A Unified Protein Embedding Model with Local and Global Structural Sensitivity

Jerry Xu

MIT PRIMES Computational Biology

Mentors: Dr. Gil Alterovitz (Harvard Medical School), Dr. Shaojun Pei (Brigham and Women's Hospital)

10/24/26

Table of Contents

1 Introduction

2 Definitions

3 Methodology

4 Results

Table of Contents

1 Introduction

2 Definitions

3 Methodology

4 Results

Why compare protein structures?

Structural comparison is relevant for many research tasks, including:

- evolutionary analysis
- functional annotation
- interpretation of mutations
- peptidomimetics

With the rapid growth of known and predicted structures in protein databanks, we need more efficient and comprehensive comparison methods.

Methods to compare protein structures

There are many ways to approximate structural similarity.

Sequence search

BLAST, MMseqs2

+ very fast
– sequence is a poor proxy for structure

Structural alignment

DALI, TM-align, Foldseek

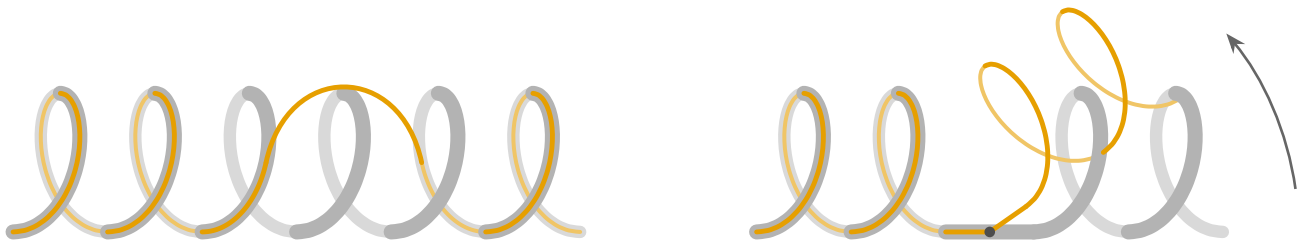
+ accurate
– needs 3D structures
– $\mathcal{O}(mn)$ per pair

Embeddings

TM-Vec

+ sequence only, $\mathcal{O}(1)$ per comparison
– predicts only one global number (TM-score)

Global vs. local structural similarity



Structure		
Global	**high** TM-score	**low** TM-score
Local	low only at the changed site	high everywhere: every domain is the same

local similarity: same similar different

Our goal

Goal: a sequence-only embedding model that incorporates both views:

- **local:** the inner product of two residue embeddings estimates their local structural similarity, for any residue pair;
- **global:** pooling the embeddings estimates the TM-score.

Table of Contents

1 Introduction

2 Definitions

3 Methodology

4 Results

TM-Score

Template Modeling score

- Metric of global structural similarity, $\in [0, 1]$
- $> 0.5 \implies$ very similar fold
- Asymmetric

Definition (TM-score)

$$\mathrm{TM}_P = \max_{\text{superpositions}} \frac{1}{m} \sum_{\text{aligned } (i,j)} \frac{1}{1 + \left(d_{ij}/d_0(m)\right)^2}$$

A local similarity label: GW and PIS

- **Patch** of residue i: its $K = 32$ nearest C_α atoms.
- Compare the two patches' **internal distance matrices** with the Gromov–Wasserstein (GW) discrepancy:

Definition (GW discrepancy and PIS)

