

# Multi-modal Knowledge-enhanced Foundation Model for Generation, Retrieval, and Reasoning of Molecules and Text

Delvin Ce Zhang, Menglin Yang, Hady W. Lauw, Hua Xu,  
and Rex Ying

Being reviewed by ICML-24

# Content

- Motivation
- Molecule Captioning
- Molecule-based Text Retrieval
- Experiments

# Motivation

- Understanding molecules' chemical properties and functions is important.
- These chemical functions are usually represented in the form of **text data**.
- It is important to automatically generate or retrieve text given molecule.

ChEBI Name	acrylamide
ChEBI ID	CHEBI: 28619
Definition	A white odorless solid, soluble in water and several organic solvents. It is a member of the class of <b>acrylamides</b> that results from the formal condensation of <b>acrylic acid</b> with <b>ammonia</b> .

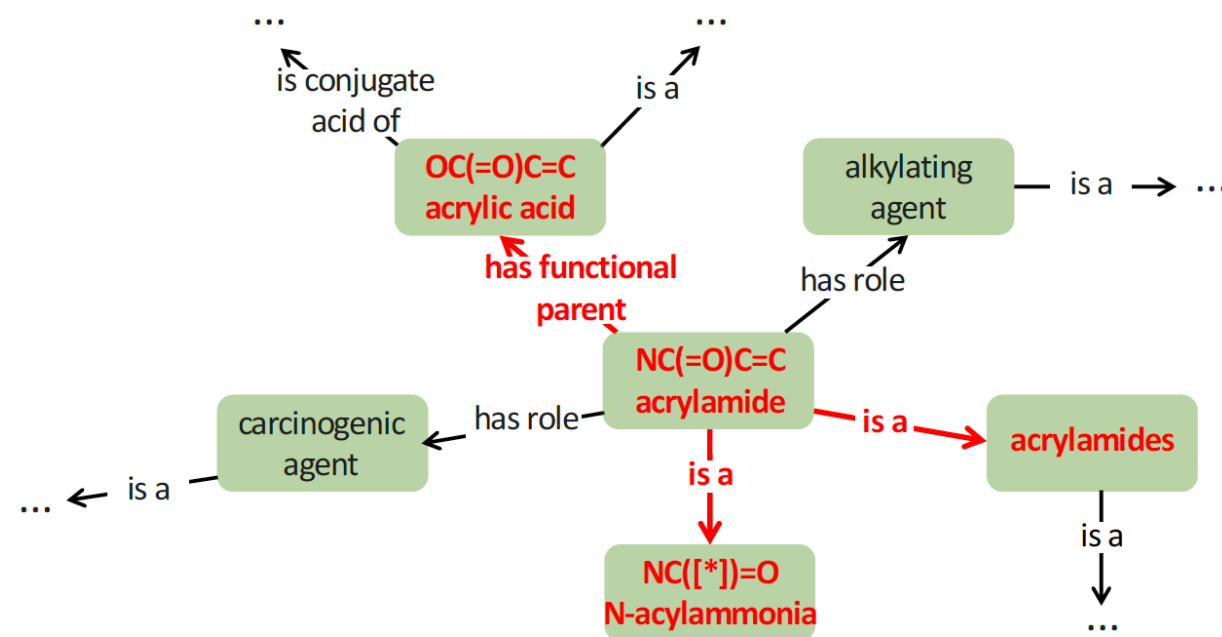
(a) Text description of acrylamide in ChEBI database

# Motivation

- In addition to molecular structures, scientific literature contains important biochemical knowledge about molecules in the form of KG.
- **Problem:** We aim to integrate **auxiliary knowledge** into molecule representation learning for molecule captioning, text retrieval, molecule retrieval.

ChEBI Name	acrylamide
ChEBI ID	CHEBI: 28619
Definition	A white odorless solid, soluble in water and several organic solvents. It is a member of the class of <b>acrylamides</b> that results from the formal condensation of <b>acrylic acid</b> with <b>ammonia</b> .

(a) Text description of acrylamide in ChEBI database

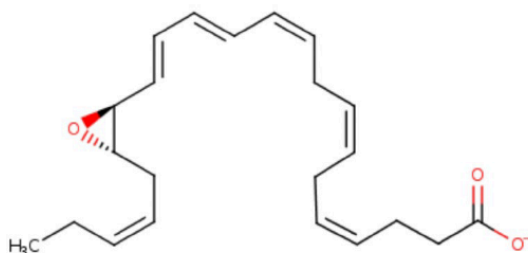


(b) Chemical knowledge graph centered at acrylamide

# Motivation

- Illustration of molecule captioning

Query molecule



Ground-truth text

This molecule is a docosanoid **anion** that is the conjugate base of (16S,17S)-epoxy-(4Z,7Z,10Z,12E,14E,19Z)-**docosahexaenoic acid**, obtained by deprotonation of the carboxy group; major species at pH 7.3.

MKMT-Sum

This molecule is a polyunsaturated fatty acid **anion** that is the conjugate base of (5Z,8Z,11Z,14Z,16Z,19Z)-**docosahexaenoic acid**, obtained by deprotonation of the carboxy group; major species at pH 7.3.

MKMT-VT

This molecule is a docosanoid **anion** that is the conjugate base of (4Z,7Z,10Z,13Z,16Z,19Z)-**docosapentaenoic acid**, obtained by deprotonation of the carboxy group; major species at pH 7.3.

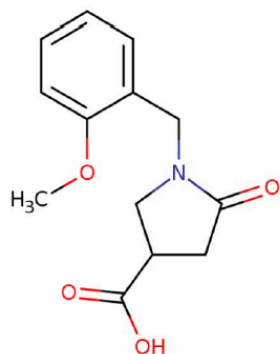
MolT5

The molecule is a **dihydroxydocosahexaenoate** that is the conjugate base of (4Z,7Z,10Z,12E,14S,16Z,19Z,21R)-**dihydroxydocosahexaenoic acid**, obtained by deprotonation of the carboxy group; major species at pH 7.3.

# Motivation

- Illustration of text retrieval

Query molecule



Retrieved text descriptions by KMKT-VT

1. This molecule is an organosulfur compound (✓).
2. this molecule is a substituted aniline and an aromatic ether (×).
3. this molecule is a carbonyl compound (×).
4. this molecule appears as white amorphous lumps or a crystalline mass with a faint odor of bitter almonds (×).
5. this molecule is an alpha - substituted cyanoacetate ester and an ethyl ester (×).

Retrieved text descriptions by MoleculeSTM

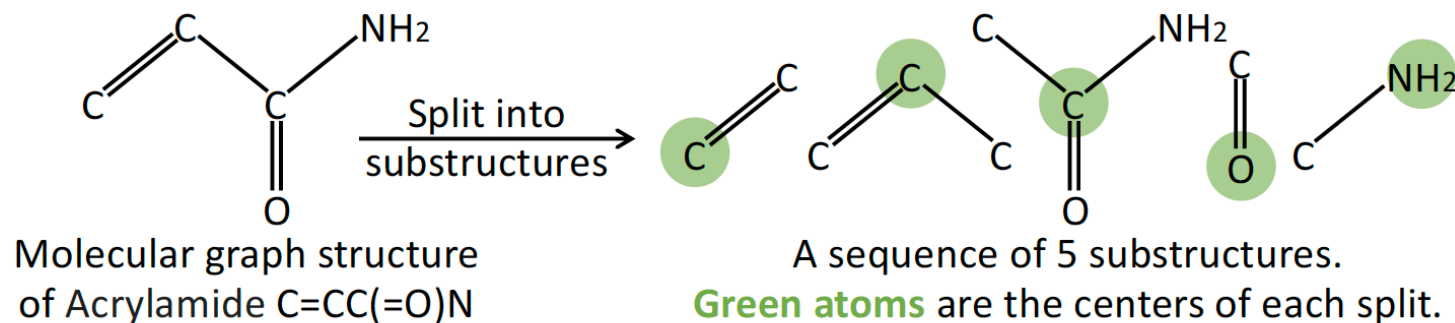
1. This molecule is an organic hydroxy compound (×).
2. this molecule is a primary amine (×).
3. This molecule is an organosulfur compound (✓).
4. this molecule is a substituted aniline and an aromatic ether (×).
5. this molecule is a carbonyl compound (×).

