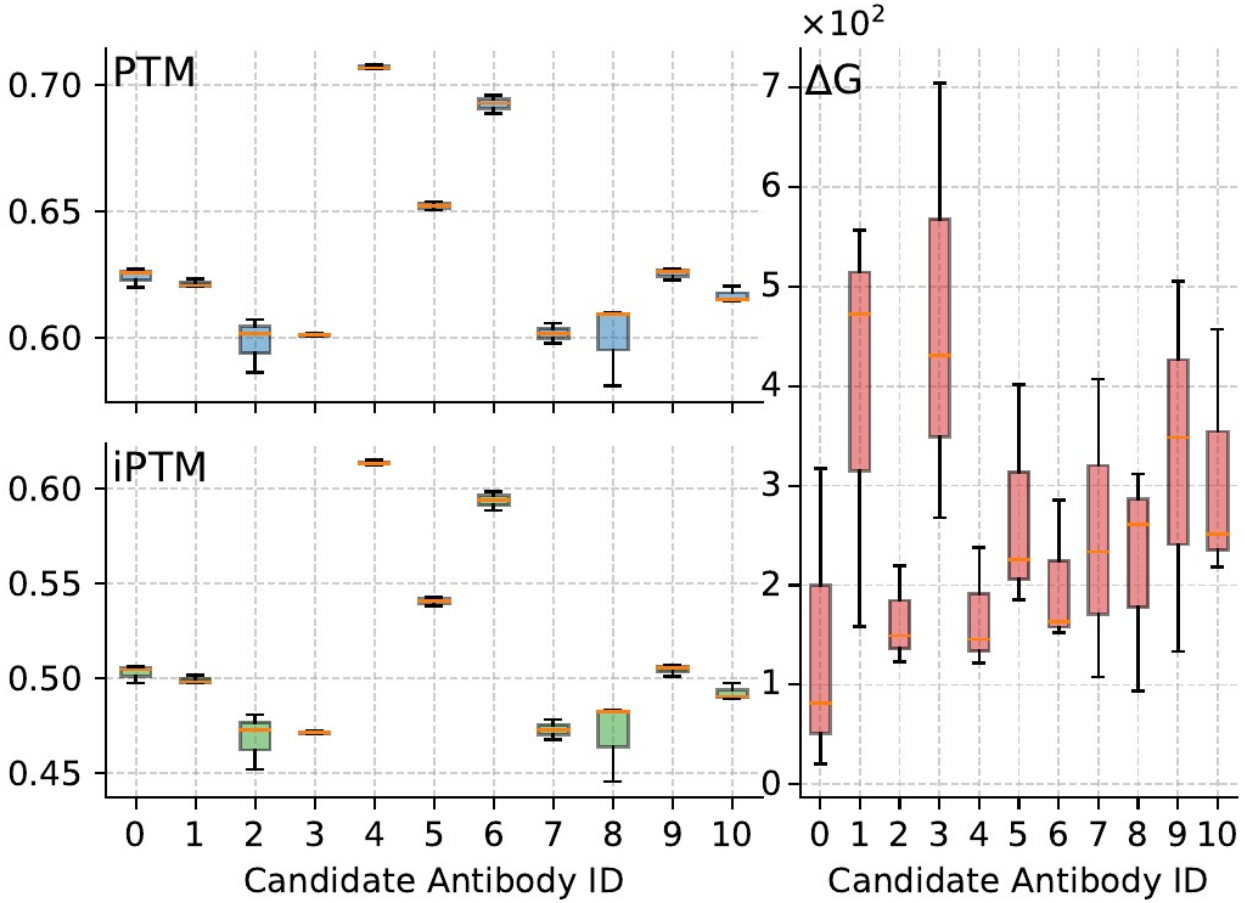
ProNet & ESM3 both collapse past severity 3 — geometry-sensitive models fail when atoms wiggle.

Rename anomaly

ESM3 collapses (sequence-reliant); ProNet stays stable (structure-focused).

Finding 4: Agentic pipelines amplify the risk

A confident upstream model does NOT guarantee a stable downstream outcome.



Antibody design case study

Pipeline: ProteinMPNN → AlphaFold3 → Rosetta

- AlphaFold3 reports **high & consistent** PTM / iPTM scores.
- Yet downstream Rosetta **ΔG variance is huge** (right boxplot).
- Stable Bio-FM confidence ≠ stable downstream objective.
- Implication: **naive trust** in upstream confidence yields brittle, non-reproducible decisions.
- Risk is *latent* — invisible until the pipeline ships.

Finding 5: Cryo-EM reconstruction stays surprisingly robust

CryoDRGN remains stable under realistic blur, rotation, and even worst-case PGD attacks.

Severity	Gauss-Blur CryoDRGN	Rotation CryoDRGN	Translation CryoDRGN	PGD CryoDRGN
1	3.50	3.50	3.50	3.50
3	3.50	3.73	7.20	3.50
5	**8.61**	4.57	**64.66**	3.50

FSC-derived resolution (Å) at the gold-standard FSC = 0.143 — lower is better.

Why is CryoDRGN robust?

- **Information aggregation** across many particle views — noise averages out.
- **Training objectives** rooted in continuous reconstruction, not discrete tokens.
- **Input continuity**: small image shifts don't flip topology like graphs do.
- Stable at PGD severity 1–3 (only 6–12 / 192 particles flip).

Implication

Continuous reconstruction objectives may inspire more robust structure / sequence models — a principled path to bridge the gap.

Takeaways & implications for deployment

What practitioners should change today — and where the field needs to go.

Audit before deploy

Run Bio-FMs through biological + ML perturbation suites. Tiny shifts can flip outputs silently.

Don't trust upstream confidence

Confidence scores (PTM, iPTM) of one Bio-FM don't predict downstream stability in agentic pipelines.

Design for continuity

Continuous / hybrid representations (à la cryo-EM) appear more robust than brittle radius-based graphs.

OUR CONTRIBUTIONS

- First **systematic, comprehensive** robustness study of Bio-FMs across sequence / structure / image.
- A **unified taxonomy** of biologically plausible perturbations and ML transformations.
- **2,128 experiments** revealing pervasive vulnerabilities — and a rare positive result for cryo-EM.
- A **practical auditing lens** for safer end-to-end biological pipelines.

THANK YOU

Questions?

Bio-Foundation Models Are Not Yet Robust to Biologically Plausible Perturbations and ML Transformations

Jinhao Duan · Ruichen Zhang · and collaborators