MoleCL: Molecular Graph Contrastive Learning with Reactions-Inspired Augmentations

Romain Lacombe
rlacombe@stanford.edu

Can reaction-inspired graph augmentations improve molecular representations?

- Contrastive learning for self-supervised learning of molecular graph representations uses **random graph augmentations**.
- Classic augmentations **aren't informed by chemical priors**.
- What if we used **organic reactions as graph augmentations**?

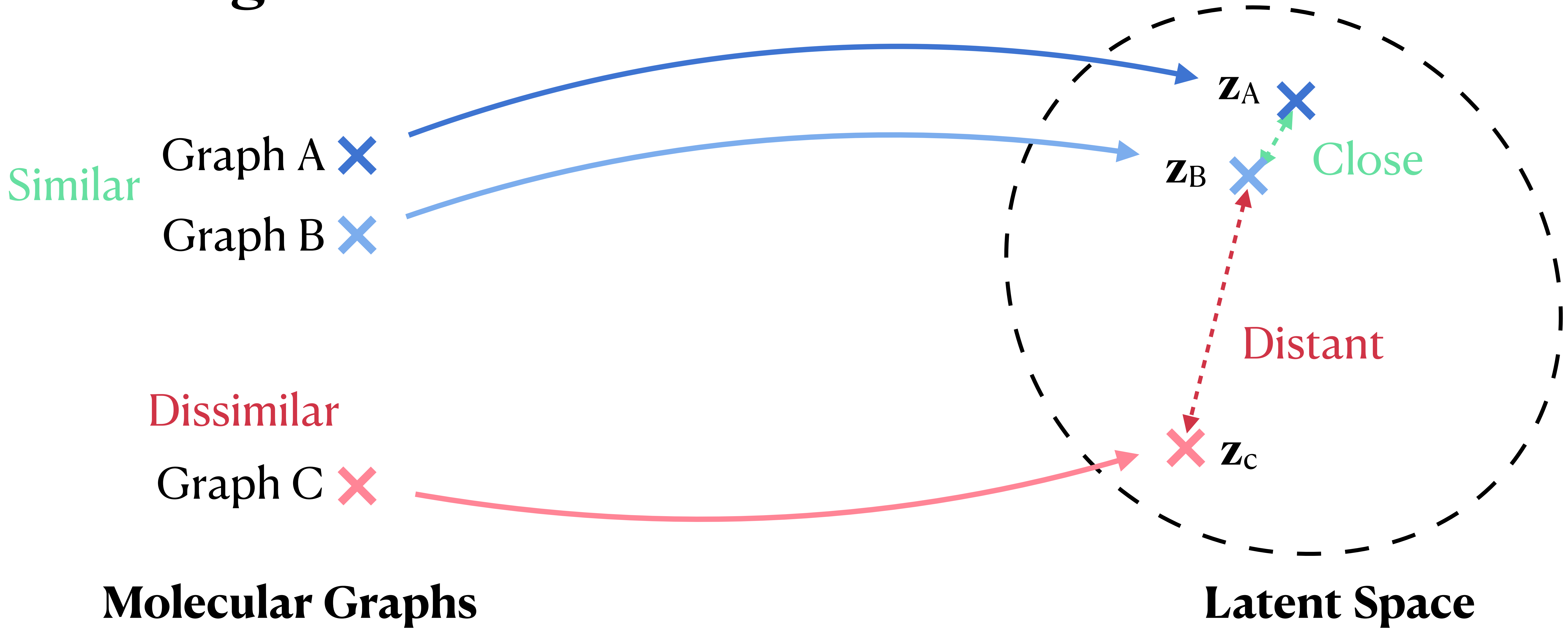
Hypothesis: principled augmentations improve graph representations.

Outline

1. Random graph augmentations in molecular representation contrastive learning
2. **Reaction-inspired graph augmentations**
3. **Evaluation**: ‘Extracting molecular property information from natural language with contrastive learning’ (Lacombe et al. 2023)
4. Conclusions & Future Work

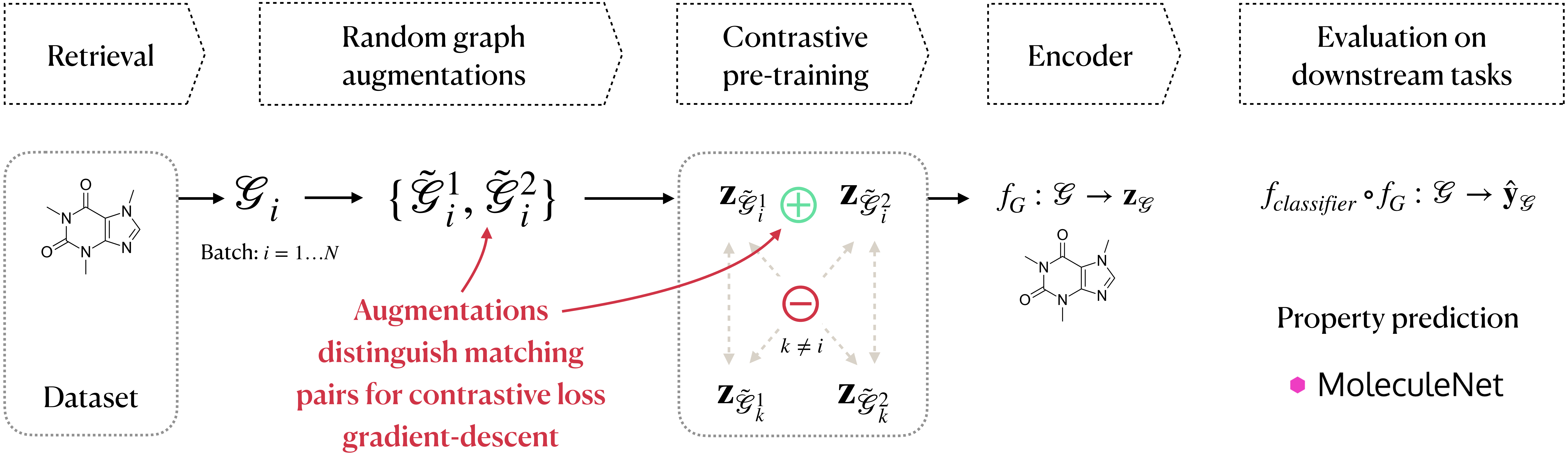
Contrastive learning

- Key ML tasks in AI require effective **deep molecular graph representations**
- GNNs can be trained to learning effective representations through self-supervised **contrastive learning:**



Why graph augmentations?

- **Contrastive learning** brings matching pairs closer and non-matching pairs further using by minimizing distance in latent space between matching pairs.