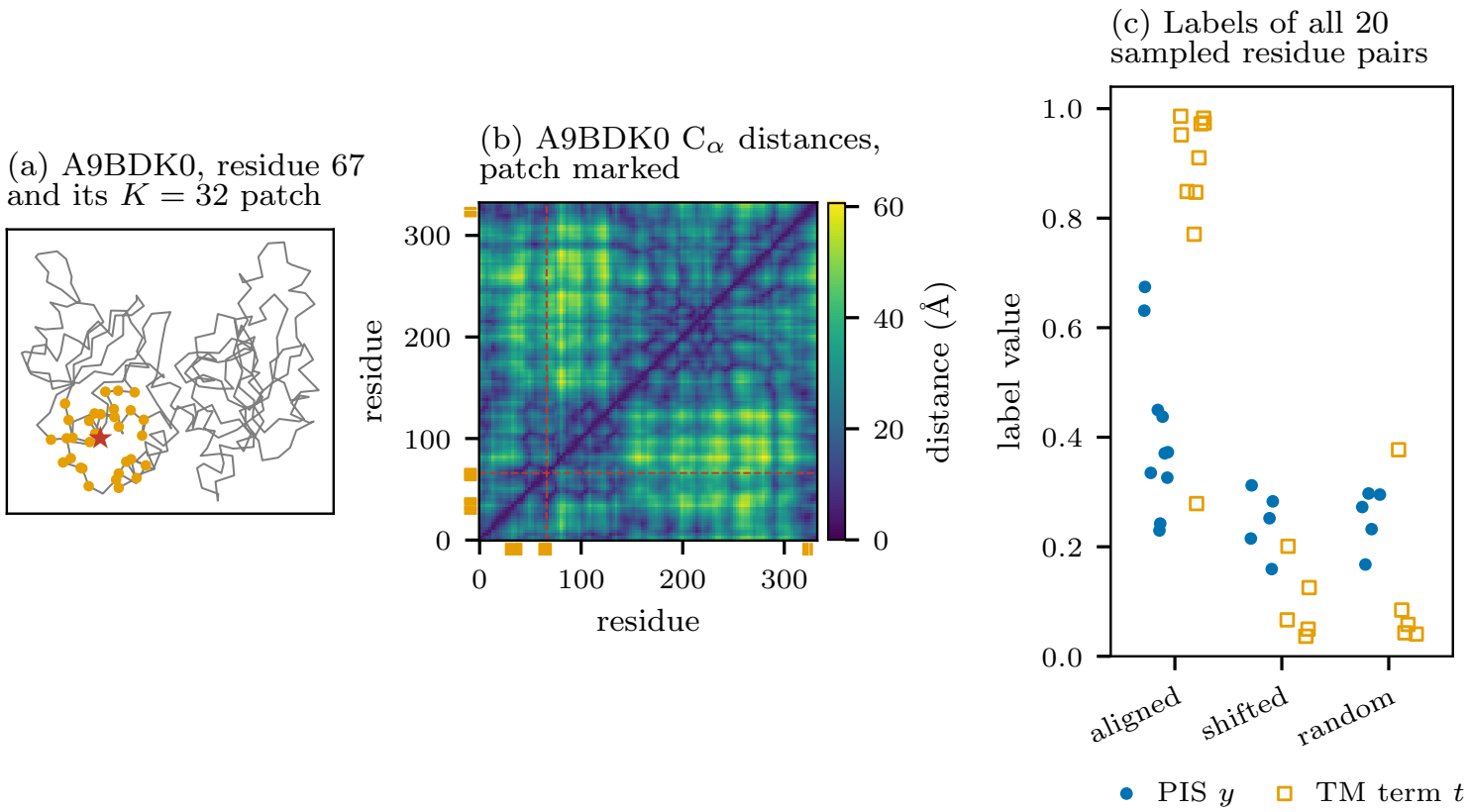
$$\mathrm{GW}(i,j) = \min_{\mathsf{T}} \sum_{a,c} \sum_{b,d} \left(C^P_{ac} - C^Q_{bd}\right)^2 T_{ab}\, T_{cd}$$

$$\text{PIS} \quad y_{ij} = \exp\left(-\mathrm{GW}(i,j)/0.05\right) \in (0,1]$$

Why GW?

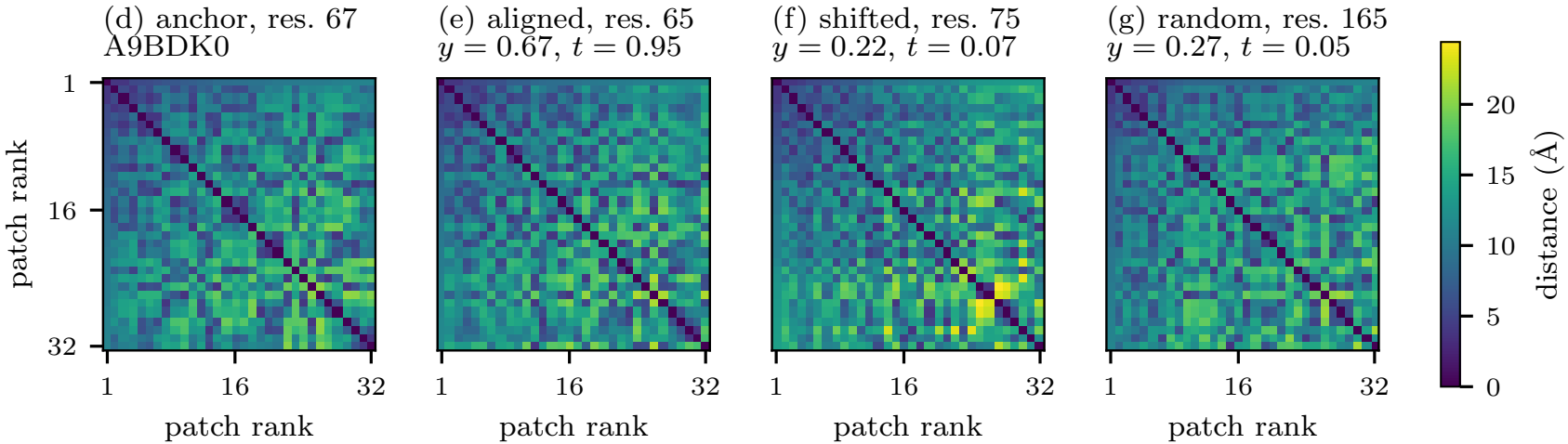
- No superposition needed
- No residue correspondence needed
- Invariant to rotations and translations
- $y = 1$ exactly when the patches are isometric
- Defined for every residue pair, so we can train on negatives

PIS example



Two carbamoyltransferases ($\mathrm{TM} = 0.79$, 22 % identity). (a) the residue patch vs the flattened 2D structure. (b) The protein's distance matrix. (c) PIS and TM values of the 20 sampled residue pairs.

