Co-evolution Transformer for Protein Contact Prediction

He Zhang, Fusong Ju, Jianwei Zhu, Liang He, Bin Shao
Nanning Zheng, Tie-Yan Liu

Protein contact prediction

- Protein contact prediction is a binary classification task for amino acid residue pairs
- A residue pair is called a contact if their distance is less than or equal to a distance threshold
- Co-evolution information is closely correlated to the contacts
- To extract the co-evolution patterns for residue pairs, multiple sequence alignments (MSAs) are generated from raw protein sequences

Protein Contact Prediction

Unsupervised Methods: They are based on estimating the inter-residue contacts using hand-crafted co-evolutionary features derived from (MSAs).

- Statistical modeling methods based on direct coupling analysis (DCA) techniques are widely used to obtain co-evolutionary features
- DCA techniques only consider single-residue and pairwise statistics of the sequences, ignoring high-order interactions among the residues within a sequence
- The information loss caused by hand-crafted features (e.g., DCA-based features)

Supervised Methods: To mitigate information loss, several models are proposed to learn residue co-evolution information directly from the sequences in the MSAs

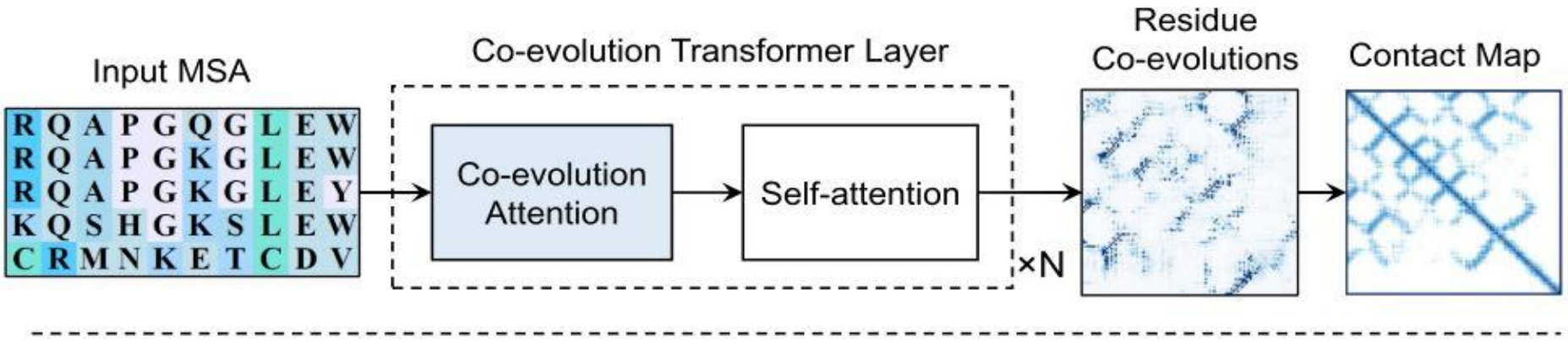
- Among them, CopulaNet, derives coevolutionary features differentially by aggregating the learned residue representations from the sequences
- They are capable of modeling the high-order interactions among the multiple residues within single sequence, the global information carried by the MSAs is ignored

Pre-training Based Methods: Pre-trained language models are adapted to representation learning for single protein sequences from the unlabeled data

- MSA transformer

Co-evolution Transformer (CoT)

- To exploit the co-evolution information from the **full MSAs effectively**
- Propose an attention-based architecture, Co-evolution Transformer (CoT)
- Co-evolution Transformer (CoT) is constructed by stacking several repeated CoT layers
- Each CoT layer is composed of two attention modules
  - co-evolution attention (CoA) module
  - self-attention module
- Given a prepared MSA, the stacked CoT layers are used to learn the residue representations.
- Residue co-evolutions are derived from the representations of the final layer and further employed to estimate the inter-residue contacts



