

Transfer learning to leverage larger datasets for improved prediction of protein stability changes

汇报人：余彪，周智泓

Background



Introduction



Brian Kuhlman

1992 — B.A. in Chemical Physics, Rice University

1998 — Ph.D. in Chemistry, Stony Brook University
(Advisor: Daniel Raleigh)

1999–2002 — Postdoctoral Fellow, University of Washington
(Advisor: David Baker; Damon Runyon Fellowship)

Rosetta Methods Development

— Energy function optimization and parallel computing

De Novo Protein Design

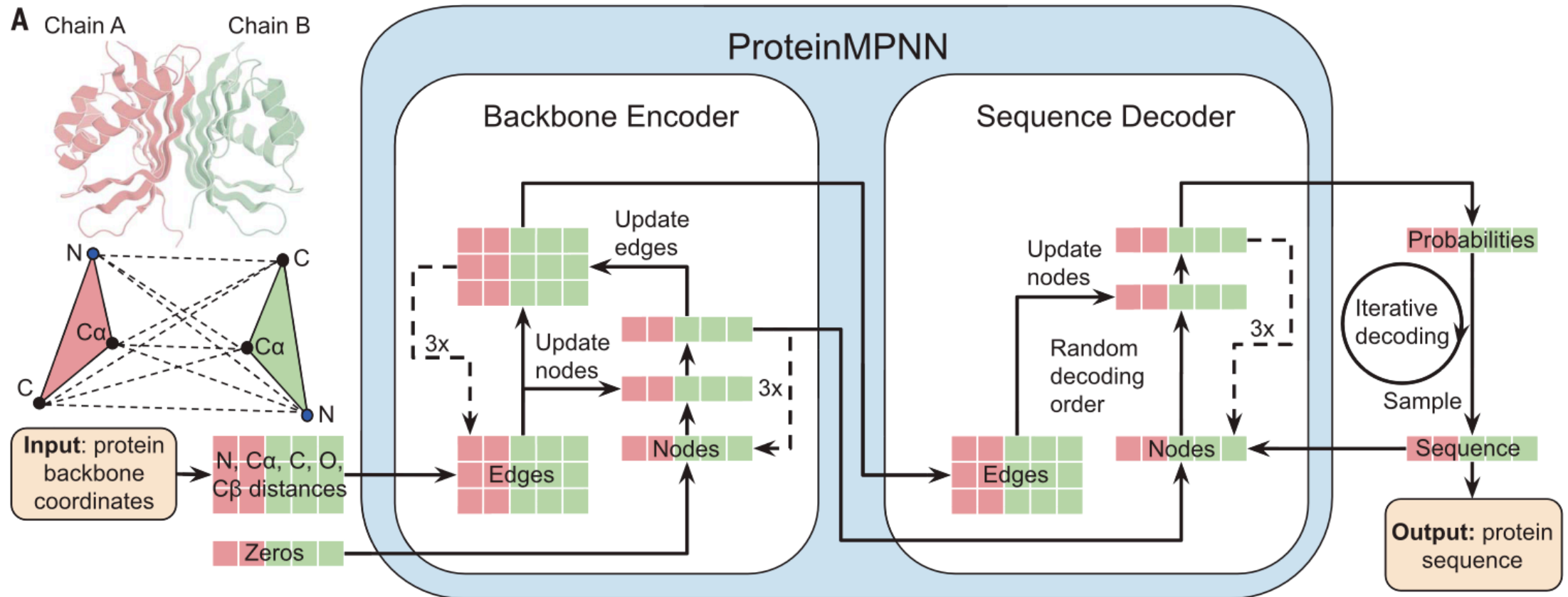
— Designing entirely new protein structures from scratch

Protein Interface Design

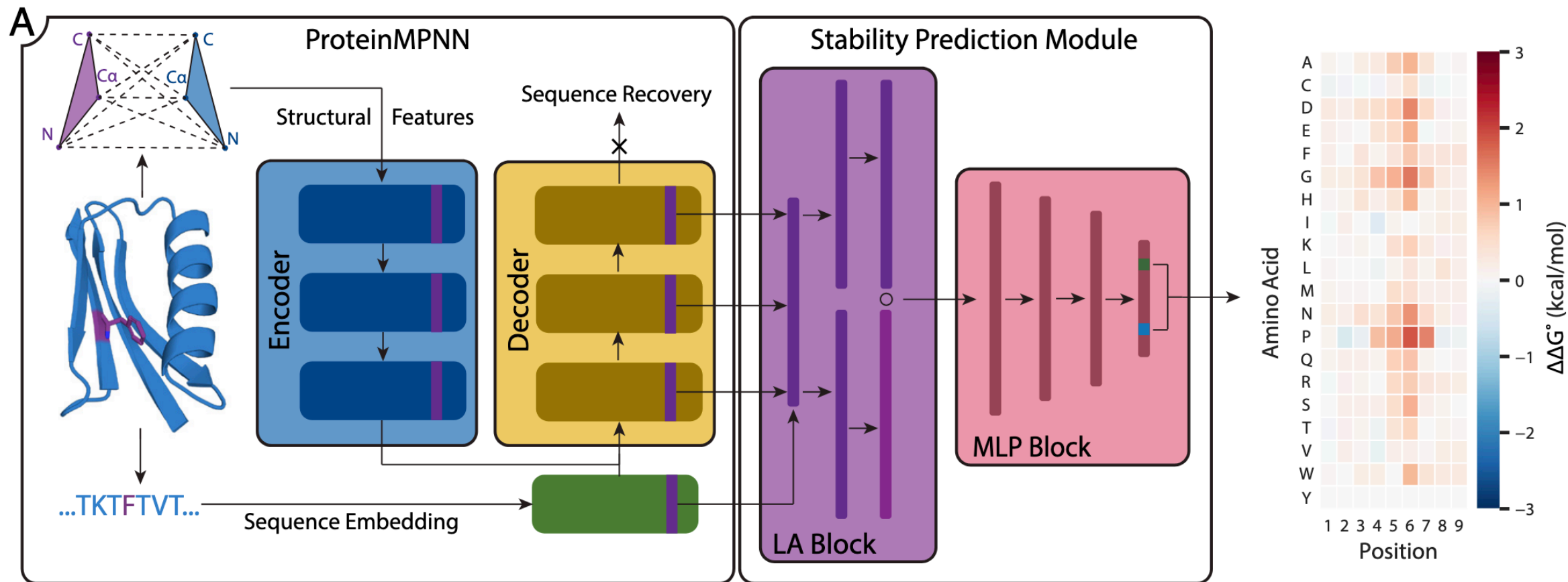
— Tuning the affinity and specificity of protein-protein interactions