PIS example



Patch distance matrices of the residue in (a) and three partner residues in the other protein: aligned, shifted, and random.

Table of Contents

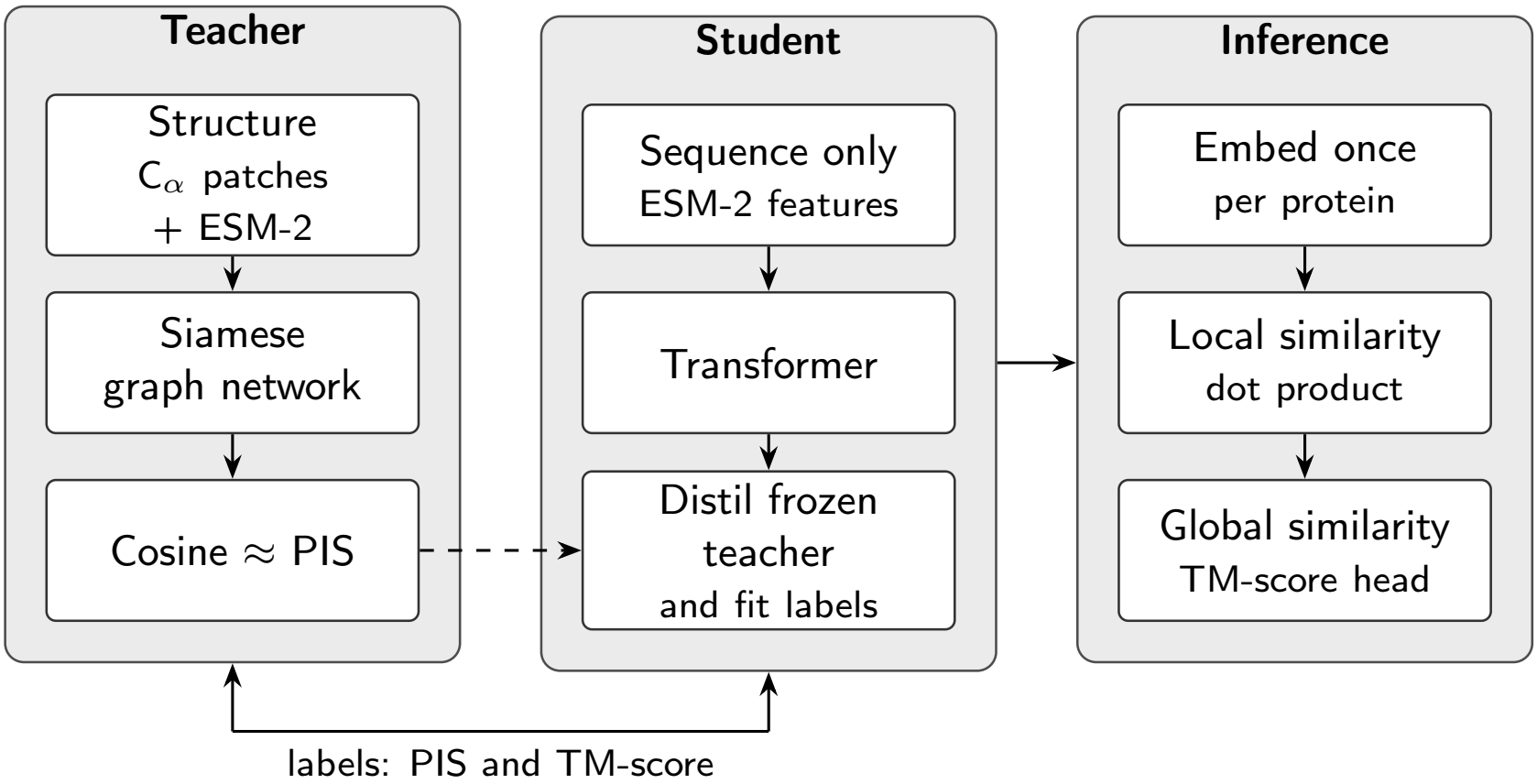
1 Introduction

2 Definitions

3 Methodology

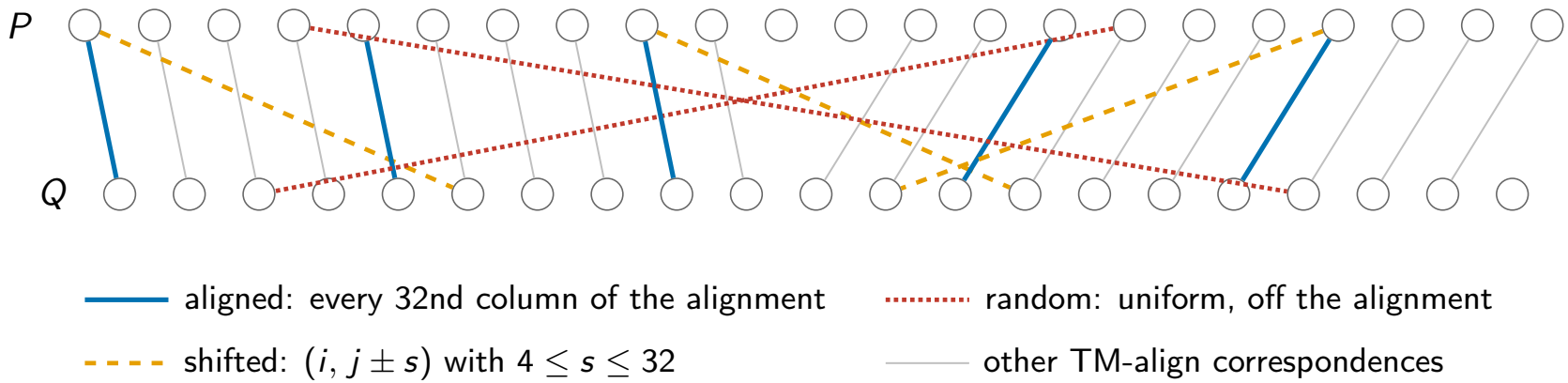
4 Results

Architecture Diagram



Labels created from structure → a structure-based teacher learns the labels → a sequence-based student distills the teacher + labels.

Which residue pairs get labels



Every 32nd aligned residue is a positive. Each positive is followed by a negative, which is either shifted or random.

Data

total pairs	985,230
distinct proteins	351,443
labelled residue pairs	12,339,034