Vanilla Transformer Encoder

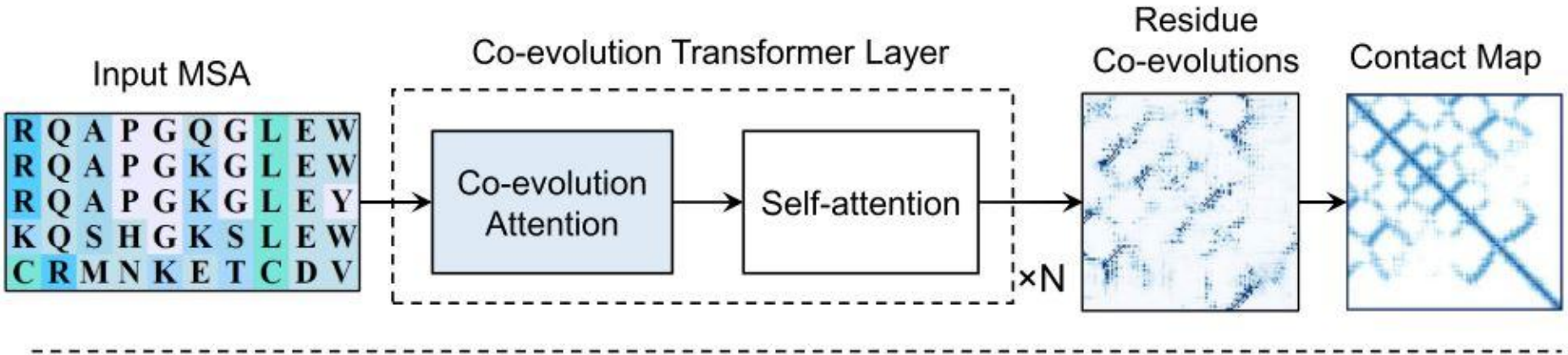
Each Transformer encoder layer consists of two modules

- multi-head self-attention (MHSA) module
- position-wise fully connected feed-forward (FFN) module

To connect these two modules, residual connections and layer normalization (LAYERNORM) are applied

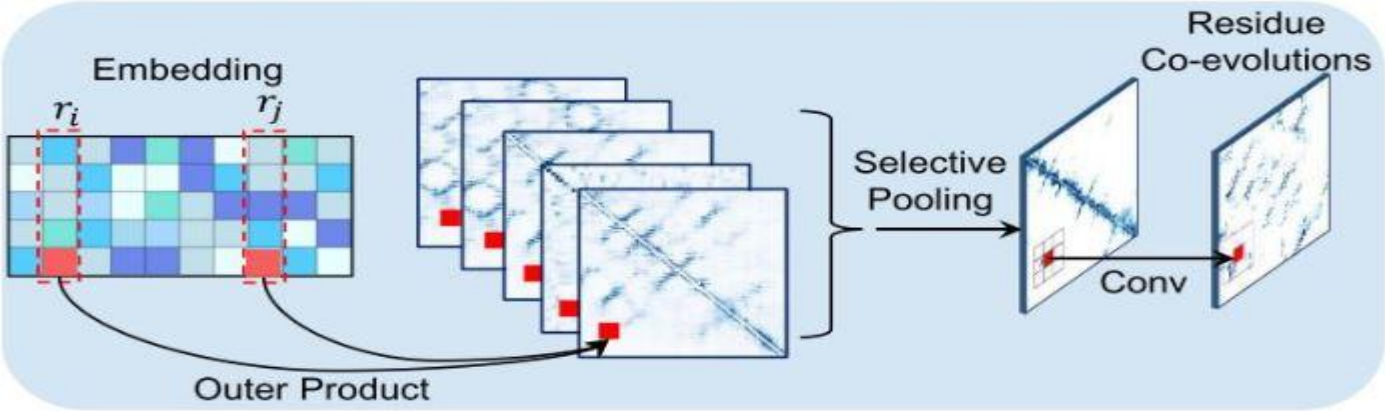
$$x = \text{LayerNorm}(x + \text{MHSA}(x)), \quad (1)$$

