GeometryShift

Naser Fadil Ali · Nf4d1l.4l1@gmail.com · Discord: Gmajxe

1 · PROBLEM

what and why

Protein machine-learning models increasingly use either sequence information or three-dimensional structure, but it remains unclear when explicit 3D geometry provides a real advantage over strong sequence-based models. Strong benchmark performance may reflect interpolation between similar proteins rather than generalization to structurally or evolutionarily different proteins. GeometryShift asks whether geometric inductive bias improves protein-model generalization under distribution shift, and which forms of geometric information are responsible. The central hypothesis: explicit 3D geometry becomes increasingly valuable as test proteins become more different from the training distribution. A negative result is also useful.

2 · APPROACH

model / method, and why this one

Compare progressively stronger representations on the same biological task: (1) pretrained protein-language-model sequence baseline; (2) residue-level GNN with sequence features and structural connectivity; (3) distance-based geometric model; (4) SE(3)/E(3)-equivariant model on coordinates; (5) sequence + geometry hybrid. Evaluate under progressively harder shifts: standard splits, low sequence similarity, structural/family separation, and combined sequence + structural shift. Geometry-specific ablations include distance-only vs equivariant geometry, removal of sequence features, coordinate perturbations, and neighborhood-size changes. Core outcome: performance retention under distribution shift.

3 · DATASET

source, size, preprocessing

Primary: ATOM3D Protein Interface Prediction (PIP) with 30% sequence-identity split plus independent docking benchmark. Secondary: ATOM3D Mutation Stability Prediction (MSP) from SKEMPI if the primary study is stable. Preprocessing includes structure parsing, residue/chain mapping, graph construction, PLM embeddings, and leakage-audited split versioning.

4 · METRICS

how you’ll know it worked

how you’ll know it worked Primary PIP metrics: AUPRC and AUROC; F1/precision/recall secondary. Scientific measurements: Performance Retention = M_OOD / M_ID and Generalization Gap = M_ID – M_OOD. Also calibration (ECE/Brier), seed variance, parameter count, training/inference time, and peak GPU memory.

5 · COMPUTE

training time, memory, dependencies

Python, PyTorch, PyTorch Geometric, ATOM3D, NumPy/SciPy, scikit-learn, Biopython, and an equivariant library such as e3nn. Pilot subsets (~2 hours), sequence/graph baselines (~1 day each), equivariant/hybrid models (~2–3 days each including ablations), with repeated seeds adding roughly 3–5× on final configs.

6 · DELIVERABLES

model, paper, demo

- · GeometryShift open-source framework: preprocessing, leakage auditing, baselines, shift generation, evaluation.
- · Reproducible sequence- and structure-shifted splits.
- · Trained checkpoints and experiment configs.
- · Systematic ablation of sequence, graph, distances, equivariance, and fusion.
- · Paper documenting methodology, results, trade-offs, and failure cases.

7 · RISK

what breaks first, and plan B

what breaks first, and plan B SE(3) may not outperform sequence-only models — test robustness, not just raw score. Geometry gains may be model-size artifacts — match parameter budgets. Distance features may explain gains — keep distance-only controls. OOD splits may hide homologs or become too small — reconstruct with stricter grouping and report sample-size effects. The project is failure-first: we will not tune until a geometric model wins. The scientific result is a characterization of when 3D geometry does — and does not — provide useful generalization.