training pairs	656,434
validation / test pairs	14,295 / 22,805

Source pairs: TM-Vec's training data.

- Structures: SWISS-MODEL homology models; sequence features: ESM-2 (650M).

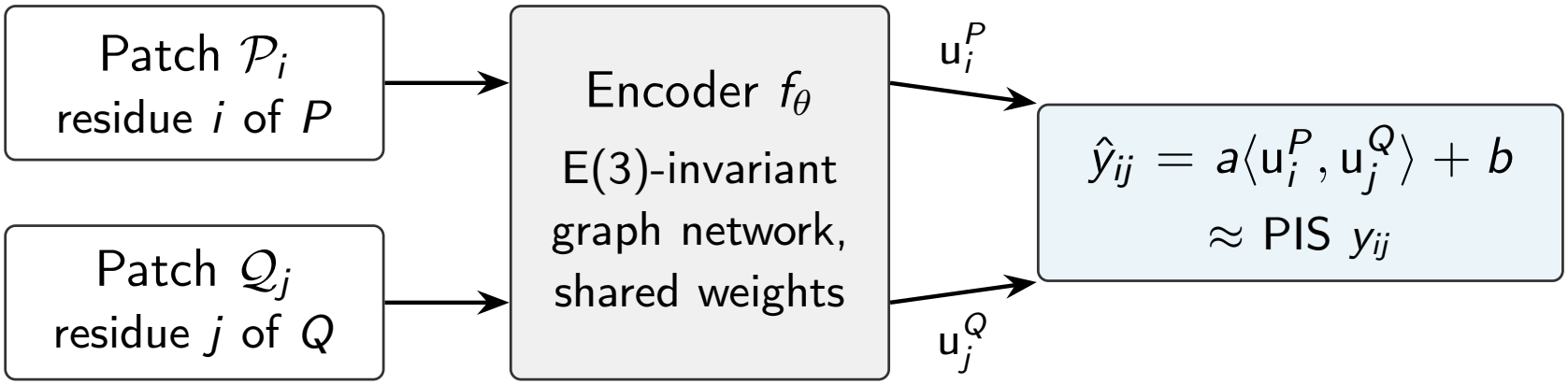
Clustering

- Proteins are clustered by 30% sequence identity (MMseqs2: 23,472 clusters), and then assigned 8:1:1 to training, validation and test.
- A pair is kept only if both proteins land in the same split (train, val, or test).

	both in training	both in validation	both in test	split apart (discarded)
proteins in different clusters	0.64	0.01	0.01	0.34
proteins in the same cluster	0.80	0.10	0.10	0
observed share of pairs	66.6 %	1.5 %	2.3 %	29.6 %

