PoinnCARE: Hyperbolic Multi-modal Learning for Enzyme Classification

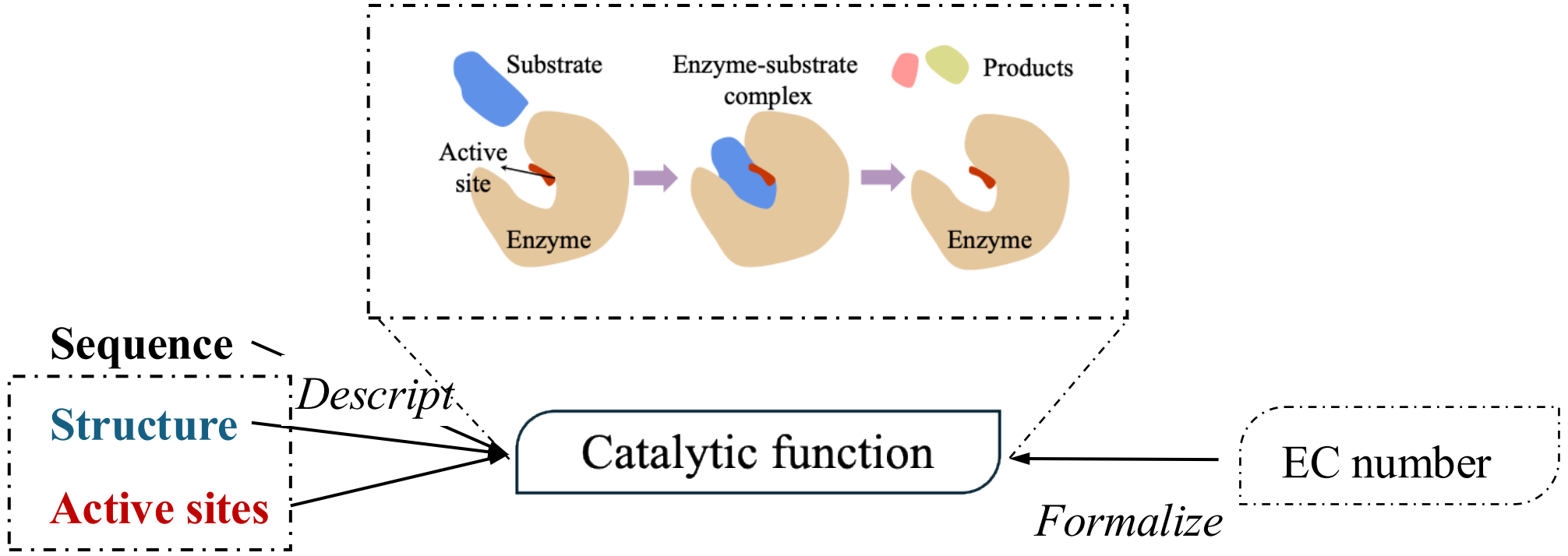
Kun Xie[1], Peng Zhou[2], Xingyi Zhang[3], Wei Liu[1],
Peilin Zhao[4], Sibo Wang[5], Biaobin Jiang[1]

[1]Tencent, [2]Hunan University, [3]MBZUAI,
[4]Shanghai Jiao Tong University, [5]The Chinese University of Hong Kong

