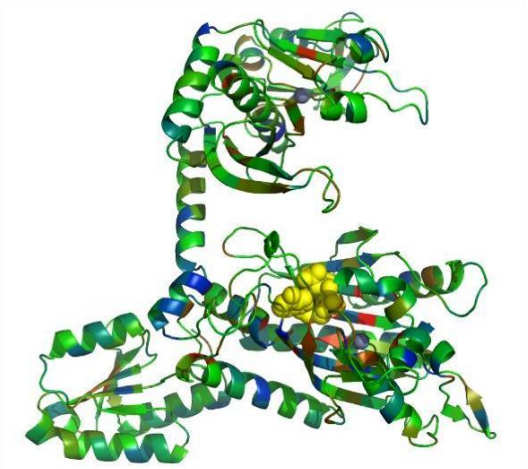
Retrieval-Augmented Language Model for Knowledge-Aware Protein Encoding

Jiasheng Zhang, Delvin Ce Zhang, Shuang Liang, Zhengpin Li, Rex Ying, Jie Shao

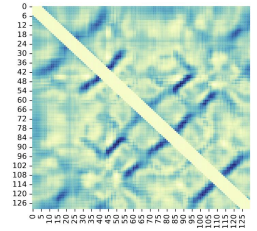
University of Electronic Science and Technology of China
Yale University