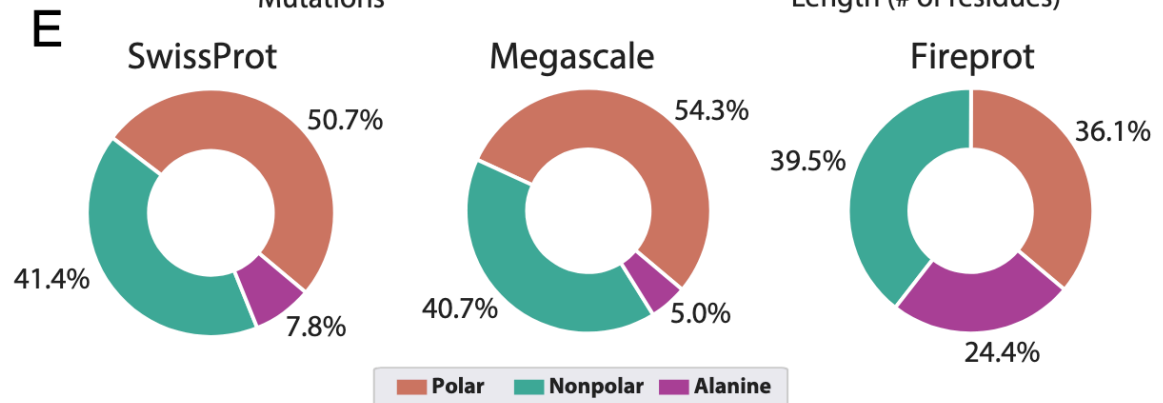
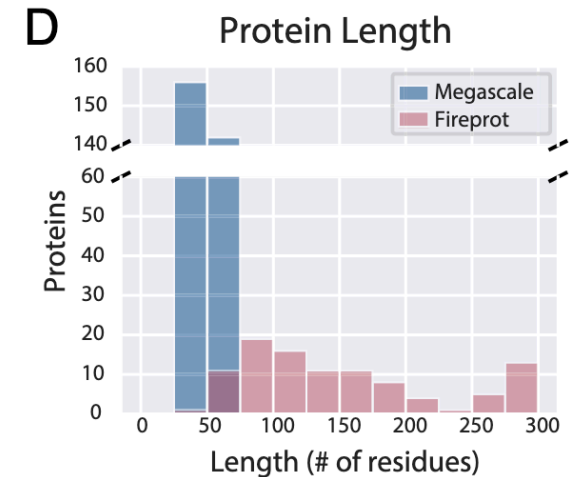
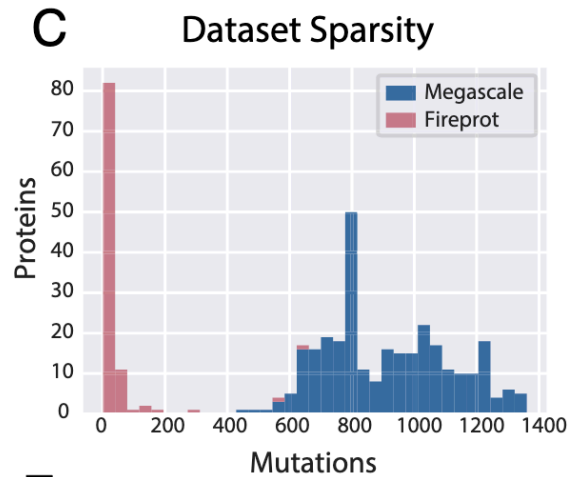
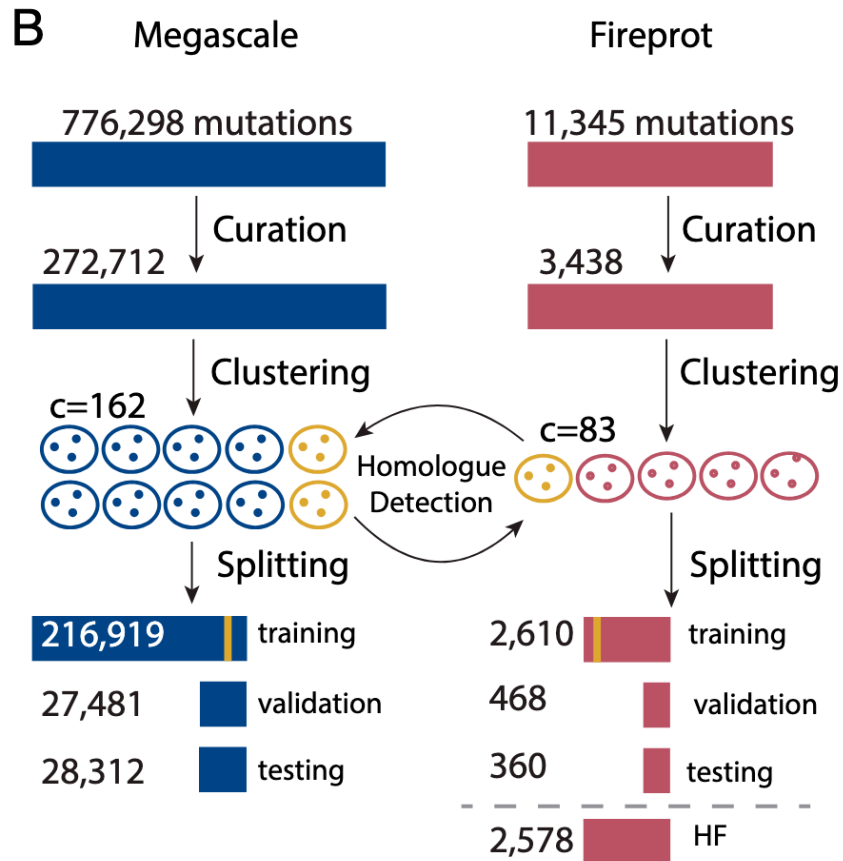
ProteinMPNN