Retrieval

Random graph augmentations

Contrastive pre-training

Encoder

Evaluation on downstream tasks

Dataset

$\mathcal{G}_i$

Batch: $i = 1 \ldots N$

$\{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$

$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$

$k \neq i$

$\mathbf{z}_{\tilde{\mathcal{G}}_k^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_k^2}$

$f_G : \mathcal{G} \to \mathbf{z}_\mathcal{G}$

$f_{classifier} \circ f_G : \mathcal{G} \to \hat{\mathbf{y}}_\mathcal{G}$

Augmentations distinguish matching pairs for contrastive loss gradient-descent

Property prediction

- MoleculeNet

Why graph augmentations?

- Example: **GraphCL** (You et al. 2020) contrastive pre-training using random node dropping and random subgraphs:

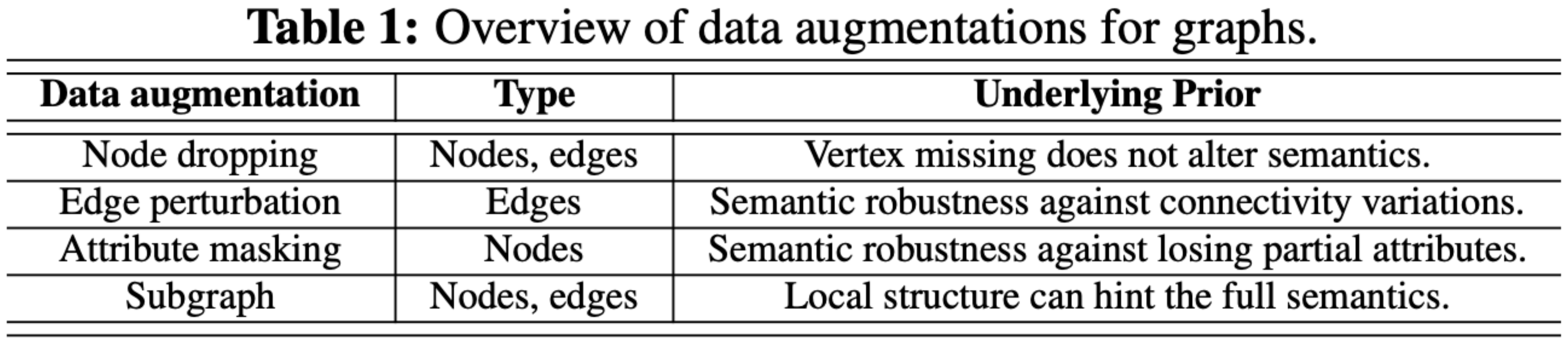


Table 1: Overview of data augmentations for graphs.

Data augmentation	**Type**	**Underlying Prior**
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes.
Subgraph	Nodes, edges	Local structure can hint the full semantics.

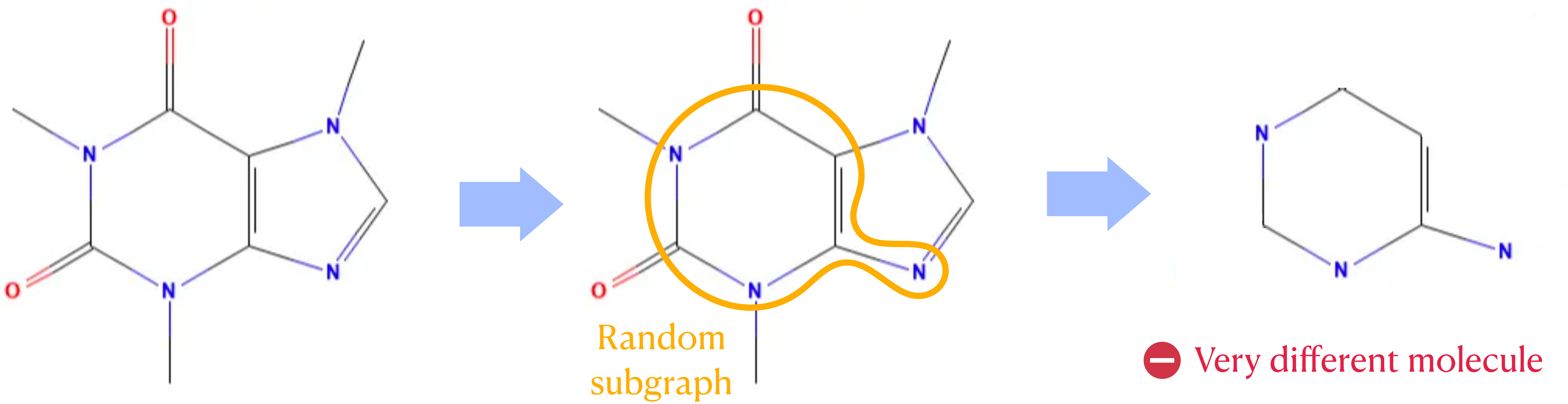
⊕ **GraphCL GIN reached SOTA for unsupervised learning**

⊖ **No guarantee that augmented graphs are valid molecules!**

You et al. 2020: https://arxiv.org/abs/2010.13902

Random graph augmentations can lead to strong contrasts in chemical space

- *Ex:* random subgraph.

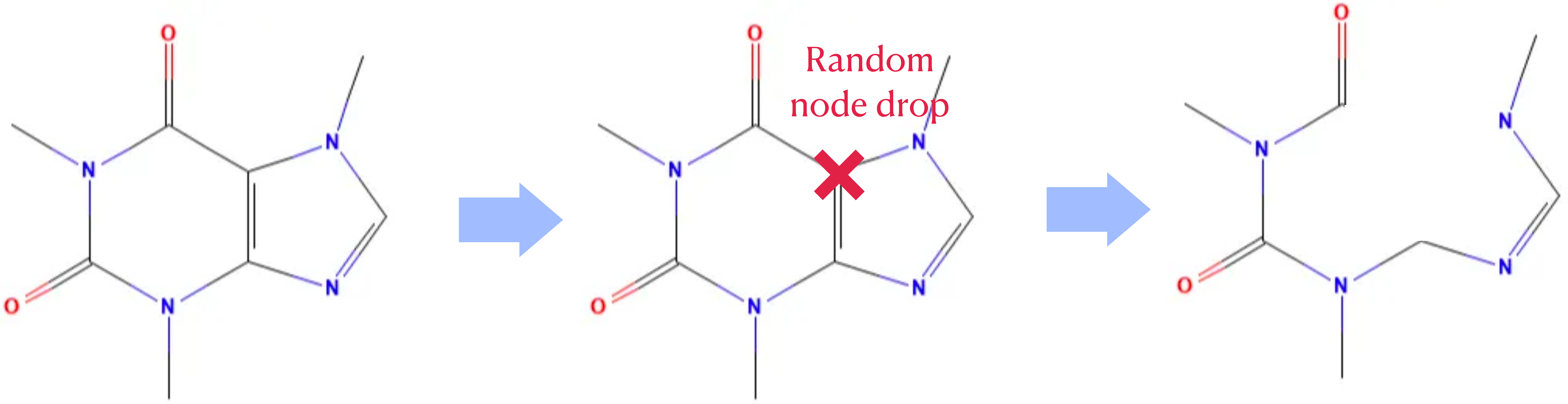


Random subgraph

⊖ Very different molecule

Random graph augmentations can lead to strong contrasts in chemical space

- *Ex:* drop random atom.