ICML 2025

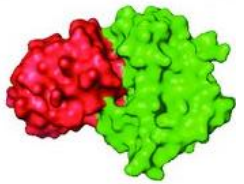
Background



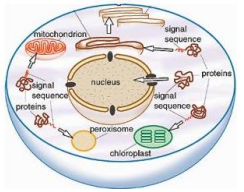
Protein Encoder



Amino Acid Contact Prediction



Protein-Protein Interaction Identification

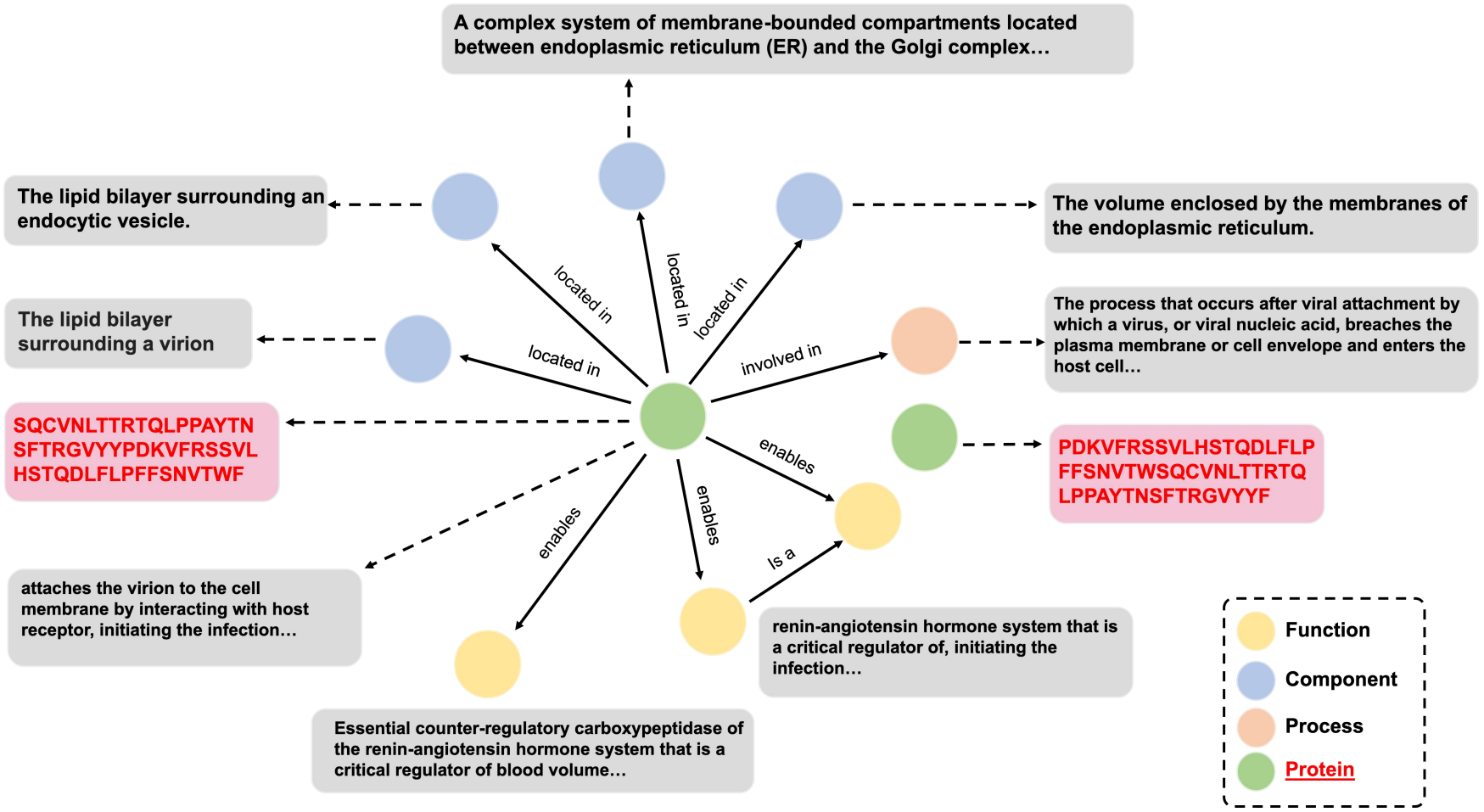


Protein Function Prediction

Protein representation learning has proven highly valuable in various application tasks such as protein-protein interaction identification and function prediction.

However, lacking factual knowledge (e.g., gene descriptions) makes existing models struggle to capture biological function encoded within protein sequences.