# Motivation

- **Challenges of existing works**
- First, existing works (e.g., Text2Mol, MolT5) only input molecular structures as queries and **ignore** the complementary abundant knowledge, resulting in inaccurate molecule captioning.
- Second, some other works, such as MoleculeSTM, fuse molecules and texts in a **coarse-grained manner**, i.e., encoding the entire molecule and text separately, then fusing them only at the output layer, leading to worse retrieval performance.

# Motivation

- **Novelty and approach of the proposed work**
- We design a foundation model for **multi-modal knowledge-enhanced** generation, retrieval, and reasoning of molecules and text.
- First, we **reason** over linked entities on **KG**, and inject entity embeddings into each molecule encoding layer with two aggregators.
- Second, we design cross-modal attention to fuse them at the **word and molecular substructure** level in a fine-grained manner.

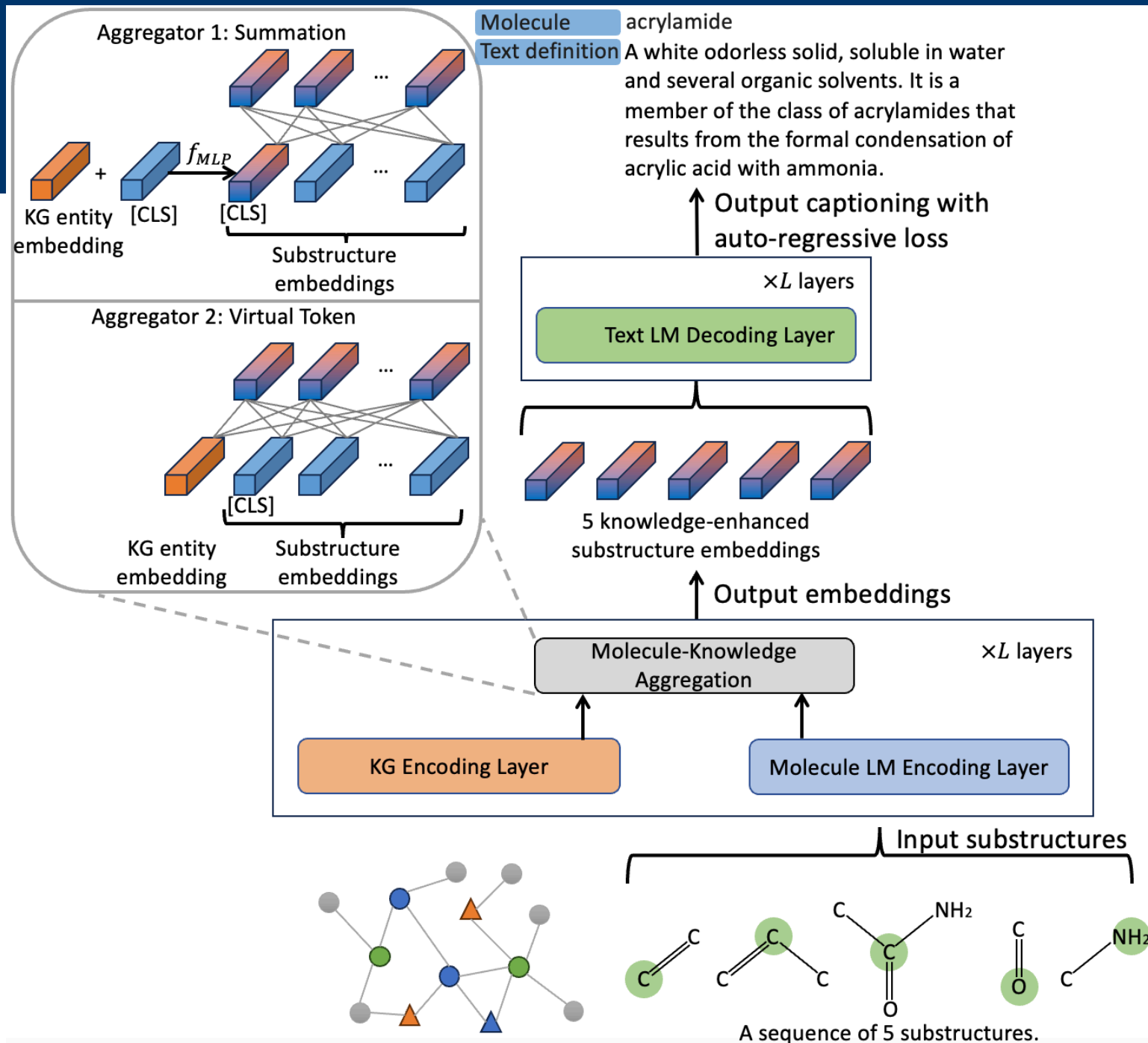


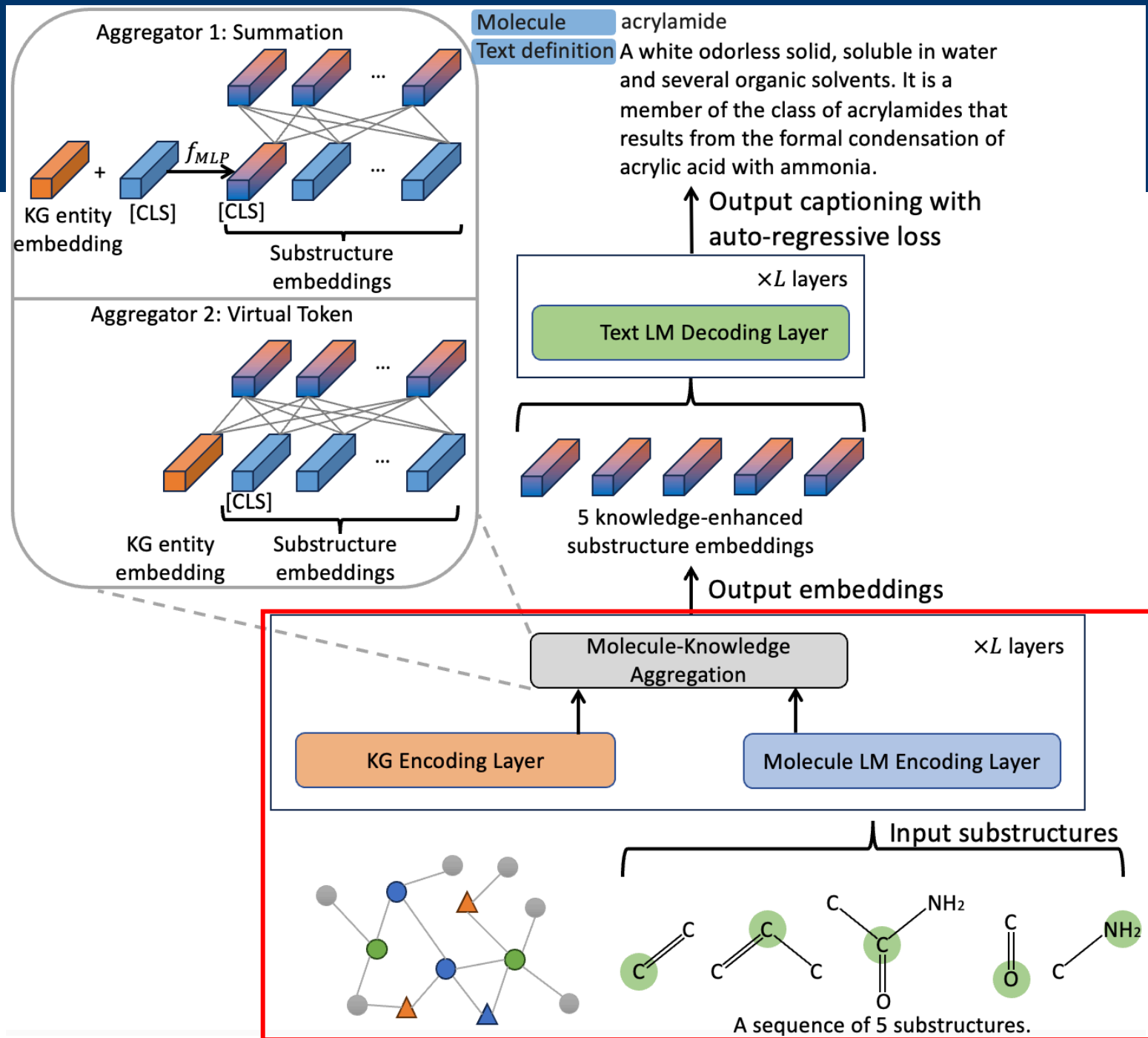
(c) Illustration of splitting a molecular graph into substructures