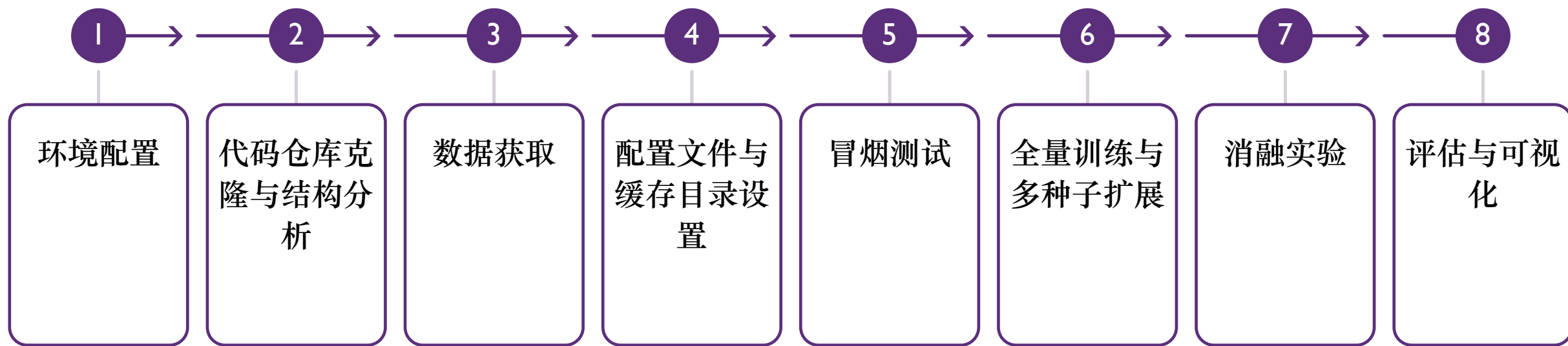
ThermoMPNN



Dataset Preparation



Workflow



Environment Configuration

- 环境调查：系统全局安装的是 CUDA 12.4且无多版本 CUDA 的模块管理系统
- 依赖安装：PyTorch 单独用 pip 安装；其余依赖（torchmetrics、pytorch-lightning、biopython、wandb、omegaconf 等）逐一用 pip 补装，mmseqs2 因含 C 扩展改用 conda 安装
- 诊断与修复（冒烟测试）
 1. PYTHONPATH 环境污染
 2. PyTorch 与 CUDA 驱动版本不匹配
 3. WandB 离线模式



Structure Analysis

- `dataset_splits/` 目录中已有作者预先完成的 MMseqs2 聚类划分文件；
- `models/` 目录中有作者训练好的 `thermoMPNN_default.pt`；
- `vanilla_model_weights/` 目录中有 ProteinMPNN 的四个预训练权重文件；
- 对 `config.yaml` 和 `local.yaml` 的分析明确了两个文件的分工：
 - 前者管理模型超参数（学习率、层数、是否启用 Light Attention 等）
 - 后者管理本地路径（数据集位置、缓存目录）



Data Acquired

- Fireprot 数据集： Zenodo API 查询发现文件名为 `fireprot_upload.zip`，下载并解压得到了 `4_fireprotDB_bestpH.csv` 和 130 个 PDB 文件
- Megascale 数据集： 先直接用论文中提到的 csv 文件名拼接 Zenodo URL 下载失败；随后通过 Zenodo API 查询记录发现文件列表中有相关压缩包，切换到 `aria2c -x 16 -s 16` 多连接下载，速度从 `wget` 的 13 KB/s 提升至约 185 KB/s；解压时发现 zip 里有 macOS 系统生成的 `__MACOSX/` 目录（包含 `._*.pdb` 元数据文件），导致 PDB 文件数量统计异常，执行 `rm -rf pdb/ __MACOSX` 清理



Config File, Cache Directory Settings and Smoke Testing

- 将 `local.yaml` 中的集群路径全部替换为本地真实路径，共更新 8 个字段：
`thermompnn_dir`、`cache_dir`、`megascale_csv`、`megascale_splits`、
`megascale_pdb`s、`fireprot_csv`、`fireprot_splits`、`fireprot_pdb`s。同时创建了 `data_all/cache/` 目录
- 正式训练前，先用 3 个 epoch 的冒烟测试验证整条流水线，并解决了新暴露的两个此前未发现的问题
- 冒烟测试通过后观察到：239 个训练蛋白 + 31 个验证蛋白全部正确加载；模型参数量 2.7M 可训练 + 1.7M 冻结 = 4.3M 总参数，与论文一致；3 个 epoch 内验证 R^2 从 0 升至 0.582，loss 单调下降，训练流水线正常



Full-Scale Training, Multi-Seed Runs and Ablation Study

- 在启动全量 100 epoch 训练之前，在 `train_thermompnn.py` 的入口处插入了一行：
`python torch.set_float32_matmul_precision('medium')`
- seed42: 100 个 epoch，约 14 小时，最佳检查点在 epoch 46（验证 Spearman 0.79）。
使用 `nohup ... &` 放入后台，通过 `tail -f` 定期查看进度
- seed1337 和 seed2024: seed1337 最佳检查点在 epoch 37（验证 Spearman 0.80），
seed2024 在 epoch 45（验证 Spearman 0.79）
- 修改配置文件副本，创建了三个消融变体的 yaml 配置: `config_no_la.yaml` (`lightattn: false`)、`config_no_pretrain.yaml` (`load_pretrained: false`)、`config_finetune.yaml` (`freeze_weights: false + mpnn_learn_rate: 0.0001`)



Results

Reproduced results vs. thesis results of **master model (ThermoMPNN)**

<i>Targets</i>	Megascale-test		Fireprot-HF	
	<i>Thesis</i>	<i>Reproduced</i>	<i>Thesis</i>	<i>Reproduced</i>
PCC	0.754 ± 0.004	0.754 ± 0.001	0.650 ± 0.005	0.649 ± 0.002
SCC	0.725 ± 0.003	0.725 ± 0.001	0.657 ± 0.003	0.652 ± 0.002
RMSE	0.708 ± 0.007	0.726 ± 0.022	1.510 ± 0.010	1.496 ± 0.019



Results

ThermoMPNN ablation study results - Megascale-test

Models	Reproduced results		Thesis results	
	PCC	SCC	PCC	SCC
ProteinMPNN	0.433	0.487	0.430 ± 0.010	0.487 ± 0.006
ThermoMPNN	0.753	0.725	0.754 ± 0.004	0.725 ± 0.003
No Light Attention	0.742	0.710	0.742 ± 0.002	0.711 ± 0.004
Finetune	0.787	0.752	0.780 ± 0.002	0.747 ± 0.001
No Pretrain	0.664	0.582	0.689 ± 0.005	0.642 ± 0.005