Presenter: XIE Kun

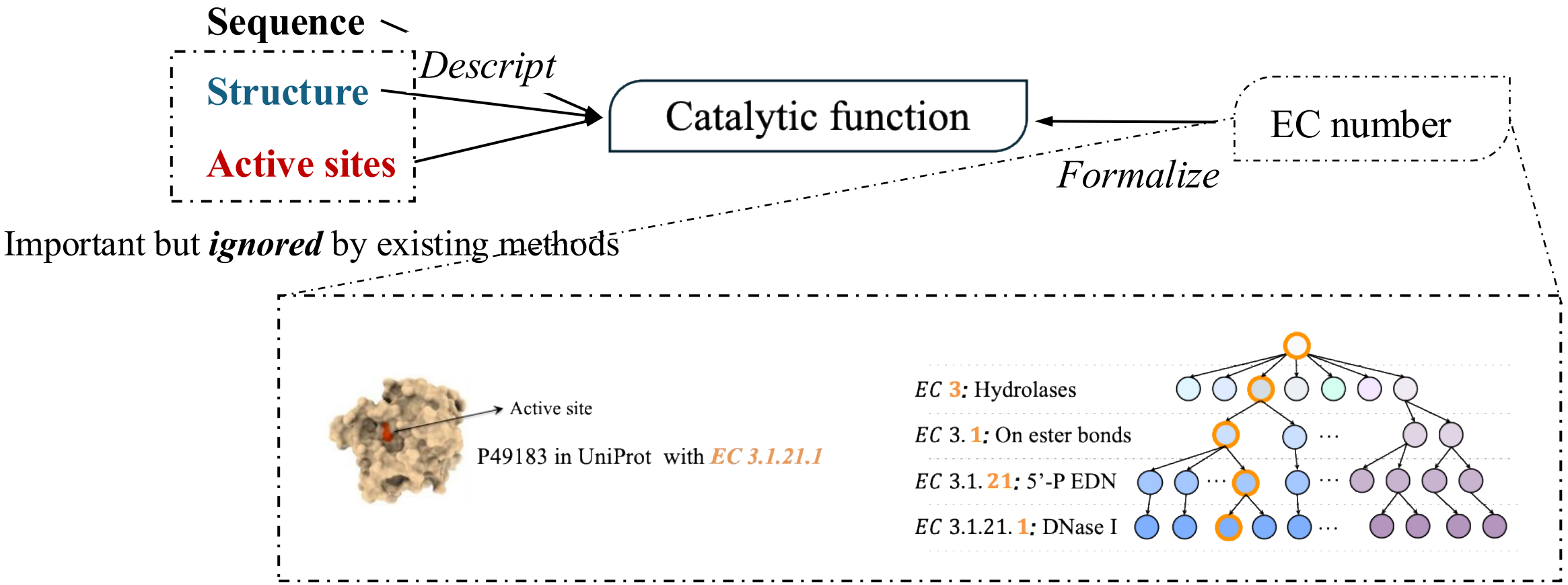
Enzyme & EC number

- Enzyme:
  - a protein catalyst
  - accelerates chemical reactions



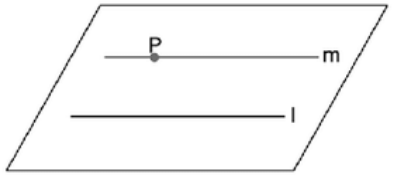
Important but ***ignored*** by existing methods

Enzyme & EC number

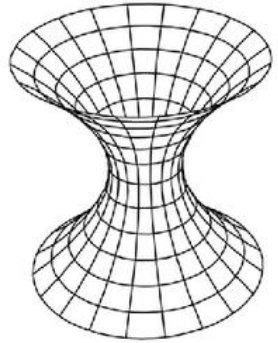


EC Number Prediction

- Input: Enzyme
- Output: EC number (≈ catalytic function)
- Challenge: ***Tree-like*** relationships among enzymes
  - Layer sizes grow ***exponentially***