Background

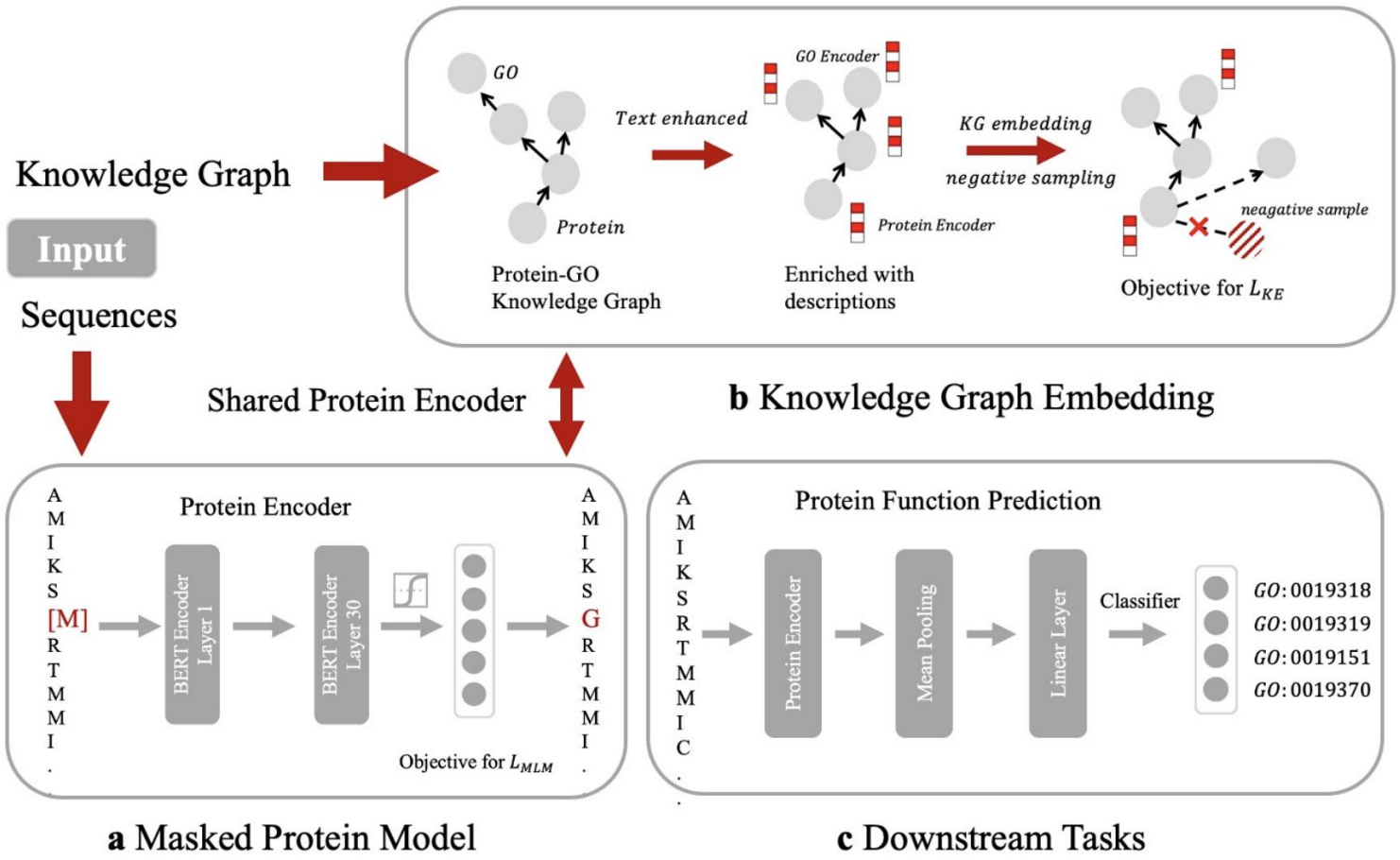


Recent efforts have introduced protein knowledge graphs, such as ProteinKG25, to incorporate prior biological knowledge into protein embeddings

PKGs describe the relationships between proteins and gene ontology (GO) entities with biological relations

Limitation of Existing Methods

OntoProtein [1]



Uses the TransE objective to optimize the alignment between protein embeddings and associated GO entity embeddings

Two pretraining objectives are used: Masked Language Modeling to learn protein sequences.
Triplet Knowledge task uses protein embedding to predict its function entities.

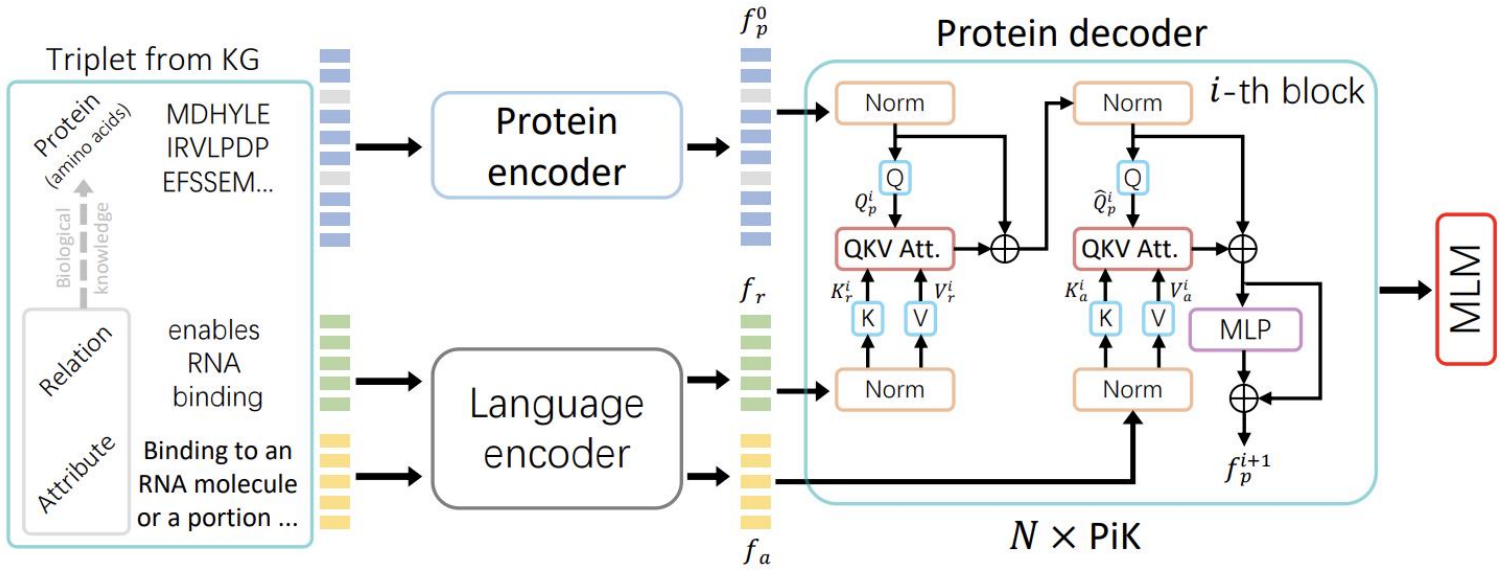
(1) Lack of structural information within PKGs, which describes high-order relationships between proteins.