# Content

- Motivation
- Molecule Captioning
- Molecule-based Text Retrieval
- Experiments

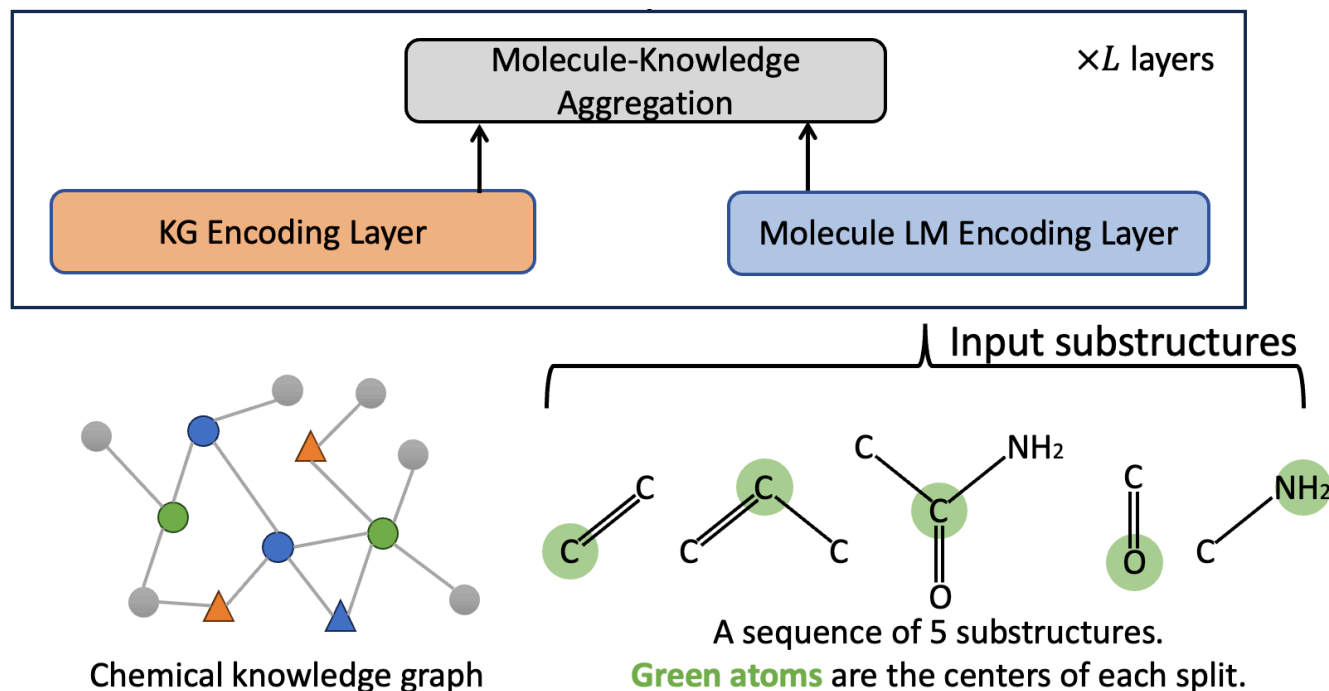


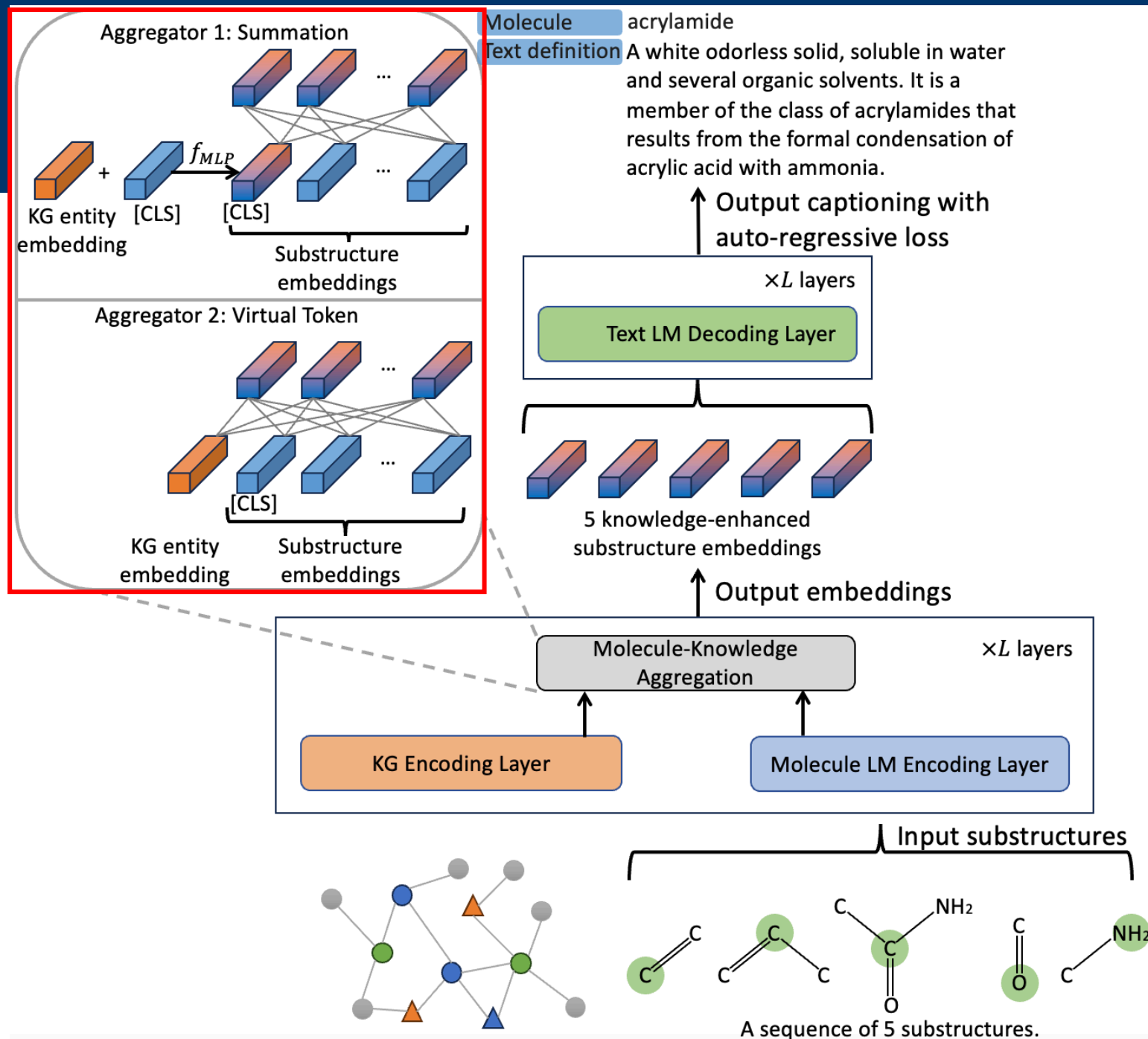


# Molecule Captioning

- 1. We first split a molecule  $m_i$  into  $S$  substructures  $\{m_{i,s}\}_{s=1}^S$ .
- 2. We use knowledge graph neural network (KGNN) to learn entity embedding.