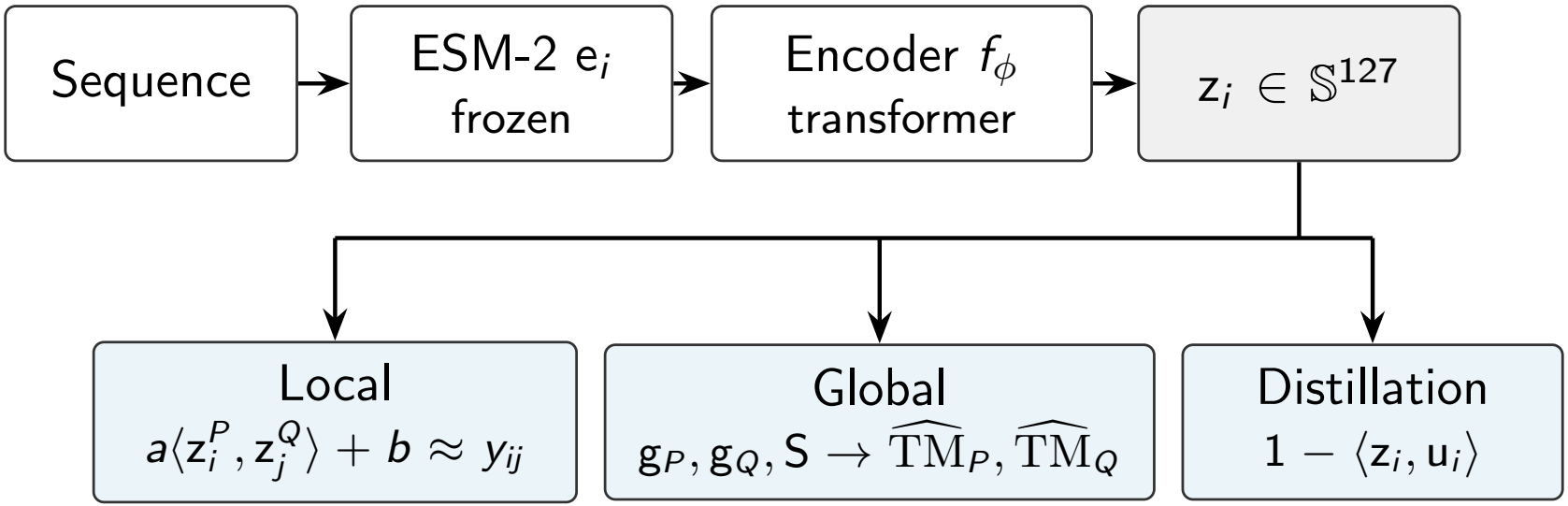
- Training gets about 95 % of the kept pairs.
- Held-out sets are enriched in related pairs (a same-cluster pair reaches the test set ten times as often).

Stage 1: the structure teacher



- E(3)-invariant graph network (4 layers) over each residue's patch, with ESM-2 node features.
- Cosine of the two patch embeddings predicts the PIS; auxiliary head predicts TM-score (not shown).
- Works with structural input, i.e. is more informed than the student: serves as an upper bound for what the student can learn.

Stage 2: the sequence-only student



- Transformer over ESM-2 sequence embeddings (one unit vector per residue)
- Learns from the frozen teacher as well as the true labels.
- Global head: pooled embeddings + residue matrix coverage statistics predicts TM-score.

Training

Training specifications:

- 656,434 training pairs, streamed in shuffled buffers
- AdamW, learning rate 10^{-4}
- One V100 GPU on PSC's Bridges-2

Ground truth information:

- TM-scores, alignments and superpositions computed with TM-align
- PIS computed with an entropic Gromov–Wasserstein solver (30 steps $\times$ 20 Sinkhorn iterations per residue pair)

Table of Contents

1 Introduction

2 Definitions

3 Methodology

4 Results

Evaluation

- Test set: 22,805 protein pairs from held-out sequence clusters.
- For PIS, results are broken down by residue-pair type: aligned, shifted, random, and compared against raw ESM-2 cosine.
- TM-score is compared against TM-Vec.

Teacher results (tentative)

	MSE	R^2	r
PIS, teacher	0.0156	0.76	0.87
PIS, raw ESM-2 cosine	0.121	−0.86	0.53
TM-score, auxiliary head	0.0067	0.93	0.97
TM-score, TM-Vec	0.0167	0.83	0.92

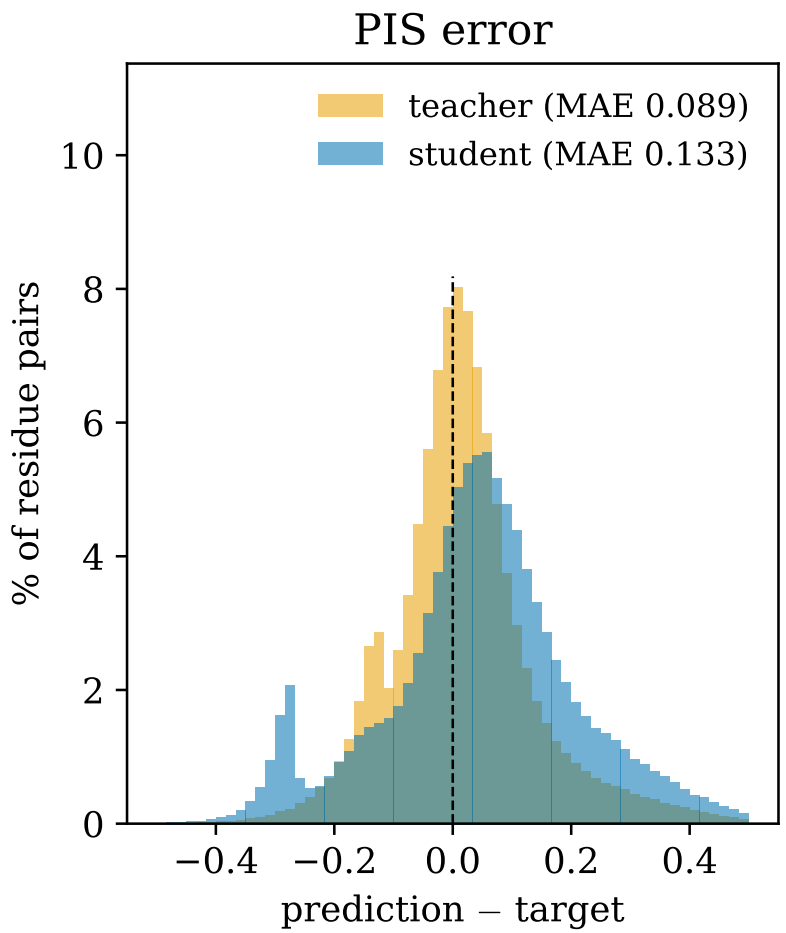
By residue-pair type, MSE (r): aligned 0.0246 (0.84), shifted 0.0072 (0.66), random 0.0061 (0.70)