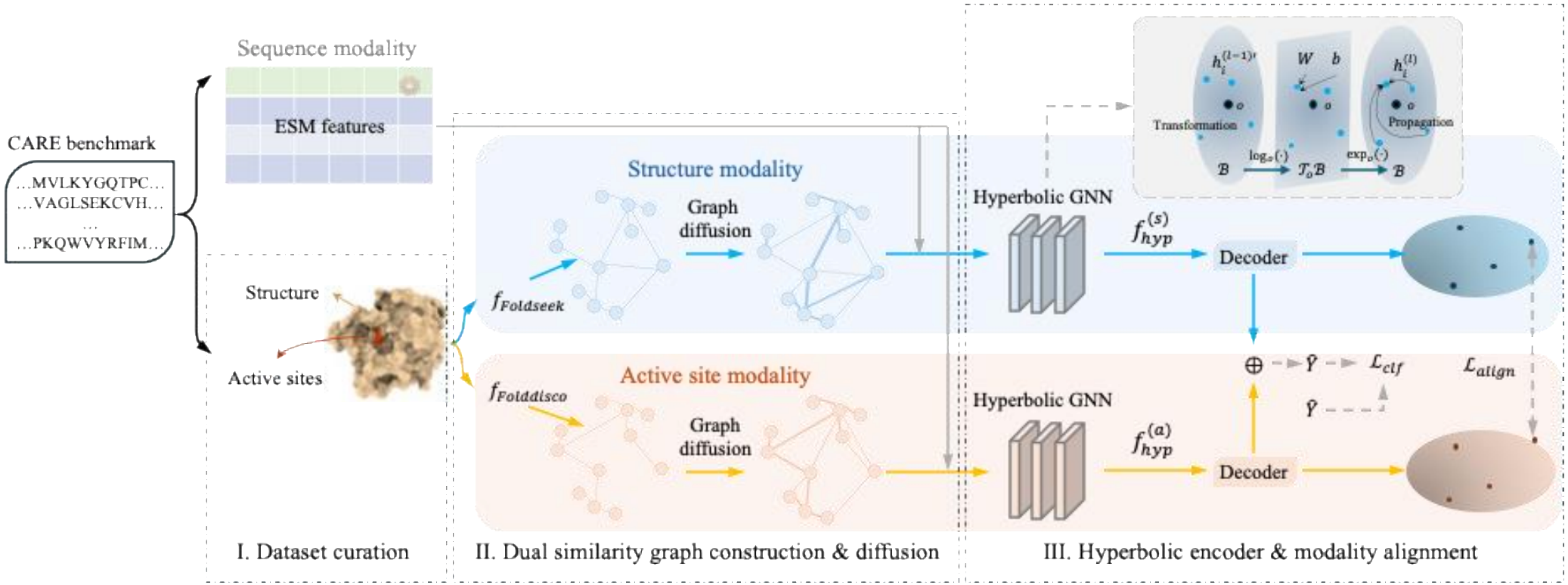
Euclidean space
In existing methods



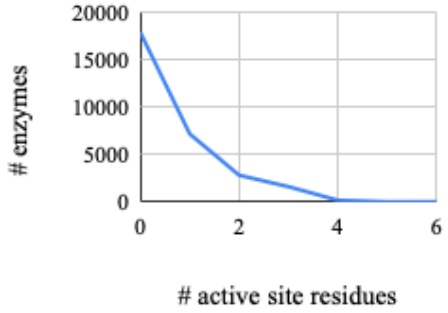
Hyperbolic space
Efficient and accurate

Method: PoinnCARE

- 1. Learning in hyperbolic space
  - GNN in Euclidean space → GNN in hyperbolic space