$$\mathbf{e}_i = \text{KGNN}(m_i, G_i)$$





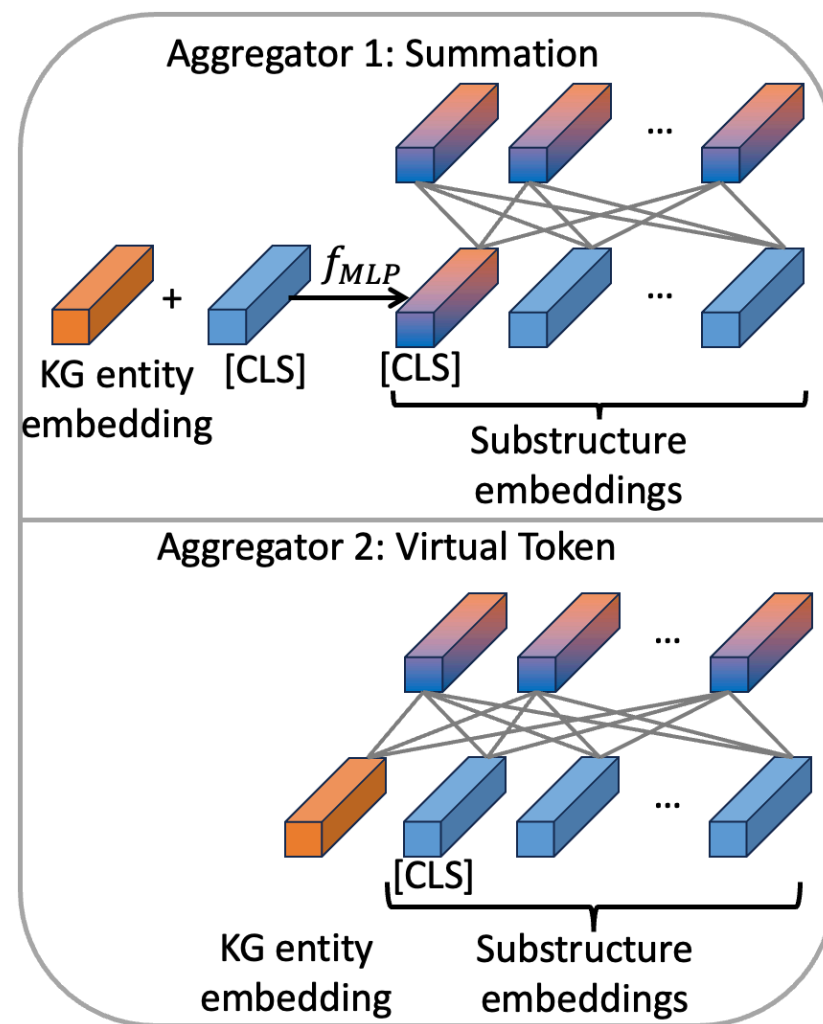
# Molecule Captioning

- 3. Knowledge-Molecule Aggregation
- 3.1. Summation aggregator

$$\tilde{\mathbf{h}}_{i,\text{CLS}}^{(l)} = \mathbf{W}_2 \left( \text{LeakyReLU} \left( \mathbf{W}_1 (\mathbf{h}_{i,\text{CLS}}^{(l)} + \mathbf{e}_i) \right) \right)$$

- 3.2 Virtual token aggregator

$$\tilde{\mathbf{H}}_i^{(l)} = (\mathbf{e}_i || \mathbf{H}_i^{(l)}) = [\mathbf{e}_i, \mathbf{h}_{i,\text{CLS}}^{(l)}, \mathbf{h}_{i,1}^{(l)}, \dots, \mathbf{h}_{i,S}^{(l)}]$$



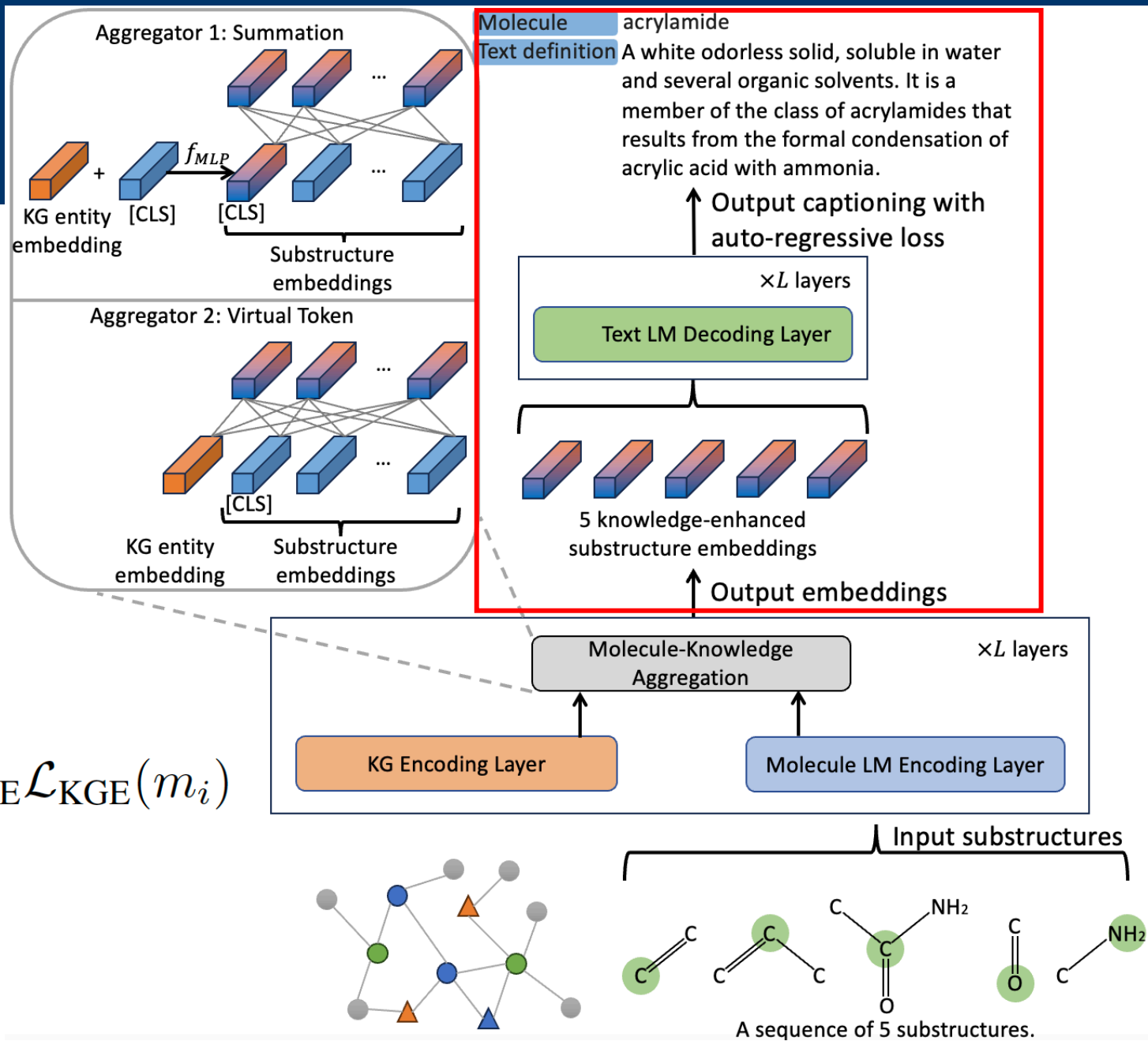
# Molecule Captioning

- 4. Text Generation. We input the fused molecule representation to decoder.

$$\hat{d}_i = f_{\text{dec}}(\mathbf{H}_i^{(L)})$$

- 5. Loss function.