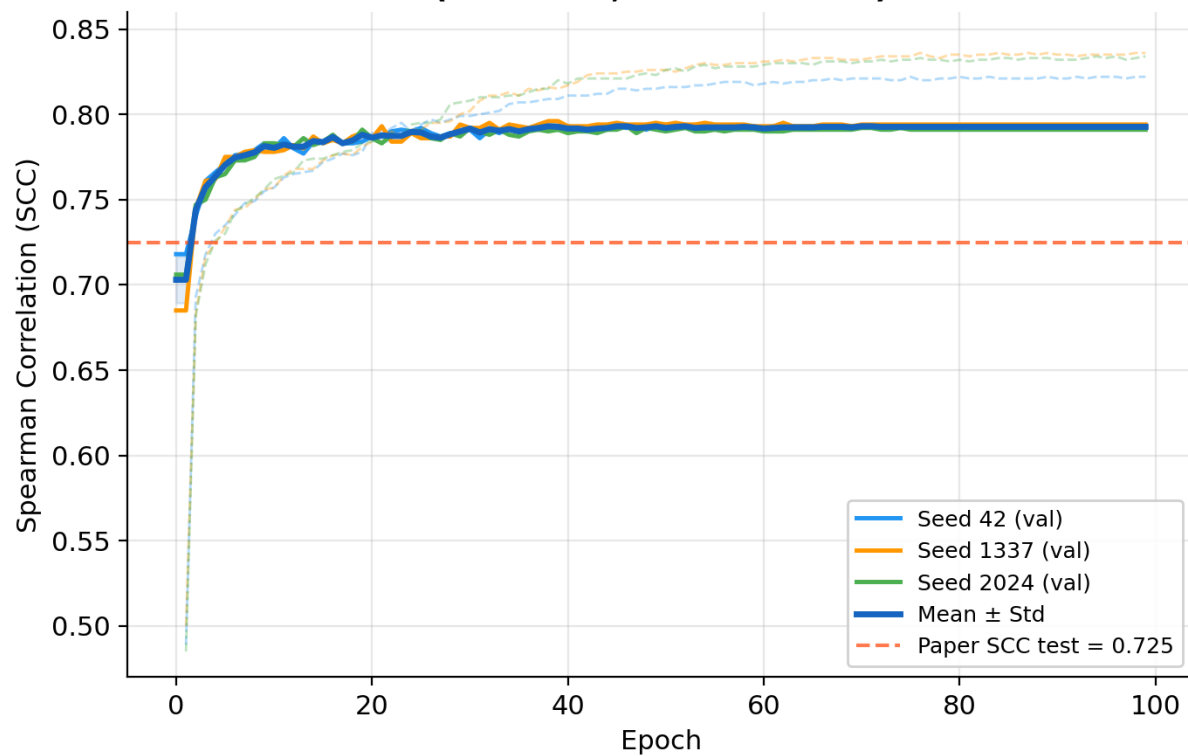
Results

ThermoMPNN ablation study results - Fireprot-HF

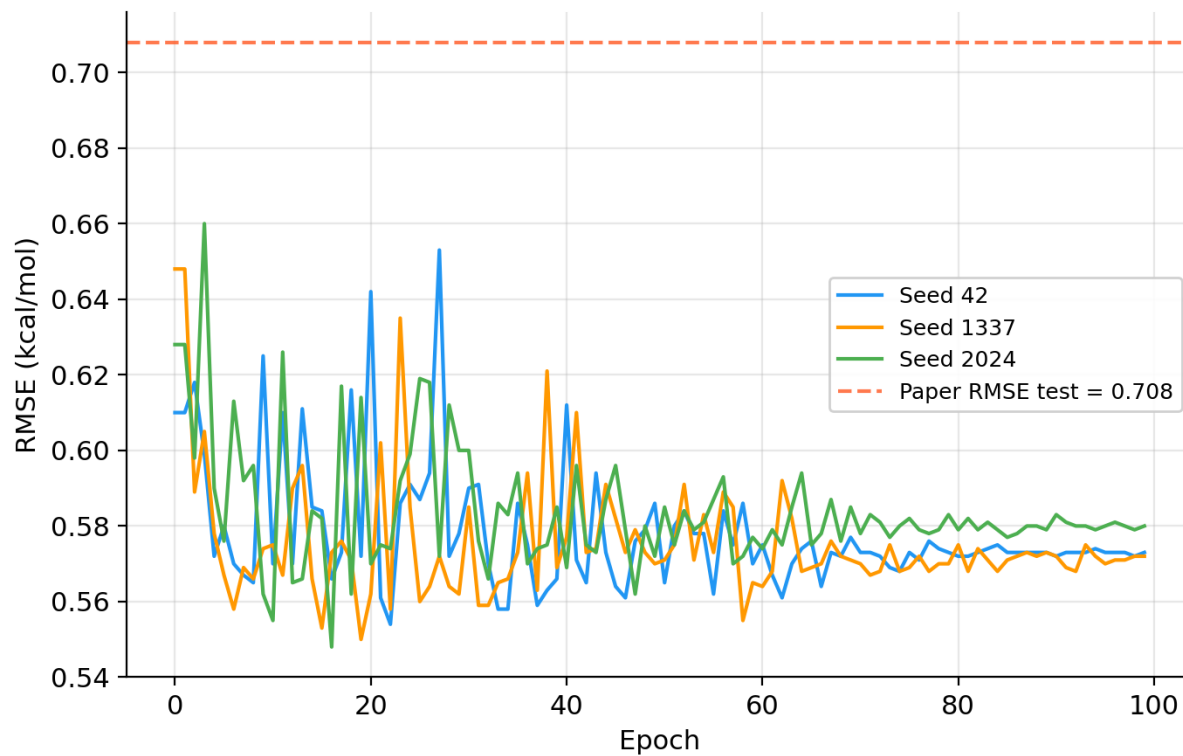
Models	Reproduced results		Thesis results	
	PCC	SCC	PCC	SCC
ProteinMPNN	0.449	0.513	0.410 ± 0.020	0.490 ± 0.010
ThermoMPNN	0.651	0.655	0.650 ± 0.005	0.657 ± 0.003
No Light Attention	0.625	0.635	0.630 ± 0.001	0.634 ± 0.001
Finetune	0.655	0.659	0.657 ± 0.006	0.666 ± 0.004
No Pretrain	0.456	0.411	0.540 ± 0.020	0.500 ± 0.020

Visualized Results

Training Curves — Three Seeds
(solid=val, dashed=train)