$$x = \text{LayerNorm}(x + \text{FFN}(x)), \quad (2)$$

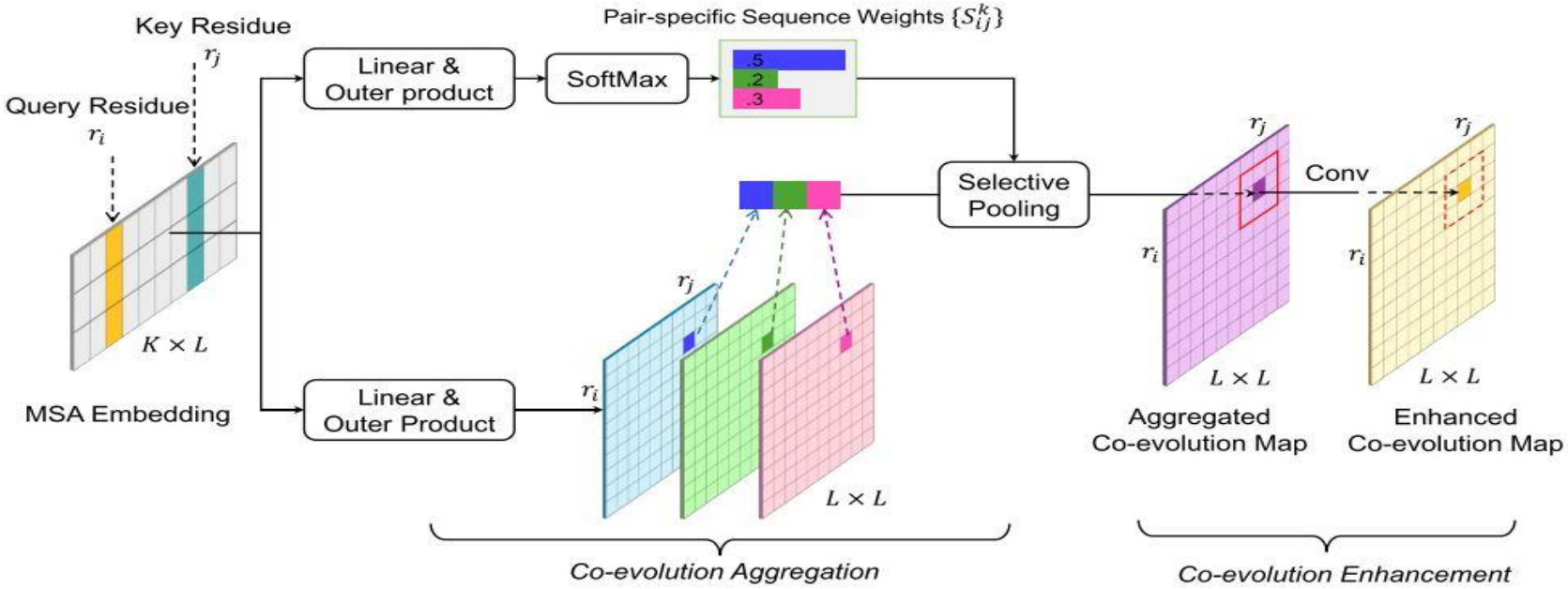


Co-evolution Attention Module (CoA)

- The goal of the Co-evolution Attention module is to leverage the whole MSA to derive pairwise inter-residue interaction features
- These features are used as attention maps to guide representation learning of each sequence in the MSA
- To achieve this goal, a CoA module employs two consequent submodules
    - co-evolution aggregation
    - co-evolution enhancement
- The co-evolution aggregation submodule is designed to generate the coevolutionary features by aggregating pairwise interactions from all homologs in the MSA
- The co-evolution enhancement submodule further enhances the coevolutionary features and derives the final co-evolution attention



Co-evolution Attention Module (CoA)



$$S_{ij}^k = \frac{1}{Z}\exp(Q_i^k \otimes Q_j^k), \qquad A_{ij} = \textsc{Proj}\left(\sum_{k=1}^{K} S_{ij}^k \odot (P_i^k \otimes P_j^k)\right) \qquad A = \textsc{Conv}([A'; A]).$$

$$\textsc{Attn}_h(X,\, A) = \textsc{SoftMax}(A\, M_h),$$

Co-evolution Attention

(AGGREGATE) and the co-evolution enhancement module (ENHANCE), can be summarized as

$$A_{ij} = \textsc{Aggregate}(X_i \otimes X_j), \tag{12}$$
$$A_{ij} = \textsc{Enhance}(A_{ij}, \textsc{Neighbor}(i, j)), \tag{13}$$
$$\mathcal{A}_{ij}^{k} \propto \textsc{SoftMax}\{A_{ij}\}, \tag{14}$$

1) CoA leverages the global information from all the homologs instead of single homolog to derive the attention, which fits better to the residue co-evolution insight

2) CoA handles the non-homologous information naturally by the selective pooling operation during aggregation, providing a solution to a widely existed but inevitable dilemma for MSAs;

3) CoA enhances the co-evolution signal by propagating information from the neighbors, making it easier to capture the high-order interactions among multiple residues.

Experiment Setup

Dataset:

- CASP14, CAMEO are two standard benchmark datasets
- CASP14 includes three kinds of protein domains
  - FM (22 domains), FM/TBM (14 domains), and TBM (50 domains)
- A domain is a protein sequence prepared by the CASP organizers.