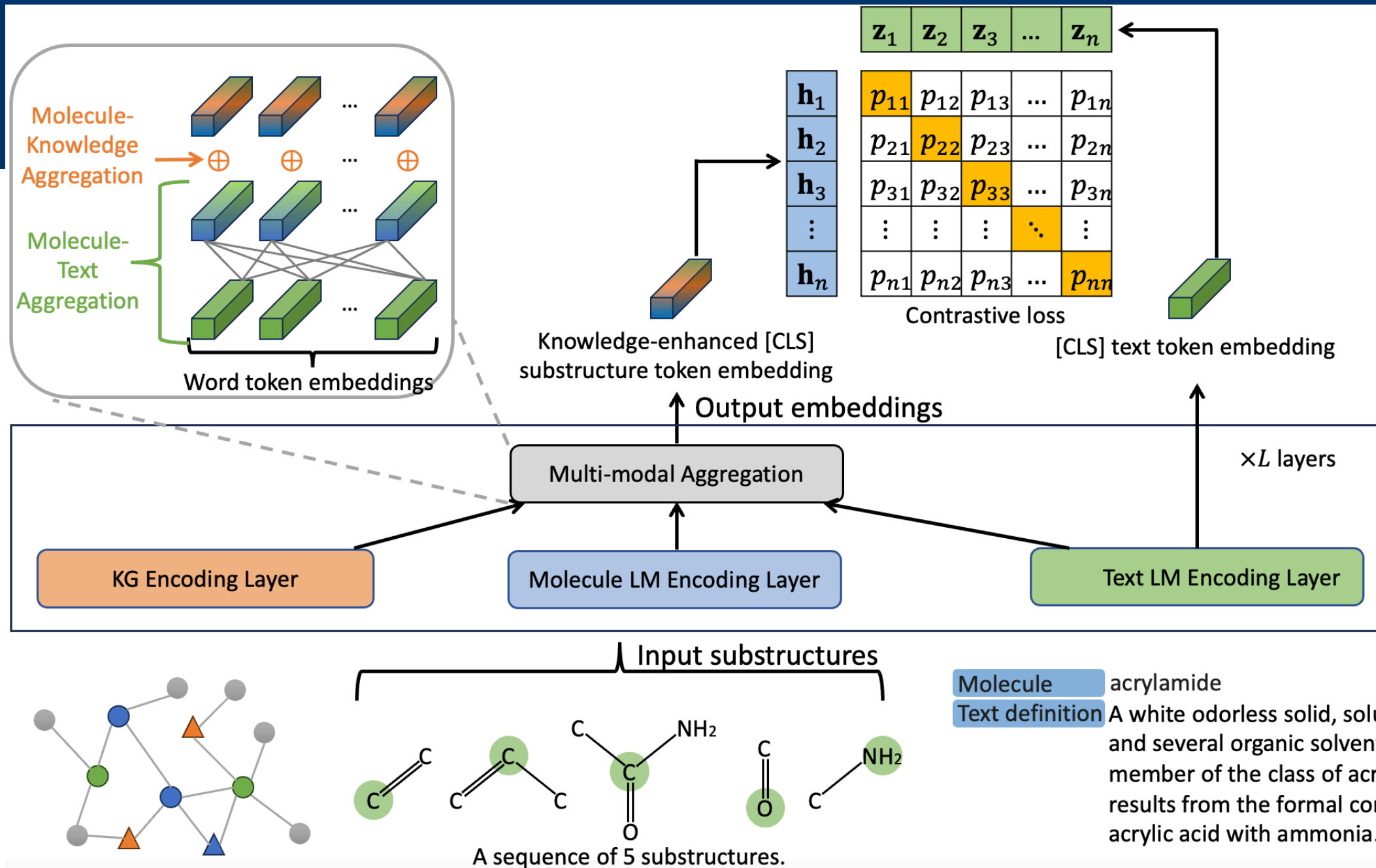
$$\mathcal{L}_{\text{MolCap}} = \sum_{m_i \in \mathcal{M}} \mathcal{L}_{\text{AutoReg}}(\hat{d}_i, d_i) + \lambda_{\text{KGE}} \mathcal{L}_{\text{KGE}}(m_i)$$

