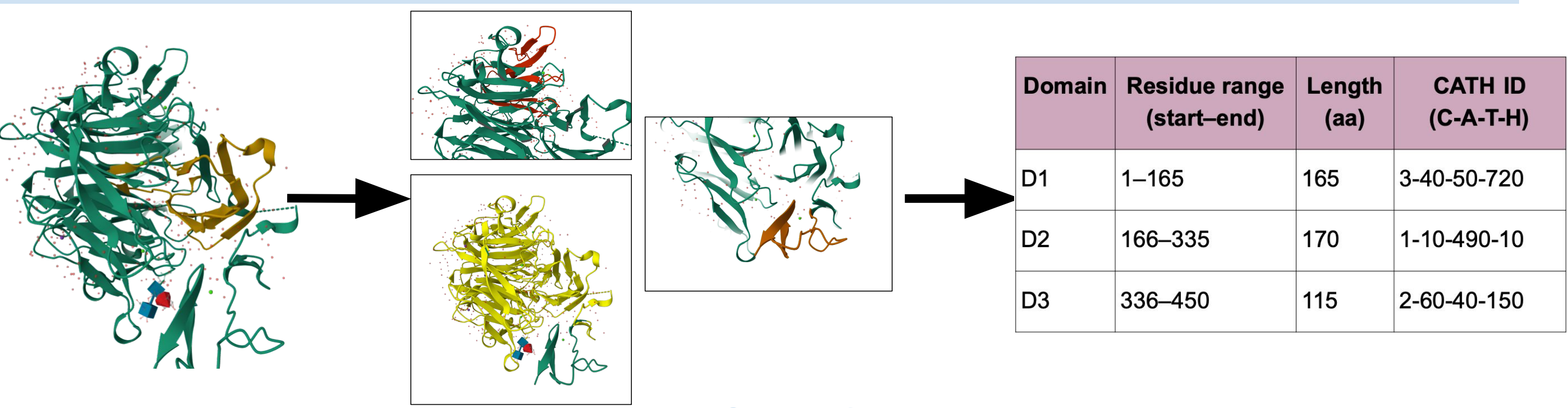


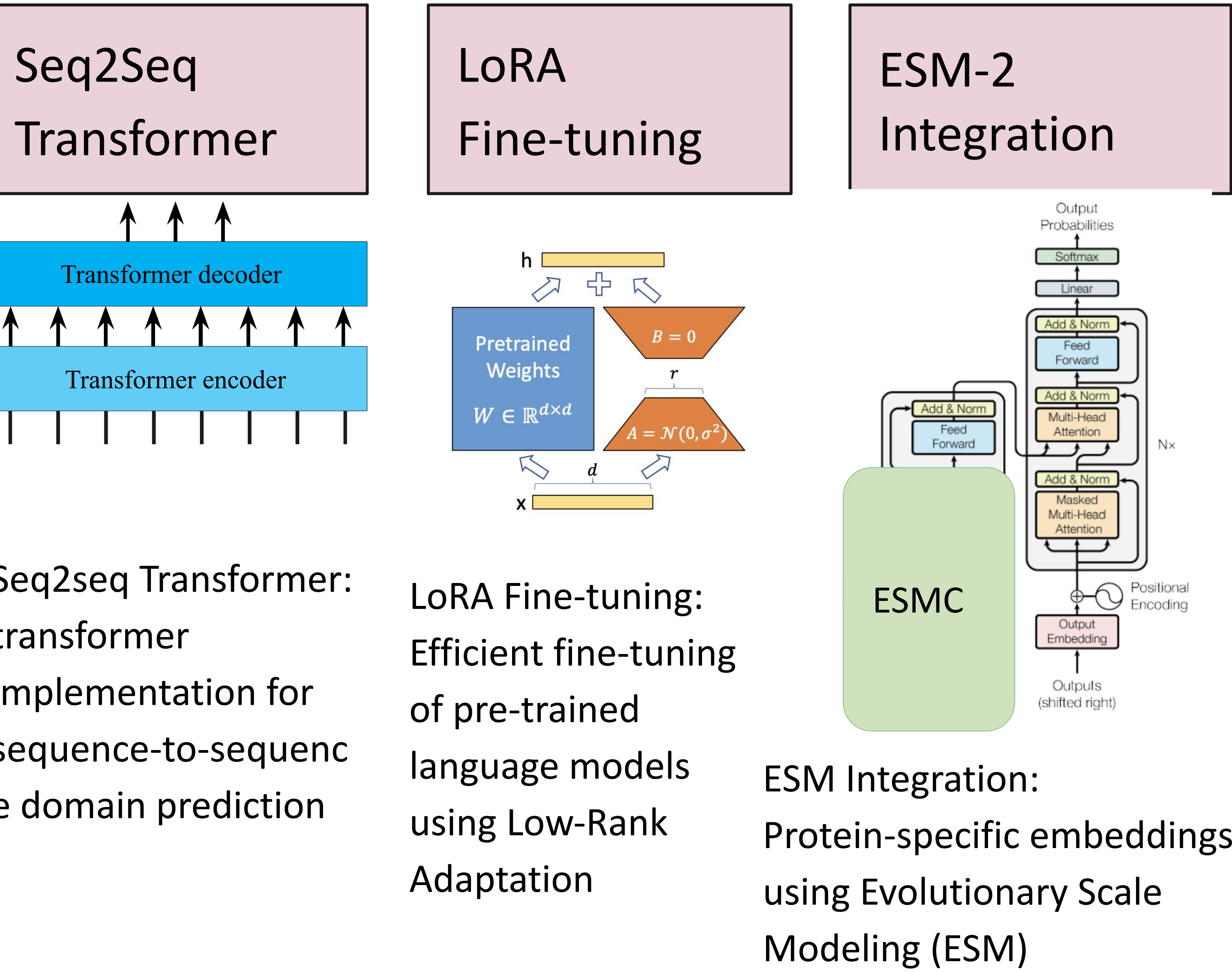
TED: Transformer-Based Protein Domain Prediction and CATH Classification