(2) Storing knowledge within model parameters fails to precisely embed knowledge information.

[1] Ontoprotein: Protein pre-training with gene ontology embedding

Limitation of Existing Methods

KeAP [1]



uses GO entity representations to guide masked token prediction of protein sequences via a cross-attention mechanism.

After feature extraction, representations of each triplet are sent to the protein decoder for reconstructing missing amino acids.

(1) Unable to incorporate knowledge modeling during task fine-tuning, leading to inconsistent optimization objectives between the pre-training and fine-tuning stages. This inconsistency can cause the knowledge learned during pre-training to be catastrophically forgotten when applied to downstream tasks

[1] Protein Representation Learning via Knowledge Enhanced Primary Structure Reasoning

Catastrophic Forgetting of Existing Methods

	After Pre-training		After Task Fine-tuning	
Models	Precision	Similarity	Precision	Similarity
OntoProtein	0.712	0.901	0.621	0.632
KeAP	0.705	0.918	0.645	0.677

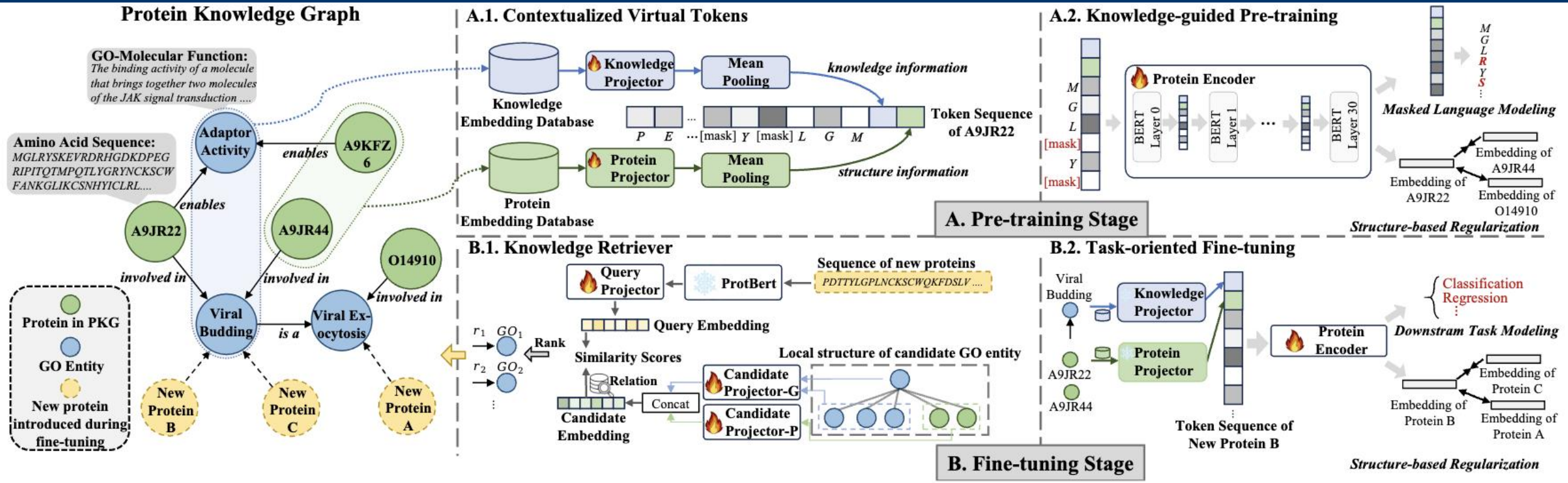
Similarity: semantic similarity between the embeddings of two proteins with the same attribute knowledge—a higher cosine similarity indicates better retention of knowledge information.