Evaluation:

- Following the procedure of trRosetta
- Contact prediction task is converted into a multi-class classification task.
- The inter-residue distance range is divided into 37 bins, i.e., (0Å, 2.5Å], (2.5Å, 3.0Å], $\cdots$, (20.0Å, $+\infty$)
- Models are trained with the bin labels. For contact, the summed probability value of the bins with distance less than 8Å are used as the final prediction

Metrics

- Precision@L, Precision@L/2, and Precision@L/5 of long-range residue contacts
- Precision@n stands for the precision score for the top-n pairs of the highest probability in the predicted contact map.
- L refers to the length of protein sequence
- long-range means there are at least 23 other residues between these two residues in the sequence.

Results

Table 1: Comparison on CASP14 and CAMEO (*Precision@L*)

Methods	CASP14			CAMEO
	FM (22)	FM/TBM (14)	TBM (50)	Hard (176)
RaptorX [14]	33.9	58.1	63.1	53.2
trRosetta [15]	31.3	57.6	61.1	50.1
CopulaNet [2]	38.5	62.2	65.5	56.5
CoT-SA (ours)	41.8	59.2	67.9	59.8
CoT (ours)	**48.2**	**66.7**	**75.6**	**66.6**

Results

Table 2: Comparison on CASP14. Gr. 368, Gr. 488, and Gr. 010 are the results of the top-3 groups in the CASP14 challenge. CoT† refers to the results of CoT with MSA selection.

	FM (22)			FM/TBM (14)			TBM (50)		
Method	L	$L/2$	$L/5$	L	$L/2$	$L/5$	L	$L/2$	$L/5$
Gr. 368	41.8	55.7	66.6	64.5	78.6	87.4	73.1	87.1	94.5
Gr. 488	40.4	52.9	65.0	63.6	78.8	88.5	72.0	86.9	93.7
Gr. 010	39.6	53.4	63.8	61.5	77.0	86.8	66.1	80.9	89.5
CoT† (ours)	**51.6**	**68.2**	**79.9**	**66.8**	**82.2**	**90.5**	**77.9**	**91.0**	**96.1**

Results

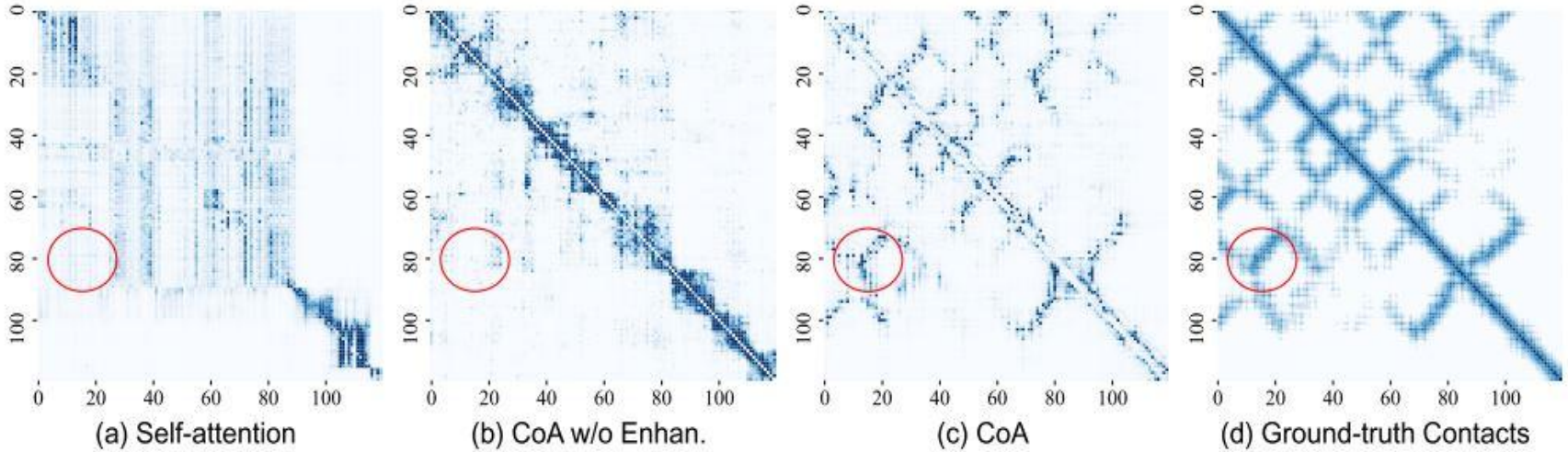


Figure 3: Comparison of the extracted attention maps for 4q2z_H. (a) CoT-SA model; (b) CoT model w/o co-evolution enhancement; (c) CoT model; (d) Ground-truth contact map. The red circle covers typical long-range contacts.