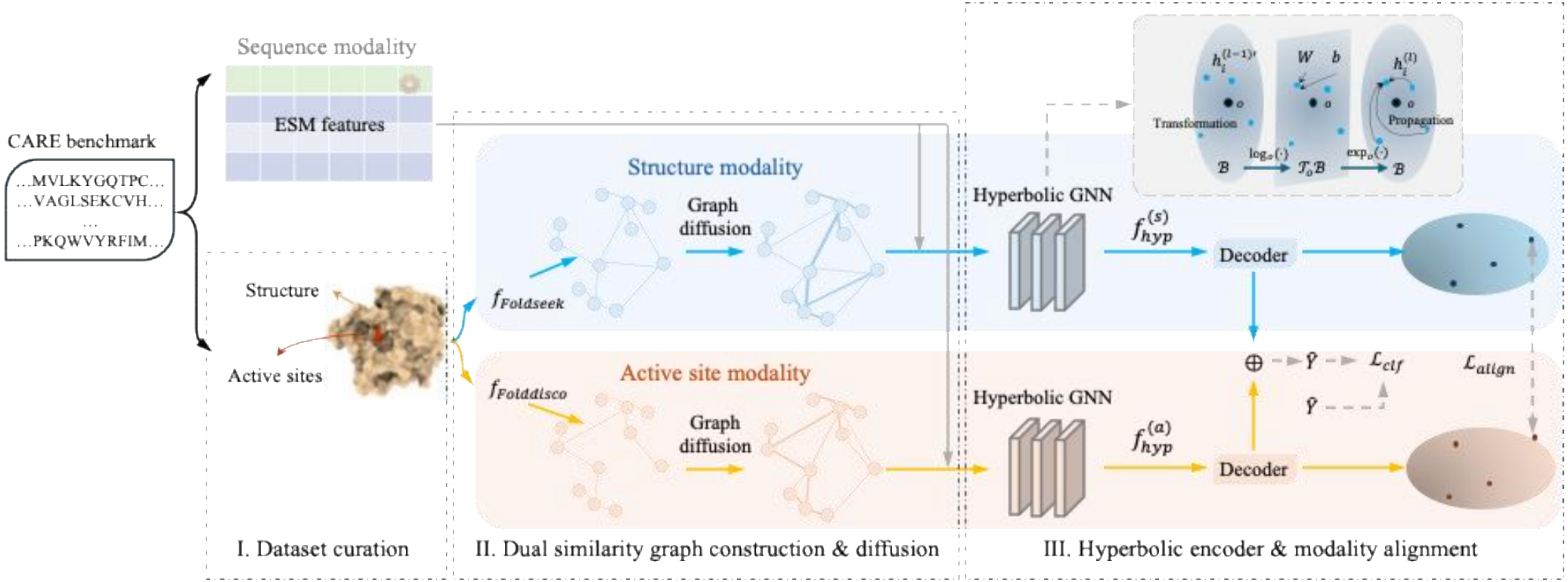


Method: PoinnCARE



- 2. Learning from multiple modalities
  - Dual similarity graphs for each modality, s.t.,
  - Graph diffusion to alleviate data sparsity
  - Multi-view graph learning to fuse different information