Precision: accuracay of the model to identify, from a set of candidate proteins, the one sharing attribute knowledge with a given protein—a higher accuracay indicates better retention of knowledge information.

OntoProtein, KeAP perform well after pretraining, confirming their ability to learn attribute knowledge.

After fine-tuning on downstream tasks, OntoProtein and KeAP show significant drops, indicating that they lose some of the knowledge acquired during pretraining.

The Proposed Framework Kara

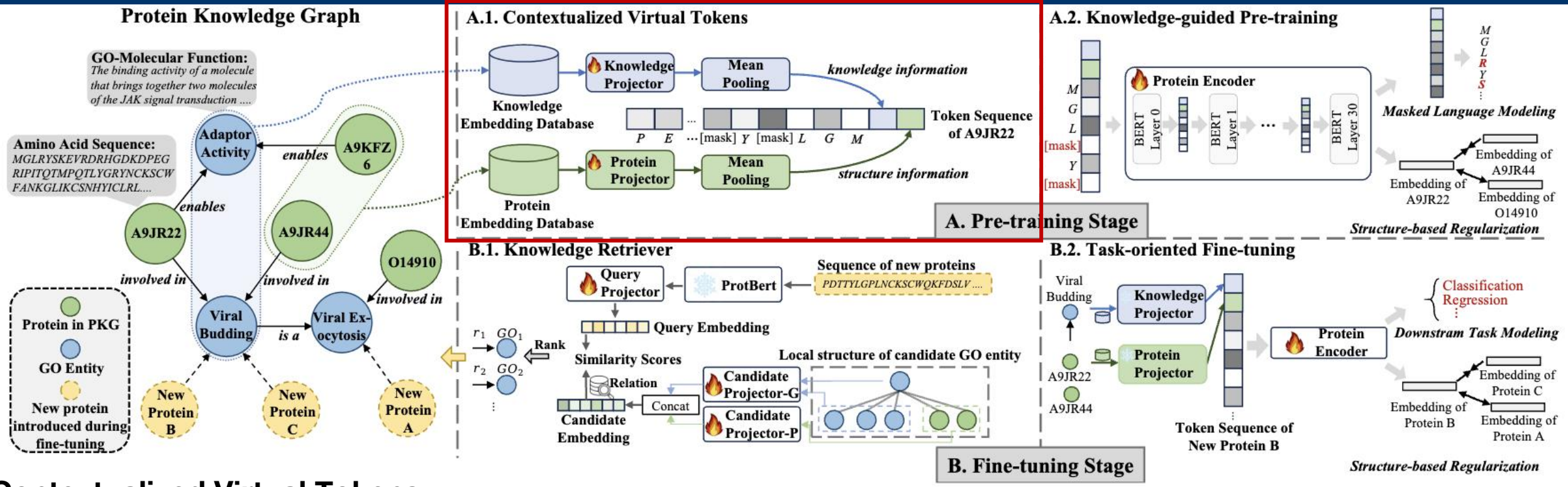


Implicitly embed knowledge information: Employ contextualized virtual tokens, achieving token-level information fusion between protein sequence and knowledge.