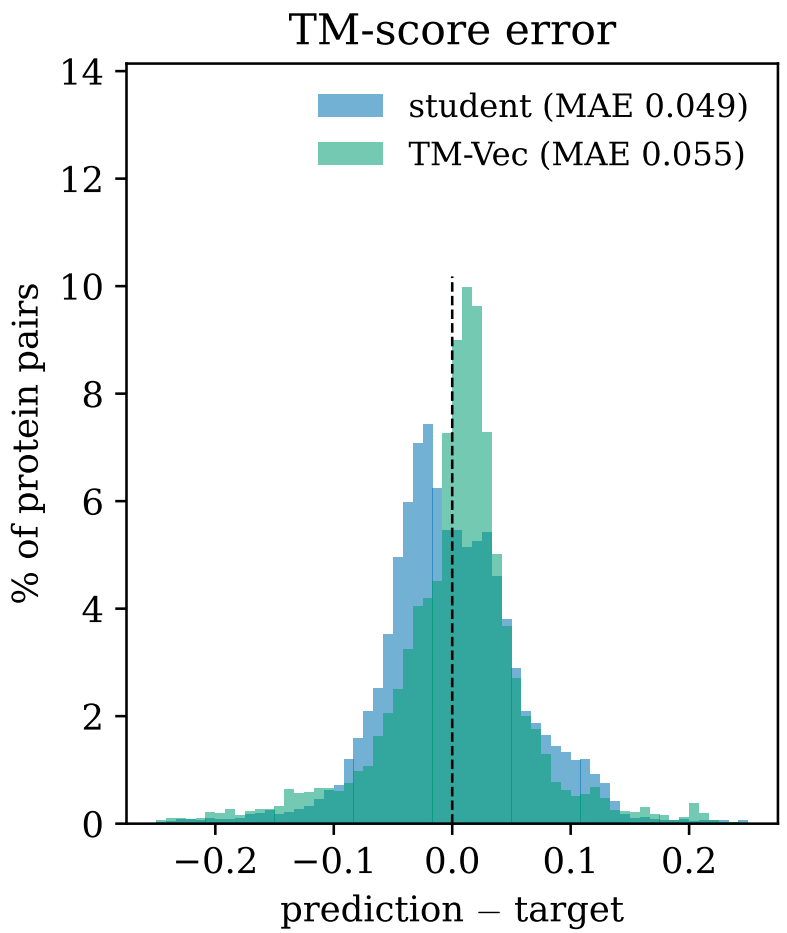
Student results (tentative)

	MSE	R^2	r
PIS, student	0.0299	0.54	0.76
PIS, raw ESM-2 cosine	0.121	−0.86	0.53
TM-score, global head	0.0053	0.93	0.96
larger TM-score†, TM-Vec	0.0104	0.86	0.93