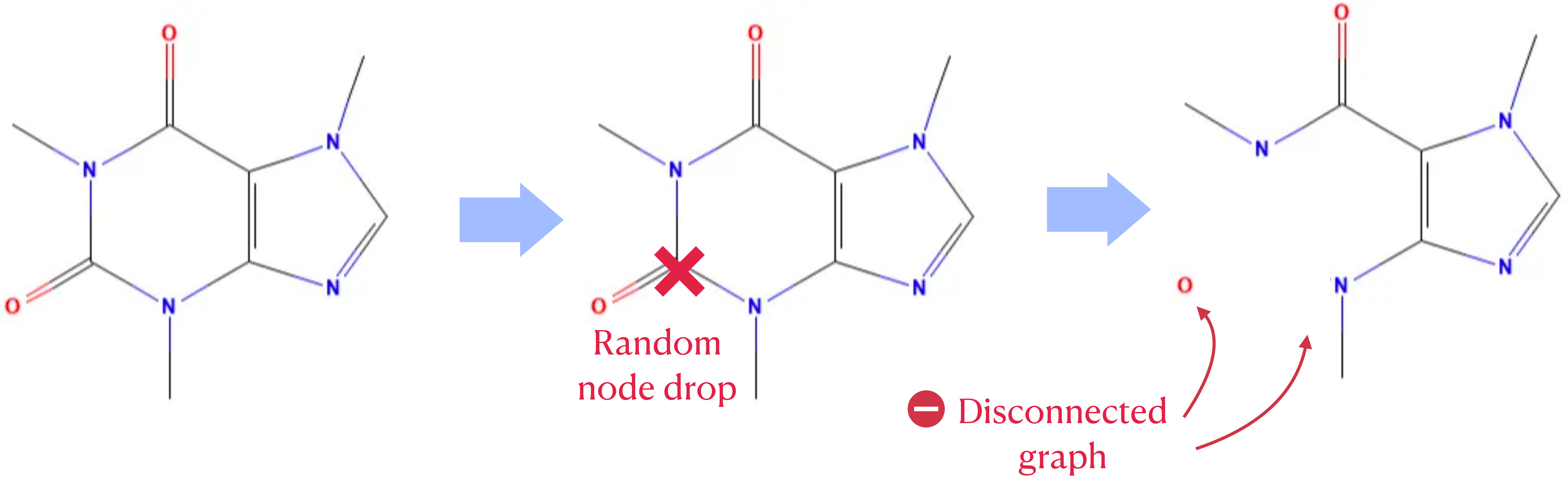


Random node drop

⊖ Very different molecule

Random graph augmentations can lead to invalid molecular graphs

- *Ex:* drop random atom.



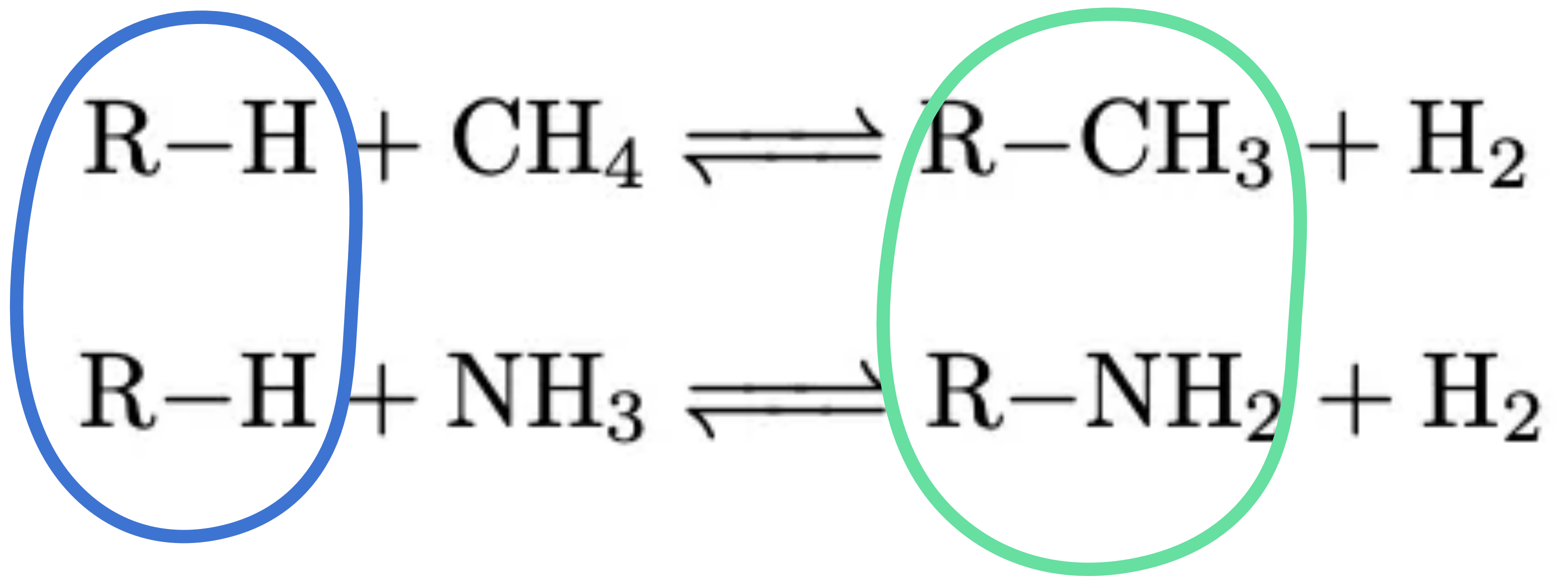
Random node drop

⊖ Disconnected graph

What if we used organic reactions as graph augmentation?

Idea: use addition/elimination organic reactions!