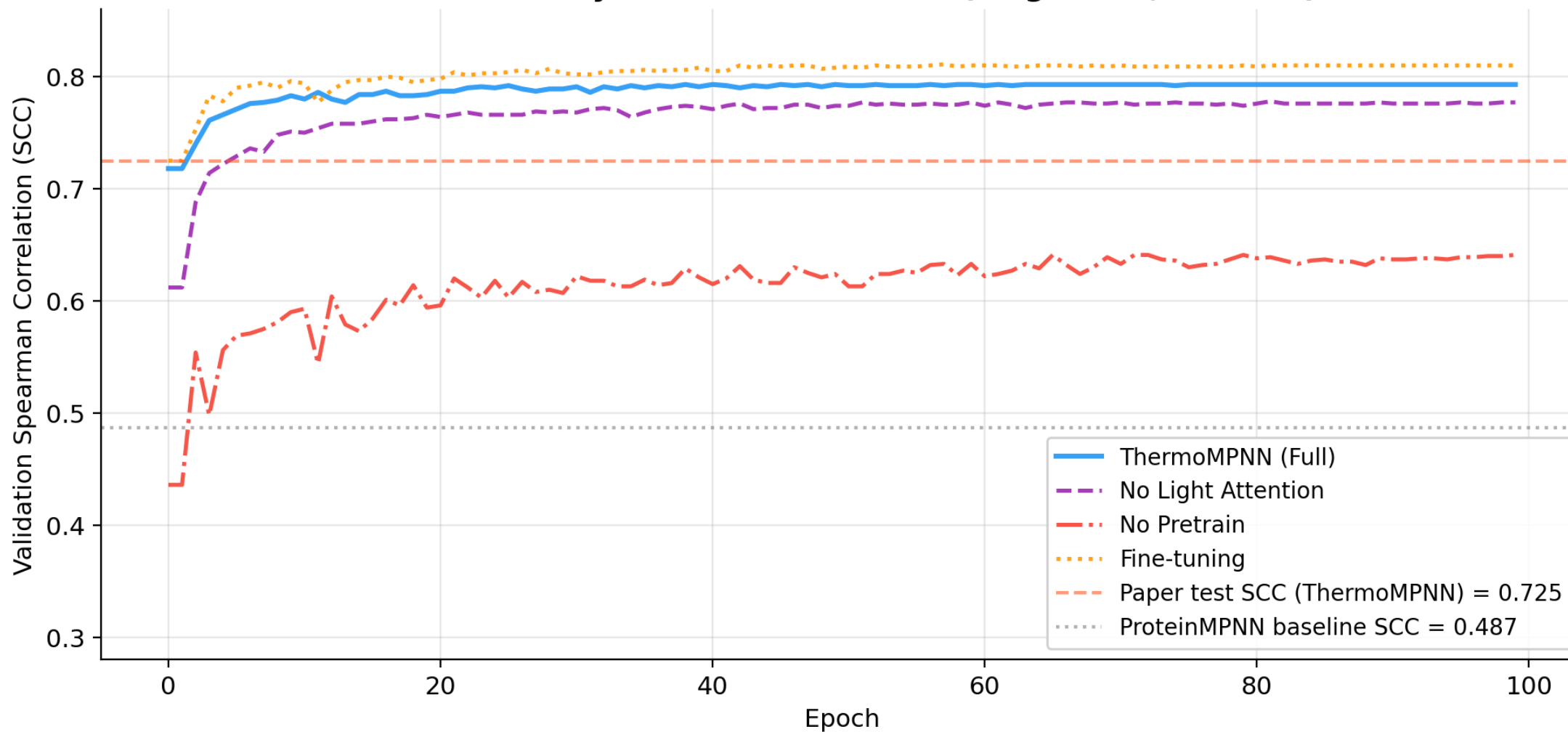


Validation RMSE — Three Seeds



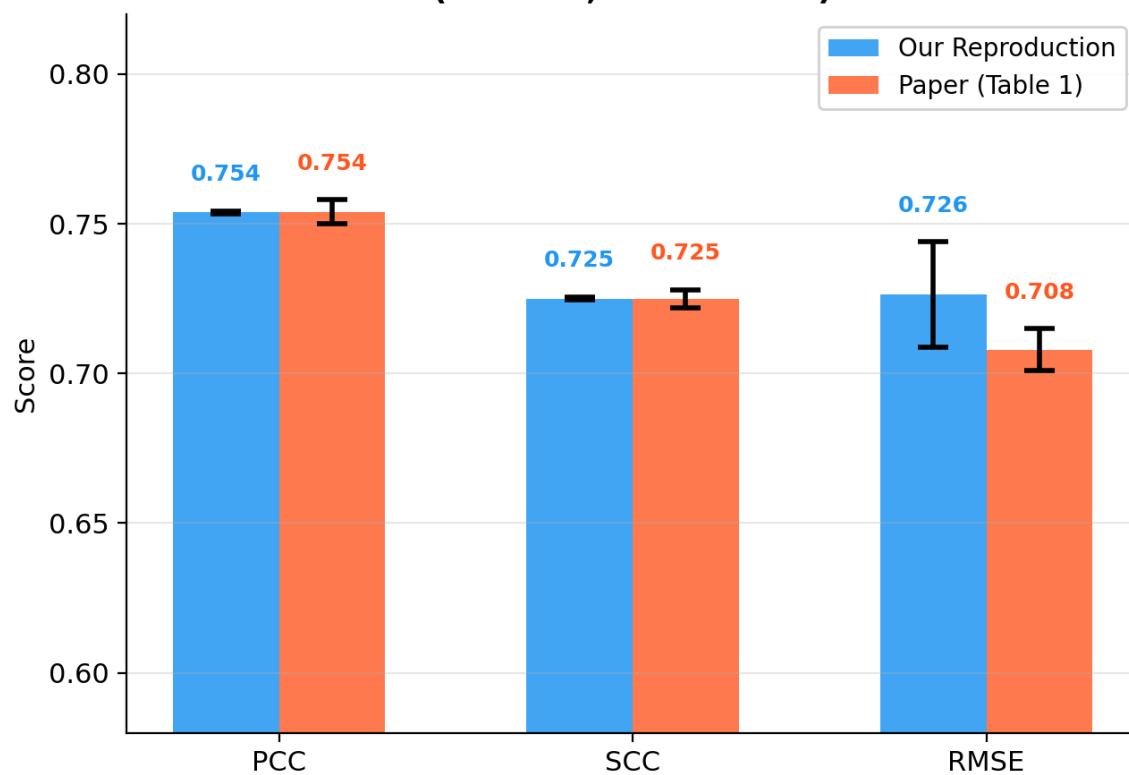
Visualized Results

Ablation Study — Validation Curves (Megascalse, Seed 42)