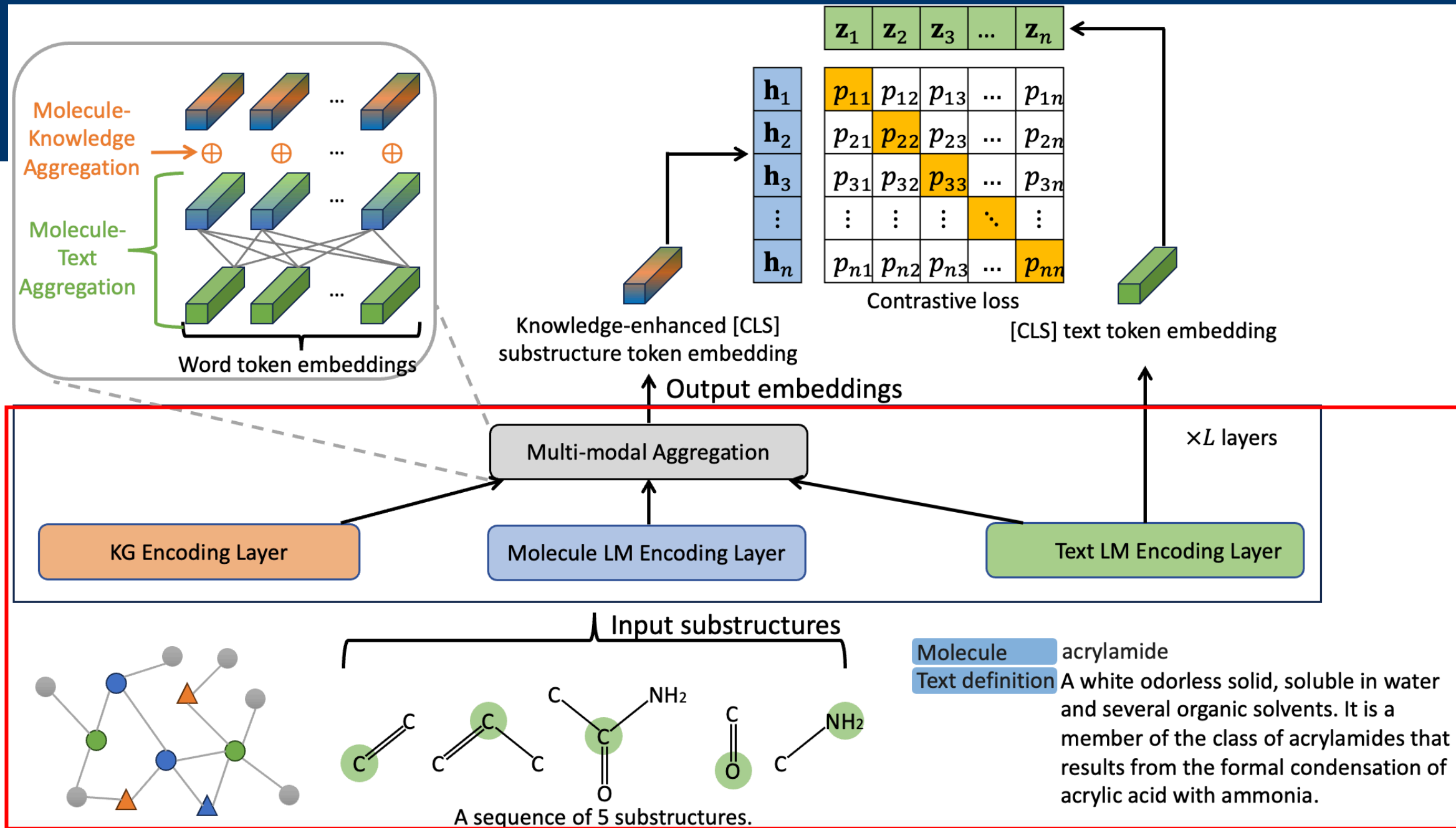


# Content

- Motivation
- Molecule Captioning
- Molecule-based Text Retrieval
- Experiments

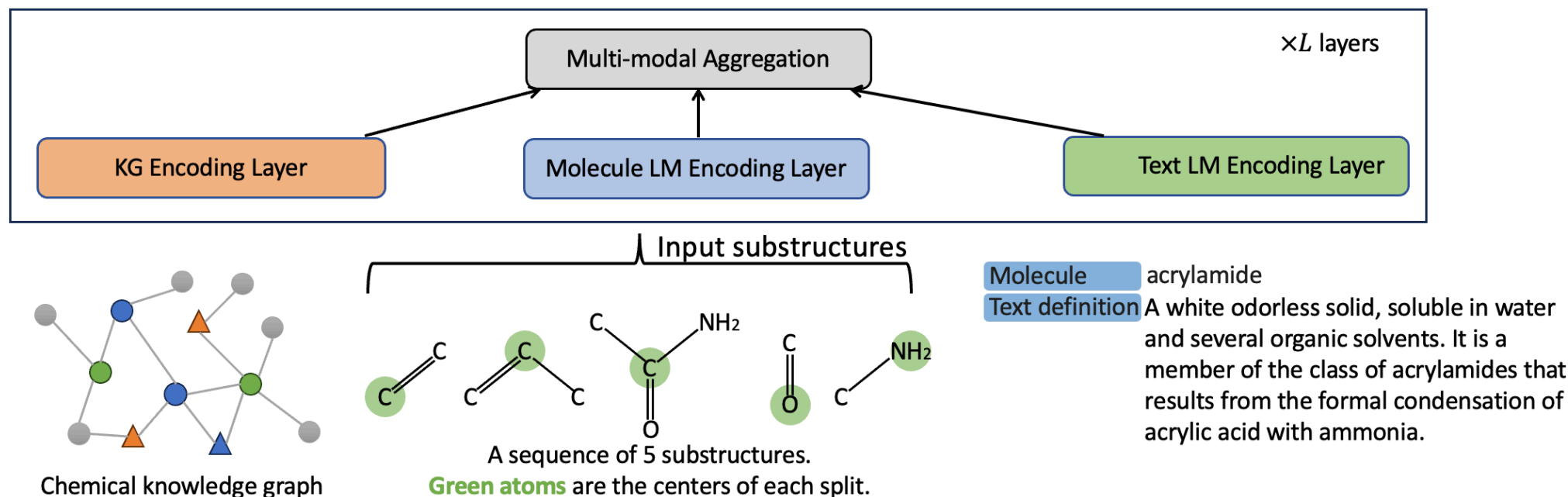


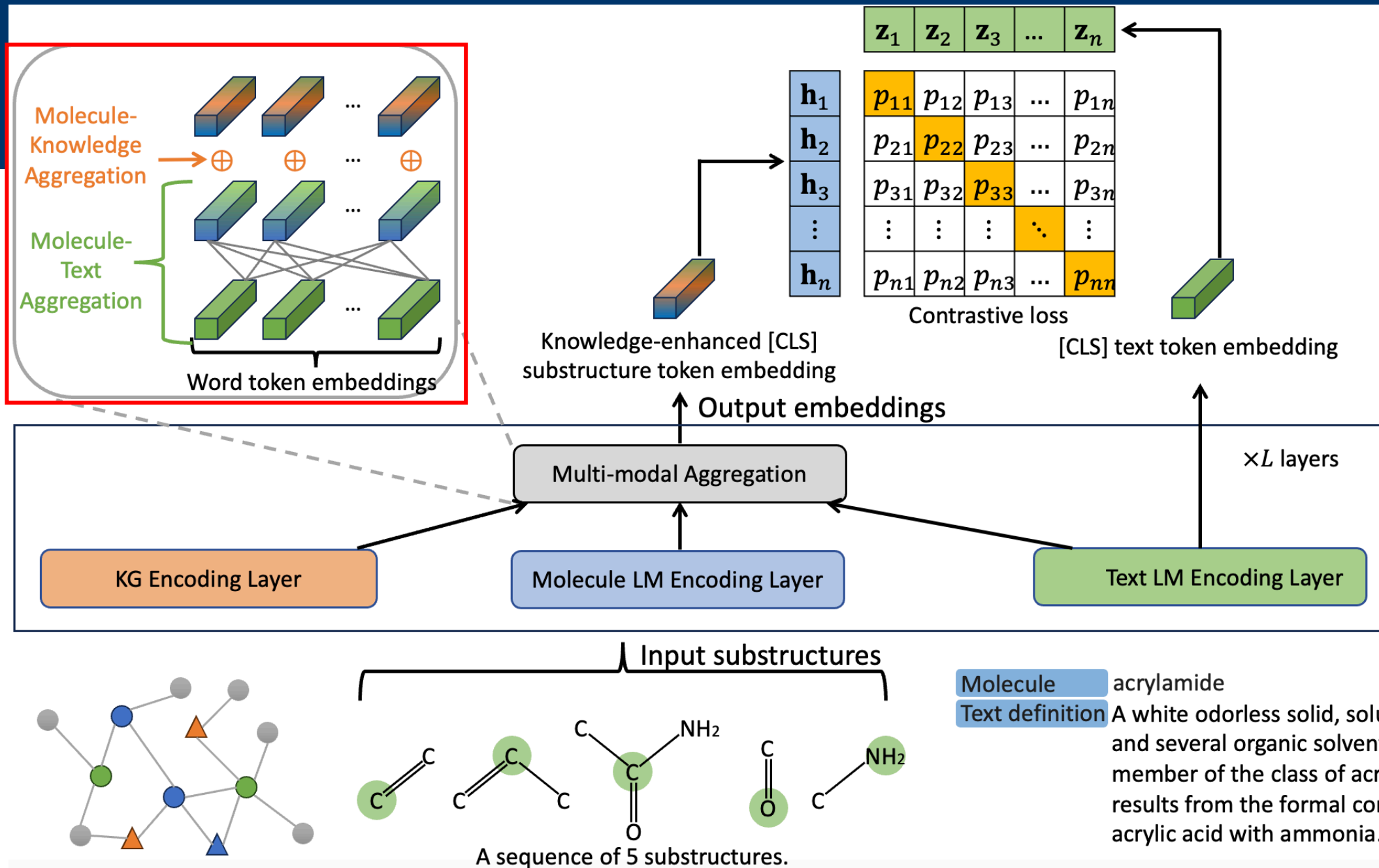




# Molecule-based Text Retrieval

- 1. We first split text  $d_i$  into  $W$  words  $\{d_{i,w}\}_{w=1}^W$ .





# Molecule-based Text Retrieval

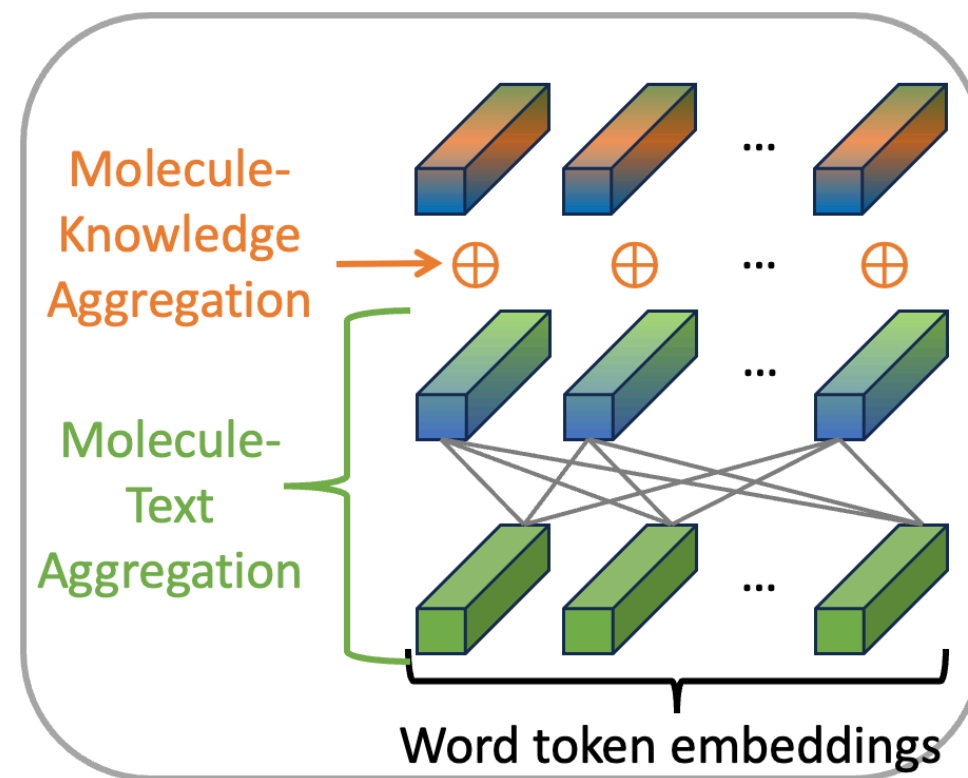
- 2. Molecule-text aggregation.
- 2.1 Attention between a substructure and words.