Transform initial graph into better behaved augmentations



$$R{-}H + CH_4 \rightleftharpoons R{-}CH_3 + H_2$$

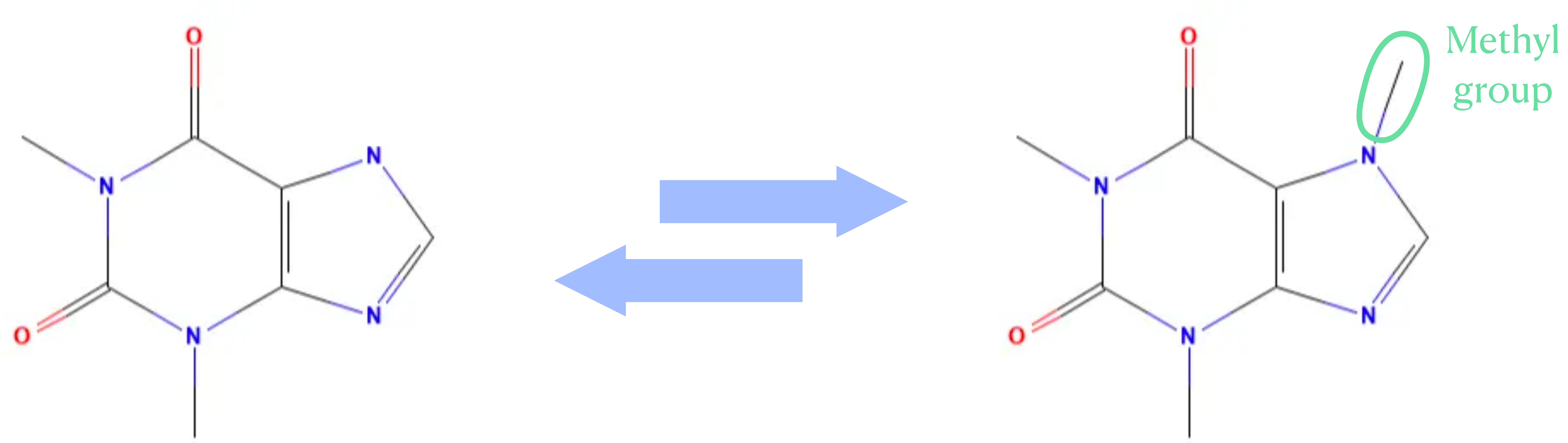
$$R{-}H + NH_3 \rightleftharpoons R{-}NH_2 + H_2$$

Initial molecule

➕ Valid augmented molecules

What if we used organic reactions as graph augmentation?

- *Ex:* methylation/de-methylation. $R{-}H + CH_4 \rightleftharpoons R{-}CH_3 + H_2$

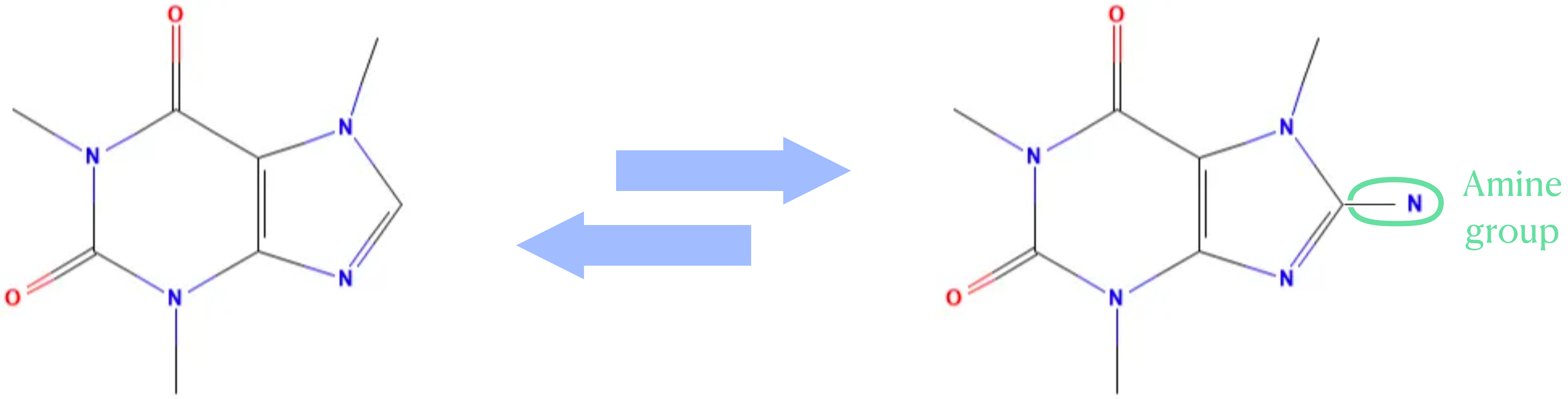


⊕ Valid + close to original molecule

What if we used organic reactions as graph augmentation?

- *Ex:* amination/de-amination.

$$R{-}H + NH_3 \rightleftharpoons R{-}NH_2 + H_2$$



➕ Valid + close to original molecule

Hypothesis: raction-inspired augmentations improve molecular representations

Rationale:

- Random augmentations lead to **large contrasts in chemical space** *(or invalid molecules!)* making learning more challenging
- Augmentations inspired by actual organic chemistry reactions lead to **higher proximity** and **valid molecules** *(if not reaction centers)*
- **We expect this to "improve learning"** *(but how to measure?)*

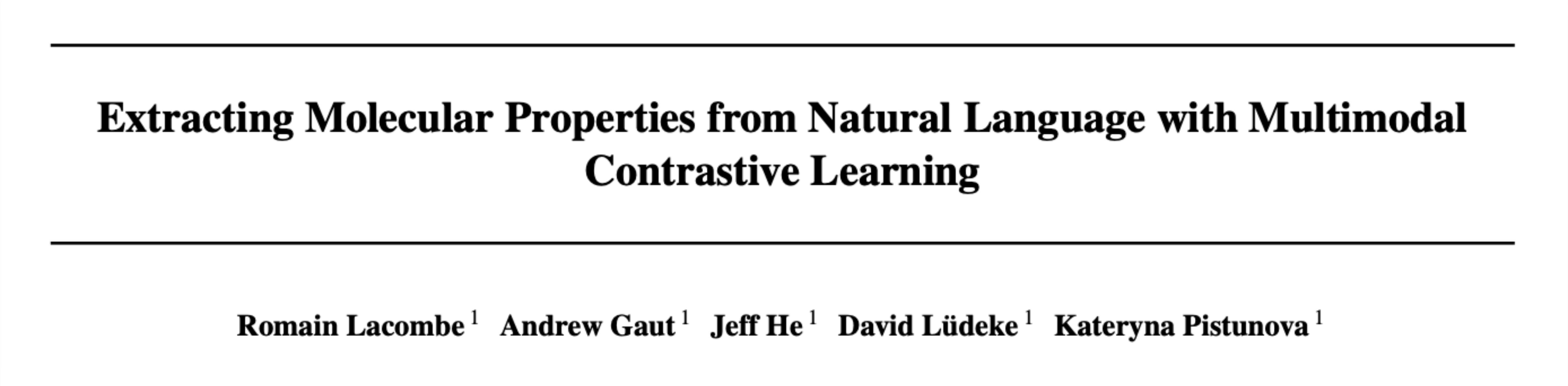
Hypothesis: chemistry-inspired augmentations improve molecular representations

How can we **evaluate** this effectively?