Motivation

Accurately predicting protein domains from raw amino acid sequences is essential for understanding a protein's functional and structural properties. We developed a method to predict domain boundaries and CATH ids in a end-to-end single model.



Method



Results

Method	num_domains_acc	iou_chain	CATH_acc
Seq2Seq Transformer	0.82	0.76	0.79
ESMC-decoder	0.68	0.61	0.64
LoRA-finetune T5	0.72	0.65	0.69

Table 2: Evaluation metrics across different methods. Higher values indicate better performance.

MDTFSFLITRNRPHLLKALNSVLNQTSINFPIVSDNSTNNETKEMVESLYLNQSNIIYKREREGYLIEHWNVKLVSEVRSEYFMIFHDDDDVMHPDMIRTMILNIIKKDHSFLAFIATNARIKKGRICYSKYGKEKIMRLKEARELFQMYAVGNWIPFPFSYIYSKSIADNMRFADKGGKYCDVAFLLDAFDLGNLAITKPYMDYYHDGQVSQFNDYFQRIKLINYYVRQKMRISPHDRPLVNFKLMNLYCWNNILVRNSDDLVSKKRFINCLKIFLKHRYMYVFKLLGCILREKMKCGKL

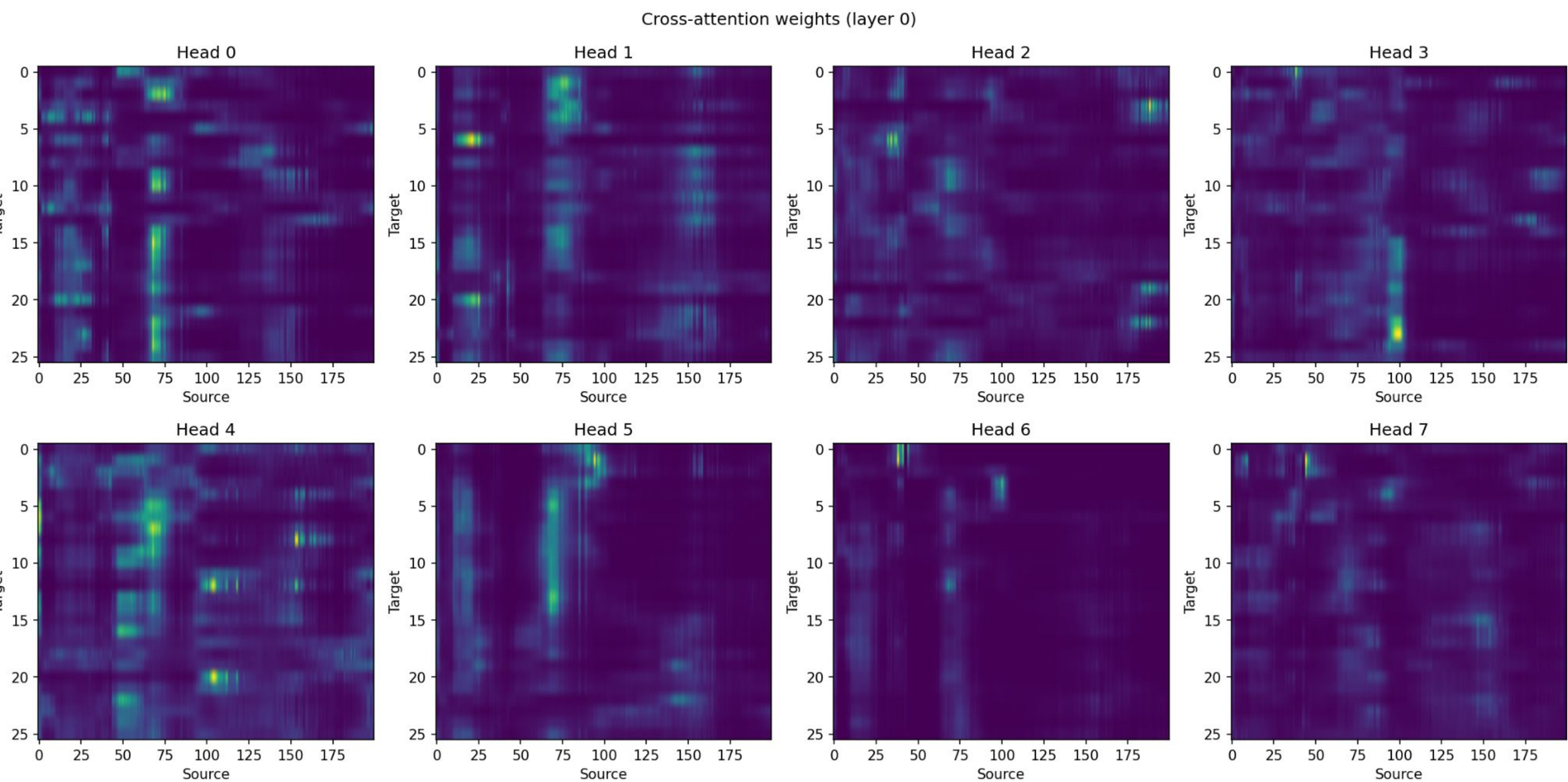
TARGET

5-218 | 3.90.550.10 * 240-305 | -

PREDICTION

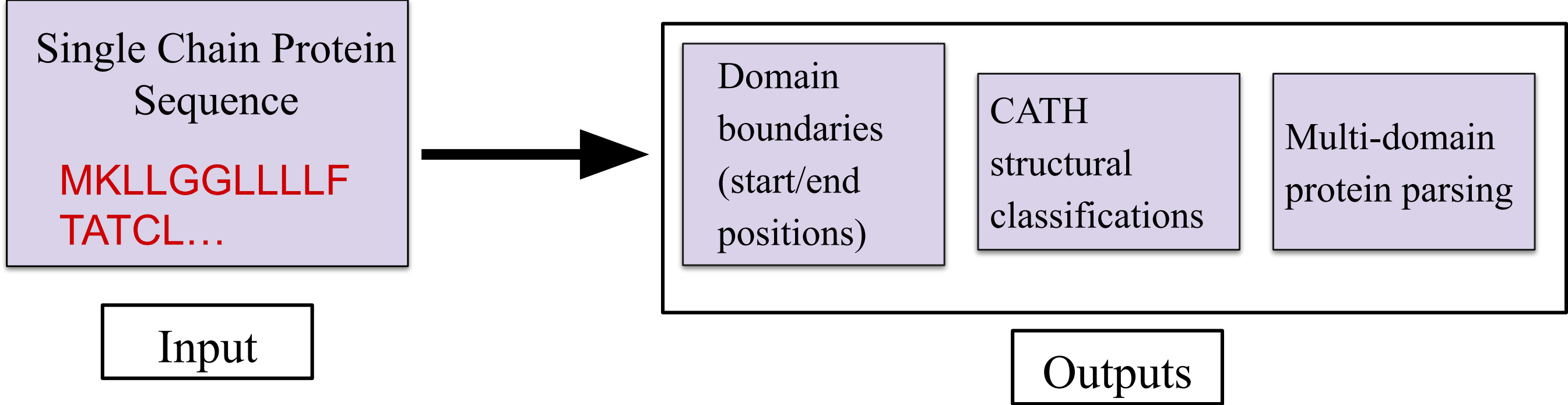
4-210 | 3.90.550.10 * 223-306 | -