Structural Information: Proposes a sructure-based regularization, bringing function similarities into protein representations.

Catastrophic Forgetting: Using a knowledge retriever to predict potential gene descriptions for new proteins. Unifying the knowledge modeling process of the pre-training and fine-tuning stages

The Proposed Framework Kara



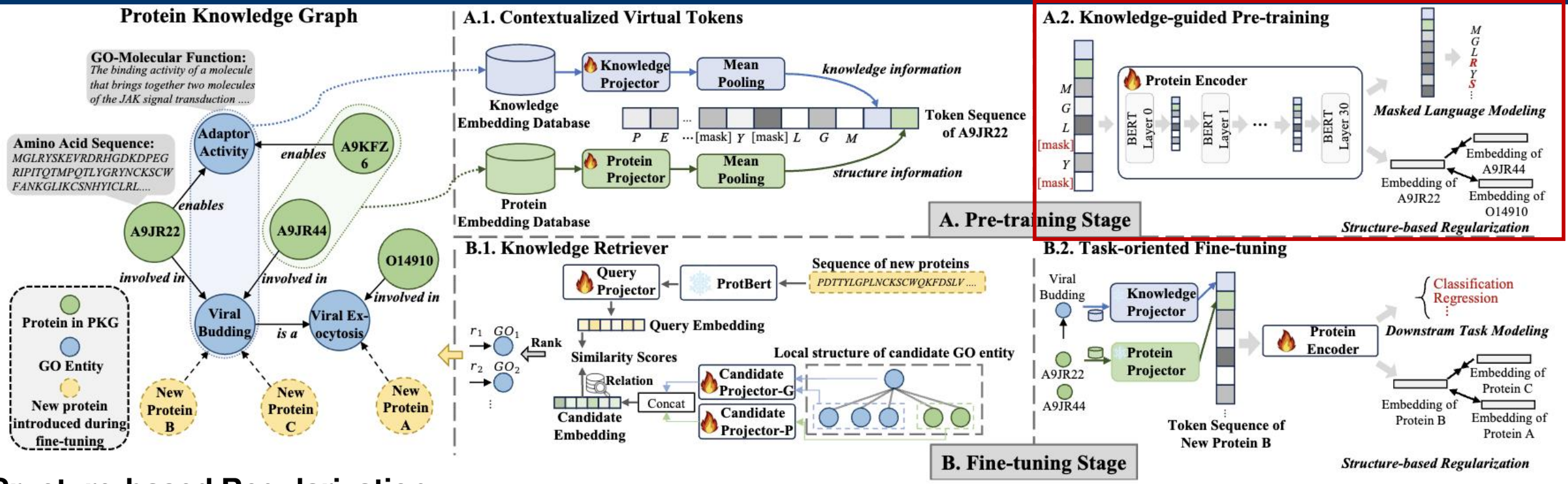
Contextualized Virtual Tokens

$$\mathbf{S}_i \leftarrow [\mathbf{v}_i^k, \mathbf{v}_i^p, \mathbf{S}_i] \in \mathbb{R}^{(2+|s_i|)\times d}.$$

Knowledge information token and struture information token

By summarizing the associated knowledge of a protein as knowledge virtual tokens and summarizing its high-order structure as structure virtual tokens, Kara can directly inject the knowledge and graph information into protein representations, instead of reserving knowledge within model parameters.