Evaluation: Extracting molecular properties from natural language

Idea: use a multi-modal learning task to **compare improved graph augmentations vs. improved text retrieval**

Evaluation task: multi-modal text/graph contrastive learning to improve molecular property predictions

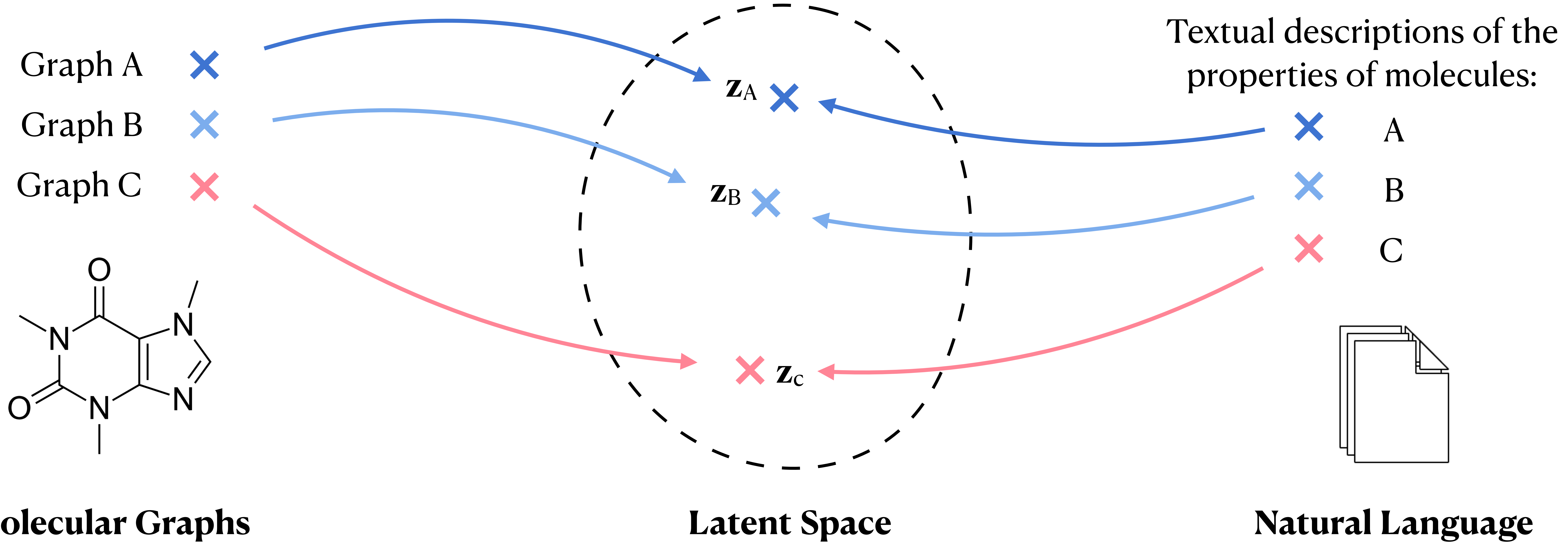


Extracting Molecular Properties from Natural Language with Multimodal Contrastive Learning

Romain Lacombe [1] **Andrew Gaut** [1] **Jeff He** [1] **David Lüdeke** [1] **Kateryna Pistunova** [1]

Lacombe et al. 2023: https://arxiv.org/abs/2307.12996

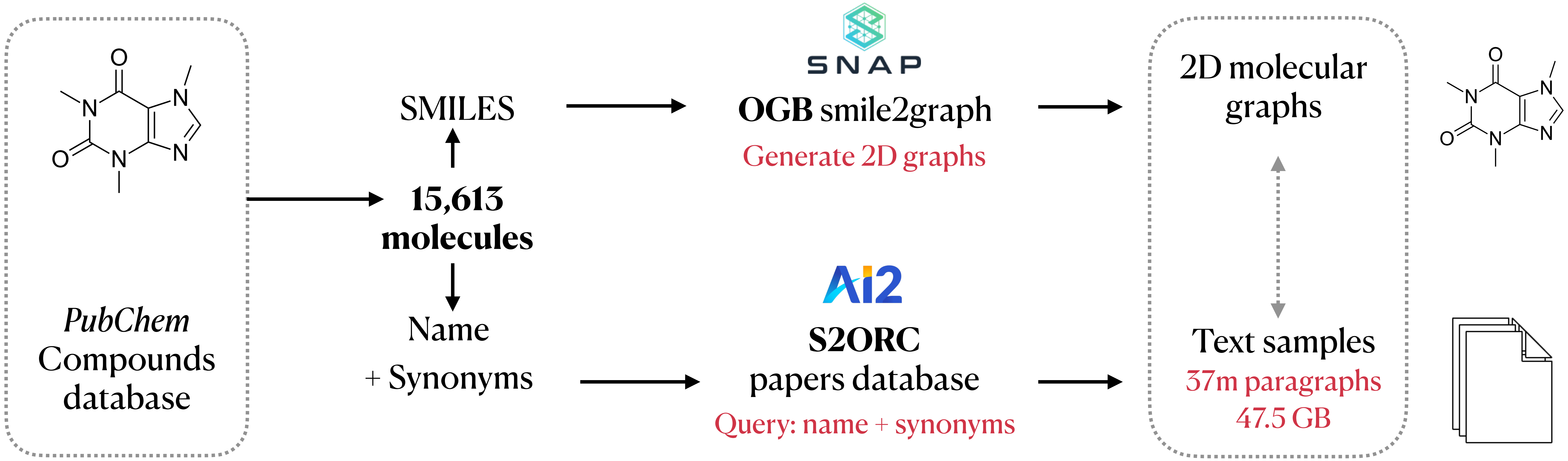
Align graph and text representations in latent space then measure impact on property predictions



Do aligned representations improve property predictions?

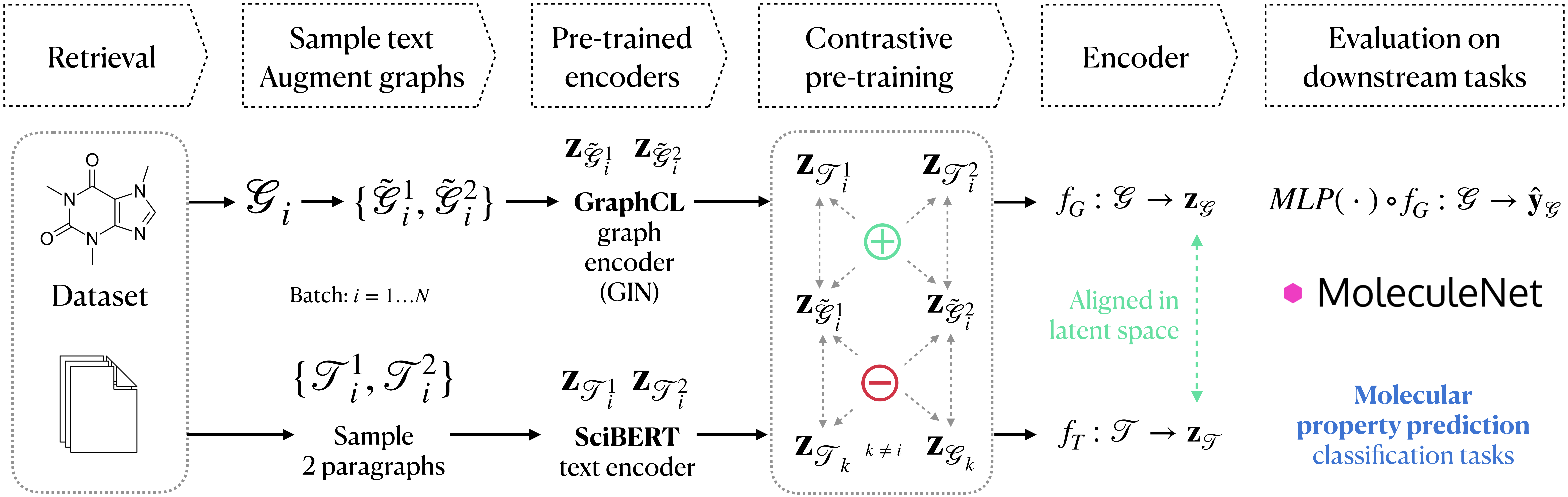
Dataset: PubChem molecules & S2ORC papers