Attention Weights and examples

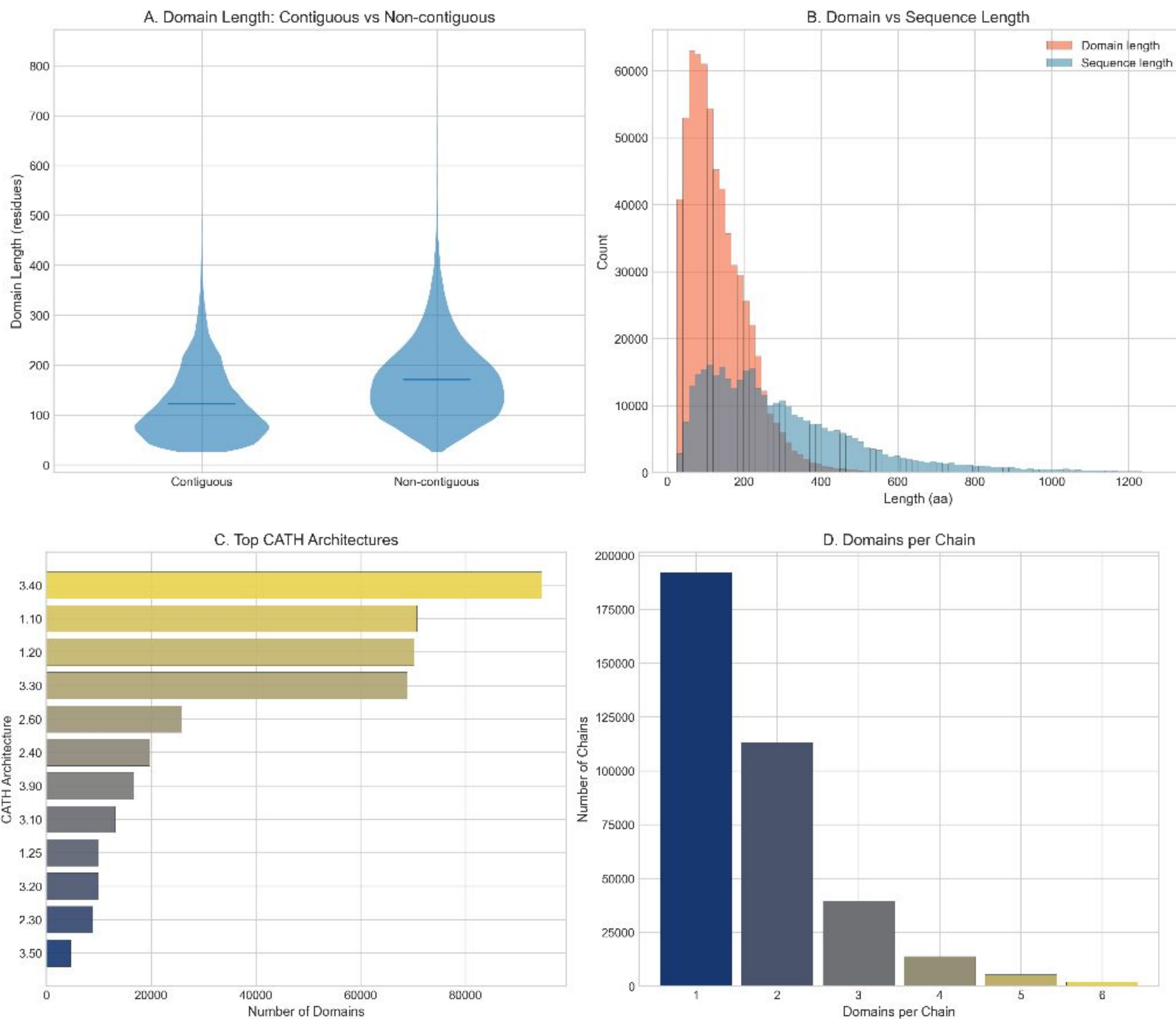


Data

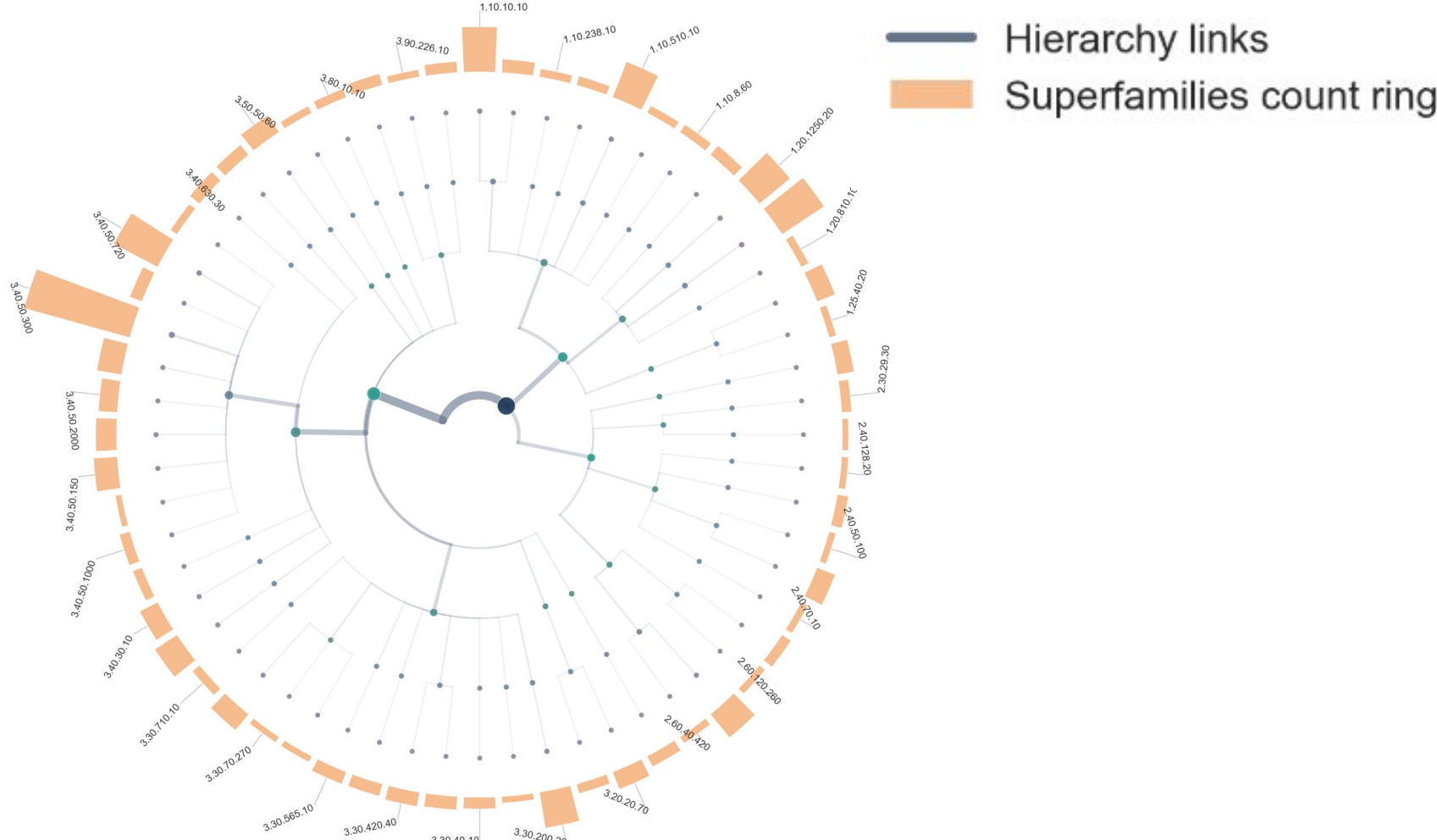
The model is trained on a dataset of approximately 186 million unique protein sequences downloaded from the TED database, which include precise annotations for domain boundaries and CATH classifications.