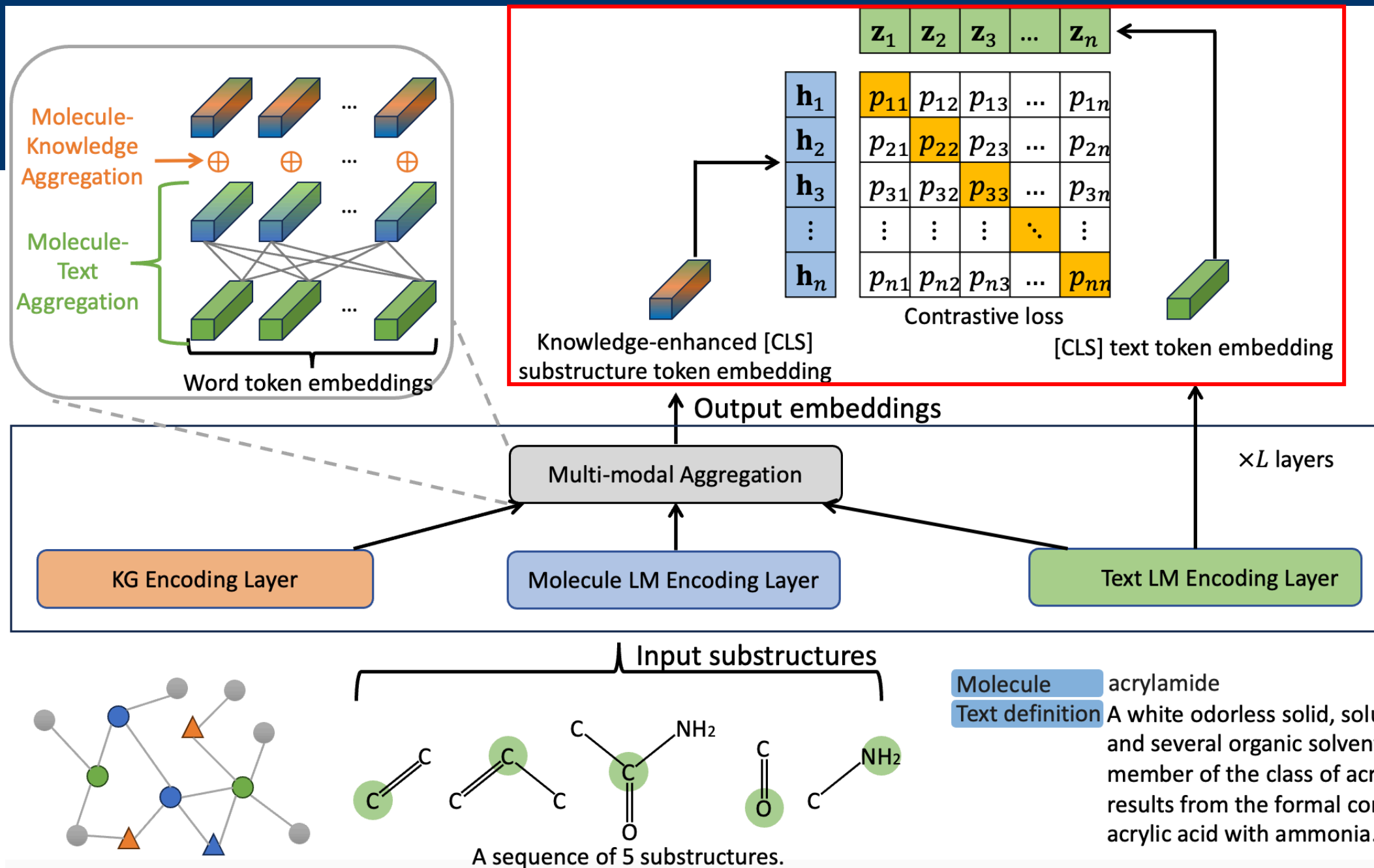
$$\tilde{\beta}(s, w) = \text{LeakyReLU}(\mathbf{b}^\top (\mathbf{h}_{i,s}^{(l)} || \mathbf{z}_{i,w}^{(l)})),$$

$$\beta(s, w) = \frac{\exp(\tilde{\beta}(s, w))}{\sum_{w'=1}^W \exp(\tilde{\beta}(s, w'))}.$$

- 2.2 Aggregation.

$$\tilde{\mathbf{h}}_{i,s}^{(l)} = \mathbf{W}'_2 \left( \text{LeakyReLU} \left( \mathbf{W}'_1 (\mathbf{h}_{i,s}^{(l)} + \sum_w \beta(s, w) \mathbf{z}_{i,w}^{(l)}) \right) \right)$$





# Content

- Motivation
- Molecule Captioning
- Molecule-based Text Retrieval
- Experiments

# Experiments

- 1. Datasets
  - ChEBIKG: 139,572 molecules, 150,414 entities, and 11 relations.
  - PubChemKG: 7,814 molecules, 13,703 entities, and 11 relations.
- 2. Implementation details
  - # KGNN layers = 2, # molecule encoding layers = 12,
  - # epochs = 50, learning rate =  $1e-5$ .



# Experiments

- 1. Molecule captioning

Model	ChEBIKG					
	BLUE-2	BLUE-4	ROUGE-1	ROUGE-2	ROUGE-L	Meteor
RNN	0.3235	0.2339	0.4310	0.3271	0.4264	0.4572
Transformer	0.0828	0.0397	0.1706	0.1048	0.1625	0.2341
MolT5	0.3072	0.2192	0.3943	0.3191	0.4123	0.4612
GPT4	0.1220	0.1056	0.1906	0.1366	0.1656	0.2346
MKMT (w/o KG)	0.3122	0.2126	0.5029	0.4338	0.5020	0.5159
MKMT-Sum	0.2948	0.2013	0.5394	0.4507	0.5392	0.5182
MKMT-VT	<b>0.3269</b>	<b>0.2384</b>	<b>0.5538</b>	<b>0.4705</b>	<b>0.5538</b>	<b>0.5200</b>

Findings: compared to baselines without KG, we outperform them due to the effective modeling of auxiliary KG.

# Experiments

- 1. Molecule captioning

Model	ChEBIKG					
	BLUE-2	BLUE-4	ROUGE-1	ROUGE-2	ROUGE-L	Meteor
RNN	0.3235	0.2339	0.4310	0.3271	0.4264	0.4572
Transformer	0.0828	0.0397	0.1706	0.1048	0.1625	0.2341
MolT5	0.3072	0.2192	0.3943	0.3191	0.4123	0.4612
GPT4	0.1220	0.1056	0.1906	0.1366	0.1656	0.2346
MKMT (w/o KG)	0.3122	0.2126	0.5029	0.4338	0.5020	0.5159
MKMT-Sum	0.2948	0.2013	0.5394	0.4507	0.5392	0.5182
MKMT-VT	<b>0.3269</b>	<b>0.2384</b>	<b>0.5538</b>	<b>0.4705</b>	<b>0.5538</b>	<b>0.5200</b>

Findings: after removing KG from our model, we further verify that KG indeed brings useful information.

# Experiments

- 1. Molecule captioning

Model	ChEBIKG					
	BLUE-2	BLUE-4	ROUGE-1	ROUGE-2	ROUGE-L	Meteor
RNN	0.3235	0.2339	0.4310	0.3271	0.4264	0.4572
Transformer	0.0828	0.0397	0.1706	0.1048	0.1625	0.2341
MolT5	0.3072	0.2192	0.3943	0.3191	0.4123	0.4612
GPT4	0.1220	0.1056	0.1906	0.1366	0.1656	0.2346
MKMT (w/o KG)	0.3122	0.2126	0.5029	0.4338	0.5020	0.5159
MKMT-Sum	0.2948	0.2013	0.5394	0.4507	0.5392	0.5182
MKMT-VT	<b>0.3269</b>	<b>0.2384</b>	<b>0.5538</b>	<b>0.4705</b>	<b>0.5538</b>	<b>0.5200</b>

Findings: virtual token aggregator is better, because virtual token has attention between entity and each substructure.