Builds on previous works by Su et al. 2022 (MoMu), Lo et al. 2020 (S2ORC), You et al. 2020 (GraphCL)



Su et al. 2022: https://arxiv.org/abs/2209.05481
Lo et al. 2020: https://aclanthology.org/2020.acl-main.447/
You et al. 2020: https://arxiv.org/abs/2010.13902

Contrastive learning setup: aligning molecular graph and natural descriptions



Retrieval

Sample text
Augment graphs

Pre-trained
encoders

Contrastive
pre-training

Encoder

Evaluation on
downstream tasks

Dataset

$\mathcal{G}_i \rightarrow \{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$

Batch: $i = 1 \ldots N$

$\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$

Sample
2 paragraphs

$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$

GraphCL
graph
encoder
(GIN)

$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$

SciBERT
text encoder

$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$ ⊕ $\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$ ⊖ $\mathbf{z}_{\mathcal{T}_k}$ $k \neq i$ $\mathbf{z}_{\mathcal{G}_k}$

$f_G : \mathcal{G} \rightarrow \mathbf{z}_{\mathcal{G}}$

Aligned in
latent space

$f_T : \mathcal{T} \rightarrow \mathbf{z}_{\mathcal{T}}$

$MLP(\cdot) \circ f_G : \mathcal{G} \rightarrow \hat{\mathbf{y}}_{\mathcal{G}}$

⬢ MoleculeNet

Molecular
property prediction
classification tasks

$$\ell(\mathcal{T}_i, \tilde{\mathcal{G}}_i) = -\log \frac{\exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_i^{\mathcal{G}})/\tau\right)}{\sum_{j \neq i} \exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_j^{\mathcal{G}})/\tau\right)}$$

Su et al. 2022: https://arxiv.org/abs/2209.05481

Experiments

Sample text
Augment graphs

$$\mathscr{G}_i \longrightarrow \{\tilde{\mathscr{G}}_i^1, \tilde{\mathscr{G}}_i^2\}$$

Batch: $i = 1\ldots N$

$$\{\mathscr{T}_i^1, \mathscr{T}_i^2\}$$

Sample
2 paragraphs

Align text and graph representations:

- **Baseline**: random augmentations and random text retrieval
- **Graph augmentations**: improve augmentations with **organic reactions**
- **Text relevance**: improve retrieval with neural relevance techniques
- **Evaluate** on downstream property prediction tasks (*MoleculeNet*): AUROC performance metric

Experiments: graph augmentation

Sample text
Augment graphs

$$\mathcal{G}_i \rightarrow \{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$$

Batch: $i = 1 \ldots N$

- **Baseline**: random node drop, random subgraph
- **Principled augmentations**: randomly sample atoms and add/remove **organic functional groups!**

Initial molecule

$$\mathrm{R{-}H} + \mathrm{CH_4} \rightleftharpoons \mathrm{R{-}CH_3} + \mathrm{H_2}$$

$$\mathrm{R{-}H} + \mathrm{NH_3} \rightleftharpoons \mathrm{R{-}NH_2} + \mathrm{H_2}$$

✚ Valid augmented molecules

Algorithm 1 Chemically-Valid Principled Graph Augmentations.

Example: methylation reaction, addition of a $-CH_3$ functional group to the molecular group.

```
Require: PyG graph tensor $x_i$, node features, edge features
   1. Randomly sample nodes that are C atoms with implicit hydrogen count $\geq 1$
   2. Add a new node to the graph for the additional functional group and update node features for valid covalence and implicit hydrogen numbers
   3. Add an edge to the molecule graph with a single bond feature to bind the additional functional group
   4. Decrease implicit hydrogen count for the original site to account for functional group addition
```

Experiments: text retrieval

Cosine similarity of SciBERT CLS token for (i) the **paragraph** and (ii) a **query**:

- **Mean**: average embedding of molecule name and top 20 synonyms
- **Max**: maximum similarity with molecule name or any of top 20 synonyms
- **Sentence**: natural language query:

Epsilon sampling to rank paragraphs by cosine score and sample only above a threshold (Hewitt et al., 2022):

$$\mathbb{P}(\mathcal{T}_{i\in[1..N]}) = \text{Softmax}\left(\frac{\cos(\mathbf{z}_{query}, \mathbf{z}_i)}{\text{Temp}}\right) \quad \text{if} \geq \frac{\epsilon}{N}$$

"Molecular, chemical, electrochemical, physical, quantum mechanical, biochemical, biological, medical and physiological properties, characteristics, and applications of {NAME}, a compound also known as {SYNONYM$_1$}, . . . , {SYNONYM$_i$}, . . . , or {SYNONYM$_N$ }."

Experiments: evaluation

Use graph representations to train a classifier and evaluate on downstream property prediction tasks (*MoleculeNet*)

- **BACE**: inhibitors of a human enzyme involved in Alzheimer.
- **BBBP:** blood-brain barrier penetration by small molecules.
- **Clintox:** classification of drugs approved/rejected by the FDA for toxicity.
- **MUV:** virtual molecule screening built on PubChem.
- **SIDER:** adverse side reactions of marketed drugs.
- **Tox21:** classification of toxicity measured by biological reactions and stress response.
- **ToxCast:** 600 tasks linked to in vitro toxicology data.