Sparse active site annotations
→ notable gap between modalities



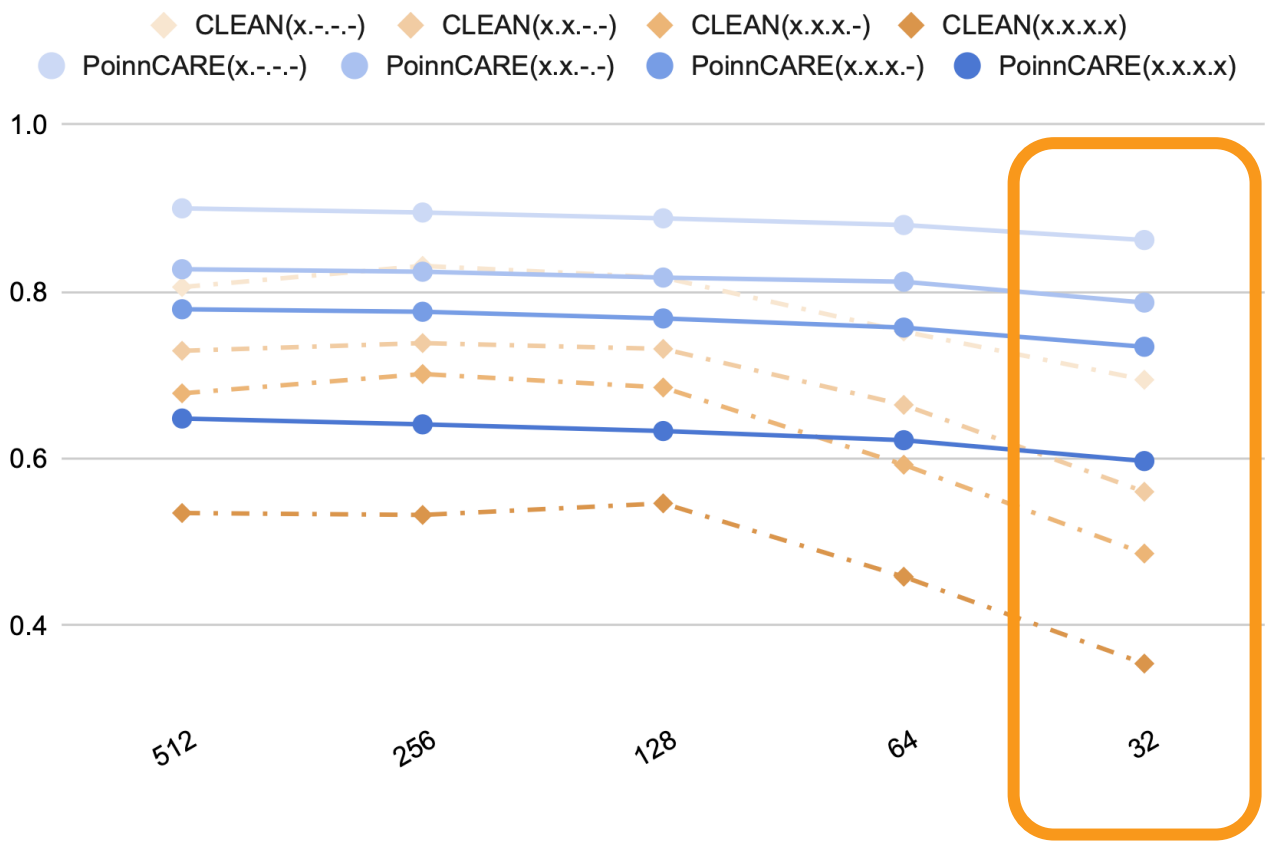
Experimental Results

- Systematic benchmark: CARE
- A broad range of SOTA baselines
- Comprehensive evaluation metrics

	<30% Identity				30-50% Identity				Avg. rank
	Level 1 (x.-.-.-)	Level 2 (x.x.-.-)	Level 3 (x.x.x.-)	Level 4 (x.x.x.x)	Level 1 (x.-.-.-)	Level 2 (x.x.-.-)	Level 3 (x.x.x.-)	Level 4 (x.x.x.x)	
Random*	0.194	0.032	0.012	0.000	0.225	0.036	0.007	0.000	13.38
BLASTp	0.697	0.590	0.569	0.475	0.923	0.879	0.850	0.773	7.00
Foldseek	0.815	0.722	0.667	0.544	0.932	0.880	0.841	0.755	4.13
Folddisco	0.756	0.600	0.511	0.378	0.798	0.755	0.723	0.564	9.88
CLEAN	0.806	0.729	0.678	0.535	0.946	0.905	0.870	0.798	2.50
CLEAN-Concat	0.810	0.704	0.646	0.507	0.946	0.893	0.859	0.777	3.63
ESM-2	0.783	0.695	0.643	0.518	0.944	0.895	0.856	0.781	4.13
ESM-c	0.691	0.574	0.527	0.436	0.911	0.848	0.808	0.745	9.25
ProtT5	0.755	0.649	0.604	0.492	0.929	0.873	0.833	0.765	6.38
ProtBert	0.672	0.546	0.502	0.410	0.874	0.804	0.767	0.709	10.38
S-PLM	0.751	0.637	0.582	0.470	0.921	0.861	0.823	0.752	7.75
ChatGPT*	0.278	0.016	0.000	0.000	0.336	0.030	0.014	0.000	13.38
Pika*	0.616	0.461	0.377	0.206	0.738	0.600	0.502	0.377	12.00
PoinnCARE	**0.900**	**0.827**	**0.779**	**0.648**	**0.961**	**0.926**	**0.887**	**0.822**	1.00

Experimental Results

- Efficient & Accurate
- High accuracy, even under *32* dimension



Conclusion

- Enhanced benchmark
  - += structural info
  - += active site info
- Similarity graphs
  - → Graph diffusion
- Hyperbolic
  - Multi-modality encoding
  - Multi-modality alignment
- Extensive experiments

Learn more ICLR