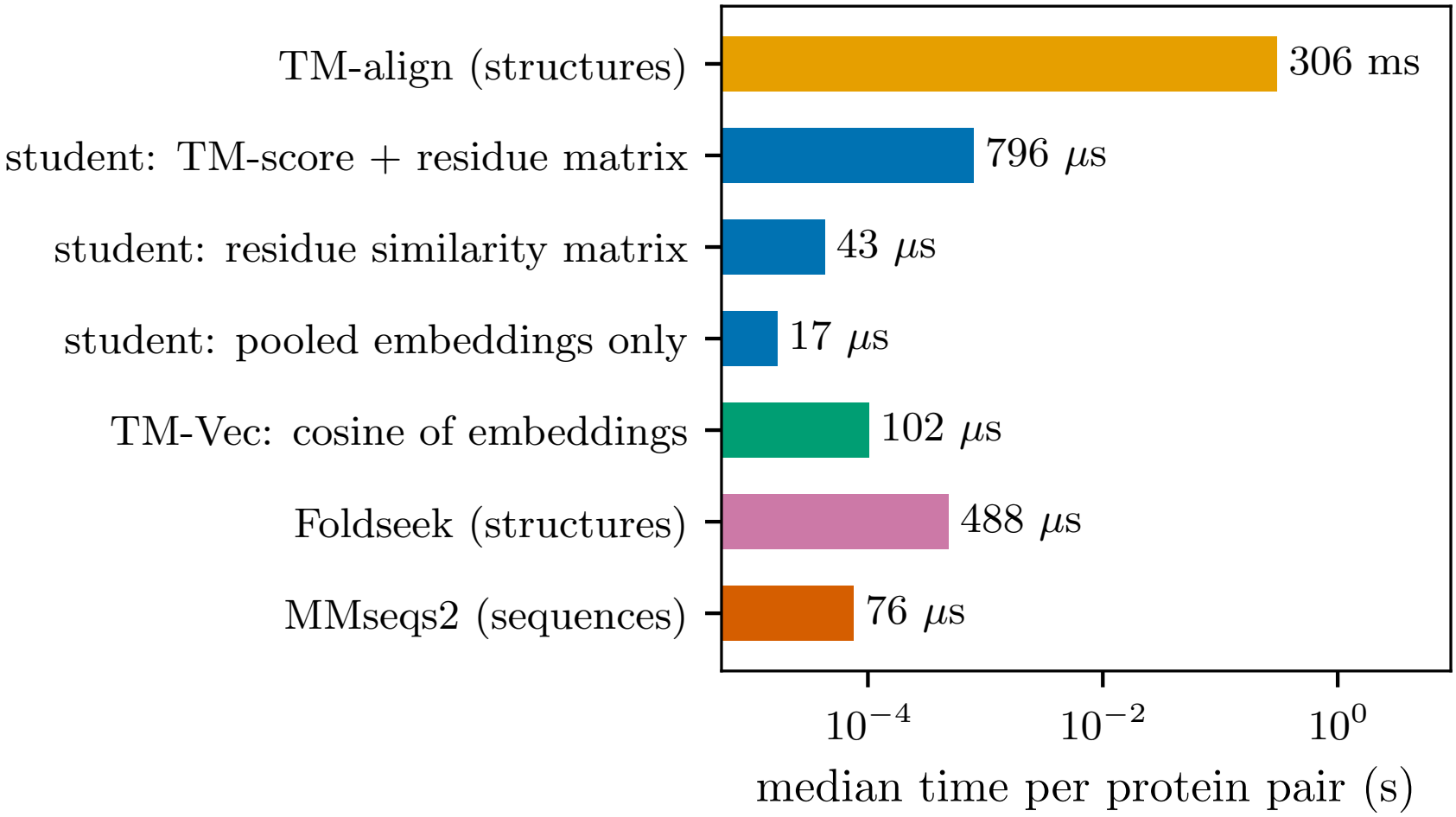
By residue-pair type, MSE (r): aligned 0.0471 (0.66), shifted 0.0137 (0.35), random 0.0117 (0.42)