Ablation study

Table 3: Ablations for CoA on CASP14 (*Precision@L*). AGGRE., ENHAN. and SA refer to the co-evolution aggregation submodule, the co-evolution enhancement submodule, and the self-attention module, respectively.

			CASP14			CAMEO
AGGRE.	ENHAN.	SA	FM (22)	FM/TBM (14)	TBM (50)	Hard (176)
Selective Pooling	✓	✓	48.2	66.7	75.6	66.6
Average Pooling	✓	✓	42.7	62.2	73.6	64.2
Selective Pooling		✓	41.6	61.2	70.3	61.8
Selective Pooling	✓		46.4	66.9	74.8	66.3

Ablation Study

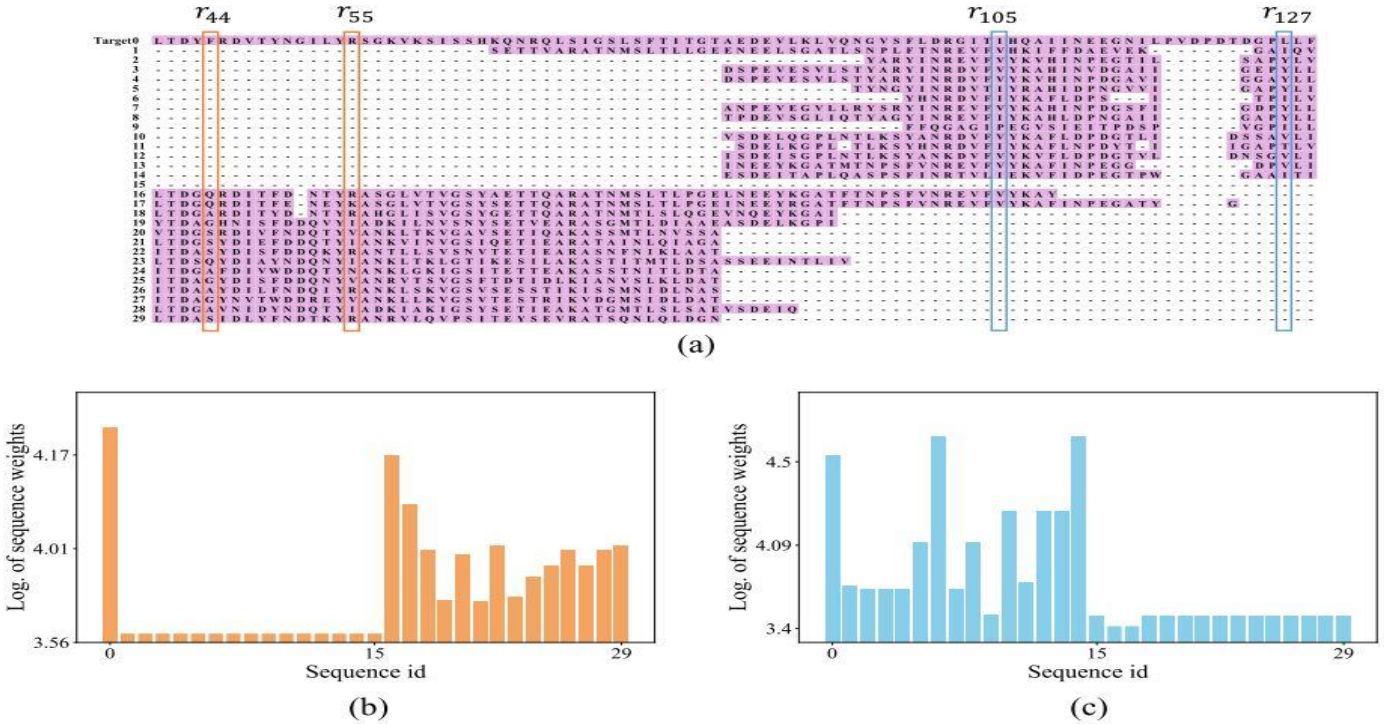


Figure 4: Learned sequence weights for T1061-D1. (a) Part of MSA for T1061-D1 (residue 40-130), where the purple region represents aligned residues and '-' stands for a gap. (b) The sequence weight distribution for the pair $\langle r_{44}, r_{55} \rangle$. (c) The sequence weight distribution for the pair $\langle r_{105}, r_{127} \rangle$.

My thoughts

- They mention the use the entire MSA efficiently, but if the length is very long, we get a lot matrices
- Formulas are not good
- They don't mention whether they train the enhancement module