The Proposed Framework Kara



Sructure-based Regularization

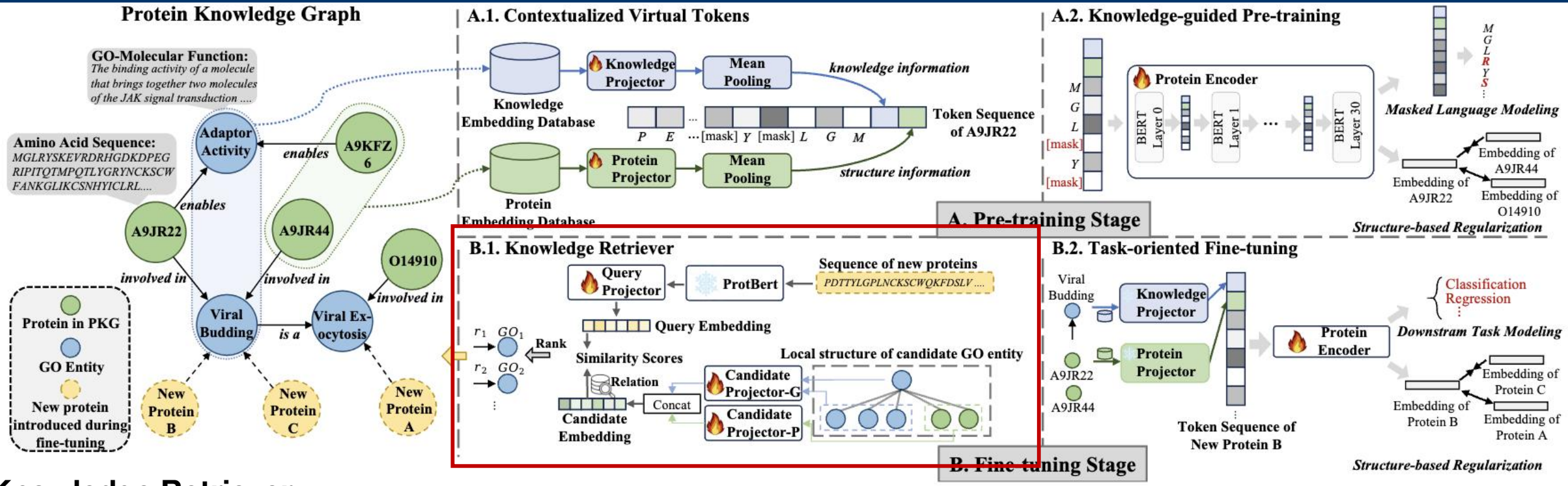
$$\mathcal{L}_{\text{reg}} = -\frac{1}{|\mathcal{N}_2(p_i)|} \sum_{p_j \in \mathcal{N}_2(p_i)} \text{MAX}(0, \text{sim}(\tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_j) - \text{sim}(\tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_k) + \gamma),$$

Protiens that 2-hop connected with the same relation.

Protiens without connections.

High-order connectivity indicates that two proteins share the same knowledge and thus should be similar in their biological functions. The structure-based regularization can integrate biological function similarities into their representations.

The Proposed Framework Kara



Knowledge Retriever

Unify the semantic space of different modalities based on multi-modal matching loss:

$$\mathcal{L}_{match} = \mathrm{MAX}(0, ||\mathrm{MLP}_G(\mathbf{g}_i) - \mathrm{MLP}_P(\mathbf{g}_i^{prot})||_1 - ||\mathrm{MLP}_G(\mathbf{g}_i) - \mathrm{MLP}_P(\mathbf{g}_j^{prot})||_1 + \gamma).$$

Reduce the retrieval complexity by finding relation-GO combinations:

$$\mathcal{E}(r_m) = \{g_m | (p_x, r_m, g_m) \in F\}$$

Experimental Results

Amino Acid Contact Prediction

	$6 \leq seq \leq 12$			$12 \leq seq \leq 24$			$24 \leq seq$		
Models	P@L	P@L/2	P@L/5	P@L	P@L/2	P@L/5	P@L	P@L/2	P@L/5
LSTM	0.26	0.36	0.49	0.20	0.26	0.34	0.20	0.23	0.27
ResNet	0.25	0.34	0.46	0.28	0.25	0.35	0.10	0.13	0.17
Transformer	0.28	0.35	0.46	0.19	0.25	0.33	0.17	0.20	0.24
ProtBert	0.30	0.40	0.52	0.27	0.35	0.47	0.20	0.26	0.34
ESM-1b	0.38	0.48	0.62	0.33	0.43	0.56	0.26	0.34	0.45
ESM-2	0.40	0.50	0.62	0.35	0.44	0.56	0.27	0.35	0.45
OntoProtein	0.37	0.46	0.57	0.32	0.40	0.50	0.24	0.31	0.39
KeAP	0.41	0.51	0.63	0.36	0.45	0.54	0.28	0.35	0.43
Kara	**0.45**	**0.55**	**0.65**	**0.39**	**0.48**	**0.59**	**0.31**	**0.39**	**0.48**

Homology Detection and Stability Prediction

Models	Homology	Stability
LSTM	0.26	0.69
ResNet	0.17	0.73
Transformer	0.21	0.73
ProtBert	0.29	0.78
ESM-1b	0.11	0.77
ESM-2	0.13	0.80
OntoProtein	0.24	0.75
KeAP	0.29	0.80
Kara	**0.32**	**0.83**

Protein-Protein Interaction Identification

	SHS27K			SHS148K			STRING		
Models	BFS	DFS	Avg	BFS	DFS	Avg	BFS	DFS	Avg
DNN-PPI	48.09	54.34	51.22	57.40	58.42	57.91	53.05	64.94	59.00
DPPI	41.43	46.12	43.77	52.12	52.03	52.08	56.68	66.82	61.75
PIPR	44.48	57.80	51.14	61.83	63.98	62.91	55.65	67.45	61.55
GNN-PPI	63.81	74.72	69.27	71.37	82.67	77.02	78.37	91.07	84.72
ProtBert	70.94	73.36	72.15	70.32	78.86	74.59	67.61	87.44	77.53
ESM-1b	74.92	78.83	76.88	77.49	82.13	79.31	78.54	88.59	83.57
ESM-2	75.05	79.55	77.30	77.19	83.34	80.26	81.32	89.19	85.30
OntoProtein	72.26	**78.89**	75.58	75.23	77.52	76.38	76.71	91.45	84.08
KeAP	78.58	77.54	78.06	77.22	84.74	80.98	81.44	89.77	85.61
Kara	**81.18**	78.85	**80.01**	**79.62**	**86.02**	**82.82**	**82.73**	**92.46**	**87.59**

Experiments in 6 representative tasks show the effectiveness of Kara. It outperforms powerful baselines across all the tasks. For instance, Kara exceeds the state-of-the-art knowledge-enhanced model KeAP by 11.6% in the long-range contact prediction and by 10.3% in the protein homology detection

Experimental Results

Catastrophic Forgetting

	After Pre-training		After Task Fine-tuning	
Models	Precision	Similarity	Precision	Similarity
OntoProtein	0.712	0.901	0.621	0.632
KeAP	0.705	0.918	0.645	0.677
w/o structure-based regularization	0.722	0.906	0.624	0.749
w/o contextualized virtual tokens	0.713	0.902	0.676	0.816
Kara	0.738	0.934	0.725	0.968

After fine-tuning, Kara's performance remains stable. Furthermore, removing the structure loss or virtual token leads to performance degradation after fine-tuning, highlighting the importance of unified knowledge integration in mitigating catastrophic forgetting.

Performance with Incomplete Protein Knowledge Graph

Models	Contact ↑ $(6 \leq seq \leq 12)$	Homology ↑	Stability ↑	Affinity ↓
OntoProtein (full KG)	0.460	0.240	0.750	0.590
KeAP (full KG)	0.510	0.290	0.800	0.520
Kara (50% KG)	0.540	0.316	0.823	0.511
Kara (70% KG)	**0.546**	**0.322**	**0.828**	**0.503**

Kara with incomplete PKG still outperforms OntoProtein and KeAP with full PKG, showing Kara's robustness. We attribute the outperformance to knowledge retriever and virtual tokens, which well integrate knowledge into model learning.