Error distributions

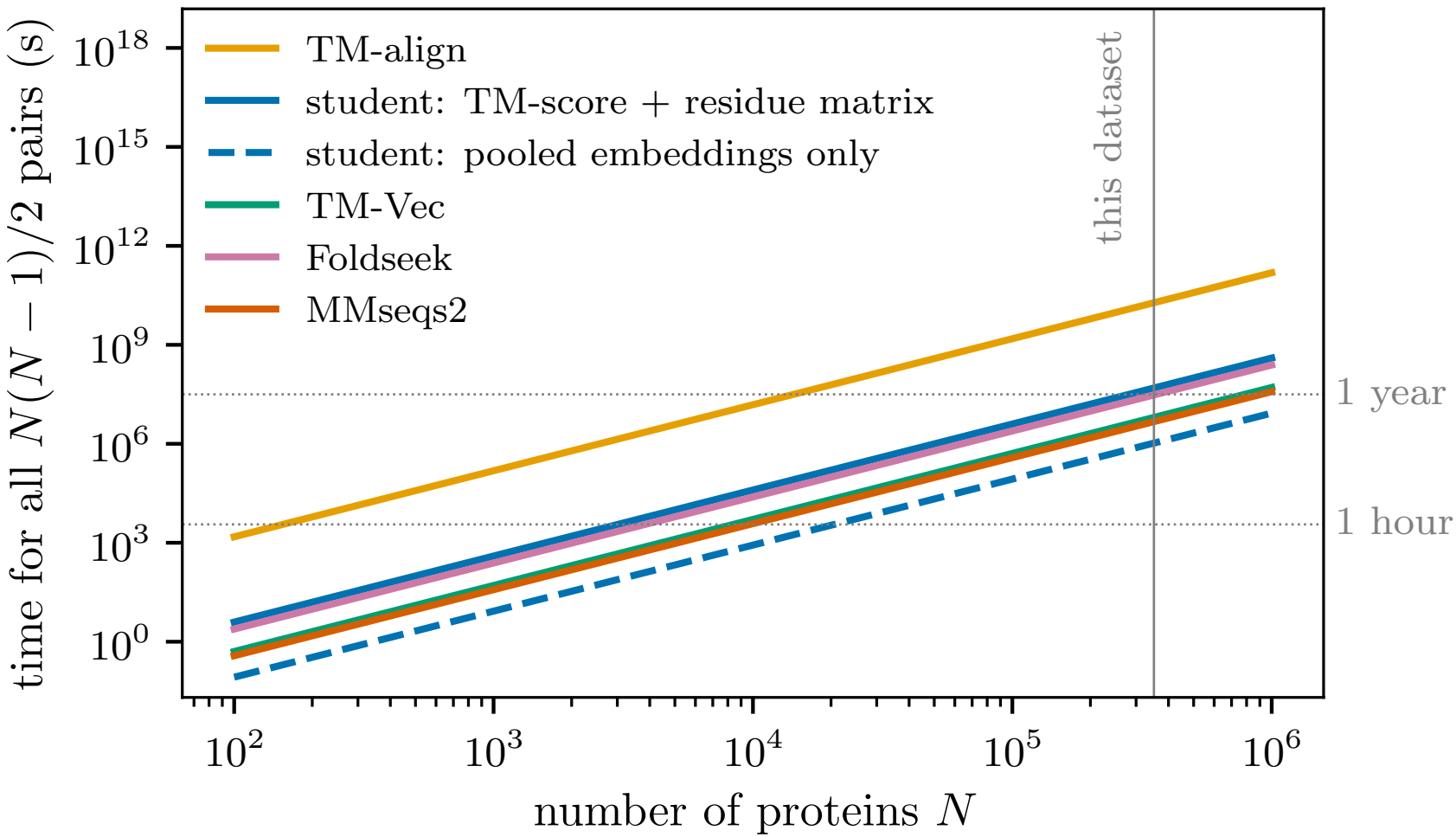




Speed: cost of one comparison



Speed: all-against-all comparison



Next steps

- Benchmark against MMseqs2, TM-Vec, DeepBLAST, pLM-BLAST and Foldseek (especially for search accuracy)
- Translate the residue similarity matrix into alignments with a decoder
- Test on mutant pairs

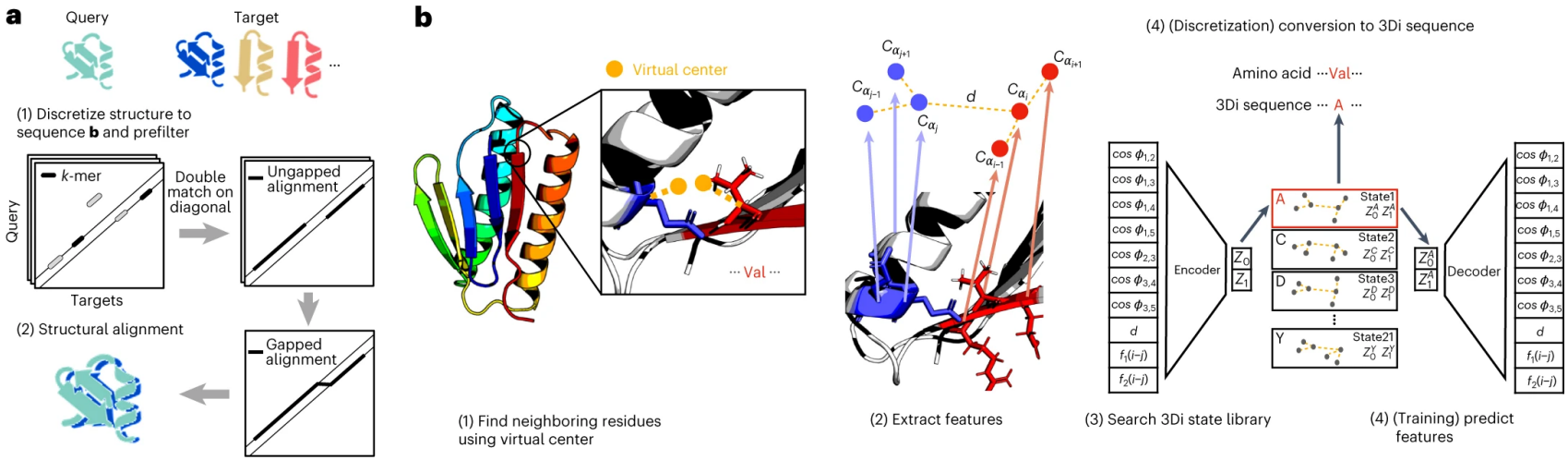


Image credit: van Kempen, M., et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2024).

Summary

- We created a local structural label, the **patch isometry score**, which doesn't require correspondence
- The **teacher** model learns to create embeddings from structural data that estimate both the local and global labels
- The **student** model utilizes only sequences, learning to distill both the teacher and the labels

Acknowledgements

I would like to sincerely thank the PRIMES program and the computational biology mentors for their guidance throughout the research process.

- Dr. Gil Alterovitz
- Dr. Shaojun Pei
- Dr. Ning Xie

End of Presentation

THANK YOU!