Exploratory Data Analysis of TED Domain Annotations



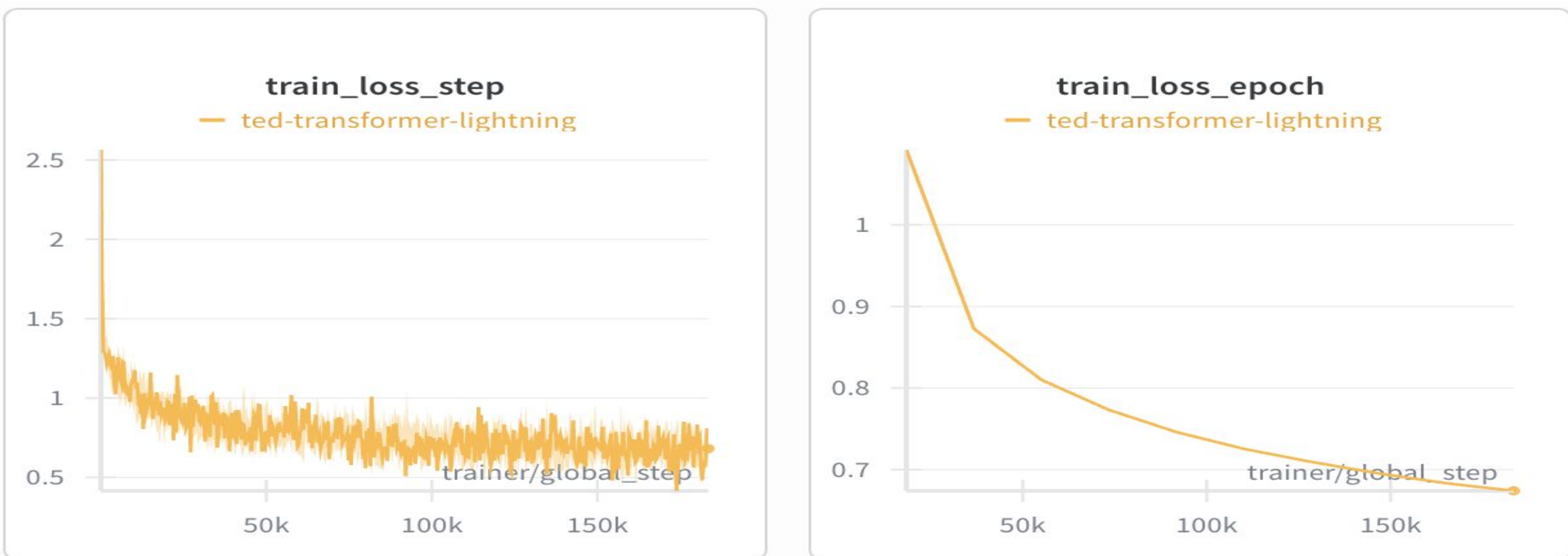
CATH Circular Hierarchy Overview - Top 60 Superfamilies



Frequency of CATH IDs ranking	CATH ID	Count
1	1-20-5	33,818
2	3-40-50-300	28,781
3	1-10-287	18,067
4	1-10-10-10	15,863
5	3-40-50-720	14,609
6	1-20-1250-20	10,893
7	3-40-190-10	10,517
8	1-10-10	8,345
9	1-20-210-10	7,918
10	3-40-50-2300	7,699

Metric	Value
# protein chains (sequences)	571,709
# domains total (all chains)	1,000,000
Mean domains per chain	1.75
Single-domain chains	293,204 (51.3%)
Multi-domain chains	278,505 (48.7%)
Max domains in a chain	14
Domains with unknown CATH	279,214 (27.9%)
Domains with known CATH	720,786 (72.1%)
# unique non-unknown CATH IDs	4,747

Training and Evaluation



Key Mission and Future Direction

Our research streamlines structural biology pipelines by predicting CATH annotations and precise domain boundaries directly from the sequence, eliminating the need for the external search tools that other models rely on.