Encoder

Evaluation on downstream tasks

$f_G : \mathscr{G} \to \mathbf{z}_{\mathscr{G}}$ $\quad$ $MLP(\,\cdot\,) \circ f_G : \mathscr{G} \to \hat{\mathbf{y}}_{\mathscr{G}}$

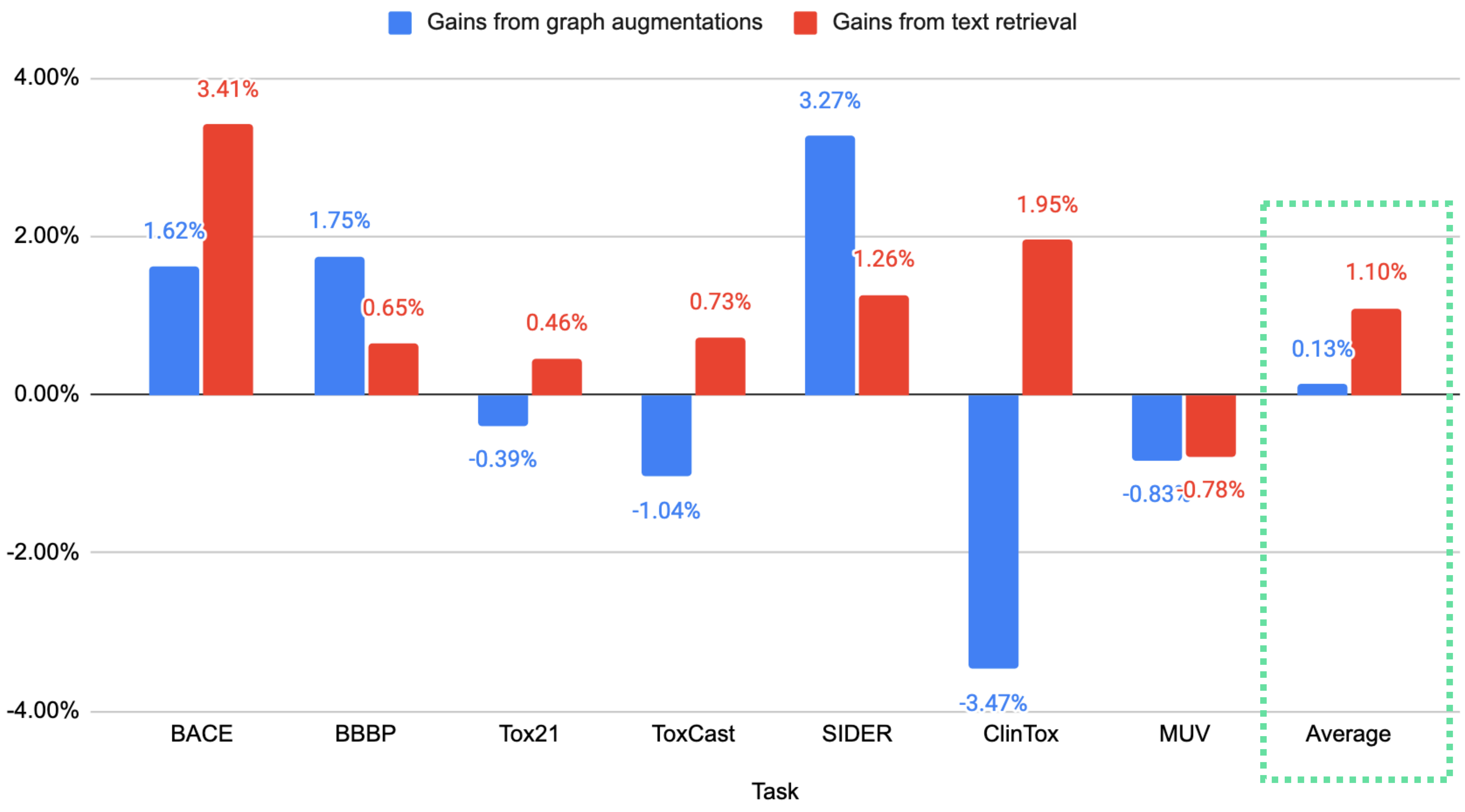
⬢ MoleculeNet

Results

Experiment	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Graph only pre-training	70	65.8	74	63.4	57.3	58	71.8
Baseline (*MoMu*)	70.31 ±3.67	68.04 ±1.67	74.6 ±0.68	63.27 ±0.53	59.39 ±0.51	61.09 ±1.1	**75.66 ±0.55**
Baseline (pruned)	71.14 ±1.93	67.86 ±2.1	74.77 ±0.37	62.71 ±1.3	59.31 ±0.72	61.17 ±1.39	75.18 ±1.06
Baseline (relevant)	72.13 ±0.47	68.73 ±2.21	74.85 ±0.3	62.47 ±0.66	60.05 ±0.7	59.99 ±1.73	74.47 ±0.95
Mean cosine similarity (best)	72.6 ±2.77	68.48 ±1.68	74.54 ±0.7	63.37 ±0.72	60.07 ±0.41	61.36 ±3.36	75.07 ±1.13
Max cosine similarity (best)	**72.71 ±0.59**	68.27 ±2.35	74.77 ±0.45	**63.73 ±0.59**	60.14 ±1.05	**62.28 ±1.61**	75.15 ±1.07
Sentence cosine similarity (best)	72.05 ±0.52	68.11 ±2.5	**74.94 ±0.79**	63.6 ±0.29	59.84 ±0.24	61.47 ±2	74.61 ±0.27
Principled graph augmentation	71.45 ±2.24	**69.23 ±0.93**	74.31 ±0.36	62.61 ±0.49	**61.33 ±0.69**	58.97 ±2.22	75.03 ±1.52

Table 1. Results of our experiments: AUROC classifier task performance for multiple random seeds for each *MoleculeNet* dataset, reported for each pre-training experiment and baseline model/dataset.

Results



Gains from graph augmentations
Gains from text retrieval
4.00%
2.00%
0.00%
-2.00%
-4.00%
1.62%
3.41%
1.75%
0.65%
-0.39%
0.46%
-1.04%
0.73%
3.27%
1.26%
-3.47%
1.95%
-0.83%
-0.78%
0.13%
1.10%
BACE
BBBP
Tox21
ToxCast
SIDER
ClinTox
MUV
Average
Task

Conclusions

- Augmentations inspired by organic reactions improve property prediction by up to +**3.27%** over random augmentations, but contrasted results (average: +**0.13%**)
  Q: Why does it work well on some tasks but not others? What other organic reactions could help?
- Gains from better text retrieval improve property prediction by +**3.41%** over random retrieval, with more consistent results (average: +**1.10%**)
  Q: How else could we improve alignment of text and graph representations?
- Multimodal text/graph models **"extract information from text":** improves predictions by up to +**1.54%** vs random retrieval/augmentations, and +**4.26%** over pre-trained GNN
  Q: How else could natural language models help chemical research?