# Experiments

- 2. Molecule-based text retrieval

Model	ChEBIKG					
	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10
Text2Mol	50.8654	0.0530	0.0121	0.0302	0.0502	0.1012
MoleculeSTM	4.7351	0.5470	0.4286	0.5324	0.6463	0.8104
KV-PLM	16.3520	0.3204	0.2095	0.3258	0.3974	0.5371
MKMT (w/o KG)	3.8111	0.6287	0.4861	0.7133	0.8155	0.9245
MKMT-Sum	3.3953	0.6566	0.5193	0.7412	0.8387	0.9388
MKMT-VT	<b>3.2317</b>	<b>0.6737</b>	<b>0.5387</b>	<b>0.7607</b>	<b>0.8544</b>	<b>0.9466</b>

Findings: similarly, compared to baselines, we verify the usefulness of KG.

# Experiments

- 2. Molecule-based text retrieval

Model	ChEBIKG					
	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10
Text2Mol	50.8654	0.0530	0.0121	0.0302	0.0502	0.1012
MoleculeSTM	4.7351	0.5470	0.4286	0.5324	0.6463	0.8104
KV-PLM	16.3520	0.3204	0.2095	0.3258	0.3974	0.5371
MKMT (w/o KG)	3.8111	0.6287	0.4861	0.7133	0.8155	0.9245
MKMT-Sum	3.3953	0.6566	0.5193	0.7412	0.8387	0.9388
MKMT-VT	<b>3.2317</b>	<b>0.6737</b>	<b>0.5387</b>	<b>0.7607</b>	<b>0.8544</b>	<b>0.9466</b>

Findings: after removing KG from our model, performance drops, verifying the positive effect of KG.

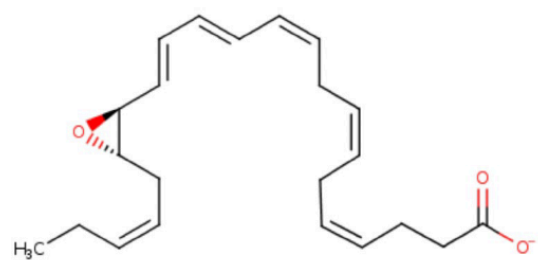
# Experiments

- 3. Molecule retrieval

Model	ChEBIKG					
	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10
Text2Mol	50.8398	0.0529	0.0116	0.0302	0.0494	0.1011
MoleculeSTM	38.5910	0.0703	0.0194	0.0521	0.0836	0.1842
KV-PLM	46.7102	0.0693	0.0184	0.0392	0.0599	0.1400
MKMT (w/o KG)	35.8724	0.0894	0.0229	0.0628	0.1029	0.2049
MKMT-Sum	<b>31.3744</b>	0.0940	0.0248	0.0685	0.1129	0.2147
MKMT-VT	33.9591	<b>0.0942</b>	<b>0.0258</b>	<b>0.0754</b>	<b>0.1176</b>	<b>0.2199</b>

# Experiments

- 4. Case study

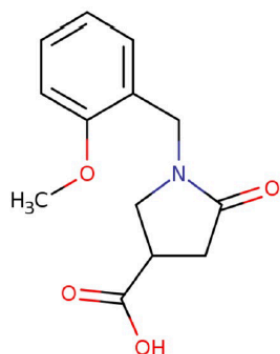
Query molecule	Ground-truth text	MKMT-Sum	MKMT-VT	MolT5
	This molecule is a docosanoid <b>anion</b> that is the conjugate base of (16S,17S)-epoxy-(4Z,7Z,10Z,12E,14E,19Z)- <b>docosahexaenoic acid</b> , obtained by deprotonation of the carboxy group; major species at pH 7.3.	This molecule is a polyunsaturated fatty acid <b>anion</b> that is the conjugate base of (5Z,8Z,11Z,14Z,16Z,19Z)- <b>docosahexaenoic acid</b> , obtained by deprotonation of the carboxy group; major species at pH 7.3.	This molecule is a docosanoid <b>anion</b> that is the conjugate base of (4Z,7Z,10Z,13Z,16Z,19Z)- <b>docosapentaenoic acid</b> , obtained by deprotonation of the carboxy group; major species at pH 7.3.	The molecule is a <b>dihydroxydocosahexaenoate</b> that is the conjugate base of (4Z,7Z,10Z,12E,14S,16Z,19Z,21R)- <b>dihydroxydocosahexaenoic acid</b> , obtained by deprotonation of the carboxy group; major species at pH 7.3.

Findings: our models generate correct chemical terms, while baseline doesn't, due to the effectiveness of KG in bringing abundant knowledge.

# Experiments

- 4. Case study

Query molecule



Retrieved text descriptions by KMKT-VT

1. This molecule is an organosulfur compound (✓).
2. this molecule is a substituted aniline and an aromatic ether (×).
3. this molecule is a carbonyl compound (×).
4. this molecule appears as white amorphous lumps or a crystalline mass with a faint odor of bitter almonds (×).
5. this molecule is an alpha - substituted cyanoacetate ester and an ethyl ester (×).

Retrieved text descriptions by MoleculeSTM

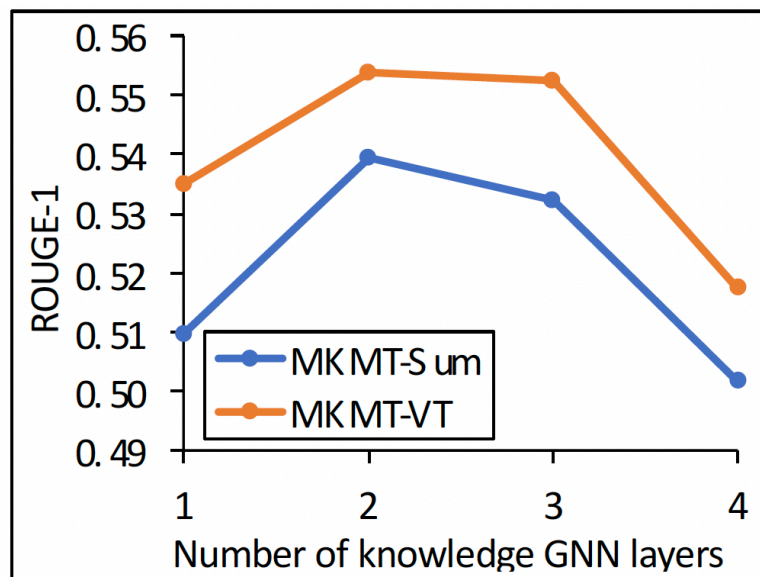
1. This molecule is an organic hydroxy compound (×).
2. this molecule is a primary amine (×).
3. This molecule is an organosulfur compound (✓).
4. this molecule is a substituted aniline and an aromatic ether (×).
5. this molecule is a carbonyl compound (×).

Findings: our models rank the correct text higher than baseline, since auxiliary knowledge complement the information in molecule and text.

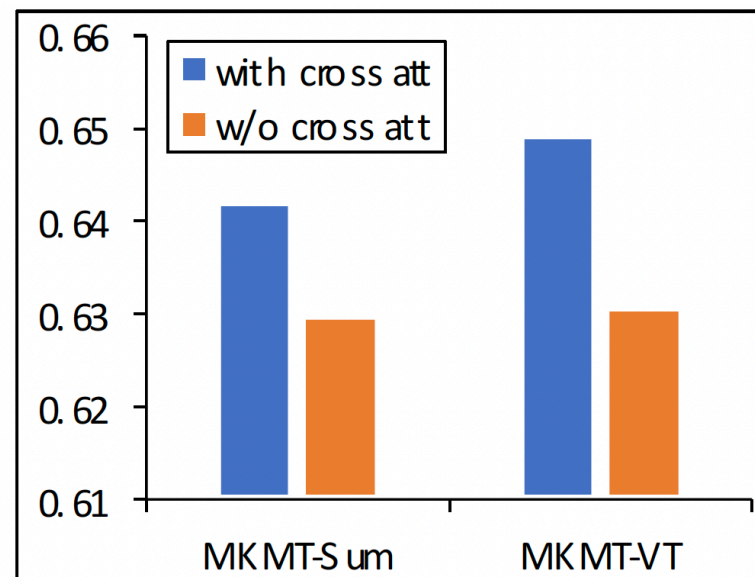


# Experiments

- 5. Ablation study



(a) Number of knowledge GNN layers



(b) Effect of cross-modal attention

Findings: (a) overly high-order knowledge brings noisy information, (b) fine-grained attention indeed helps retrieval.

Yale

Thank You