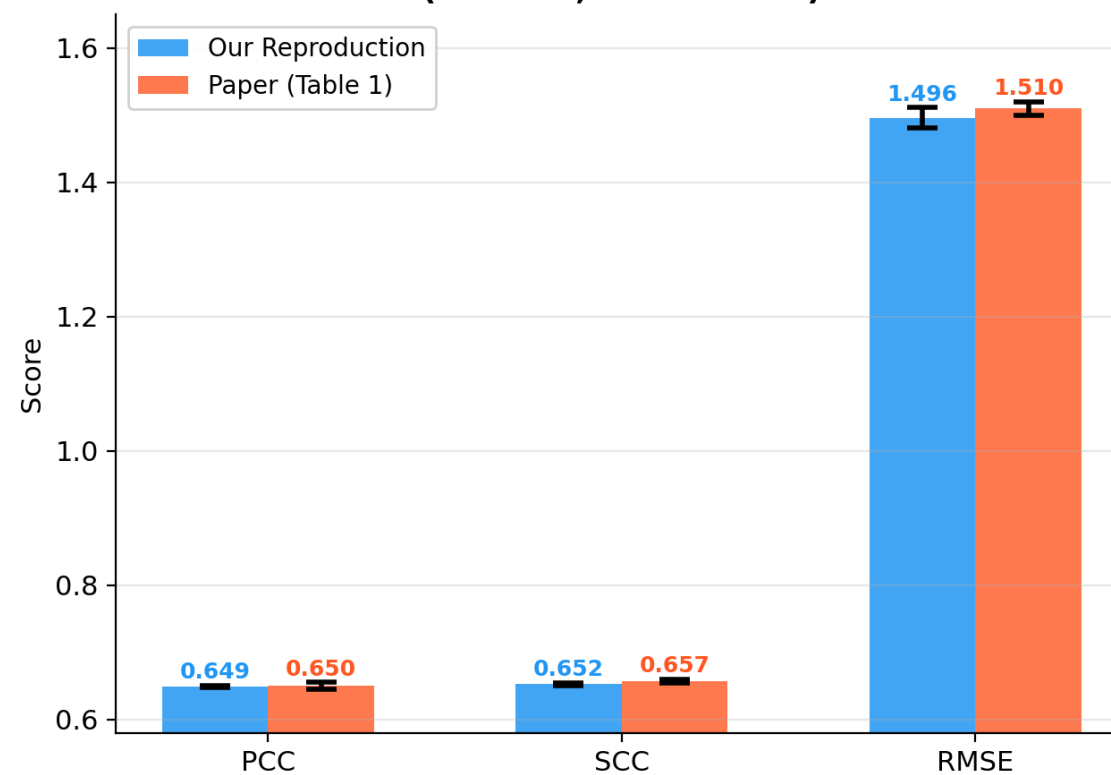


Visualized Results

**Main Model — Megascale-test
(3 seeds, mean $\pm$ std)**



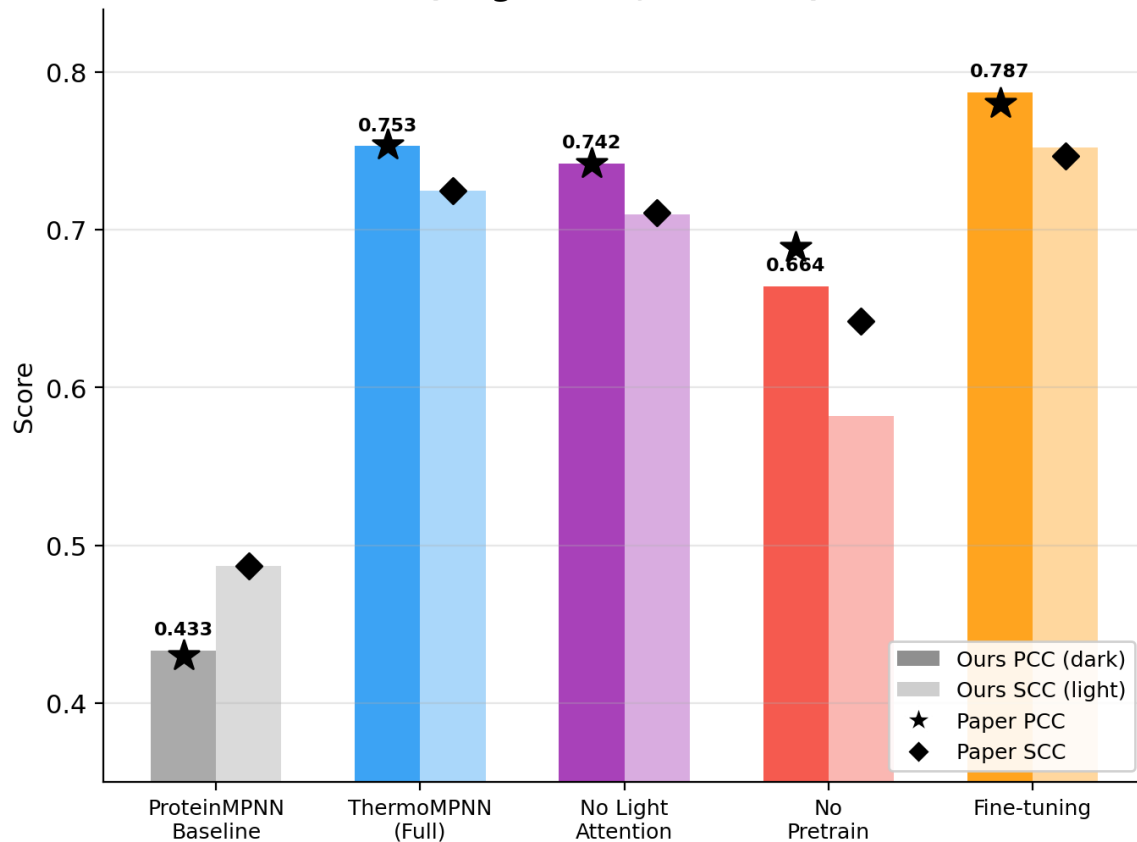
**Main Model — Fireprot-HF
(3 seeds, mean $\pm$ std)**