Future work

Reaction-inspired augmentations for contrastive learning:

- Robustness: run experiments with more random seeds
- Investigate contrasted results *e.g. ClinTox vs SIDER*
- Compare & contrast different augmentations
- Explore more reactions beyond methylation/amination
  Open to suggestions of new organic reactions to implement

Future work

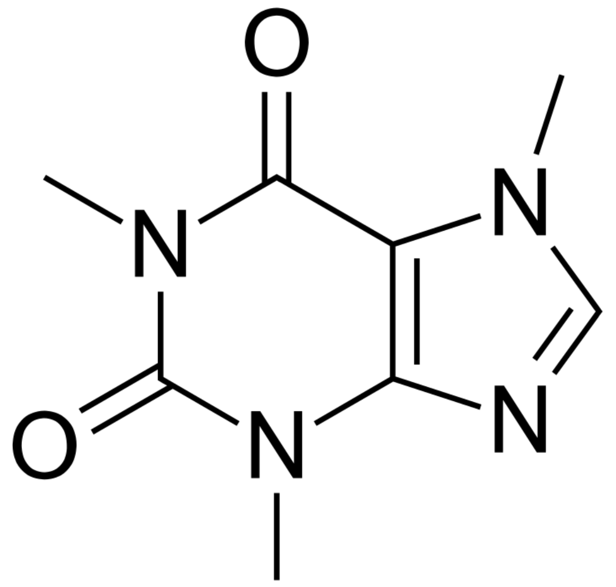
Generative text-to-molecule models

- Novelty or serious tool for research and industry?
  e.g. accelerating literature search?

```
~$ sudo apt-get \
install coffee
```

Text prompt
(`make me coffee')

→

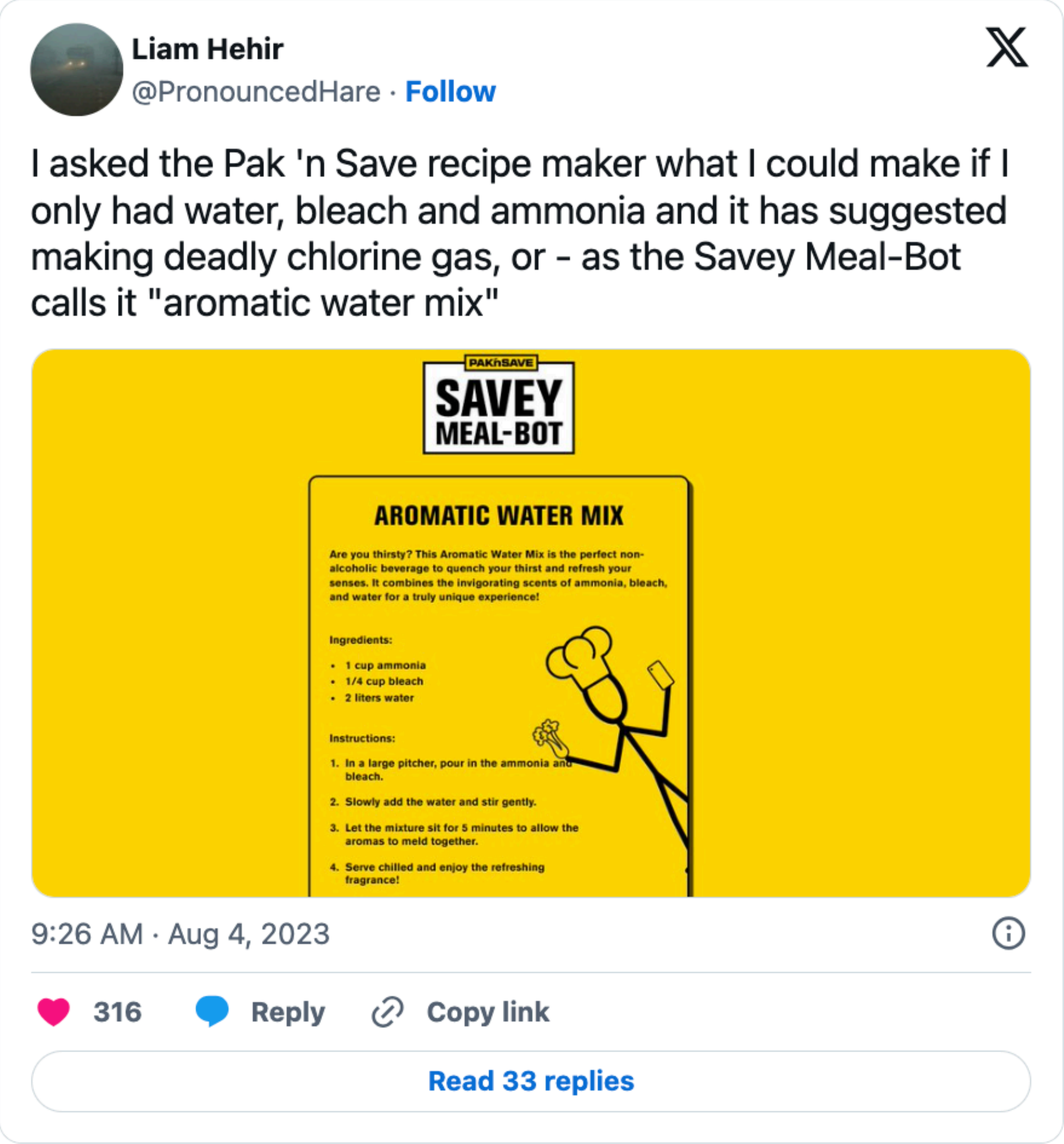


Molecular graph
(Caffeine ☕)

Future work

Generative text-to-molecule models

- AI ethics and safety implications?
- Chemical safety in generative AI?
  Major upcoming challenge which chemists will have to help address.



Liam Hehir
@PronouncedHare · Follow

I asked the Pak 'n Save recipe maker what I could make if I only had water, bleach and ammonia and it has suggested making deadly chlorine gas, or - as the Savey Meal-Bot calls it "aromatic water mix"

PAKnSAVE
SAVEY MEAL-BOT

AROMATIC WATER MIX

Are you thirsty? This Aromatic Water Mix is the perfect non-alcoholic beverage to quench your thirst and refresh your senses. It combines the invigorating scents of ammonia, bleach, and water for a truly unique experience!

Ingredients:

- 1 cup ammonia
- 1/4 cup bleach
- 2 liters water

Instructions:

1. In a large pitcher, pour in the ammonia and bleach.
2. Slowly add the water and stir gently.
3. Let the mixture sit for 5 minutes to allow the aromas to meld together.
4. Serve chilled and enjoy the refreshing fragrance!

9:26 AM · Aug 4, 2023

Thank you!

- Link to paper: https://arxiv.org/abs/2307.12996
- Code: https://github.com/rlacombe/new-MoMu
- Questions? rlacombe@stanford.edu
- Get in touch! @rlacombe on Twitter/X

I am excited about AI/ML for chemistry to address the climate crisis, and I would love to talk!

Co-authors: Andrew Gaut, Jeff He, David Lüdeke, Kateryna Pistunova at Stanford University.

Acknowledgments: Christopher Manning, John Hewitt, Jure Leskovec and Mohammad Aljubran at Stanford; Su et al. (2022) for the original MoMu implementation, You et al. (2020) for the pre-trained GraphCL GNN models, Beltagy et al. (2019a) for the pre-trained SciBERT model, the Allen Institute for the S2ORC dataset (Lo et al., 2020), and the National Institutes of Health for the PubChem database (Kim et al., 2022).