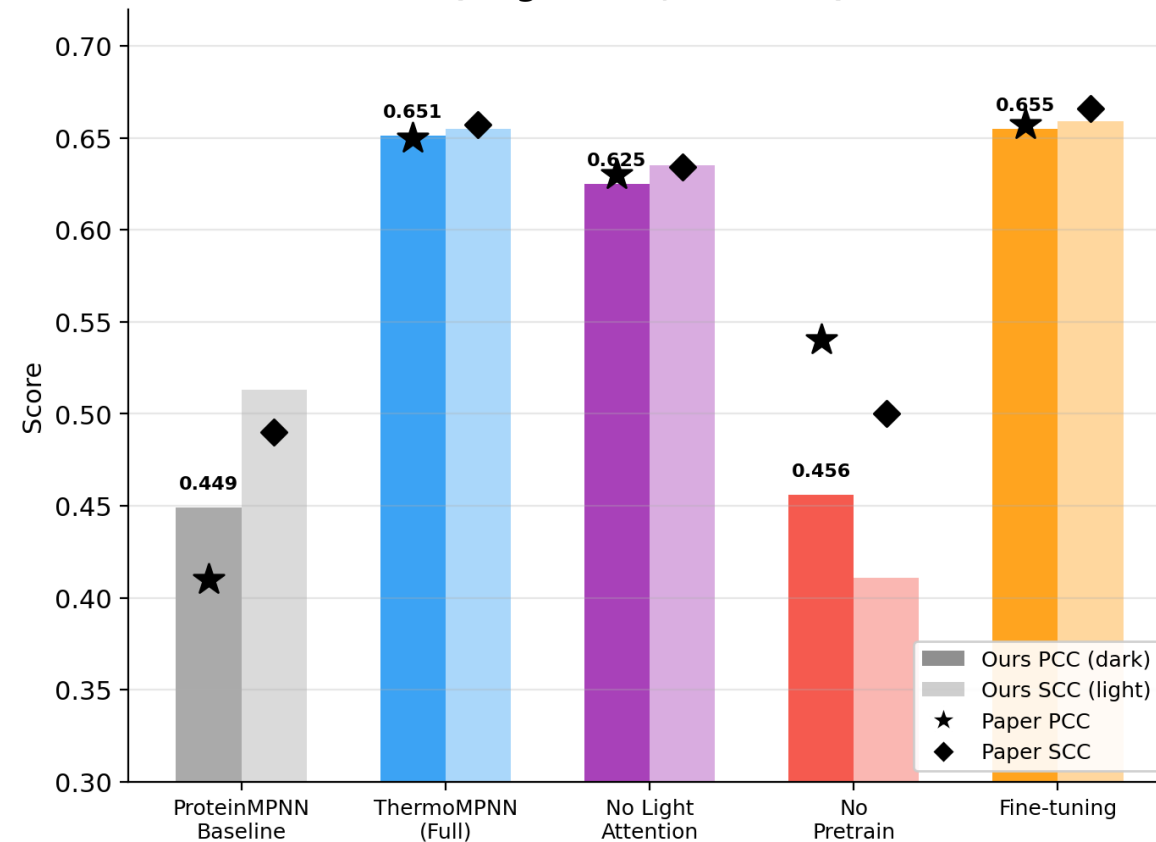
Visualized Results

Ablation Study: Impact of Each Design Component

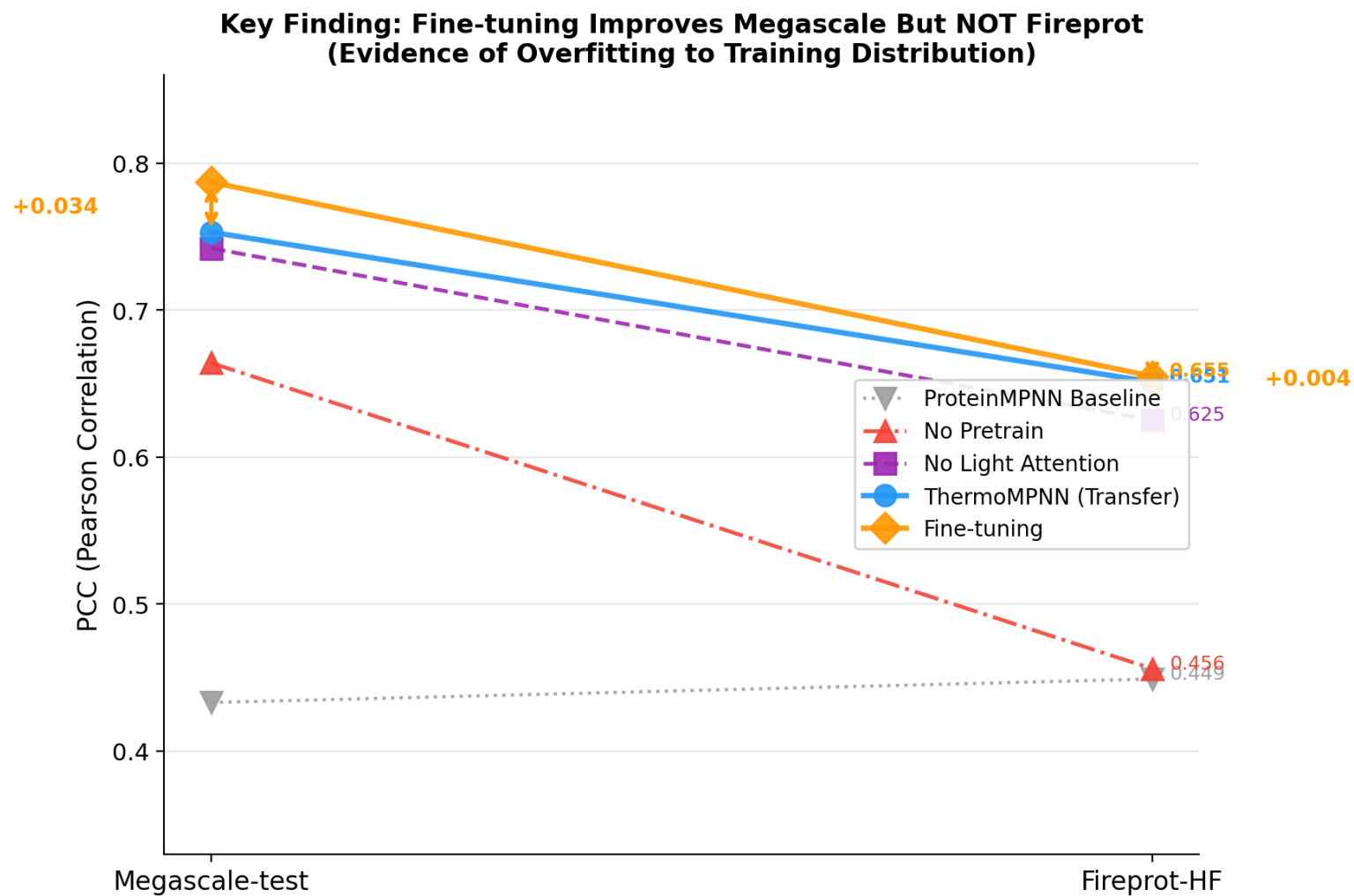
**Ablation Study — Megascale-test
(single seed, seed=42)**



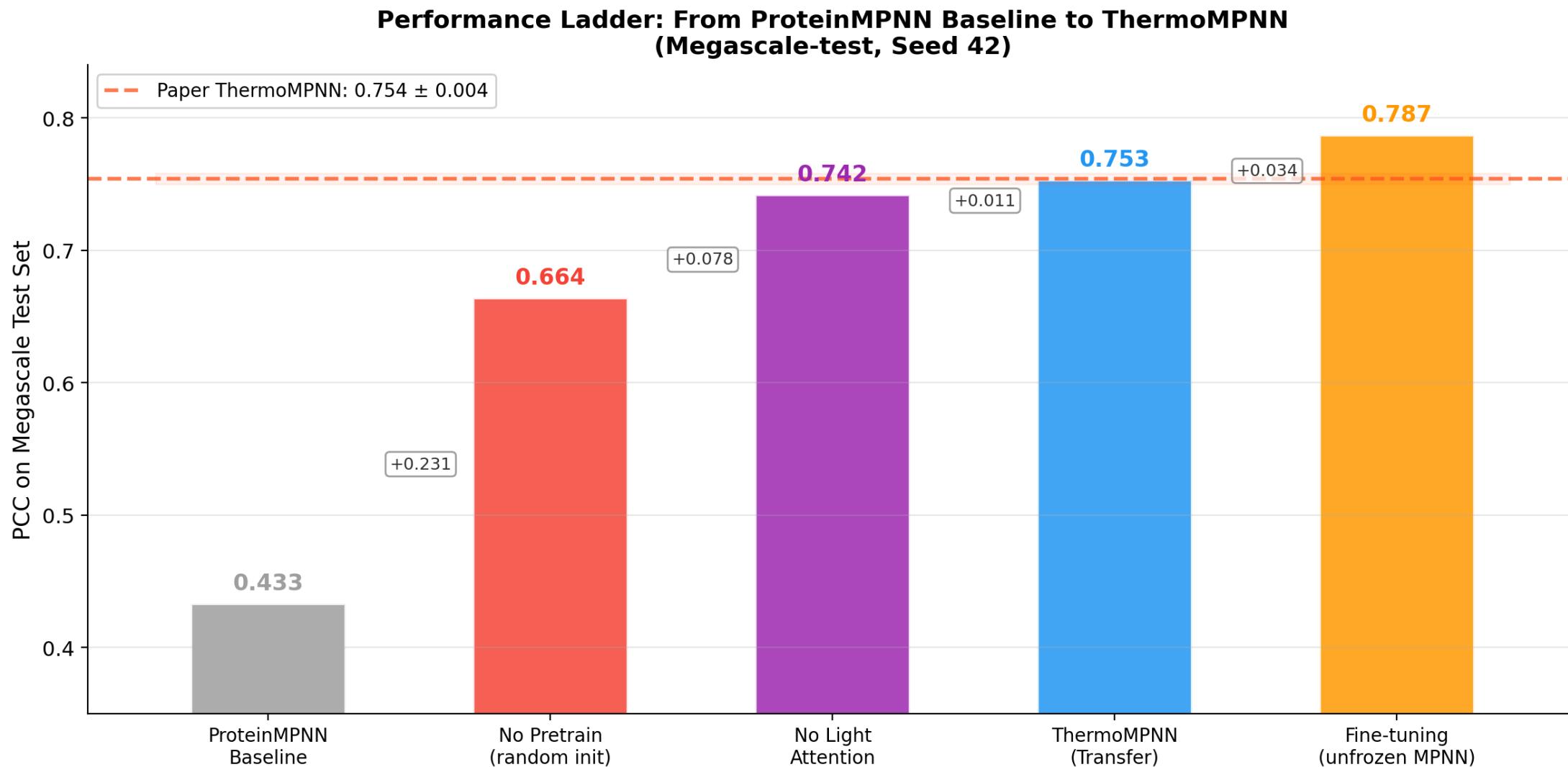
**Ablation Study — Fireprot-HF
(single seed, seed=42)**



Visualized Results

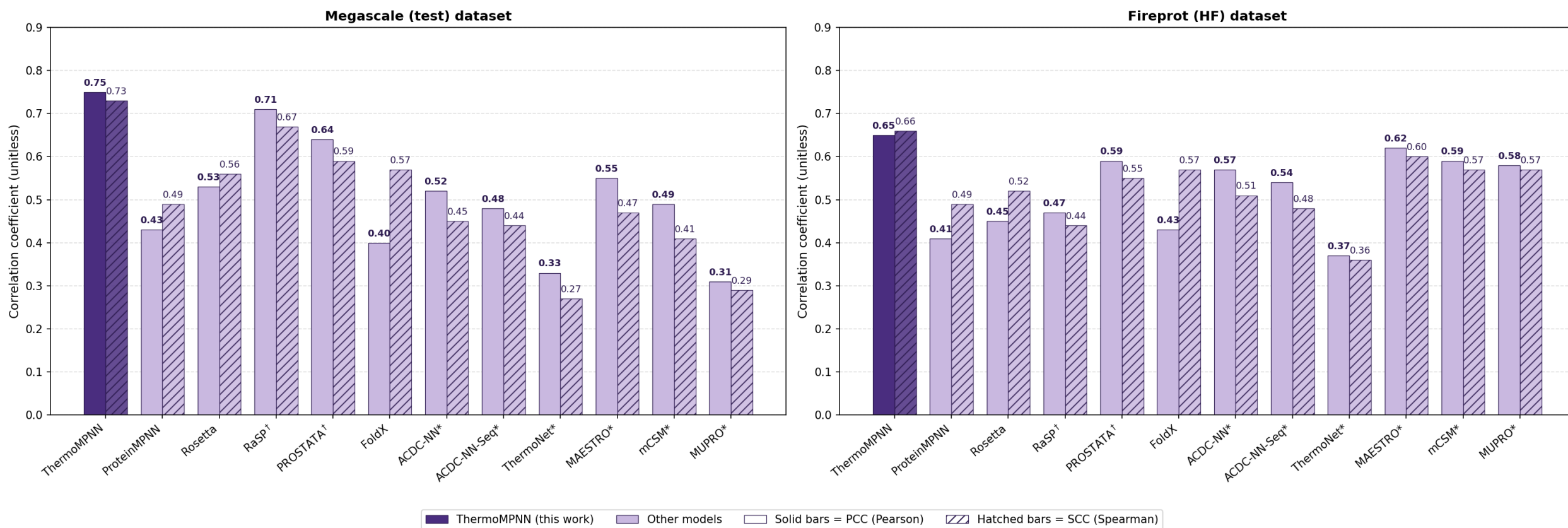


Visualized Results



Results

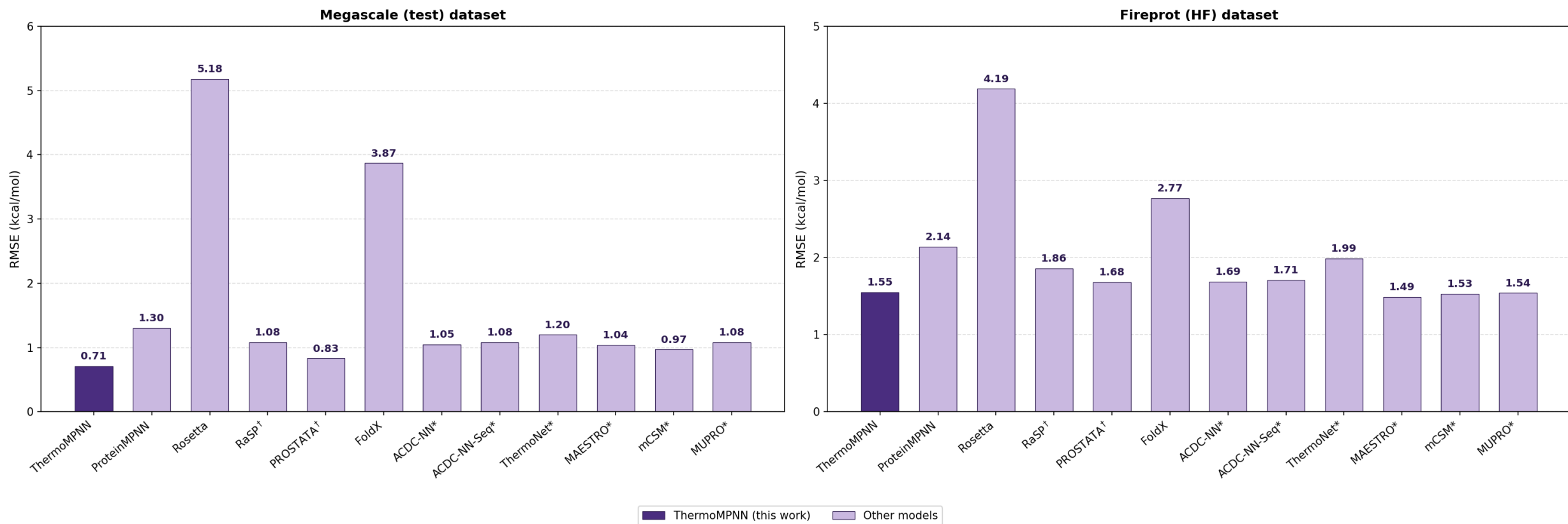
Correlation metrics — PCC and SCC (higher is better ↑)



[†] Retrained on Megascale dataset * Score may be inflated due to close homologues (>25% sequence identity) in training data

Results

RMSE comparison (lower is better ↓)



† Retrained on Megascale dataset * Score may be inflated due to close homologues (>25% sequence identity) in training data

Acknowledgement

感谢Claude倾情指点